TÜRKİYE BİLİMSEL VE TEKNİK ARAŞTIRMA KURUMU

THE SCIENTIFIC AND TECHNICAL RESEARCH COUNCIL OF TURKEY

Elektrik Elektronik ve Enformatik Araştırma Grubu (EEEAG)

Electrical and Electronics and Informatics Research Grant Group

ORTA DOĞU TEKNİK ÜNİVERSİTESİ – BİLGİ TOPLUMU TEKNOLOJİLERİ – MÜKEMMELİYET MERKEZ PROJESİ

(MIDDLE EAST TECHNICAL UNIVERSITY - INFORMATION SOCIETY TECHNOLOGIES - EXCELLENCY CENTER)

PROJE NO: 105E068

PROF. DR. ASUMAN DOĞAÇ

OCAK 2008
ANKARA

ÖNSÖZ

Bilgi teknolojileri, çok hızlı bir şekilde ilerleyen teknolojik gelişmelerle birlikte öngörülemez fırsatlar yaratmaktadır. Buna paralel olarak, Türkiye'nin Avrupa seviyesindeki rekabet edebilirliği, bilgi toplumu teknolojilerini geliştirmesine ve bunları ekonomik gelişiminde kullanabilme yeteneğine bağlıdır. Bu sebeple, bilgi toplumu teknolojilerinin araştırma ve teknolojik gelişimin yeteneklerinin acil bir şekilde güçlendirmesine ihtiyacı vardır.

"Orta Doğu Teknik Üniversitesi – Bilgi Toplumu Teknolojileri – Mükemmeliyet Merkez Projesi (METU-ISTEC)" projesi Türkiye'nin lider teknik devlet üniversitesindeki bilgi toplumu teknolojilerinin araştırma ve teknolojik gelişimi çalışmalarını yeniden yapılandırmayı amaçlamıştır.

Bu projede yapılan çalışmalar sonucu 7. Çerçeve Programı bilgi toplumu teknolojileri alanına, birinci çağrıda, METU-ISTEC proje grupları toplam 17 proje önerisi yollamışlardır. Bu projelerden 8'i Türkiye koordinatörlüğünde yollanmıştır. Türkiye'nin yer aldığı projelerin 14.78%'i, Türkiye koordinatörlüğünde yollanmış projelerin 27.59%'u METU-ISTEC kapsamında hazırlanmıştır. Ayrıca eşik değeri geçen 6 proje, 2'si Türkiye koordinatörlüğünde olmak üzere METU-ISTEC kapsamında hazırlanmıştır. İlk çağrı kapsamında koordinatörlüğünü Orta Doğu Teknik Üniversitesi'nin üstlendiği iSURF projesi ve Orta Doğu Teknik Üniversitesi'nin ortak olarak yer aldığı ROSSI projesi destek almıştır.

METU-ISTEC gruplarının yer aldığı proje önerilerinin toplam bütçesi 51.490.540 Euro olup, 9.224.016 Euro'luk bütçe METU-ISTEC gruplarına aittir. Destek alan projelerin toplam bütçesi 5.929.337 Euro olup, Türkiye'ye ait toplam bütçe 1.077.001 Euro'dur.

METU-ISTEC projesi kapsamında hazırlık aşamasında yeni araştırmacılara destek verilmiş, ODTÜ bünyesinde yeni çalışma grupları oluşturulmuştur. Çalışma grupları, araştırma alanlarında proje hazırlarken öncelikle mevcut projeleri, gelişmeleri, akademik çalışmaları ve standartları inceleme fırsatı bulmuşlardır. Ardından oluşturulan çekirdek gruplar aktif olarak proje hazırlama aşamasında görev almışlardır. Böylece proje boyunca yeni genç araştırmacılar yetiştirilmeye başlanmıştır. Ayrıca, METU-ISTEC kapsamında hazırlanan projelerde 9 yerli şirketin 7. Çerçeve Programı proje önerilerine katılımı sağlanmıştır.

METU-ISTEC projesi TÜBİTAK EEEAG (Elektrik Elektronik ve Enformatik Araştırma Grubu) tarafından desteklenmiştir.

İÇİNDEKİLER

TABLO ve ŞEKİL LİSTESİ

ÖZ

Bilgi teknolojileri, çok hızlı bir şekilde ilerleyen teknolojik gelişmelerle birlikte öngörülemez fırsatlar yaratmaktadır. Buna paralel olarak, Türkiye'nin Avrupa seviyesindeki rekabet edebilirliği, Bilgi Toplumu Teknolojilerini geliştirmesine ve bunları ekonomik gelişiminde kullanabilme yeteneğine bağlıdır. Bu sebeple, Türkiye'nin Avrupa Ülkelerini yakalama adına, Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimin yeteneklerinin acil bir şekilde güçlendirmesine ihtiyacı vardır. Bu açılım da Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimiyle ilgili insan kapitaline ve aktivitelere yatırımı gerektirmektedir. Buna ilaveten, Avrupa Birliği üyesi ülkelerle Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi konusunda işbirliğinin güçlendirilmesi kaçınılmazdır. Bu proje önerisinin ana amacı; Avrupa doğrultusunda Türk araştırma alanlarının, Türkiye Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi yetenekleri doğrultusunda destekleyip, gerekli entegrasyon ve güçlendirmenin sağlanmasıdır.

"Orta Doğu Teknik Üniversitesi – Bilgi Toplumu Teknolojileri – Mükemmeliyet Merkez Projesi (METU-ISTEC)" projesi Türkiye'nin lider teknik devlet üniversitesindeki Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi çalışmalarını ve sıralanan FP6 Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi stratejik hedef alanlarını yeniden yapılandırmayı amaçlamaktadır; e-İş ve e-Devlet, anlamsal-tabanlı bilgi sistemleri, kavramsal sistemler ve yazılımlarla servisler için açık geliştirme ortamları, gömülü sistemler. Bu yeniden yapılandırma, uygulamalı ve genel teknoloji araştırmalarını birleştirmek ve Avrupa'da öncelik taşıyan uygulamalı araştırma alanlarına odaklanmak için fırsat yaratacaktır. Bu amaç doğrultusunda, METU-ISTEC projesi çok önemli kaynakları seferber etmektedir, bunlar; birçok akademisyen (20 profesör), mezun öğrenciler (39 doktora ve 32 yüksek lisans öğrencisi) dir. METU-ISTEC projesi sonuç olarak en az dokuz adet koordinatörlüğünü ODTÜ öğretim üyelerinin yapacağı FP7 proje önerileri çıkarmayı hedeflemiştir.

Anahtar kelimeler: Türk Araştırma alanlarının Avrupa doğrultusunda entegrasyonu ve geliştirilmesi, e-İş ve e-Devlet, anlamsal-tabanlı bilgi sistemleri, kavramsal sistemler ve yazılımlarla sevisler için açık geliştirme ortamları, gömülü sistemler.

ABSTRACT

The Information Technology is creating unforeseen opportunities through very rapid technological developments and the European-level competitiveness of Turkey strongly depends on her ability to develop the information society technologies to improve its economic growth. Therefore, Turkey has an urgent need for strengthening her IST RTD capabilities to catch up with EU countries. This requires investing on IST-RTD related human capital and activities in Turkey, and strengthening collaboration with IST-RTD efforts in EU Member States Countries. The ultimate goal of this proposal is to serve this purpose of integrating and strengthening the Turkish Research Area towards the Europe by reinforcement of the Turkey's Information Society Technologies (IST) Research and Technological Development (RTD) capabilities.

METU-ISTEC project aims to re-structure IST RTD research in the leading public technical university of Turkey in particular, and in the following FP6 IST RTD strategic objective areas: e-Business and e-Government, Semantic-based Knowledge Systems, Cognitive Systems and Open development platforms for software and services, and Embedded Systems in Turkey, in general. This restructuring will create an opportunity to combine applied and generic technology research and will help focus on the applied research areas which are European priority. For this purpose, METU-ISTEC project is mobilizing very significant amount of resources including many academicians (20 professors) as well as very many graduate students (39 PhD students and 32 MS students). METU-ISTEC will disseminate the results obtained to much wider audiences through workshops and conferences and will diffuse and exploit project results stimulating the community. The project has laid out plans to achieve networking with EU counterpart institutes. The actions planned to achieve this objective are study visits and visiting fellows to facilitate communication and collaboration between METU and other centers having compatible scientific interests. As a result, METU-ISTEC project will produce at least nine FP7 project proposals coordinated by METU Faculty.

Keywords: Integrating and Strengthening the Turkish Research Area towards the Europe, e-Business and e-Government, Semantic-based Knowledge Systems, Cognitive Systems and Open development platforms for software and services, Embedded Systems

PROJE ANA METNİ

GİRİŞ

Bilgi teknolojileri, çok hızlı bir şekilde ilerleyen teknolojik gelişmelerle birlikte öngörülemez fırsatlar yaratmaktadır. Buna paralel olarak, Türkiye'nin Avrupa seviyesindeki rekabet edebilirliği, Bilgi Toplumu Teknolojilerini geliştirmesine ve bunları ekonomik gelişiminde kullanabilme yeteneğine bağlıdır. Bu sebeple, Türkiye'nin Avrupa Ülkelerini yakalama adına, Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimin yeteneklerinin acil bir şekilde güçlendirmesine ihtiyacı vardır. Bu açılım da Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimiyle ilgili insan kapitaline ve aktivitelere yatırımı gerektirmektedir. Buna ilaveten, Avrupa Birliği üyesi ülkelerle Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi konusunda işbirliğinin güçlendirilmesi kaçınılmazdır. Bu proje önerisinin ana amacı; Avrupa doğrultusunda Türk Araştırma Alanlarının, Türkiye Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi yetenekleri doğrultusunda destekleyip, gerekli entegrasyon ve güçlendirmenin sağlanması olmuştur.

Proje önerisinin sahibi Orta Doğu Teknik Üniversitesi, Türkiye Endüstrisi, Mühendisliği, Üretimi ve yeni e-İş aktiviteleri için referans noktası oluşturmaktadır. Hızlı bir gelişim gösteren Türk Ekonomisi ve Endüstrisi; liderler, danışmanlar ve problem çözücü işbirliği için geleneksel olarak ODTÜ gibi ün sahibi enstitülere bakmaktadır. Fakat, ODTÜ'de Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi çabaları zayıf bir şekilde yapılanmıştır; bir taraftan ODTÜ en iyi öğrencileri çekip dünya çapında rekabet edebilecek bir eğitim sunarken, diğer taraftan Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi aktiviteleri, potansiyeline göre çok düşük bir seyir izlemektedir. Avrupa'daki diğer lider araştırma kurumlarının aksine ODTÜ'nün yetişmiş insan kapasitesi ve teknik altyapısı gerekli şekilde organize edilememiştir. Bazı aktif araştırma alanları endüstriyel uygulamalar yerine merak odaklı yönlenmektedir. ODTÜ'deki araştırma, teknik altyapı ve insan kaynaklarıyla beraber ilgili araştırma alanlarına yeniden yapılandırılmalıdır. Bu amaç için METU-ISTEC projesi ODTÜ'deki Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi çalışmalarını ve stratejik hedef alanlarını yeniden yapılandırmayı amaçlamıştır.

GENEL BİLGİLER

METU-ISTEC projesi bünyesinde bilgi toplumu teknolojilerinin farklı araştırma alanlarıyla ilgili dokuz adet çalışma grubu oluşturulmuştur. Tüm gruplar kendi araştırma alanlarıyla ilgili "gereç ve yöntem" bölümünde detaylandırılmış on üç adet görevi yerine getirmekle yükümlü olmuşlardır. Gruplar öncelikle kendi alanları ile ilgili var olan son teknolojik durumu, mevcut başarılı uygulamaları, akademik çalışmaları ve sonrasında da alanla ilgili standartları araştırıp incelemiştir. Yapılan araştırmalar ve incelemeler çerçevesinde detaylı bir araştırma planı çıkarılmıştır. Belirtilen tüm çalışmalar ayrı dokümanlar halinde ilk gelişme raporunda sunulmuştur.

İkinci altı aylık dönemi kapsayan süreçte, 7. Çerçeve Programı "Information and Communications Technology (ICT)" alanına ait stratejik alanlar açıklanmıştır. METU-ISTEC kapsamında planlanan ilk atölye çalışması olarak düzenlenen 09.01.2007 tarihli toplantıda çalışma grupları, 7.Çerçeve Programı stratejik hedeflerini incelemiş ve her grup öncelikli stratejik hedefini belirlemiştir. Bu adımdan sonra tüm gruplar, birbirinden bağımsız olarak aşağıda belirtilen stratejik hedeflere yönelik proje hazırlama çalışmalarına yoğunlaşmışlardır:

1. Grup: Objective 4.1 Digital libraries and technology-enhanced learning
2. Grup: Objective 5.1 Personal Health Systems for Monitoring and Point-of-Care diagnostics
3. Grup: Objective 3.7 Networked Embedded and Control Systems
4. Grup: Objective 1.2 Service and Software Architectures, Infrastructures and Engineering
5. Grup: Objective 4.2 Intelligent Content and Semantics
6. Grup: Objective 1.2 Service and Software Architectures, Infrastructures and Engineering
7. Grup: Objective 2.1 Cognitive systems, interaction, and robotics
8. Grup: Objective 4.2 Intelligent Content and Semantics
9. Grup: Objective 1.2 Service and Software Architectures, Infrastructures and Engineering

GEREÇ VE YÖNTEM

Proje süresince başvuru aşamasında belirlenmiş ve kabul edilmiş olan çalışma takvimi uygulanmıştır. Bazı iş paketlerinin zamanlamalarında değişiklikler yaşanmıştır. Bu değişikliklerin çoğu METU-ISTEC projesi takvimi hazırlandığında henüz 7. Çerçeve Program'ı takviminin belirlenmemiş olmasından kaynaklanmıştır. Örneğin, tüm çalışma gruplarına ait görev olan "(2-10.13): Topluluk Çerçeve Programlarına konuyla ilgili proje teklifinin hazırlanması" görevi, 7. Çerçeve, 1. Çağrı kapanış tarihinin daha önce olması sebebiyle daha erken bitirilmiştir. Ek olarak çalışma grupları METU-ISTEC projesi süresince teklif hazırlama çalışmalarına devam etmişlerdir. Bulgular bölümünde listelenmiş görevler kapsamında yapılan çalışmalar detaylı biçimde incelenmiştir.

İzlenmiş olan proje aşama ve zamanlama planı aşağıdaki gibidir:

Başlıca Aşamalar	Ayrıntılı Bilgi	Zamanlama
Görev 1.1: İdari Yönetim	Bu iş kapsamında proje idari olarak yönetilecektir. Bu kapsamda 6. ayın sonunda 1. gelişme raporu, 12. ayın sonunda ise 2. gelişme raporu hazırlanacaktır. Ayrıca, grupların harcamaları incelenecek, ve raporlanacaktır.	1-18
Görev 1.2: Teknik Yönetim	Bu iş kapsamında proje teknik açıdan yönetilecektir. Ayrıca proje bitiminde "Proje Bitirme Raporu" bu iş çerçevesinde hazırlanacaktır. Bu raporda proje sonucu elde edilen bilgi birikimi, tecrübe ve başarıları kapsayacaktır. Ayrıca, hedef çerçevesinde karşılaşılan olumsuzluklar irdelenip, sonuçlar çıkarılacaktır.	1-18
Görev 1.3: Stratejik Yönetim	Bu iş kapsamında proje stratejik açıdan yönetilecektir.	1-18

	Bu iş kapsamında projede verimliliği arttırmak ve iletişimi kolaylaştırmak için bir web portalı oluşturulacaktır.	
Görev (2-10).1 Konuyla ilgili var olan son durumun Avrupada önceki topluluk projelerini ve en iyi uygulamaları da kapsayacak şekilde araştırılması	Bu iş kapsamında seçilen alanla ilgili var olan son teknolojik durum ve eğilimler incelenmelidir. Durumda gerekirse eksiklikler ve aksaklıklar tespit edilmeli, ve kullanımdaki teknolojiler incelenerek proje için gerekli olanların öğrenilmesi yoluna gidilmelidir. Ayrıca bu alanda daha önce yapılmış ya da yapılmakta olan Avrupa projeleri incelenmeli, Avrupadaki en iyi uygulamalar tespti edilmelidir. Yapılan araştırma bir raporda toplanıp araştırma planına eklenmelidir.	1-3
Görev (2-10).2 Konuyla ilgili var olan son durumun akademik alandaki son gelişmeleri de kapsayacak şekilde araştırılması	Bu iş kapsamında seçilen alanla ilgili var olan son teknolojik durum ve eğilimler incelenmelidir. Durumda gerekirse eksiklikler ve aksaklıklar tespit edilmeli, ve kullanımdaki teknolojiler incelenerek proje için gerekli olanların öğrenilmesi yoluna gidilmelidir. Alanla ilgili son akademik gelişmeler de incelenmeli ve konuyla ilgili bir literatür taraması da yapılmalıdır. Yapılan araştırma bir raporda toplanıp araştırma planına eklenmelidir.	1-3
Görev (2-10).3 Konuyla ilgili standartların araştırılması	Kullanımdaki standartlar incelenmeli, gelecekte yürürlüğe girmesi beklenen standartlar araştırılmalı ve yazılacak proje teklifinde bu standartlardan bahsedilmeli ve istifade edilmelidir. Bu kapsamda standardizasyon kurumlarının çalışmaları da incelenmeli, ve bu standartlara katkıda bulunmanın olası yolları irdelenmelidir. Bu çalışmalar tamamlandığında sonuçlar bir raporda toplanmalıdır.	1-3
Görev (2-10).4 Seçilen alanla ilgili bir araştırma ve yenilik stratejisi planının çıkarılması	Yapılan araştırma ve incelemeler çerçevesinde bundan sonra izlenecek yol tayin edilmeli ve detaylı bir araştırma planı olarak raporlanmalıdır. Bu raporda ayrıca alanla ilgili bir doktora tez konusu belirlenmelidir.	1-4
Görev (2-10).5 Konuyla ilgili yerel bir proje teklifinin hazırlanması	Bu iş Avrupa Komisyonuna verilecek proje teklifine ek olarak, benzer bir projenin yerel olarak da üretilmesini kapsamaktadır. Bu kapsamda TÜBİTAK ya da benzer bir kurumla işbirliği yapmaya veya bu kurumdan destek almaya yönelik bir proje teklifi üretilmelidir.	1-6
Görev (2-10).6 Avrupa ülkelerinden işbirliği yapılacak olası ortakların tespit edilmesi	Bu iş Avrupa ülkelerinden proje teklifinde projenin ortağı olacak ve projede etkin görev alacak ortakların belirlenmesini kapsamaktadır. Bu kapsamda projede yapılacak işler tespit edilmeli ve bu işlere uygun ortaklar aranmalıdır. Olası ortaklarla iletişime geçilip projede yer almak isteyip istemeyecekleri	1-8

	tartışılmalıdır.	
Görev (2-10).7 Seçilen konuda çalışmak üzere genç araştırmacıların göreve getirilmesi	Yapılan araştırmalarda görev almak üzere Yüksek Lisans ve/veya Doktora öğrencilerinin bulunması ve proje kapsamında görevlendirilmesi beklenmektedir. Böylece, projede görev dağılımıyla birlikte, ayrıca genç araştırmacıların da çerçeve programlarında deneyim kazanması mümkün olacaktır. Projenin ilk ayında seçilen alanda araştırma yapmaya uygun en iyi nitelikteki bir doktora öğrencisi belirlenmelidir. Bu öğrenci tezini projeyle birlikte seçilen konuyla yürütmeli ve proje sonucunda tezini teslim etmelidir.	1
Görev (2-10).8 Türk şirketlerini proje tekliflerine almak	Hazırlanacak olan proje tekliflerine yönelik bir Türk şirketinin tespit edilip, projede bu şirketin yetenekleri kapsamında projede uygun bir görevin verilmesi beklenmektedir. Böylece Türkiye'nin Çerçeve Programlarına katılımı arttırılmış olacaktır.	1-4
Görev (2-10).9 Topluluk Çerçeve Programlarına konuyla ilgili proje teklifinin ilk taslağının hazırlanması	Bu iş kapsamında Topluluk Çerçeve Programlarına konuyla ilgili sunulacak proje teklifinin ilk taslağı hazırlanacaktır. Bu taslak 1. atölyede sunulacaktır. Bu taslak proje teklifine yönelik ilk adım olacaktır.	1-10
Görev (2-10).10 Topluluk Çerçeve Programlarına konuyla ilgili proje teklifinin ikinci taslağının hazırlanması	Bu iş kapsamında Topluluk Çerçeve Programlarına konuyla ilgili sunulacak proje teklifinin ikinci taslağı hazırlanacaktır. Bu taslak 2. atölyede sunulacaktır. Bu taslakla birlikte sunulacak proje teklifi biraz daha şekillenmiş olacaktır.	11-14
Görev (2-10).11 Topluluk Çerçeve Programlarına konuyla ilgili proje teklifinin hazırlanması	Projenin en önemli amaçlarından biri olan bu iş çerçevesinde Avrupa Komisyonu Çerçeve Programlarına yönelik proje teklifi hazırlanacak ve sonrasında sunulacaktır. Bu sayede Türkiye'nin bu programlara katılımı arttırılmış olacak ve araştırma desteği almanın yanısıra, Avrupa ülkeleri ile işbirliği yapılacak, tecrübe ve bilgi paylaşımından yararlanılacaktır. Bu iş sonucunda ortaya çıkacak son proje teklifi 3. atölyede sunulacaktır.	15-18
Görev (2-10).12 Avrupa Komisyonu Uzman veritabanına özgeçmişlerin girilmesi	Bu iş çerçevesinde Avrupa Komisyonu projelerini değerlendirmek amacıyla Avrupa Komisyonunun Uzman veritabanına tüm öğretim görevlilerinin özgeçmişlerini girmesi beklenmektedir. Bu işlem Avrupa Komisyonunun http://emmfp6.cordis.lu sayfasından yapılabilmektedir. Özgeçmişler veritabanına girildikten sonra, Avrupa Komisyonu ihtiyaç duyduğunda bu veritabanından uygun uzmanlarla irtibata geçerek, elindeki proje tekliflerinin değerlendirilmesini istemektedir. Böyle bir tecrübe ile Çerçeve Programlarını daha iyi anlamak mümkün olacaktır.	1
Görev (2-10).13 Başka proje tekliflerinde olası ortaklıklar için	Gerek TÜBİTAK'ın info@fp6.org.tr ile ist@tubitak.gov.tr ve fp6_ist@tubitak.gov.tr listelerine	1-18

idealist duyurularını takip etmek	yaptığı duyurular aracılığı ile, gerek TÜBİTAK'ın TR-ACCESS (http://traccess.tubitak.gov.tr/) projesinin haberlistesi aracılığı ile, gerekse ideal-ist'te (http://www.ideal-ist.net/) ortak arama duyurularının takip edilmesi. Bu iş çerçevesinde ortak arama duyurularının incelenip, başka proje tekliflerinde olası ortaklıkların elde edilmesi amacıyla bu projelerle irtibata geçilmesi beklenmektedir. Bu projelerin bu iş paketinde tanımlı proje ve alanlarda olması gerekmemektedir.	

Tablo 1 – Proje Takvimi

BULGULAR

Bu bölümde öncelikle proje yönetimini kapsayan ilk üç görev altında yapılan çalışmalarla ilgili bilgi verilmiştir. Ardından, çalışma takviminde yer alan ilk iki gelişme raporunda detaylı olarak anlatılmış olan diğer görevlerden kısaca bahsedilmektedir. İkinci gelişme raporu değerlendirme sonucunda, hakemler 7. Çerçeve Programı'na sunulmuş olan proje önerileri hakkında ve çekirdek ekip yetiştirme konusunda yapılan çalışmalar hakkında bilgi istemişlerdir. Bu bölümde bu konulara ağırlık verilmiştir.

Görev 1.1- İdari Yönetim:

Bu görev tabloda belirtildiği gibi proje boyunca sürdürülmüştür. Proje süresince iki gelişme raporu hakem görüşüne sunulmuş ve onay almıştır. Proje süresince çalışma gruplarının yapmış oldukları harcamalar incelenmiş ve her dönem raporlanmıştır.

Görev 1.2- Teknik Yönetim:

Bu iş kapsamında proje teknik açıdan yönetilmiştir. İlk altı aylık dönem incelendiğinde; projenin başarıyla sürdürülmesi amacına yönelik olarak, çalışma gruplarıyla iki adet toplantı yapılmıştır. Buna ek olarak Orta Doğu Teknik Üniversitesi'nde yüze yakın firmanın da katılım yaptığı 6.çerçeve programının son çağrısına ve 7. çerçeve programına yönelik bilgilendirme günü düzenlenmiştir. Bu bilgilendirme günüyle ilgili detaylı bilgi ilk gelişme raporunda sunulmuştur.

İkinci altı aylık dönem incelendiğinde; projenin başarıyla sürdürülmesi amacına yönelik olarak, çalışma gruplarıyla 09.01.2007 ve 28.03.2007 tarihlerinde iki adet toplantı yapılmıştır. METU-ISTEC kapsamında planlanan ilk atölye çalışması olarak düzenlenen 09.01.2007 tarihli toplantıda çalışma grupları, 7.Çerçeve Programı stratejik hedeflerini incelemiş ve her grup öncelikli stratejik hedefini belirlemiştir. Bu adımdan sonra tüm gruplar, birbirinden bağımsız olarak seçilen stratejik hedeflere yönelik proje hazırlama çalışmalarına yoğunlaşmıştır.

Yukarıda belirtilen toplantılara ek olarak, çalışma grupları bağımsız olarak çeşitli bilgilendirme günleri, ortak arama günleri ve konferanslara katılmışlardır. Bu toplantılar hakkında detaylı bilgi ikinci gelişme raporunda sunulmuştur.

Üçüncü ve son altı aylık dönemde proje bitirme raporu hazırlanmıştır.

Görev 1.3- Stratejik Yönetim:

Bu görev tabloda belirtildiği gibi proje boyunca sürdürülmüştür. Bu iş kapsamında projede verimliliği arttırmak ve iletişimi kolaylaştırmak için bir web portalı oluşturulmuştur ve proje başlangıç tarihinden bu yana hizmet vermektedir. Bu portala http://www.srdc.metu.edu.tr/webpage/projects/istec/ adresinden ulaşılabilir.

Görev (2-10).(1-13) Kapsamında Yapılan Çalışmalar:

Proje bünyesinde dokuz adet çalışma grubu vardır. Tüm gruplar kendi araştırma alanlarıyla ilgili on üç adet görevi yerine getirmekle yükümlüdür. Gruplar öncelikle kendi alanları ile ilgili var olan son teknolojik durumu, mevcut başarılı uygulamaları, akademik çalışmaları ve sonrasında da alanla ilgili standartları araştırıp incelemiştir. Yapılan araştırmalar ve incelemeler çerçevesinde detaylı bir araştırma planı çıkarılmıştır. Belirtilen tüm çalışmalar ayrı dokümanlar halinde ilk gelişme raporunda sunulmuştur.

Yukarıda belirtilen teknik çalışmalara paralel olarak tüm gruplar, kendi alanlarında, Avrupa ülkelerinden proje ortağı olacak ve projede etkin görev alacak ortakların belirlenmesi konusunda çalışmalar yapmıştır. Bu kapsamda proje ihtiyaçları belirlenmiş ve bu işlere uygun ortaklar aranmıştır. METU-ISTEC projesinin bir diğer hedefi ise Türk şirketlerinin Avrupa Komisyonu projelerine katılımını arttırmak olmuştur. Bu kapsamda, hazırlanmış olan proje tekliflerine yönelik Türk şirketleri de tespit edilmiştir. Yerli ve yabancı ortak tespiti konusunda yapılan tüm çalışmalar ilk gelişme raporunda sunulmuştur. Ayrıca çalışma gruplarının önermiş olduğu aşağıdaki on proje Ideal-IST'te (http://www.ideal-ist.net/) ortak arama duyuruları olarak yayınlanmıştır:

Kod	Proje Adı	Çağrı No	Hedef Alan
PS-TR-618	iLTL: Intelligent Cooperation Platform for Intermodal Freight Logistics	ICT Call 1 (FP7-2007-ICT-1)	6.1 ICT for the intelligent vehicles and mobility services
PS-TR-402	I-MAPS: Identity Management, Privacy and Security for Networked Applications	ICT Call 1 (FP7-2007-ICT-1)	1.4 Secure, dependable and trusted infrastructures
PS-TR-336	iHygia: Intelligent Proactive Personal Care and Monitoring Platform	ICT Call 1 (FP7-2007-ICT-1)	5.1 Personal health systems for monitoring and point-of-care diagnostics
PS-TR-190	I-SURF: A Dynamic Interoperability Service Utility Middleware for Collaborative Supply Chain Planning across Multiple Industry Domains supported by RFID devices	ICT Call 1 (FP7-2007-ICT-1)	1.3 ICT in support of the networked enterprise
PS-TR-955	3A-Net: resource Aware Application Adaptability in pervasive mobile Networks	ICT Call 1 (FP7-2007-ICT-1)	1.2 Service and software architectures, infrastructures and engineering
PS-TR-369	PerMIMS – A Framework for Personalised MultiMedia Information Management Systems	ICT Call 1 (FP7-2007-ICT-1)	4.2 Intelligent content and semantics

PS-TR-474	QASTEL-A Framework for Ontology-based Question Answering Systems used for Technology Enhanced Learning	ICT Call 1 (FP7-2007-ICT-1)	4.1 Digital libraries and technology-enhanced learning
PS-TR-240	A Framework for Intelligent Content Creation (FRICOC)	ICT Call 1 (FP7-2007-ICT-1)	4.2 Intelligent content and semantics
PS-TR-1061	Development of a Service Oriented Middleware for Real-Time Applications on Pervasive Networks-SORAPNet	ICT Call 1 (FP7-2007-ICT-1)	1.2 Service and software architectures, infrastructures and engineering
PS-TR-142	MASHUP: MigrAtion to Service Harmonization compUting Platform	ICT Call 1 (FP7-2007-ICT-1)	1.2 Service and software architectures, infrastructures and engineering
PS-TR-1985	Critical Infrastructure Crisis and Protection Simulation	FP7-ICT-SEC-2007-1	1.0-02 Modelling and simulation for training

Tablo 2 – Yayınlanmış Ideal-IST Duyuruları

METU-ISTEC proje grupları, koordinatörlüğünü ODTÜ'nün üstlendiği projelere ek olarak, ortak olarak da Avrupa Komisyonu projelerine katılmayı hedeflemiştir. Bu kapsamda Ideal-IST duyuruları takip edilmiş ve uygun proje önerilerine başvurulmuştur. Her grubun bu konuda yapmış olduğu detaylı çalışma ikinci gelişme raporunda belirtilmiştir.

İkinci altı aylık dönem, proje hazırlama çalışmalarıyla devam etmiştir. Bu konuyla ilgili görevlerin kapsamında proje öneri taslakları ikinci gelişme raporunda sunulmuştur. İkinci gelişme raporu hakem görüşünden yola çıkarak, 7. Çerçeve Programı'na yollanmış olan projelerden bu bölümde bahsedilecektir.

7. Çerçeve Programı "Information and Communications Technology (ICT)" alanına ilk çağrıda Türkiye'den toplam 115 proje başvurusu olmuştur. Bu projelerin 29'u Türkiye koordinatörlüğünde yollanmıştır. METU-ISTEC proje grupları toplam 17 proje önerisi yollamışlardır. Bu projelerden 8'i Türkiye koordinatörlüğünde yollanmıştır. Türkiye'nin yer aldığı projelerin 14.78%'i, Türkiye koordinatörlüğünde yollanmış projelerin 27.59%'u METU-ISTEC kapsamında hazırlanmıştır.

Aynı çağrıda Türk şirketlerin yer aldığı proje başvurularının 30'u eşik değeri geçmiştir. Eşik değeri geçen projelerden 7'si Türkiye koordinatörlüğünde yollanmıştır. Bu projelerden 6'sında METU-ISTEC grupları yer almış, 2'sinde koordinatör olarak başvurmuştur. Türkiye'nin yer aldığı eşik değer üstü projelerin 20%'si, Türkiye koordinatörlüğünde yollanmış eşik değer üstü projelerin 28.57%'si METU-ISTEC kapsamında hazırlanmıştır.

METU-ISTEC kapsamında hazırlanan projelerde aşağıdaki 12 yerli şirketin katılımı sağlanmıştır:

- Başarı Mobile Bilişim Ürünleri ve Hizmetleri A.Ş.
- Mobilera
- Matriks Data A.Ş.
- Siemens EC Kurumsal İletişim Hiz. San. ve Tic. A.S.
- SRDC Yazılım Araştırma ve Geliştirme ve Danışmanlık Tic. Ltd. Sti.
- VTEK Bilisim ve İletisim Teknolojileri San. Ve Tic. Ltd. Sti.
- Innova IT Solutions

- Turkcell
- Cybersoft C/S Information Technologies Ltd. Co.
- Alcatel-Lucent
- Türkiye Elektrik İletim A.Ş.
- Botaş

İlk çağrı kapsamında koordinatörlüğünü Orta Doğu Teknik Üniversitesi'nin üstlendiği iSURF projesi ve Orta Doğu Teknik Üniversitesi'nin ortak olarak yer aldığı ROSSI projesi destek almıştır.

METU-ISTEC gruplarının yer aldığı proje önerilerinin toplam bütçesi 51.490.540 Euro olup, 9.224.016 Euro'luk bütçe METU-ISTEC gruplarına aittir. Destek alan projelerin toplam bütçesi 5.929.337 Euro olup, Türkiye'ye ait toplam bütçe 1.077.001 Euro'dur.

METU-ISTEC gruplarının hazırlamış oldukları proje önerilerine ait özet bilgiler aşağıda sunulmuştur.

1. 1. Gruba Ait FP7 Proje Önerileri

1. PerMIMS

Proje Adı:	PerMIMS – A Framework for Personalised MultiMedia Information Management Systems	
Proje Akronimi:	PerMIMS	
Stratejik Alan:	FP7-1-4.2 Intelligent Content and Semantics	
Proje Türü:	STREP	
Projedeki Rol:	Koordinator	
Bütçe Bilgisi:	**Toplam Bütçe:**	3,365,963 Euro
	ODTÜ Payı:	703,600 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
Universidad Carlos III de Madrid	İspanya
Westfalische Wilhelms – Universitat Münster	Almanya
Center for Research and Technology Hellas	Yunanistan
Orca Interactive Ltd.	İsrail
Basarı Mobile Bilişim Ürünleri ve Hizmetleri A.Ş.	Türkiye
Agencia EFE, S.A.	İspanya

Proje Özeti:

Günümüzde çoklu ortam verisi ticaret, güvenlik, eğitim ve eğlence gibi çok çeşitli alanlarda yer almaktadır. Özellikle cep telefonları, dijital kameralar ve kayıt cihazlarının artması ile

oluşturulan kişisel çoklu ortam verisinde çok fazla artış oluşmaktadır. Bireylerin sahip oldukları bu çok büyük veri yığınını yönetmek önemli bir sorun haline gelmiştir. PerMIMS projesi kişisel çoklu ortam verisinin tanımlanması, organizasyonu, kişiselliştirilmesi ve dağıtımını sağlayan bir bilgi yönetim çatısının oluşturulmasını hedef alır.

PerMIMS projesinde oluşturulacak çatı sayesinde kullanıcıların çoklu ortam verisini daha anlamsal şekilde saklama, arama ve gösterme imkanları olacaktır. Bunların anlamsal bir şekilde yapılabilmesi için veriyi tanımlayan yardımcı veriye ihtiyaç vardır. Yardımcı veri çoklu ortam verisinden de elde edilebileceği gibi verinin bulunduğu ortamdan da çıkarılabilir. Verinin bulunduğu Internet adresindeki tanımlayıcı dizinsel bilgiler, TV yayınından kaydedilen bir film için yayına gömülü gelen dizinsel veriler (Elektronik Program Rehberi EPG) yardımcı verilerin oluşturulması için bir kaynak olabilir. Proje çerçevesinde Anlambilimsel Ağ (Semantic Web) teknolojileride kullanılacaktır. Çoklu ortam verisi üzerindeki aramların anlamsal seviyesini arttırmak için ontolojilerde kullanılacaktır. Genel olaylar ve varlıklar hakkında bilgi veren üst seviye ontolojilerin yanında daha alt seviyede çoklu ortam verisindeki olayla ilgili özel ontolojilerin de (gezi veya spor ontolojileri gibi) oluşturulması ve kullanılması proje kapsamında öngörülmüştür. Proje hedeflerinde konuyla ilgili en son gelişmeleri içermek ve bu seviyeyi araştırma faaliyetleriyle aşmak bulunmaktadır.

(Ek 1.1: Proje Değerlendirme Formu)

2. MEDIBLE

Proje Adı:	Multimedia Enhanced Digital Books for Interactive Learning	
Proje Akronimi:	MEDIBLE	
Stratejik Alan:	FP7-1-4.2 Intelligent Content and Semantics	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	3,747,348 Euro
	ODTÜ Payı:	448,000 Euro

Ortak Listesi:

Ortak Adı	Ülke
Systema Technologies S.A.	Yunanistan
Univeristy of Surrey	İngiltere
eZ Systems AS	Norveç
Scopus Video Networks	İsrail
Technisce University Delft	Hollanda
Middle East Technical University	Türkiye
Universitat Augsburg	Almanya
BIT Media GmbH & Co KG	Avusturya
Jyvaskylan yliopisto	Finlandiya

Proje Özeti:

Medible projesinin amacı dijital kitapların düz metin içeriklerini, diğer kitap ve kütüphanelerdeki interaktif ve kişisel çoklu ortam içeriği kullanarak, otomatik olarak iyileştirme amacını gütmektedir. Diğer bir deyişle dijital kitaplara yeni boyutlar katmaktır. Böylece her kitap dijital dünyada öğrenmeye yeni bir pencere olacaktır. Bu yaklaşım sayesinde dijital kaynaktan birşeyler öğrenmeye çalışmak daha kolay ve hızlı hale gelecektir. Kaynakların başka kaynaklardaki düz metin içeriğin yanı sıra video, ses ve resim gibi kaynaklarla da zenginleştirilmesi işin görsel boyuta taşınmasını da sağlayacaktır.

Bu bağlamda kullanılan kaynakların içerikleri belirlenecek ve seçilen bölümler anlamsal çoklu ortam verisi ve diğer kaynaklara dinamik bağlantılar atılarak metinler genişletilecektir. Bu işleme kullanıcıların da katılımı mümkün kılınılacaktır. Kullanıcıların profillerine ve geri beslemelerine göre metinler farklı şekilde de zenginleştirilebilecektir. Bu proje içerik analizi, konuya özel veri toplama ve bilgi yönetimi kavramlarını pedagoji uzmanları danışmanlığı altında öğrenme kavramında birleştirecektir.

Burada hedeflerden biri de düz metin içerikten bilgisayarların kullanabileceği kaynaklara erişimi sağlayacak bilginin çıkarılması ve bunun öğrenen kişilere uygun bir düzende sunulmasıdır. İyileştirilmiş dijital kitaplar var olan gösterim teknolojilerine ve okuma araçlarına uyumlu hale getirilecektir. Ayrıca bu projenin sunacağı yeni servis müzeler, yayımcılar ve yayıncılar, kütüphaneler ve eğitim enstitüleri tarafından da kullanılabilecektir.

(Ek 1.2: Proje Değerlendirme Formu)

2. 3. Gruba Ait FP7 Proje Önerileri

1. SORAPNet

Proje Adı:	Development of a Service Oriented Middleware for Real-Time Applications on Pervasive Networks-SORAPNet	
Proje Akronimi:	SORAPNet	
Stratejik Alan:	FP7-1-1.2 Service and Software Architectures, Infrastructures and Engineering	
Proje Türü:	STREP	
Projedeki Rol:	Koordinator	
Bütçe Bilgisi:	**Toplam Bütçe:**	2,501,700 Euro
	ODTÜ Payı:	468,900 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
EDISOFT	Portekiz
National Institute For R&D in Informatics	Romanya
TIGA Technologies SARL	Fransa

Universidad de Valladolid	İspanya
Infosistema	Portekiz
Mobilera	Türkiye

Proje Özeti:

Son yıllarda iletişim teknolojilerindeki gelişmeler bu tür teknolojilerin kullanımlarının dramatik olarak yaygınlaşacağını işaret etmektedir. Bu gelişmelerin en çok gerçek zaman uygulamalar üzerinde dramatik etki bırakacağını düşünüyoruz. Bu kapsamda daha önce tahayyül edilmeyen uygulamalar artık gerçek olabilecektir. Örneğin, doğal afet kapsamında olan deprem, sel, terör, ve bulaşıcı hastalıkların gerçek zamanda anında gözlenmesi ve denetlenmesi artık gerçekleşebilmelidir. Diğer gerçek zaman uygulamalar boyutunda olan misyon özellikli uygulamalardan tahıl ekimi, çeşitli üretim fabrikaları veya kritik sosyal olayların takipleri ve/veya denetimleri de bu gelişmelerden yararlanabilecek duruma gelinmiştir. Buna rağmen bu tür uygulamalara hizmet eden etkin sistemlerin ya olmadığı yada sınırlı olduğu görülmektedir. Olanlarda da servis kalitesinin günün teknolojisinin gerisinde kaldığı görülmektedir.

Doğal olarak, iletişim altyapısının yaygınlığı, uyarlanabilirliği ve dayanaklığı yukarıda geçen gerçek zaman uygulamaların vazgeçilmez nitelikleridir. Bu altyapıda kablo, radyo ve uydu iletişim altyapısı ile sabit veya hareketli iş istasyonlarının entegrasyonu söz konusudur. Bu araştırmanın amacı böylesi bir ortamda gerçek zaman özellikli uygulamalarda servis kalitesini artırıcı iletişim ve ilişki kurallarıyla birlikte ara-yüzlerin de entegre bir şekilde geliştirilmesidir.

(Ek 3.1: Proje Değerlendirme Formu)

3. 4. Gruba Ait FP7 Proje Önerileri

1. 3A-Net

Proje Adı:	3A-Net: resource Aware Application Adaptability in pervasive mobile Networks	
Proje Akronimi:	3A-Net	
Stratejik Alan:	FP7-1-1.2 Service and Software Architectures, Infrastructures and Engineering	
Proje Türü:	STREP	
Projedeki Rol:	Koordinator	
Bütçe Bilgisi:	**Toplam Bütçe:**	2,835,420 Euro
	ODTÜ Payı:	656,220 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
Aratos Technologies S.A	Yunanistan
University of Mannheim	Almanya

Matriks Data A.Ş.	Türkiye
University of Antwerp	Belçika
Onion S.p.A.	İtalya

Proje Özeti:

İletişim ve bilişim teknolojilerindeki hızlı değişim göz önüne alındığında, yakın gelecekte, günümüzde sahip olduğumuzdan çok daha farklı hesaplama ve bilgi biçimlerine bir dönüşüm olacağı öngörülmektedir. Birçok araştırmacı tarafından ifade edildiği üzere, hesaplama ve bilginin, pil, elektrik prizi, telefon prizi ve hatta havadaki oksijen gibi, her yerde ve ücretsiz erişebilir olması beklenmektedir. Yaygın mobil aygıtlarının, ağların ve yazılım teknolojilerinin, kullanıcıların evde, işte yada hareket halinde ihtiyaç duydukları çeşitli servisleri sunarak, seçeneklerimizi çarpıcı bir şekilde arttıracağı tahmin edilmektedir.

Ancak, günümüzde gerçek anlamda yaygın ve her zaman her yerde beliren hesaplamanın ortaya çıkışı, bu alanlardaki mobil platformlar üzerinde çalışan uygulamaların tasarım ve geliştirilmesi için yeterli desteğin eksikliği sebebiyle engellenmektedir. Yeni kullanıcıların yeni ihtiyaçlarına cevap verebilmek için uygulama geliştirenlerin mobil etmenlerden gelen bilgileri birleştirebilmelerine olanak sağlayan etkili sistemlere önemle ihtiyaç duyulmaktadır.

Mobil platformlar, her biri uygun yazılım sistemleri tarafından desteklenen kablosuz yerel ağlar (LAN), Wi-Fi ve WiMAX ağları, hücresel sistemler ve İnternet'le birlikte kablosuz sensör ağları olarak düşünülebilir. Tabandaki yaygın altyapıyı azami derecede kullanabilmek için, yüksek derecede yaygın uygulamalara ihtiyaç duyulmaktadır. Ancak, mobil platformlar yaygın altyapı olmalarından dolayı sürekli ve düzenli bir kaynak kümesi sunamamaktadırlar. Örneğin, hareket halindeyken, bir kablosuz ağ ulaşılamaz olabilir, veya bir mobil aygıtta farklı bir çalışma ortamıyla karşılaşabiliriz yada işbirliği içerisinde yürütülen bir uygulamanın bir ortağı ulaşılamaz olabilir yada o anda ulaşılamaz hale gelebilir. Alternatif olarak, başka bir kablosuz ağ yada WiMAX ağının kapsam alanına girebiliriz, ya da ortağımız ulaşılabilir konuma gelebilir. Bu sebeplerden böyle bir mobil platformda çalışacak olan yüksek derecede yaygın olan her uygulama, bu tür değişikliklere uyum sağlayabilir olmalıdır, başka bir deyişle, her türlü kaynağın kullanılabilirliğine uyum sağlamalıdır.

Bu projenin amacı, yukarıda bahsedilen eksiklikleri giderebilmek için, uygulama ve ihtiyaç duyduğu kaynaklar arasında bir orta katman geliştirilmesidir. Bu yaklaşımda, içinde bulunulan duruma bağlı olarak uygulamanın belirli özellikleri ve işlevleri aktif veya pasif hale getirilmekte, veya farklı seçeneklere yönlendirilmektedir. Proje kapsamında geliştirilmesi planlanan yapı, afet yönetimi alanında hazırlanacak bir pilot uygulama aracılığı ile sınanacaktır.

(Ek 4.1: Proje Değerlendirme Formu)

4. 5. Gruba Ait FP7 Proje Önerileri

1. QASTEL

Proje Adı:	QASTEL-A Framework for Ontology-based Question Answering Systems

	used for Technology Enhanced Learning	
Proje Akronimi:	QASTEL	
Stratejik Alan:	FP7-1-4.1 Digital libraries and technology-enhanced learning	
Proje Türü:	STREP	
Projedeki Rol:	Koordinator	
Bütçe Bilgisi:	**Toplam Bütçe:**	4.093.720 Euro
	ODTÜ Payı:	615.600 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
Siemens EC Kurumsal İletişim Hiz. San. ve Tic. A.S.	Türkiye
Rheinisch-Westfalische Technische Hochschule Aachen	Almanya
FUNDACION ROBOTIKER	İspanya
Ort Braude College of Engineering	İsrail
Institute for Applied Informatics (InfAI) e.v.	Almanya
GIUNTI Labs S.r.l.	İtalya
Language and Computing	Belçika

Proje Özeti:

Günümüz toplumlarında bireyler, artan bilgi ve içerik yükü ile bilgi ve yeteneklerin geliştirilmesi konusunda artan taleplerle karşı karşıyadır. Bu talepleri karşılamak için de öğrenme sürecini kolaylaştıracak sistem ve araçlara ihtiyaç duyulmaktadır. Bu ihtiyaçlar doğrultusunda teknolojiyle desteklenen öğrenme konusunda daha duyarlı ortamların geliştirilmesi hedefiyle, Avrupa Birliği 7. Çerçeve Programının "Bilgi ve İletişim Teknolojileri" temasında "Sayısal Kütüphaneler ve Teknolojiyle Desteklenen Öğrenme" konusunda birinci çağrıya küçük veya orta ölçekli araştırma projesi (STREP) teklifinde bulunulmuştur. Proje başlığı QASTEL kısa adıyla, "Teknolojiyle Desteklenen Öğrenmede Kullanılan Ontoloji Tabanlı Soru Cevaplama Sistemleri İçin Bir Çatı" olarak belirlenmiştir.

Teknolojiyle desteklenen öğrenme, öğrenciler ve eğitmenler arasındaki bilgi paylaşımı ve iletişimin sağlanmasıyla ilgilenmektedir. Bununla birlikte ders materyallerinin etkili kullanımı da teknolojiyle desteklenen öğrenmenin hedefleri arasındadır. Bu hedef doğrultusunda proje teklifinde, ontoloji tabanlı bir soru cevaplama sisteminin bir öğrenme yönetim sistemiyle entegrasyonunu sağlayacak bir çatı önerilmiştir. Önerilen mekanizmada, yalnızca öğrenciler sorularına cevap almayacak, aynı zamanda eğitmenler de öğrencilerin öğrenme sürecinde karşılaştıkları problemler hakkında bilgi sahibi olacaklardır. Kullanıcılar sisteme doğal dilde sorular soracak ve uygun cevapları da sistemden alacaklardır. Bunlara ek olarak sistem, kullanıcı davranışlarını izleyerek öğrenecek ve kendini zaman içinde bu davranışlara göre yapılandıracaktır. Bu çatının kullanımı, bilgi kazancı için hızlı ve verimli yollar sağlayacak, kullanıcı yetenek ve becerilerini artıracaktır. Sistemin temel aktiviteleri, pedagojik yaklaşımları, etkileşimliliğin kullanımı ve bağlam farkındalığını içerecektir.

Oluşturulacak çatının geliştirme sürecinde, makine tarafından anlaşılabilir içerik gösterimini ve soru cevaplama sistemlerinin teknolojiyle desteklenen öğrenme sürecine kolay entagrasyonunu sağlayacak standartların geliştirilmesi de hedeflenmektedir. Bu bağlamda akla gelecek olan içerik üretim safhasının sonuçlandırılması ise kullanıcıya, oluşturulan standartlar doğrultusunda içerik yaratılması konusunda yardım edecek özel araçların kullanımı ile gerçekleştirilecektir.

(Ek 5.1: Proje Değerlendirme Formu)

2. ReMPos

Proje Adı:	Regional Memory Portal	
Proje Akronimi:	ReMPos	
Stratejik Alan:	FP7-1-4.2 Intelligent Content and Semantics	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	2,468,800 Euro
	ODTÜ Payı:	272,000 Euro

Ortak Listesi:

Ortak Adı	Ülke
Turku University of Applied Sciences	Finlandiya
Mihailo Pupin Institute	Karadağ
OSJC The Centre of Information in the Sphere of Culture (Centre PIC)	Rusya Federasyonu
University of Reading	İngiltere
Industrial Research Institute for Automation and Measurements PIAP	Polonya
METU	Türkiye
University Leipzig	Almanya

Proje Özeti:

Proje kütüphane, müze ve arşivlerde yer alan resmi kültürel miras içeriğinin saklandığı bölgesel hafıza portallerinin oluşturmasını amaçlar. Genel içeriğin yanında vatandaşlardan gelebilecek daha kişisel ve özel içerikte portalde gösterilebilecektir. Portaldeki verinin organizasyonu ve yayımlanması geleneksel yöntemlerin yanında güvene dayalı paylaşım ve etkileşim gibi daha katılımcı yeni yöntemler kullanılarak yapılabilecektir. Portaller bölgesel hafızanın arşivlenmesi, organizasyonu, zenginleştirilmesi ve bu gibi birçok Avrupa bölgesel hafıza verilerinin birleşmesinden oluşan bir ağ için gerekli araç ve materyalleri sağlayacaktır.

Teknolojik ve sosyal bir ilerleme sağlayacak olan Web 2.0 teknolojileriyle iletişimin, dağıtık karar verme yetisinin, bilgiyi serbest yaymanın ve tekrar kullanmanın önü açılacaktır. Doğal dil işleme teknikleriyle doğal dilde anlama, arama, cevap verme ve bilgiye ulaşma özellikleri sağlanacaktır.

Pilot portal Finlandiya Turku'da kurulacak ve bölgenin mirası olan dijital içeriğe ulaşımı sağlayacak. Pilot portal sonrasında diğer üye ülkelerde kuralacak portallerle birlikte çalışabilir bir ağ kurulacaktır.

(Ek 5.2: Proje Değerlendirme Formu)

5. 6. Gruba Ait FP7 Proje Önerileri

1. iSURF

Proje Adı:	An Interoperability Service Utility for Collaborative Supply Chain across Multiple Domains Supported by RFID Devices	
Proje Akronimi:	iSURF	
Stratejik Alan:	FP7-1-1.3 ICT in support of the networked enterprise	
Proje Türü:	STREP	
Projedeki Rol:	Koordinator	
Bütçe Bilgisi:	**Toplam Bütçe:**	3,129,337 Euro
	Türkiye Payı:	647,100 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
Software Research, Development, Consultation Company	Türkiye
INTEL Performance Learning Solutions Limited, Ireland	İrlanda
TXT e-solutions S.P.A	İtalya
Fraunhofer-Gesellschaft Institute Manufacturing Engineering and Automation	Almanya
Uninova – Instituto de Desenvolvimento de Novas Tecnologias.	Portekiz
Fratelli Piacenza S.p.A.	İtalya

Proje Özeti:

Günümüzde rekabete dayalı olan iş dünyası, farklı endüstrilerden şirketleri birbiriyle anlaşabilir hale getirecek uygulamalara ihtiyaç duymaktadır. iSURF projesi, planlama konusunda aşağıda listelenen özelliklere sahip "akıllı tedarik zinciri ağı" kuracaktır:

- Bağlanan şirketlerin kapasiteleri doğrultusunda kullanabileceği, bilgiye dayalı, farklı uygulamalar arasında çalışabilen bir platform gerçekleştirilecektir.
- Farklı standartlar kullanılarak yaratılmış ve kullanılmakta olan şirket faaliyetlerinin planlanması ve tahmini ile ilgili dokümanların paylaşımını sağlayacaktır.
- RFID teknolojisi ile ürün bilgilerinin otomatik işlenmesini sağlayarak kobilerin yeteneklerini artıracaktır.
- Şirketler arası faaliyeti planlaması ve tahmini aşamalarını otomatik hale getiren ve iş alanına göre uyarlanabilir iş süreci modelleme platformu sağlayacaktır.

- Kurulacak olan platformla şirketlere ait özel uygulamaların otomatik entegrasyonunu sağlamak amacıyla web servis teknolojilerini kullanacaktır.
- Ürün bilgisinin dinamik olarak tüm ortaklar tarafından alınabilmesi ve beklenmeyen durumlarda dinamik olarak ortak arama yapılabilmesini sağlayacak uygulamalar gerçekleştirilecektir.
- Platformun güvenliği ve gizliliğini sağlayacak ek uygulamalar geliştirilecektir.

(Ek 6.1: Proje Değerlendirme Formu)

2. iLTL

Proje Adı:	iLTL: Intelligent Cooperation Platform for Intermodal Freight Logistics	
Proje Akronimi:	iLTL	
Stratejik Alan:	FP7-1-6.1 ICT for Intelligent Vehicles and Mobility Services	
Proje Türü:	STREP	
Projedeki Rol:	Koordinator	
Bütçe Bilgisi:	**Toplam Bütçe:**	2.791.015 Euro
	ODTÜ Payı:	621.980 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.	Almanya
National Institute for Transport and Logistics	İrlanda
METTLE sarl	Fransa
ETRA Investigacion y Desarrollo, S.A.	İspanya
VTEK Bilisim ve İletisim Teknolojileri San. Ve Tic. Ltd. Sti.	Türkiye

Proje Özeti:

iLTL Projesi, kesintisiz ve verimli intermodal lojistik zincirlerinin kurulumunu destekleyen Web tabanlı açık kaynak kodlu işbirliği platformu sağlayacaktır. Bu platformda aşağıdaki bileşenler bulunacaktır:

- İntermodal Lojistik Elektronik Pazar Yeri: Bu pazar yeri sayesinde farklı modlarda çalışan taşımacılar yada terminaller birbirlerini kolayca bularak işbirliği sağlayacaktır. Katılımcılar kendilerinin taşıma yeteneklerini ve yük kapasitelerini devamlı olarak GNSS sunucularından ve RFID okuyucularından gelen gerçek zamanlı bilgileri kullanarak pazar yerine bildireceklerdir.
- İntermodal Lojistik Planlayıcısı: Bu planlayıcı, taşınacak yük için o an ortamda bulunan taşıyıcıların yük kapasitelerini ve konumlarını göz önünde bulundurarak en uygun rotayı çıkaracaktır.
- Lojistik İç Uygulamalarının Ortak Bilgi Modeli Adaptörü: Bu adaptör sayesinde farklı modlarda çalışan taşımacılar birbirleri ile mesajlaşıp doküman alışverişinde bulunabileceklerdir. Böylece birlikte işlerlik sağlanacaktır.

- Konum ve Çevre İzleme Uygulaması: Bu uygulama GNSS Sunucularını ve RFID Okuyucularınından gelen bilgileri kullanarak taşınan yükün o an nerede olduğunu ve çevresel bilgilerini yetkili kişilere bildirecektir. Böylece taşınan yükün güvenli bir şekilde taşınmasına katkıda bulunacaktır.
- GNSS (EGNOS / GALILEO) Sunucusu: Bu sunucu o an taşınan yükün konum bilgisini iLTL platformunun diğer bileşenlerine güvenli bir şekilde iletecektir.

RFID Servis Bileşeni: Bu bileşen sayesinde o an taşınan yük üzerinde bulunan RFID çiplerinden gelen bilgiler iLTL platformuna aktarılabilecektir. Böylece yükün etrafındaki sıcaklık, nem ve sarsıntı gibi çevresel faktörler öğrenilebilecektir.

(Ek 6.2: Proje Değerlendirme Formu)

3. CASHMERE

Proje Adı:	Context-Aware Services for Higher-Mobility and Emergency-Recovery Environments	
Proje Akronimi:	CASHMERE	
Stratejik Alan:	FP7-1-6.1 Intelligent Vehicles and Services	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	2.691.807 Euro
	ODTÜ Payı:	353.920 Euro

Ortak Listesi:

Ortak Adı	Ülke
Universite Libre de Bruxelles	Belçika
Consiglio Nazionale delle Ricerche	İtalya
Ecole Polytechnique Federale de Lausanne	İsviçre
IONIC Software S.A.	Belçika
VOICE-INSIGHT S.A. / N.V.	Belçika
Daimler Chrysler Customer Assistance Center	Hollanda
Synergiums SA	Lüksemburg
Middle East Technical University	Türkiye

Proje Özeti:

CASHMERE projesi içerik-farkındalığına yeni bir yaklaşım geliştirmeyi, ve konum-farkında ve içerik-farkında gezici destek için genişletilebilir servis-etkinleştiren platform entegre etmeyi amaçlamaktadır. Bu yeni yaklaşım; profil modellemesi ve yönetimi, homojen olarak kapsayan kullanıcı profilleri, servis tanımları (bilgi istemlerine cevap vermek için hazırda bulunan veriler hakkında bilgi sahibi olmak ta dahil), ve içerik tanımları (bir bilgi isteminin yorumlanmasını veya isteme en uygun cevap olarak döndürülecek bilgilerin ve servislerin seçimini etkileyebilecek diğer bilgileri de tanımlayan içerik tanımları) genel teorilerinin tanımlanmasına dayanmaktadır. Bu teori; profilleri, onları kullanan uygulamaları ve servisleri yeniden düzenlemeye gerek kalmadan olabilecek en fazla esneklik ve gelişebilirlik sağlayan bir yolla

tasarlamak için yönergeler sunacaktır. Aynı zamanda bu teori, platformu birleştiren yeni servislerin ilgili içerik-farkındalığını kolayca elde etmesini sağlayacaktır. Bu, şu andaki, içeriğin bir defaya mahsus tanımlandığı ve uygulamanın içine gömülü kodlandığı içerik-farkındalığı servisleriyle zıt düşmektedir. İçerik-farkındalığı özellikle, gezici sistemlerde bulunan, kullanıcıların yabancı bir çevrede, planlanmamış anormal durumlarla karşılaştığı ve bu yüzden kişisel ve konum-belirli desteğe ihtiyaç duyduğu acil durumların ele alınması için önemlidir. Proje, üç farklı senaryo şeklinde çeşitlendirilen, kullanıcı geziciliğinden kaynaklanan çeşitli engellere cevap vermeyi amaçlamaktadır. Kısaca, projenin kendine özgü karakteristikleri; çok-biçimli (özellikle ses-tabanlı çok-dilli) kullanıcı arayüzlerinin kullanımını, ayrıntılı profillendirme tekniklerinin yardımıyla servislerin kolayca yapılandırılan kişiselleştirilmesini ve içeriklendirilmesini, gerekli gizliliğin ve güvenliğin işlevsellik yitirilmeden uygulanmasını, en son standartlara uyan ve yeni servislerin kolayca eklenmesini sağlayan temel bir anlamsal ağ çerçeve-uygulamasını, ve çok çeşitli iyi-tanımlanmış test senaryolarını içermektedir.

(Ek 6.3: Proje Değerlendirme Formu)

4. CASA

Proje Adı:	Computer Aided Semantic Annotation	
Proje Akronimi:	CASA	
Stratejik Alan:	FP7-1-4.2 Intelligent Content and Semantics	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	2.884.364 Euro
	ODTÜ Payı:	613.440 Euro

Ortak Listesi:

Ortak Adı	Ülke
University of Szeged	Macaristan
University of Patras	Yunanistan
Midle East Technical University	Türkiye
BIT Media GmbH & Co KG	Avusturya
Innova IT Solutions	Türkiye

Proje Özeti:

Anlamsal Ağ'in ana kapısı, kullanıcıların (içerik sağlayıcıların) dijital içeriklerini eklemeleri konusunda ciddi bir motivasyonsuzluk problemi yaşamaktadır. Bu problemin iki kaynağı olduğunu düşünüyoruz: birincisi, kullanıcının gözünden bakıldığında açıklamaların kullanıcıya direkt bir faydasının bulunmaması; ikincisi ise açıklama ekleme işlemlerinin normal kullanıcılar için çok karışık olması. CASA projesinin amacı, içerik sağlayıcılar için akıllı ve kullanıcı dostu bir açıklama ekleme uygulaması geliştirilmektir. Bizim vakamızda, kullanıcı kendi etiketlerini ve bağlantılarını kullanabilir(belirteç olarak sadece doğal dil kelimelerini kullanması şartıyla). Arkaplanda çalışan ve örnekleri -elle eklenmiş önceki açıklamaları baz alarak- anında etiketleyen bir özdevimli öğrenme algoritması(aktif öğrenme), kullanıcıya elle düzenleyebileceği

bir örnek açıklama sunar(öğrenme algoritması için etiketlediği hangi örnek daha faydalı olur). Bu araştırmanın temel amacı; kullanıcının tecrübelerini kullanmasını, kontrolünü ve etkileşimini sağlayarak harcayacağı eforu minimuma indirmektir(en ileri seviye kullanıcılar dahil olmak üzere). Açıklama standart OWL biçeminde kaydedilmektedir. Bu açıklamayı faydalı hale getirmek için iki yazılım geliştiriyoruz(bir arama önyüzü ve içeriğin bir ağaç yapısında görüntülenmesi). Bu otomatik açıklama çıkarma aracının ve iki yazılımımızın(ki bunlar içeriğin takibini daha da kolaylaştırıyor), kullanıcıya dijital içeriğini CASA aracını kullanarak açıklaması konusunda yeterince fayda sağladığına inanıyoruz. Küçük ontolojileriden belirli kullanıcıların kişisel etiketleridir. Eğer özdevimli öğrenme işlemi, normal dildeki etiketler arasındaki ilişkiyi orataya çıkarabilirse, bunlar OWL örnekleri olarak tanımlanabilir. Böylece, bir açıklama eklendiği zaman, otomatik olarak bir OWL örneği olarak da tanımlanacaktır. İlgilendiğimiz ikinci araştırma konusu ise, bu küçük ontolojilerin SUMO/MILO gibi mevcut resmi ontolojilere doğal dil tanımlayıcılarını ve ilişkilerini kullanarak dönüştürülmesidir.

(Ek 6.4: Proje Değerlendirme Formu)

5. SEMODICS

Proje Adı:	Semantically enhanced Modeling Tools for the Design of Complex Software Systems	
Proje Akronimi:	SEMODICS	
Stratejik Alan:	FP7-1-1.2 Services and Software Architectures, Infrastructures and Engineering	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	4.654.812 Euro
	ODTÜ Payı:	283.800 Euro

Ortak Listesi:

Ortak Adı	Ülke
Centre for Research and Technology Hellas	Yunanistan
Alcatel-Lucent Deutschland AG	Almanya
TELETEL Telecommunications & Information Technology SA	Yunanistan
SOLINET GmbH Telecommunications	Almanya
Middle East Technical University	Türkiye
Intrasoft International SA	Belçika
Institut National de Recherche en Informatique et en Automatique	Fransa
Hewlett-Packard Italiana	İtalya
Thales SA	Fransa

Proje Özeti:

Mühendislikte en kritik süreçlerden bir tanesi sistem tasarımı iken, üretim sürecinin en önemli parçalarından bir tanesi de sunum modellemesidir. Şu anda telekominikasyon, havacılık ve uzay sanayii, imalat alanlarında kullanılan çok çeşitli modelleme teknelojileri bulunmaktadır. Ancak

bu teknelojiler ileri düzeyde modelleme olanakları sunmakla beraber modellerin paylaşımı ve yeniden kullanımı için olan destek hala sınırlıdır ve bu sebeple yoğun uğraşlara neden olmaktadırlar.

Bu projenin temel amacı var olan modelleme araçlarına anlamsal eklentiler yapabilecek bir yazılım platformu sunmaktır. Böylece geleneksel modelleme araçları ve tasarım uygulamaları ile anlamsal teknolojiler arasındaki açık kapatılabilecektir. Son kullanıcı için faydaları kulllanıcının bilgi paylaşma ve tekrar kullanabilme yeteneğinin geliştirilmesi ve modelleme sürecinin gerçek zamanlı model tutarlılğının kontrol ve onayının sağlanması olarak özetlenebilir.

SEMODICS projesinin bilimsel ve teknik hedefleri şu şekilde özetlenebilir:

· Modelleme temel yapı taşlarının anlamsal olarak zenginleştirilmiş bir şekilde tanımlanması, ve bu yolla komplex dinamik sistemlerin davranışlarını ifade edebilecek yapıların oluşturulması

· Dağıtık Model Deposunun oluşturulması. Bu model deposu daha önce farklı modelleme dillerinde yaratılmış modellerin anlamsal olarak indekslenerek tutulmasını sağlayacaktır.

· Modelleme yapılırken bu anlamsal olarak zenginleştirilmiş modellere üzerinde çıkarımlar yapabilecek algoritmaların ve yazılımların geliştirilmesi

(Ek 6.5: Proje Değerlendirme Formu)

6. MARVEL

Proje Adı:	Manage Adverse Reactions Verified ELectronically	
Proje Akronimi:	MARVEL	
Stratejik Alan:	FP7-1-5.2 Risk Assessment and Patient Safety	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	3,652,640 Euro
	ODTÜ Payı:	417,200 Euro

Ortak Listesi:

Ortak Adı	Ülke
Fondazione Centro San Raffaele del Monte Tabor	İtalya
Info World	Romanya
TXT e-Solutions	İtalya
CANTA H	Romanya
Middle East Technical University	Türkiye
Kadris	Fransa
European Society of Clinical Pharmacy	Belçika

Proje Özeti:

Sağlık sistemleri çeşitli güçlüklerle karşılaşmaktadır. İlaca ve prosedürlere dayalı hatalar, hasta güvenliği ve bilgisinin özelliği ile ilgili hatalar ve güvenilir bilginin eksikliği bu güçlüklerin arasında öne çıkanlardır. Olguları ve hataları belgelendirmek üzere düzgün bir biçimde işleyen raporlama sistemleri, sağlık toplumunun toplam kaliteyi arttırmada, sosyal ve finansal yükümlülükleri azaltmada birincil önceliklerindendir. Halihazırda klinik riskler ve ilaç yan etkileri ile ilgili bilgiler az ve de yerel, ulusal ve uluslararası seviyelerde farklı veritabanlarında dağılmış durumdadır. MARVEL'in amacı, bu veritabanlarındaki bilgileri, gerek duyulduğunda, uzmanlar için erişebilir hale getirmektir. Hastaya reçete yazılması esnasında, doktorlara destek sağlamak için, farklı kaynaklardan alınan bilgilerin sağlık bilgi sistemlerine entegrre edilmesi mümkündür. Bu doğrultuda, MARVEL klinik, lojistik ve organizasyonel risklerle ilgili bilgileri entegre ederek uzmanlar için bilgi ve uyarılar sağlamayı; bu sayede ilaç yan etkilerinden kaynaklanan riskleri en aza indirmeyi amaçlamaktadır. Belirtilen amaçları gerçekleştirmek için, MARVEL, hastanelerdeki elektronik reçete sistemlerine entegre edilmek üzere tasarlanmış bir dizi modüler uygulama olarak geliştirilecektir. Bu tip bir mimari ile ölçeklendirilebilir ve esnek bir sistemin hayata geçirilmesi mümkün olacak ve büyük çapta hasta güvenliği çalışmaları için varolan elektronik reçete sistemleriyle arayüzler oluşturulabilecektir.

(Ek 6.6: Proje Değerlendirme Formu)

6. 7. Gruba Ait FP7 Proje Önerileri

1. ROSSI

Proje Adı:	Emergence of communication in RObots through Sensorimotor and Social Interaction	
Proje Akronimi:	ROSSI	
Stratejik Alan:	FP7-1-2.1 Cognitive Systems, Interaction, Robotics	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	2.800.000 Euro
	ODTÜ Payı:	430.000 Euro

Ortak Listesi:

Ortak Adı	Ülke
University of Bologna	İtalya
University of Parma	İtalya
University of Lubeck	Almanya
University of Skovde	İsveç
Middle East Technical University	Türkiye
University of Wales	İngiltere

Proje Özeti:

Bu proje hem robotlar arasında ve hem de robotlarla insanlar arasındaki iletişimin geliştirilmesi ve oluşturulmasına odaklıdır. Özel olarak proje iletişim kuran etmenler arasında dünyanın ortak bir izleniminin paylaşılmasının ne kadar kritik olacağını inceleyecektir. Değişik algı ve motor sistemlerine sahip olan etmenler değişik dünya izlenimlerine sahip olacak ve iletişimde sorunlar yaşayabilecektir.

Biz insanlarla robotlar arasındaki iletişimin kolaylaştırılabilmesi için dünya modellerinin hangi yönlerinin paylaşılması gerektiğini inceleyeceğiz. Sağlarlık tabanlı dünya modellerinin gereksinimlerini daha iyi anlamayı ve bu tür paylaşılmış modeller üzerinden iletişimin nasıl yapılabileceğini incelemek istiyoruz. Yakın zamanda ortaya çıkan nörofizyolojik veriler premotor kortekste iki grup nöron olduğunu göstermiştir: bir grup nesnelere yönelik hareketler sırasında ve aynı zamanda bu hareketlerin kullanıldığı nesnelerin görsel olarak algılanması ile aktif olmakta (kanonik nöronlar), diğer grup ise diğerlerinin aynı hareketleri yapması durumunda ve bunlara karşılık gelen sözlü kelimelerin duyulması ile aktif olmaktadır (ayna nöronlar). Bu nörofizyolojik kanıt platformu üzerinde robotların çevre ve ortamları anlayışlarına dayanarak insanlarla iletişim kurabilecek ve beraber çalışabilecek robotlar geliştirilmesi mümkün olabilir.

Bu projenin odağı basit komutları (değişik nesnelerle etkileşime yönelik kelimeler) yerine getirebilecek ve etrafındaki değişik nesnelera bağlı olarak makul şekilde davranabilecek robotlar geliştirilmesidir. Bu amaca yönelik olarak psikolojiden, nörobilimden, bilgisayar bilimlerinden ve robot mühendisliğinden gelen disiplinlerarası bir takım kurulmuştur.

(Ek 7.1: Proje Değerlendirme Formu)

2. DISCOVER

Proje Adı:	Distributed Cognitive Robotic Systems	
Proje Akronimi:	DISCOVER	
Stratejik Alan:	FP7-1-2.1 Cognitive Systems, Interaction, Robotics	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	2,408,238 Euro
	ODTÜ Payı:	393,236 Euro

Ortak Listesi:

Ortak Adı	Ülke
Orebro Universitet	İsveç
Instituto Superior Técnico	Portekiz
Bonn-Rhein-Sieg University of Applied Sciences	Almanya
Instituto Gulbenkian de Ciencia	Portekiz
Middle East Technical University	Türkiye

Proje Özeti:

Bu projenin ana amacı Dağıtık Bilişsel Robotik Sistemleri (DBRS) için yeni bir teorinin temellerini geliştirmektir. DBRS ile birbiriyle ortaklaşa çalışan bir grup cihazlardan oluşmuş bir sistemi kastediyoruz. Bu sistem, bir bütün olarak, ortamla algı ve aksiyon üzerinden etkileşip bilişsel davranışlar sergileyebilecektir.

Bir DBRS ortamdaki robotları ve görünmez şekilde ortama yerleştirilmiş cihazları entegre edebilir. Robotik alanının DBRS'e göstermeye başladığı ilgi üç nedene bağlanabilir: birincisi ortama geçmiş ağa bağlı cihazlar artık gündelik hayata girmiştir, dolayısıyla her otonom robot bu cihazlara bağlanmaya ve bunlarla ortak çalışmaya hazırlıklı olmalıdır; ikincisi birbirine bağlı birden fazla robotik cihazları tek başına çalışan bir robot sisteminden daha esnek, gürbüz ve yetenekli olabilir; üçüncüsü bir DBRS, modülerlik, genişletilebilirlik ve konfigüre edilebilirlik yönlerinde bariz avantajlar içerir.

Bütün bu cazip özelliklerine rağmen DBRS'lerin geliştirilmesi, bu sistemlerin tasarımı, sentezi ve analizi konularındaki eksikliklerinde ötürü yeterli olmamaktadır. Bu proje, bu türden geliştirmeler için gereken genel araştırma problemlerini inceleyecek ve varolan yöntemlerin bu problemlerin çözümünde nasıl kullanılabileceğini araştıracaktır. Bu alanlar, yapay zeka, bilişsel sistemler, robot sürüleri, sistem teorisi ve bağışıklık sistemlerini içermektedir. Çalışmalarımız, yaşlıların ev ortamındaki sağlık ve güvenliklerini konu alan bir dizi demonstrasyon senaryoları ile yönlendirilecektir. Ancak, projenin ana çıktısı, yukarıdaki alanlardan alınan metotların tekniklerin birleştirilmesiyle ortaya çıkacak olan sağlam genel prensiplerin ortaya çıkarılması olacaktır. Özel olarak bir DBRS'nin efektiflik, adaptiflik ve gürbüzlük yönlerinden varolan robot sistemlerinin nasıl ötesine geçilebileceği incelenecektir.

(Ek 7.2: Proje Değerlendirme Formu)

3. THRIL

Proje Adı:	Thinking Human-Robot Ecologies in the Language of Interaction	
Proje Akronimi:	THRIL	
Stratejik Alan:	FP7-IDEAS	
Proje Türü:	ERC Starting Grant Stage 1	
Projedeki Rol:	Koordinatör	
Bütçe Bilgisi:	**Toplam Bütçe:**	1,200,000 Euro
	ODTÜ Payı:	1,200,000 Euro

Proje Özeti:
Bu proje başarılı bir insan-robot ekolojisi kurmanın yolunun, birbirinden bağımsız tasarlanmış robotik yapıların ortamlarımıza koymaktan ibaret olmadığını iddia etmektedir. İnsan-robot ekolojilerinde önem, robotik yapıların ayrı olarak tasarlanmalarından, bunların ortamları paylaşacakları insanlar ve diğer cihazlarla olacak etkileşimlerine kaydırmıştır. Biz bu kaymanın, düşünce düzleminde de bir perspektif değişimi gerektirdiğini düşünüyoruz.

Bu projenin temelinde bir dobra fikir yatmaktadır: Başarılı bir insan-robot ekolojisinin geliştirilmesi sistem seviyesinde bir etkileşim dilinde yapılmalıdır ve robotik yapıların tasarım

ve geliştirilmesi hedeflenen ekolojideki etkileşim gereksinimleriyle şekillendirilmelidir. Önerilen perspektif değişimi, canlıların laboratuvar ortamlarında değil kendi doğal ortamlarında incelenmesini savunan etoloj bilim dalının ortay çıkmasına benzetilebilir. Sistem seviyesinde düşünmede derken, yeni bir robotik yapının bir ortama girişinin sadece onun algı ve hareket yetenekleri yönünden değil, bu yapının o ekolojide tetikleyeceği tepkiler cinsinden düşünülmesi kastedilmiştir. İletişim dili olarak ise. ekolojinin etkileşimlerini destekleyen sembolik bir dilden çok, bir robotik yapıya ekolojinin geri kalanın vereeceği dağıtık tepkiler anlaşılmalıdır.

(Ek 7.3: Proje Değerlendirme Formu)

7. 8. Gruba Ait FP7 Proje Önerileri

1. FRICOC

Proje Adı:	A Framework for Intelligent Content Creation	
Proje Akronimi:	FRICOC	
Stratejik Alan:	FP7-1-4.2 Intelligent Content and Semantics	
Proje Türü:	STREP	
Projedeki Rol:	Koordinatör	
Bütçe Bilgisi:	**Toplam Bütçe:**	2,723.500 Euro
	ODTÜ Payı:	590,100 Euro

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
University of Thessaloniki	Yunanistan
Sirma Group	Bulgaristan
Turkcell	Türkiye
University of Manchester	İngiltere
Ort Braude College of Engineering	İsrail

Proje Özeti:

Günümüzde, Web tarafından sağlanan bilgi miktarı arttıkça, web sitelerinin yapıları karmaşık bir hal almakta ve bunun sonucu olarak verimli bilgi erişimi daha da zor bir iş haline gelmektedir. Kullanıcıların web sitelerini ziyaret şekli, kullanıcın ihtiyaç, bilgi birikimi ve ilgi alanlarıyla sıkı sıkıya bağlıdır. Ancak, bu ihtiyaçlar web sitesinin oluşturulması sırasında tasarımcılarının varsaydığı ihtiyaçlardan önemli ölçüde farklı olabilirler. Oysa ki, bir web sitesinin, kullanıcılarının tercihlerini yansıtan bir yapıda olması gereklidir. Bu hedef, sitenin kulllanıcıların tercihlerine kendini devamlı adapte etmesi ile sağlanabilir. Ideal olarak, web siteleri kendi kendine adapta edebilir yapıda olmalıdır, diğer bir deyişle, bir web sayfası, kullanıcıların sayfayı kullanış şeklıne göre ihtiyaçları öğrenerek, yapısını ve tasarımını geliştirebilmelidir. Günümüzde, kendi kendine adapte olabilen web sayfaları üretmek için çeşitli teknik ve yaklaşımlar mevcuttur. Hatta bu teknik ve yaklaşımlar Google, Amazon gibi bilinen web

sitelerince kullanılmaktadır. Bu teknikler genellikle kayıt ve sayfaların gezilme bilgilerine dayanmakta ve adaptasyon işlemi, sayfaların, web bağlantıları gibi, sadece belli kısımlarıyla sınırlı kalmaktadır. Bu projede, mevcut adaptasyon tekniklerini, web gezintilerinin anlamsal bilgilerini kullanarak geliştirmeyi ve adaptasyonu sayfaların bütün alt kısımlarına uygulamayı planlıyoruz. Önerilen projenin bir başka hedefi de, hem tasarım hem de kütük yapısında, Web2..0 standartlarını kendi kendine adaptasyon teknikleriyle birleştirmektir. Yakın gelecekte, kullanıcıların Web'e, dizüstü ve masaüstü bilgisayarlar yerine, daha çok mobil araçlar kullanarak bağlanmaları beklenmektedir. Adaptasyon ve kişiselleştirme teknikleri mobil araçlar için farklı gereksinimlere ihtiyaç duymakatdır. Bu projede, hedeflene problemin mobil araçlar açısından da ihtiyaçları incelenecek ve bu ihtiyaçlara çözümler önerilecektir. Projede hedeflenen problem için önerilecek yaklaşım telekommünikasyon alanında pilot bir uygulama üstünde gösterilecektir.

(Ek 8.1: Proje Değerlendirme Formu)

8. 9. Gruba Ait FP7 Proje Önerileri

1. MASHUP

Proje Adı:	MigrAtion to Service Harmonization compUting Platform	
Proje Akronimi:	MASHUP	
Stratejik Alan:	FP7-1-1.2 Software and Service Architectures and Infrastructures and Engineering	
Proje Türü:	STREP	
Projedeki Rol:	Ortak	
Bütçe Bilgisi:	**Toplam Bütçe:**	3.613.876 Euro
	ODTÜ Payı:	508.920 Euro

Ortak Listesi:

Ortak Adı	Ülke
Cybersoft C/S Information Technologies Ltd. Co.	Türkiye
Middle East Technical University	Türkiye
German Research Centre for Artificial Intelligence	Almanya
Liverpool John Moores University	İngiltere
Fraunhofer IPA	Almanya
Vienna University of Technology	Avusturya
Dajeil Ltd.	İrlanda
Syros Shipyards Ltd.	Yunanistan

Proje Özeti:

Servis Temelli Mimari (STM), bir mimari model sunmanın dışında servisleri yapıtaşı kabul ederek iş süreç yönetimi için anahtar bir yaklaşım olarak görülmektedir. Farklı

kaynaklardan alınan birbirinden farklı, yapısal ve yapısal olmayan verileri ve fonksiyonları bir araya getirmek ve mantıksal olarak uyumlu bir biçimde kullanıcıya sunmak servis kalitesinin sağlanması açısından engelleri de beraberinde getirmektedir. Bunun yanında trilyonlarca yatırımın yapıldığı eski teknoloji ile geliştirilmiş uygulamaların servis temelli mimariye adapte edilmesi bu geçiş döneminin en büyük ve önemli sorunlarındandır. Tahminlere göre şu an Bilgi Teknoloji (BT) sistemlerinin %80'i kalıt(legacy) platformlarda çalışmaktadır. Ayrıca bugün 200 milyar kalıt kod satırı on birlerce büyük anabilgisayarlarda (IBM 360 gibi) işlemleri gerçekleştirmektedir. Forrester'ın yaptığı anket sonuçlarına göre bilim teknoloji bütçesinin sadece %20 si yeni projelerde kullanılırken, %80 i mevcut kalıt projelerin bakım ve onarım çalışmalarında harcanmaktadır. Hatta BT bütçesinin ortalama %30 u bakım ve onarım çalışmaları yapan personel maaşları için ayrılmaktadır.

Eski teknoloji ile geliştirilmiş, kalıt uygulamalar yaş, kullanılan programlama dili, geliştirme ortamı ve mimari açıdan farklılık göstermektedir. Servis temelli yaklaşım göz önüne alınarak tasarlanıp geliştirilmediğinden örneğin tek katmanlı ve kullanıcı/sunucu mimariye sahip tek parça kalıt yapılarının STM odaklı bir altyapıya geçişi sıkıntı yaratmaktadır. Mevcut sistem ile STMnin yapısal farklılıkları da göz önüne alındığında adaptasyonun maliyetli olacağı aşikârdır. Kalıt modernizasyonunu sağlayan araçlar gün geçtikçe artmaktadır. STMye geçiş için ortaya atılan yaklaşımlar, yazılımın uygulama ve veri katmanında kodların yeniden göz geçirilip yapılandırılması, mimarinin değiştirilmesi ya da yeniden düzenlenmesini içermektedir. Ayrıca bu süreç içerisinde ileri ve geri mühendislik süreçlerinin uygulandığı yeniden yapılanma(reengineering) tekniği kullanılmaktadır.

MASHUP (MigrAtion to Service Harmonization compUting Platform) bu yaklaşımlar yerine eski teknoloji ile geliştirilmiş uygulamaların alana özel kalıt sargılar yardımıyla dinamik olarak bir araya gelmesini sağlayarak yeni geliştirilen servislerle birlikte esnek iş süreçlerini gerçekleştirmeyi amaçlamaktadır. Bunun yanında Web 2.0 ve STM birlikteliği ile sadece veri ve uygulama katmanında değil gösterim katmanında da uygulamaları eşgüdümlü hale getiren ve servis yönetimini sağlayan bir platform geliştirmek hedef olarak seçilmiştir. Bu amaçla, MASHUP servis sargılarını yaparken kalıtlar için "Web Okuma", "Ekran Yakalama" ve "İş Süreci Bütünleştirme" gibi uygulamaları gerçekleştirecektir. MASHUP Sistemi 3 ana parçadan oluşmaktadır:

(1) Servis Harmonizasyon(Uyumlaştırma) Platformu için Referans Mimari: Servis varlıkları için tanımlanmış bir meta-model kullanarak MASHUP Koreografi Motoru yardımıyla Alana Özel Motorların dinamik bütünleşmesini sağlayan, servis kalite gereksinimlerini karşılayan ve servis seviyesinde anlaşmalarla ilgilenen bir altyapıdır.

(2) Diller, Motorlar ve Araç Takımları: Kalıt Servis Sargıları için yapı taşlarıdır. Alana Özel Takım Aracı (a) alana özel gösterim ve yapıları içeren *alana özel dil*, (b) bu dili işleyecek *motor* ve (c) varlıkları tasarlama, geliştirme ve yönetme işlerini gerçekleştiren *araç takımları*ndan oluşur. Her biri kalıt uygulamanın platform ve satıcıya özgü durumlarını izole ederek, onları alana özel sargılarla içyapıdan soyutlayarak genel(generic) bir altyapıda eşgüdümlü hale gelmeleri

(3) Yöntem: MASHUP gösterimleri, teknikleri ve süreçlerini kapsayan detaylı tanımlamaları içerir. MASHUP yöntemi, STMye geçişte Koreografi Motoru ve Alana Özel Takım araçlarını içeren MASHUP sisteminin nasıl kullanılacağını anlatmaktadır.

MASHUP açık kaynak servis temelli altyapısı ile dinamik servis yaratımı, bu servislerin eski teknoloji ile geliştirilmiş uygulamalarla birlikte çalışabilirliği, eski teknoloji ile geliştirilmiş, kalıt uygulamaların STMye kalite özelliklerini eksiksiz yerine getirecek bir şekilde adapte edilmesi sağlanacaktır.

(Ek 9.1: Proje Değerlendirme Formu)

9. FP7 - 3.Çağrı Proje Çalışmaları

METU-ISTEC grupları FP7-ICT alanına aşağıdaki proje önerisini sunmayı hedeflemektedirler:

1. MASTERY

Proje Adı:	Mining Actionable Cases by means of Temporal, Probabilistic Reasoning mechanisms and Semantic Interoperability
Proje Akronimi:	MASTERY
Stratejik Alan:	FP7-ICT-2007-3 – 4.4: Intelligent Content and Semantics
Proje Türü:	STREP
Projedeki Rol:	Koordinatör

Proje Özeti:

Vaka tabanlı muhakeme (Case-based reasoning (CBR)) vakalar şeklinde ifade edilmiş geçmiş tecrübelerden faydalanarak yeni problemleri anlamayı ve çözmeyi amaçlar. Vaka tabanlı muhakeme özellikle bilginin toplanması ve formal bir şekilde ifade etmenin zor olduğu tıp ya da hukuk gibi alanlarda uygulanmaktadır. Başarılı uygulamaları bulunmakla beraber Vaka tabanlı muhakeme sistemlerinin temel varsayımı uzmanlar tarafından hazırlanan vaka tanımlarının vaka sunucularında bulunuyor olmasıdır. Ancak bu şekilde tanımlı vakalar olmasa bile, problem çözümünün farklı evreleri hakkında bilgi sunan biligi kaynaları bulunmaktadır. Örneğin sağlık alanında elektronik sağlık kayıtları tıp vaka tanımlarını oluşturmak için gereken bilgiyi içermektedir. MASTERY projesinin amacı bu problemi ele alıp, dağıtık ve heterojen olarak bulunan bu çeşit veri kaynaklarında bilgi çıkarım algoritmaları kullanarak otamatik olarak vaka tanımlarını oluşturasbilecek bir altyapı hazırlamaktır.

Birden fazla dokümandan vaka tanımlarını çıkarabilmek için tek bir kaynaktan anlam çıkarabilmek gereklidir fakat tek başına yeterli değildir. Bununla beraber, bilgisayarlala işlenebilen farklı kaynaklardan elde edilmiş bu bilgilerin içinde saklı bulunan ilişkileri açığa çıkarabilmesi için belirli bir içerikle yorumlanması gerekmektedir. Bunun için de zamansal, olasılıksal ve anlamsal muhakeme gereklidir. Bu amaçla, MASTERY projesi yeni vakalar yaratmak için farklı veri kaynaklarında bulunan bilgileri alan bilgisi ve zamansal bilgiyle harmanlayıp bu bilgeleri anlamlayan ve aralarındaki ilişkileri muhakeme edebilen bir platform

10. FP7-Security Alanına Ait Proje Çalışmaları

METU-ISTEC grupları FP7-Security alanına yönelik iki proje önerisinde bulunmuşlardır:

"Development of an Advanced Training Framework for Critical Infrastructure Protection, Using MOdelling and Simulation" (CIPMOS) adlı proje önerisi "ICT-SEC-2007-1.0-02: Modelling and simulation for training" stratejik alanına yollanmıştır.

1. CIPMOS

Proje Adı:	Development of an Advanced Training Framework for Critical Infrastructure Protection, Using MOdelling and Simulation
Proje Akronimi:	CIPMOS
Stratejik Alan:	ICT-SEC-2007-1.0-02: Modelling and simulation for training
Proje Türü:	STREP
Projedeki Rol:	Koordinatör

Ortak Listesi:

Ortak Adı	Ülke
Middle East Technical University	Türkiye
Global Security Intelligence	İngiltere
Deep Blue	İtalya
Alcatel-Lucent	Türkiye
Multimedia Campus	İtalya
Turkish Electricity Transmission Company	Türkiye
The Fraunhofer-Institute for Intelligent Analysis and Information Systems	Almanya

Proje Özeti:

Bilgi ve iletişim teknolojileri, enerji, su ve gıda, sağlık, kimyasal ve nükleer endüstri, ulaşım, finans, vb. sektörler kritik altyapılara sahiptirler. Bu tip yapılar, kazalar, terörist saldırılar, doğal afetler, siber saldırılar gibi olaylar sonucunda zarar görebilir. Bu gibi olaylar sonucunda kritik altyapılarda meydana gelebilecek krizlerin etki alanı büyük olur. Kriz altyapılarının artan karmaşıklığı ve karşılıklı bağlılıkları, gelişen teknolojiler sonucu geliştirilen iç sistemlerin çeşitliliği bu etkiyi daha da artırır. İletişim, koordinasyon ve işbirliği kritik altyapılar için çok önemli ve vazgeçilmez kavramlardır. Kritik yapıların karmaşıklığı, farklı altyapıların karşılıklı

bağlılıkları ve sosyal ekonomik ve çevresel faktörler bu prensipleri sağlamayı zorlaştırır. CIPMOS projesinde elektrik iletim ve telekomünikasyon altyapıları yapıları için güvenlik sorumlu aktörlerin eğitimi amaçlı ileri bir simülasyon çerçevesi geliştirilecektir. Sistem ayrıca, kritik yapıların güvenlik önlemleri açısından analizi, kriz durumlarında karar destek ve güvenlik önlemlerinin standartlaştırılması gibi amaçlar için de kullanılabilecektir.

2. ADOPPS

Proje Adı:	Development of a Modelling & Simulation Framework for the Analysis & Design of Optimised Pipeline Protection Systems
Proje Akronimi:	ADOPPS
Stratejik Alan:	ICT-SEC-2007-1.0-03: Optimised Situational Awareness Through Intelligent Surveillance of Interconnected Transport or Energy Infrastructures.
Proje Türü:	Collaborative Project
Projedeki Rol:	Ortak

Ortak Listesi:

Ortak Adı	Ülke
SELEX Sensors and Airborne Systems Limited	İngiltere
Middle East Technical University	Türkiye
Elsag Datamat Spa	İtalya
Optibase Limited	İsrail
Waterfall Solutions Limited	İngiltere
Botaş Petroleum Pipeline Corporation	Türkiye
DESFA (Hellenic Gas Transmission System Operator)	Yunanistan

11. Çekirdek Grup Yetiştirme Çalışmaları

METU-ISTEC projesinin bir amacı da proje boyunca yeni genç araştırmacıları projeye dahil etmek ve değişik alanlarda hizmet edecek çalışma grupları yaratmak olmuştur.

Birinci ve beşinci gruplar ODTÜ Bilgisayar Mühendisliği Bölümü bünyesindeki Akıllı Sistemler Labı'nı (Intelligent Systems Lab – ISL http://isl.ceng.metu.edu.tr/) METU-ISTEC projesi sayesinde kurmuşlardır. ISL'nin kurulmasıyla proje dahilinde bir çekirdek grubun oluşturulması sağlanmıştır. 7. Çerçeve programının ICT 1. çağrısında verilen proje teklifi ISL üyeleri tarafından hazırlanmıştır. Bunun yanında daha sonraki çağrılara proje teklifinde bulunabilmek için gerekli alt yapı ve bilgi birikimi lab içinde oluşturulmuştur. Çağrıya spesifik araştırma konularının dışında lab bünyesinde otomatik muhakeme, uzman sistemler, bilgi gösterimi ve muhakeme, dijital kütüphaneler, anlambilimsel ağ teknolojileri, doğal dil işleme ve robotik konularında faaliyet gösterilmektedir. Samet Akpınar, Özgür Alan, Gencay K. Evirgen, Orkunt Sabuncu, Barış Tanrıkulu proje çalışmalarında aktif olarak görev almaktadırlar.

İkinci grubun araştırma konularında aktif olarak görev alan dört kişilik çekirdek grup oluşturulmuştur. Ömer Sinan Saraç, M. Ersan Topaloğlu, Zerrin Sökmen doktora çalışmalarına devam etmekte olup Ayşegül Yaman yüksek lisans çalışması yapmaktadır.

METU-ISTEC üçüncü ve dördüncü grupları birlikte "Distributed and Mobile Computing" konusu üzerinde çalışmaya başladıktan sonra aşağıdaki ekibi oluşturmuştur:

- Demet Aksoy, Assoc. Prof.
- Muslim Bozyiğit, Prof.
- Veysi İşler, Assoc. Prof.
- Cevat Şener, Dr.
- Ghalib Asadullah Shah, Dr.
- Khaled Banafa, Doktora öğrencisi
- Ali Bokar, Doktora öğrencisi
- Swaleh Kintu Kavuma, Yüksek lisans öğrencisi
- Taner Kurtuluş

Proje çerçevesinde Ghalib Asadullah Shah, Prof. Bozyiğit'in danışmanlığında, "Wireless Sensor Networks" konusundaki çalışmalarına devam etmiş ve doktora çalışmasını tamamlamıştır.

Bu grup Intel'den "WiMAX networks" konusunda altyapı oluşturmak ve araştırma çalışmaları yapmak üzere destek almıştır.

Altıncı grubun araştırma konularında üç öğrenci aktif olarak görev almıştır. Özgür Kılıç ve Yalın Yarımağan doktora çalışmalarına devam etmektedirler. Ayrıca İbrahim Taşyurt yüksek lisans çalışmasını sürdürmektedir.

Yedinci grubun araştırma konularında bir doktora öğrencisi, Sertan Girgin, bir de yüksek lisans öğrencisi, Fatih Gökçe, aktif olarak görev almışlardır.

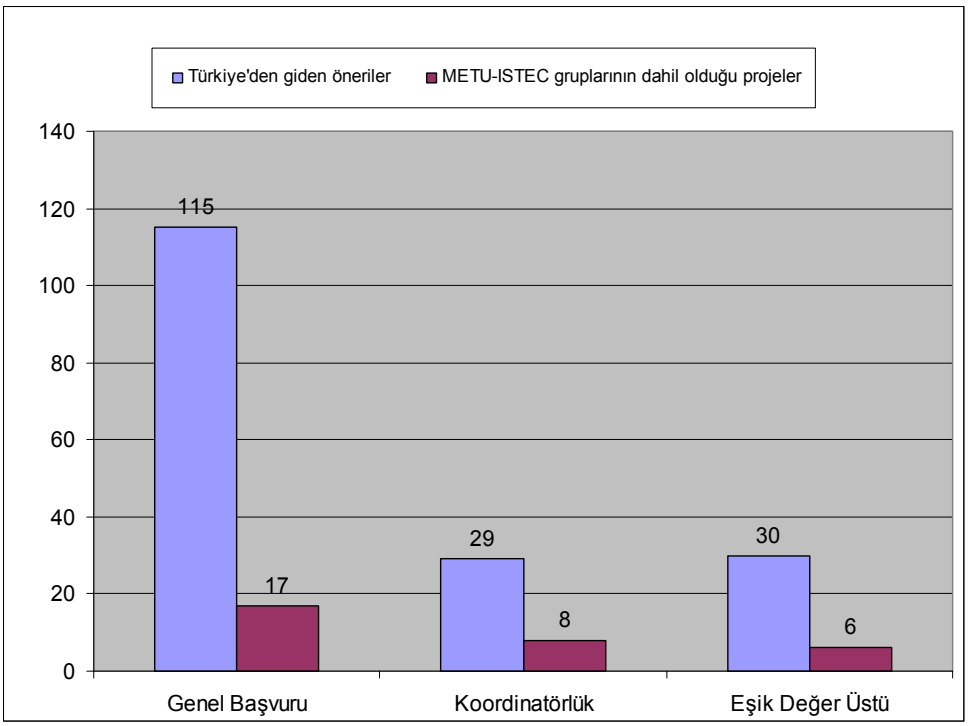
Sekizinci grubun araştırma konularında iki öğrenci aktif olarak görev almıştır. Özgür Alan doktora çalışmasına, Çağlar Ata yüksek lisans çalışmasına devam etmektedir.

Dokuzuncu grubun araştırma grubu içerisinde doktora öğrencisi, Selma Süloğlu bulunmaktadır. METU-ISTEC desteği ile proje önerisinin hazırlanmasına aktif olarak katılmıştır ve yayınlanan bildirilere yardımcı yazar olarak katılmıştır. Öneri hazırlama sürecinde ortak arama ve iletişim, iş paketlerini belirleme, proje bütçesi hazırlama gibi konularda tecrübe kazanmıştır. Bu süreçte yapılan araştırmaların doktora konusuna yansıtılması planlanmaktadır. Ayrıca edinilen bu tecrübe ile ileride açılacak çağrılarda uygun konu başlıklarına proje önerisi gönderilmesi planlanmaktadır. Bunun yanında gruba ait proje önerisinin bilimsel içeriği ile ilgili N. İlker Altıntaş, “Feature-Based Software Asset Modeling with Domain Specific Kits” başlıklı tezini 10.08.2007 sunmuş ve tezi kabul edilmiştir.

TARTIŞMA/SONUÇ

Proje başarıyla tamamlanmıştır. Projenin temel amacı Türkiye'nin Bilgi Toplumu Teknolojilerinin araştırma ve teknolojik gelişimi programlarına katılımını artırmak olduğu göz önünde tutulduğunda, projenin olumlu etkisi 7. Çerçeve Programı ilk çağrı sonuçlarında görülmektedir.

7. Çerçeve Programı ICT alanına ilk çağrıda Türkiye'den toplam 115 proje başvurusu olmuştur. Bu projelerin 29'u Türkiye koordinatörlüğünde yollanmıştır. METU-ISTEC proje grupları toplam 17 proje önerisi yollamışlardır. Bu projelerden 8'i Türkiye koordinatörlüğünde yollanmıştır. Türkiye'nin yer aldığı projelerin 14.78%'i, Türkiye koordinatörlüğünde yollanmış projelerin 27.59%'u METU-ISTEC kapsamında hazırlanmıştır.



Aynı çağrıda Türk şirketlerin yer aldığı proje başvurularının 30'u eşik değeri geçmiştir. Eşik değeri geçen projelerden 7'si Türkiye koordinatörlüğünde yollanmıştır. Bu projelerden 6'sında METU-ISTEC grupları yer almış, 2'sinde koordinatör olarak başvurmuştur. Türkiye'nin yer aldığı eşik değer üstü projelerin 20%'si, Türkiye koordinatörlüğünde yollanmış eşik değer üstü projelerin 28.57%'si METU-ISTEC kapsamında hazırlanmıştır.

İlk çağrı kapsamında koordinatörlüğünü Orta Doğu Teknik Üniversitesi'nin üstlendiği iSURF projesi ve Orta Doğu Teknik Üniversitesi'nin ortak olarak yer aldığı ROSSI projesi destek almıştır.

METU-ISTEC gruplarının yer aldığı proje önerilerinin toplam bütçesi 51.490.540 Euro olup, 9.224.016 Euro'luk bütçe METU-ISTEC gruplarına aittir. Destek alan projelerin toplam bütçesi 5.929.337 Euro olup, Türkiye'ye ait toplam bütçe 1.077.001 Euro'dur.

7. Çerçeve Programı "Security" alanına ilk çağrıda METU-ISTEC projesinin yer aldığı iki proje teklifi hazırlanmıştır. Bu projelerden biri ODTÜ koordinatörlüğünde yollanmıştır.

METU-ISTEC kapsamında hazırlanan projelerde 12 yerli şirketin 7. Çerçeve Programı proje önerilerine katılımı sağlanmıştır.

METU-ISTEC projesi kapsamında hazırlık aşamasında yeni araştırmacılara destek verilmiş, ODTÜ bünyesinde yeni çalışma grupları oluşturulmuştur. Çalışma grupları, araştırma alanlarında proje hazırlarken öncelikle mevcut projeleri, gelişmeleri, akademik çalışmaları ve standartları inceleme fırsatı bulmuşlardır. Ardından oluşturulan çekirdek gruplar aktif olarak proje hazırlama aşamasında görev almışlardır. Böylece proje boyunca yeni genç araştırmacılar yetiştirilmeye başlanmıştır.

Proje süresince yapılan bilimsel araştırmalar sonucunda aşağıdaki yayınlar hazırlanmıştır:

- Dogac A., Kabak Y., Namli T., Okcan A., Collaborative Business Process Support in eHealth: Integrating IHE Profiles through ebXML Business Process Specification Language, IEEE Transactions on Information Technology in Biomedicine, Accepted for publication
  (Science Citation Index Core, Impact Factor: 1.575)

- Kabak Y., Dogac A. "A Survey and Analysis of Electronic Business Document Standards", ACM Computing Surveys, Submitted for publication.

- G. A. SHAH, Ö. B. AKAN, M. BOZYIGIT, "Multi-Event Adaptive Clustering (MEAC) Protocol for Heterogeneous Wireless Sensor Networks", in Proc. Fifth Annual Mediterranean Ad Hoc Networking Workshop (MedHoc-Net), also in Lecture Notes in Computer Science 4003: pp365-383, June 2006.

- G. A. SHAH, M. BOZYIGIT, DEMET AKSOY, "RAT: Routing by Adaptive Targeting in Wireless Sensor/Actor Networks", in Proc. Second IEEE/ACM International Conference on COMmunication System softWAre and middlewaRE (COMSWARE), January 2007.

- G. A. SHAH, M. BOZYIGIT, "Exploiting Energy-aware Spatial Correlation in Wireless Sensor Networks", 2nd International Workshop on Software for Sensor Networks (SensorWare 2007), January 2007

- Semih Cetin, N. Ilker Altintas, Halit Oguztuzun, Ali H. Dogru, Ozgur Tufekci and Selma Suloglu, "A Mashup-Based Strategy for Migration to Service-Oriented

Computing", ICPS'07, IEEE International Conference on Pervasive Services, p.169-172, July 15-20, 2007, İstanbul

- "Z.Sokmen, M.Ozturk, V.Atalay, R.Cetin Atalay, "A Hybrid Method For The Identification of Expression Patterns From Microarray Data" , 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2007) and 6th European Conference on Computational Biology (ECCB), Vienna, Austria, July 21-25, 2007.

- Semih Cetin, N. Ilker Altintas, Halit Oguztuzun, Ali H. Dogru, Ozgur Tufekci and Selma Suloglu, "Legacy Migration to Service-Oriented Computing with Mashups", ICSEA 2007, International Conference on Software Engineering Advances, 25-31 August, 2007, Cap Esterel, French Riviera, Fransa

- Sabuncu, O. and F.N. Alpaslan, "Computing Answer Sets Using Model Generation Theorem Provers", Proc. 4th International Workshop on Answer Set Programming (ASP'07), Porto, Portugal, 8-13 September 2007.

- Dogac A., Namli T., Okcan A., Laleci G., Kabak Y., Eichelberg M. Key Issues of Technical Interoperability Solutions in eHealth and the RIDE Project eChallenges Conference, The Hague, The Netherlands, October 2007

- Doğandağ S. And F.N. Alpaslan, "Ontology on Semantic Web", Proc. e-Challenges 2007, The Hague, The Netherlands, 24-26 October, 2007.

- Yarimagan Y., Dogac A., Semantics Based Customization of UBL Document Schemas Journal of Distributed and Parallel Databases, Springer-Verlag, Volume 22, Numbers 2-3 / December 2007, pp. 107-131.

- G. A. SHAH, M. BOZYIGIT, Energy-efficient Real-time Coordination and Routing in Wireless Sensor and Actor Networks, ACM WiNET, 2007, submitted for publication

- G. A. SHAH, M. BOZYIGIT, D. AKSOY, Adaptive Pull-Push based Event Tracking in Wireless Sensor Actor Networks, IEEE Mobile Computing, 2007, submitted for publication

PROJE ÖZET BİLGİ FORMU

Proje Kodu : 105E068

Proje Başlığı :

Orta Doğu Teknik Üniversitesi – Bilgi Toplumu Teknolojileri – Mükemmeliyet Merkez Projesi

(Middle East Technical University - Information Society Technologies - Excellency Center)

Proje Yürütücüsü ve Yardımcı Araştırıcılar :

Prof. Dr. Asuman DOĞAÇ

Projenin Yürütüldüğü Kuruluş ve Adresi :

Yazılım Araştırma ve Geliştirme Merkezi (Software Research and Development Center)
SRDC, Bilgisayar Mühendisliği Bölümü, ODTÜ, İnönü Bulvarı,
06531 Ankara TÜRKİYE

Destekleyen Kuruluş(ların) Adı ve Adresi :

Projenin Başlangıç ve Bitiş Tarihleri : 01/05/2006 – 01/02/2008

Öz : (en çok 70 kelime)

Bilgi teknolojileri, çok hızlı bir şekilde ilerleyen teknolojik gelişmelerle birlikte öngörülemez fırsatlar yaratmaktadır. Buna paralel olarak, Türkiye'nin Avrupa seviyesindeki rekabet edebilirliği, bilgi toplumu teknolojilerini geliştirmesine ve bunları ekonomik gelişiminde kullanabilme yeteneğine bağlıdır. Bu sebeple Türkiye'de, bilgi toplumu teknolojilerinin araştırma ve teknolojik gelişimin yeteneklerinin acil bir şekilde güçlendirmesine ihtiyaç vardır.

METU-ISTEC projesi Türkiye'nin lider teknik devlet üniversitesindeki bilgi toplumu teknolojilerinin araştırma ve teknolojik gelişimi çalışmalarını yeniden yapılandırmayı amaçlamıştır.

Anahtar Kelimeler:

Türk Araştırma alanlarının Avrupa doğrultusunda entegrasyonu ve geliştirilmesi, e-İş ve e-Devlet, anlamsal-tabanlı bilgi sistemleri, kavramsal sistemler ve yazılımlarla sevisler için açık geliştirme ortamları, gömülü sistemler

Projeden Kaynaklanan Yayınlar :

1. Dogac A., Kabak Y., Namli T., Okcan A., Collaborative Business Process Support in eHealth: Integrating IHE Profiles through ebXML Business Process Specification Language, IEEE Transactions on Information Technology in Biomedicine, Accepted for publication
(Science Citation Index Core, Impact Factor: 1.575)

2. Kabak Y., Dogac A. “A Survey and Analysis of Electronic Business Document Standards”, ACM Computing Surveys, Submitted for publication.

3. G. A. SHAH, Ö. B. AKAN, M. BOZYIGIT, “Multi-Event Adaptive Clustering (MEAC) Protocol for Heterogeneous Wireless Sensor Networks”, in Proc. Fifth Annual Mediterranean Ad Hoc Networking Workshop (MedHoc-Net), also in Lecture Notes in Computer Science 4003: pp365-383, June 2006

4. G. A. SHAH, M. BOZYIGIT, DEMET AKSOY, “RAT: Routing by Adaptive Targeting in Wireless Sensor/Actor Networks”, in Proc. Second IEEE/ACM International Conference on COMmunication System softWAre and middlewaRE (COMSWARE), January 2007.

5. G. A. SHAH, M. BOZYIGIT, “Exploiting Energy-aware Spatial Correlation in Wireless Sensor Networks”, 2nd International Workshop on Software for Sensor Networks (SensorWare 2007), January 2007

6. Semih Cetin, N. Ilker Altintas, Halit Oguztuzun, Ali H. Dogru, Ozgur Tufekci and Selma Suloglu, "A Mashup-Based Strategy for Migration to Service-Oriented Computing", ICPS'07, IEEE International Conference on Pervasive Services, p.169-172, July 15-20, 2007, İstanbul

7. “Z.Sokmen, M.Ozturk, V.Atalay, R.Cetin Atalay, “A Hybrid Method For The Identification of Expression Patterns From Microarray Data” , 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2007) and 6th European Conference on Computational Biology (ECCB), Vienna, Austria, July 21-25, 2007.

8. Semih Cetin, N. Ilker Altintas, Halit Oguztuzun, Ali H. Dogru, Ozgur Tufekci and Selma Suloglu, “Legacy Migration to Service-Oriented Computing with Mashups”, ICSEA 2007, International Conference on Software Engineering Advances, 25-31 August, 2007, Cap Esterel, French Riviera, Fransa

9. Sabuncu, O. and F.N. Alpaslan, “Computing Answer Sets Using Model Generation Theorem Provers”, Proc. 4th International Workshop on Answer Set Programming (ASP’07), Porto, Portugal, 8-13 September 2007.

10. Dogac A., Namli T., Okcan A., Laleci G., Kabak Y., Eichelberg M. Key Issues of Technical Interoperability Solutions in eHealth and the RIDE Project eChallenges Conference, The Hague, The Netherlands, October 2007

11. Doğandağ S. And F.N. Alpaslan, “Ontology on Semantic Web”, Proc. e-Challenges 2007, The Hague, The Netherlands, 24-26 October, 2007.

12. Yarimagan Y., Dogac A., Semantics Based Customization of UBL Document Schemas Journal of Distributed and Parallel Databases, Springer-Verlag, Volume 22, Numbers 2-3 / December 2007, pp. 107-131.

13. G. A. SHAH, M. BOZYIGIT, Energy-efficient Real-time Coordination and Routing in Wireless Sensor and Actor Networks, ACM WiNET, 2007, submitted for publication

14. G. A. SHAH, M. BOZYIGIT, D. AKSOY, Adaptive Pull-Push based Event Tracking in Wireless Sensor Actor Networks, IEEE Mobile Computing, 2007, submitted for publication

Bilim Dalı :
Doçentlik B. Dalı Kodu :

Collaborative Business Process Support in eHealth: Integrating IHE Profiles through ebXML Business Process Specification Language

Asuman Dogac *Member, IEEE,* Yildiray Kabak, Tuncay Namli, Alper Okcan

***Abstract*—IHE specifies Integration Profiles describing selected real-world use cases to facilitate the interoperability of healthcare information resources. While realizing a complex real world scenario, IHE Profiles are combined by grouping the related IHE actors. Grouping IHE Actors implies that the associated business processes (IHE Profiles) that the Actors are involved must be combined, that is, the choreography of the resulting collaborative business process must be determined by deciding on the execution sequence of transactions coming from different profiles. There are many IHE profiles and each user or vendor may support a different set of IHE Profiles that fits to its business need. However, determining the precedence of all the involved transaction manually for each possible combination of the profiles is a very tedious task.**

In this paper, we describe how to obtain the overall business process automatically when IHE Actors are grouped. For this purpose, we represent the IHE Profiles through a standard, machine processable language, namely, OASIS ebXML Business Process Specification (ebBP) Language. We define the precedence rules among the transactions of the IHE Profiles, again, in a machine processable way. Then through a graphical tool, we allow users to select the actors to be grouped and automatically produce the overall business process in a machine processable format.

***Index Terms*—Clinical and Health Care Information Systems, Collaborative eHealth Processes, Integrating Healthcare Enterprise (IHE), Grouping IHE Actors**

I. Introduction

Standards are necessary both for integration and for interoperability. However, any actual implementation of a standard requires some form of tailoring. Therefore, in developing practical and effective interoperability solutions, the industry relies on integration profiles which are business processes describing selected real-world use-cases. An important industry initiative, Integrating Healthcare Enterprise (IHE) [1], has taken this profiling approach to achieve interoperability in the eHealth domain.

IHE is a not-for-profit initiative founded by the Radiological Society of North America (RSNA) [2] and the Healthcare Information and Management Systems Society (HIMSS) [3] and now supported by a wide range of healthcare professional societies world-wide. Through the IHE Profiles, the interactions between the IT systems in healthcare are described and the details of interfaces are fixed based on the standards. The approach taken in developing the integration profiles is first to define the basic transactions describing the interactions between the IT systems and then to define the workflows describing the real life business processes by using these transactions together with the standard interfaces.

Since IHE Profiles describe the specific use cases, there is a need to combine more than one IHE Profile to achieve the required functionality in realizing a real world scenario. For example, one profile specifies how patient clinical data can be shared among different healthcare enterprises. This profile called, *IHE Cross-Enterprise Document Sharing (IHE-XDS) Profile* [4], allows Electronic Healthcare Records (EHRs) to be shared through a common ebXML Registry/Repository architecture. Another profile specifies how patient identifiers used in different healthcare institutes should be mapped to each other. This profile, called *Patient Identifier Cross-referencing Integration (IHE-PIX) Profile* [5], supports the cross-referencing of patient identifiers from multiple patient identifier domains. It is clear that to be able to share patient data created by different applications through IHE-XDS, it is also necessary to map the patient identifiers used by these applications to each other. Therefore, most of the time, it is necessary to use the IHE-XDS Profile together with the IHE-PIX Profile. IHE provides for this by grouping the relevant actors of the involved profiles.

When IHE Profiles are combined by grouping the relevant IHE Actors, the result is a collaborative healthcare business process integrating the workflows defined for each profile and the sequence of the transactions coming from different profiles must be decided.

However, given the large number of IHE Profiles (currently over sixty profiles), determining the precedence of all the involved transaction manually for each possible combination of the profiles is a very a tedious task. Note that each user or vendor may support a different set of IHE Profiles as they choose and hence the overall business process for each user or vendor may be different. And secondly, the overall business process must be described in a machine processable way to be used by the collaborating applications.

In this paper, we describe how to obtain the overall machine processable multi party collaborative business process automatically when IHE Actors are grouped. For this purpose, we first describe IHE Profiles in OASIS ebXML Business Process Specification Language (ebBP) [6]. Then we show that by representing the IHE Profiles through directed graphs and by using the rules giving the precedence information among

This work is supported by the European Commission, Project No: IST-027065 RIDE and in part by the Scientific and Technical Research Council of Turkey (TÜBİTAK), Project No: EEEAG 105E068 and is realized as a proposal to the OASIS ebXML Business Process Technical Committee.

the IHE Transactions, it is possible to automatically order the transactions in the overall business process when the Actors are grouped. The graphical tool we provide based on this mechanism produces the overall business process automatically when a user selects the IHE Actors to be grouped.

The paper is organized as follows: In Section II, we briefly summarize the related standards. Section III describes how IHE Profiles can be represented through ebBP. In Section IV, we present the details of how the overall business process can be automatically obtained when IHE Profiles are combined by grouping IHE Actors. Section V gives the implementation status of the tool developed. Section VI describes the related work. Finally, Section VII concludes the paper and presents the future work.

II. An Introduction to the Related Standards

In this section, we present a brief introduction to the standards and frameworks used in this work, namely, the OASIS ebXML Business Process Specification Language (ebBP) and the IHE Integration Profiles.

A. OASIS ebXML Business Process Specification Language: ebBP

The eBusiness eXtensible Markup Language (ebXML) Business Process Specification Schema (ebBP) technical specification defines a standard, machine processable language by which business systems can be configured [6].

A *Business Collaboration* consists of a set of roles that collaborate by exchanging *Business Documents*. The roles and the documents they exchange are defined in *Business Transactions*. The roles in *Business Transactions* are always generic and labelled as *Requesting* and *Responding* roles. When a *Business Transaction* definition is used for a specific purpose in a *Business Collaboration*, it becomes a *Business Transaction Activity*. The specific roles (e.g. DocumentSource, DocumentConsumer) are specified at the *Business Transaction Activity* level. Similarly when *Business Collaborations* are used for a specific purpose, they are termed as *Collaboration Activity*. A *Business Collaboration* is defined as a choreography of *Business Transaction Activities* and/or *Collaboration Activities*. The purpose of a *Choreography* is to specify which *Business Transaction Activity* and/or *Collaboration Activity* should happen. There are a number of *States* that facilitate the choreographing of *Business Activities*. These include a "Start" state and a "Completion" state (which is either a "Success" or a "Failure") as well as a series of gateways: a "Fork gateway", a "Join gateway" and a "Decision gateway". There are two types of "Fork gateway": "OR" and "XOR" [6].

The ebBP defines a *Business Transaction* as an abstract entity and provides six concrete business patterns to be used in real life business collaborations. Each of the concrete business patterns describes whether a response document is required and which business signals are required in a specific interaction:

- *Commercial Transaction (previously represented as Business Transaction)*: This pattern defines a formal obligation between parties and requires a response, a request receipt acknowledgement and a response receipt acknowledgement or an exception.
- *Notification*: The pattern is used for business notifications such as a *Notification of Failure* and it does not require a response document.
- *Information Distribution*: The pattern represents an informal information exchange between parties.
- *Query/Response*: This pattern is used by a requester for an information query of which the responding party already has. In this pattern, on the receiver side there is no backend processing; otherwise Request/Response pattern should be used.
- *Request/Response*: This pattern is used when an initiating party requests information that a responding party already has and when the request for business information requires a complex interdependent set of results and backend processing.
- *Request/Confirm*: The pattern is used when an initiating party requests confirmation about its status with respect to previous obligations or a responder's business rules.

B. IHE Integration Profiles

IHE Integration Profiles are business processes describing selected real-world use-cases. Each IT system or application involved in the use cases is called an "Actor". The interactions between IHE Actors are defined through "Transactions". More specifically, IHE Transactions define how IT systems or applications communicate by using existing standards such as HL7 [7] or DICOM [8] to accomplish a specific task. In this way, IHE Integration Profiles define a collection of business processes.

Currently there are about sixty IHE Profiles addressing different use cases in the healthcare IT domain including how to share patient clinical information (IHE-XDS), how to map patient identifiers from one domain into another (IHE-PIX) and how to provide authentication and audit trail (IHE-ATNA) [9]. IHE continuously defines more profiles to address the remaining integration issues in the healthcare IT domain.

In the following section, IHE Cross-Enterprise Document Sharing (XDS) Profile is briefly summarized to provide an insight to the IHE Profiles.

C. IHE Cross-Enterprise Document Sharing (XDS) Profile

In the IHE XDS Profile, the repository is used for storing the clinical documents and the related metadata stored at the registry is used to facilitate the discovery of the documents.

In this profile, the group of healthcare enterprises that agree to work together for clinical document sharing is called an "XDS Affinity Domain". Such enterprises agree on a common set of policies such as how the consent is obtained, the access is controlled, and the common set of coding terms to represent the metadata of the documents. The metadata defined is used for searching the registry to locate the documents in the repository. The IHE XDS Profile Actors and Transactions are depicted in Figure 1:

1) "Patient Identity Source" Actor provides the patient identifier to the "Document Registry" Actor with *Patient Identity Feed (IHE-ITI-8)* Transaction.

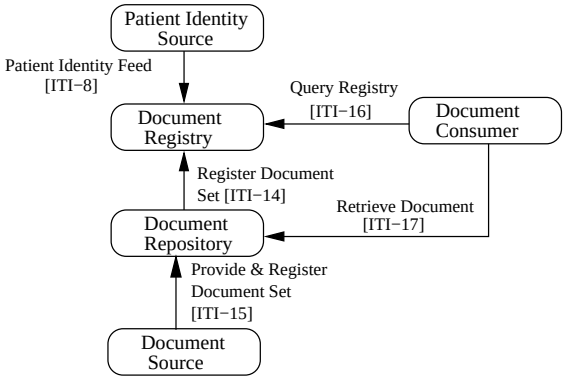


Fig. 1. IHE XDS Profile

TABLE I
THE CORRESPONDANCE OF IHE XDS TRANSACTIONS TO EBBP BUSINESS TRANSACTION PATTERNS

IHE Concepts	ebBP Concepts
IHE Actor	ebBP Role
IHE Transaction (generic)	ebBP Business Transaction
IHE Transaction (specific to a Profile)	ebBP Business Collaboration
IHE Message, Document	ebBP Business Document
IHE Sequence Diagram	ebBP Choreography
IHE Profile	ebBP Business Process

2) The "Document Source" Actor, by using the *Provide and Register Document Set (IHE-ITI 15)* Transaction, sends the documents to the "Document Repository" Actor.
3) The "Document Repository" Actor stores these documents at the persistent storage and sends the metadata of these documents to appropriate "Document Registry" Actor by using *Register Document Set (IHE ITI-14)* Transaction.
4) The "Document Registry" Actor maintains metadata about each registered document in a document entry. This includes a link to the document in the Repository where it is stored.
5) The "Document Consumer" Actor queries a "Document Registry" Actor for documents meeting certain criteria on metadata by using *Query Registry (IHE-ITI 16)* Transaction, and retrieves selected documents from one or more "Document Repository" Actors through *Retrieve Document (IHE-ITI 17)* Transaction.

III. REPRESENTING IHE PROFILES THROUGH EBBP

ebBP defines generic business transaction patterns and specifies how to organize them into business processes in a standard and machine processable way. IHE Profiles, on the other hand, define healthcare domain specific business transactions and specify some of the business processes in the healthcare domain. The main advantage of representing IHE Profiles through ebBP is that, in this way IHE Profiles become machine processable in a standard way and this in turn introduces a new ability to combine various healthcare processes automatically.

The correspondence between the ebBP concepts and the IHE concepts are shown in Table I. IHE Actors correspond to ebBP Roles. A "generic" IHE Transaction, that is, an IHE Transaction that is not bound to an IHE Profile yet, is represented through ebBP Business Transaction. When an IHE Transaction is bound to an IHE Profile through a Business Collaboration, the concrete "Requesting Role" and the "Responding Role" as well as the choreography of the transactions are determined. The IHE messages and the documents used in a Business Transaction correspond to ebBP Business Documents. IHE Profile Sequence Diagrams determine the choreography of the related Business Collaboration.

TABLE II
CORRESPONDENCE OF IHE XDS TRANSACTIONS TO EBBP BUSINESS TRANSACTION PATTERNS

IHE Transaction	ebBP Pattern	Requesting Role	Responding Role	Business Document
Provide/Register Document Set	Business Transaction	Document Source	Document Repository	Document, Document Metadata
Query Registry	Query/ Response	Document Consumer	Document Registry	XDS Query
Retrieve Document	Request/ Response	Document Consumer	Document Repository	URI
Register Document Set	Request/ Confirm	Document Repository	Document Registry	Document Metadata, Document Metadata Ack.
Admit/Register or Update Patient (Patient Identity Feed)	Information Distribution	Patient Identity Source	Document Registry, Patient Identity Cross-reference Manager	HL7 ADT A01/ A04/ A05

The following steps are involved in representing the IHE Profiles in ebBP:

- We map the IHE transactions to the ebBP transaction patterns. ebBP defines six concrete business transaction patterns with well defined semantics and we demonstrate how to represent the IHE Transactions through the ebBP Transaction patterns.
- Once IHE Transactions are defined through ebBP, we show how to specialize them to IHE Business Collaborations. When an IHE Transaction is bound to an IHE Profile through a Business Collaboration, the concrete "Requesting Role" and the "Responding Role" as well as the choreography of the transactions are determined. For example, the generic "Patient Identity Feed" Transaction [IHE-ITI-2] between two generic roles such as "Patient Identity Supplier" and "Patient Identity Receiver" can be specialized to "Patient Identity Feed Transaction in IHE XDS" or to "Patient Identity Feed Transaction in IHE PIX" by setting the concrete roles accordingly, that is, for the PIX Profile the "Patient Identity Receiver" is set to the "PIX Manager Actor"; for the XDS Profile it is set to the "Document Registry Actor".
- The related Business Collaborations are combined to make up the IHE Profile. At this step, it may be necessary to introduce new business collaborations by using the previously defined Business Collaborations to be able to express further interactions among them as explained in Section III-C.

A. Defining IHE Transactions through ebBP Business Transaction Patterns

ebBP patterns describe certain generic semantics as described in Section II-A. For example, the "Commercial Transaction" (also termed as the "Business Transaction") defines a formal obligation between parties and requires a response,

a request receipt acknowledgement and a response receipt acknowledgement or an exception. This pattern matches with the “Provide/Register Document Set (IHE-ITI-15)” Transaction semantics. As another example, the “IHE Query Registry (IHE-ITI-16)” Transaction is a specialization of the “ebBP Query/Response Transaction Pattern” since the “Document Consumer” is in need of information that the Registry may have. In Table II, the correspondences between IHE-XDS Transactions and ebBP patterns are given.

In the following, we provide an example on how an IHE Transaction can be represented through ebBP. The ebBP definitions of other IHE Transactions are available from our Web site [10]. “Patient Identity Feed Transaction (IHE-ITI-8)” communicates patient information, including corroborating demographic data, after a patients identity is established, modified or merged.

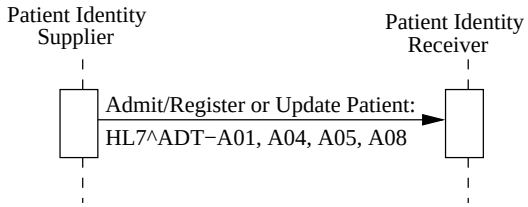


Fig. 2. “Patient Identity Feed Transaction (IHE-ITI-8)” Sequence Diagram

The IHE Sequence Diagram for “Patient Identity Feed Transaction (IHE-ITI-8)” is shown in Figure 2. In this figure, the “Patient Identity Supplier” Actor provides the patient identifier when one of the following events occur: creation, update or merge. The messages used for this purpose are HL7 Admit Patient (ADT) messages such as A01 for admission of an in-patient into a facility or A04 for registration of an outpatient for a visit of the facility.

Figure 3 gives the code listing of the “IHE Patient Identity Feed” transaction in ebBP. This transaction is a specialization of ebBP Notification pattern since it does not require a response. It defines the generic “Requesting Role” and “Responding Role” as “PIDFeedBT-Initiator” and “PIDFeedBT-Responder” respectively. The RequestingBusinessActivity defines the business document reference as “HL7-2.3.1-ADT-Document”. Furthermore, the necessary signals for state alignment are defined.

B. Defining IHE Business Collaborations through ebBP Business Collaborations

An IHE Business Collaboration is obtained by specializing a generic IHE Transaction to a specific IHE Profile that it takes part. Furthermore, the choreography (the ordering and transitions between Business Transactions) of the collaboration is described through ebBP standard choreography constructs such as “Start”, “Transition”, “ToLink” and “FromLink”.

For example, the IHE “Patient Identity Feed” Transaction is specialized to the IHE PIX Profile by setting the “PIDFeedBT-Initiator” role defined in “Patient Identity Feed” Transaction (Figure 3) to “PIDSource” and the “PIDFeedBT-Responder” role to “PIXManager” as shown in [11]. Furthermore, the

```
<Notification nameID="IHE-ITI-8" name="Patient Identity
Feed Business Transaction"
        isGuaranteedDeliveryRequired="true">
  <RequestingRole nameID="PIDFeedBT-Initiator"
  name="Initiator"/>
  <RespondingRole nameID="PIDFeedBT-Responder"
  name="Responder"/>
  <RequestingBusinessActivity
    nameID="PIDFeedBT-RBA"
    name="Patient Identity Feed Requesting Business
    Activity" isAuthorizationRequired="true"
    isNonRepudiationRequired="true">
    <DocumentEnvelope
      nameID="HL7-v2.3.1-ADT"
      businessDocumentRef="HL7-2.3.1-ADT-Document"
      name="HL7 v2.3.1 ADT Message"
      isAuthenticated="persistent"
      isConfidential="persistent"/>
    <ReceiptAcknowledgement name="ra2" nameID=
    "PIDFeedBT-RBA-RA" signalDefinitionRef="ra2"/>
    <ReceiptAcknowledgementException name="rae2" nameID=
    "PIDFeedBT-RBA-RAE" signalDefinitionRef="rae2"/>
    <AcceptanceAcknowledgement name="aa2" nameID=
    "PIDFeedBT-RBA-AA" signalDefinitionRef="aa2"/>
    <AcceptanceAcknowledgementException name="aae2"
    nameID="PIDFeedBT-RBA-AAE" signalDefinitionRef=
    "aae2"/>
  </RequestingBusinessActivity>
  <RespondingBusinessActivity nameID="PIDFeedBT-ResBA"
    name="Patient Identity Feed Responding Business
    Activity">
  </RespondingBusinessActivity>
</Notification>
```

Fig. 3. The Code Segment Defining IHE-ITI-8 through ebBP “Notification” Transaction Pattern

collaboration defined in [11] references the “Patient Identity Feed” Transaction (IHE-ITI-8) defined in Figure 3 through “businessTransactionRef” element. As already mentioned, the choreography of the collaboration is described through ebBP standard choreography constructs. For example, in [11], after the start state, a link to the “PIDFeedBTA” activity is specified and the transitions from this state to the two possible states (Success and Failure) are stated. Failure state is omitted from the code listing for the sake of simplicity.

C. Defining IHE Profiles through ebBP Business Processes

In defining the IHE Profiles through ebBP, previously defined collaborations are reused and new collaborations are introduced as needed. For example, in IHE XDS there are collaborations among the “Document Source”, the “Document Repository” and the “Document Registry” Actors: when a “Document Repository” Actor receives a “Document set” from a “Document Source”, it stores this document set at the persistent storage and sends its metadata to the appropriate “Document Registry” by using the “IHE Register Document Set” Transaction. In order to express this choreography between three IHE Actors, it is necessary to introduce a new business collaboration by using the previously defined Business Collaborations between two IHE Actors as shown in Figure 4. The previously defined Business Collaborations between two IHE Actors, that is “XDS-IHE-ITI-15” and “XDS-IHE-ITI-14” are re-used by providing collaboration references to them and the additional choreography is defined by stating the transitions, first from “CA-IHE-ITI-15” to “CA-IHE-ITI-14”, and then from “CA-IHE-ITI-14” to “ITI14-15-Success”. Failure state is omitted from the code listing for the sake of simplicity.

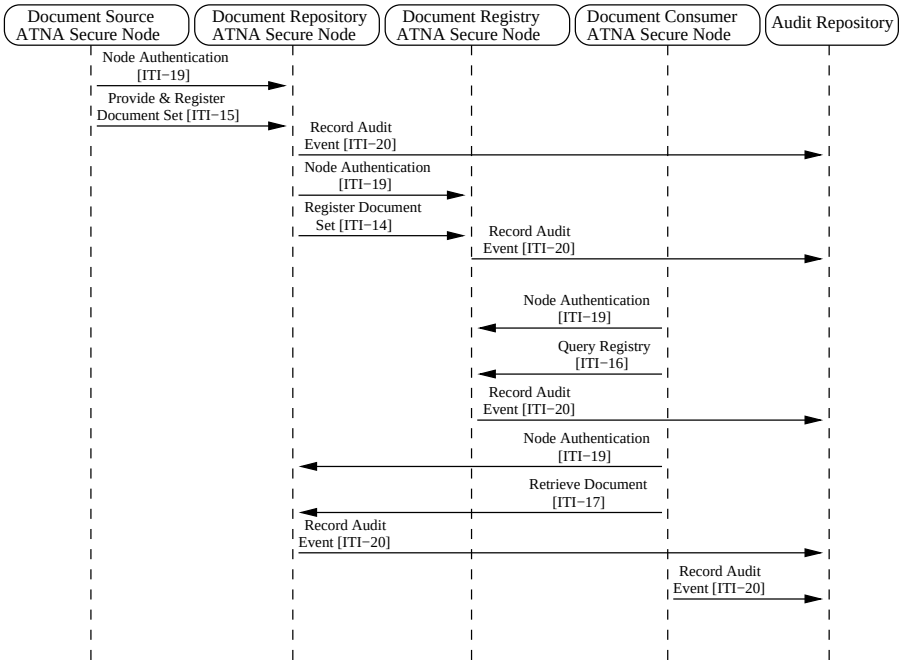


Fig. 5. The Sequence Diagram Showing the Ordering of Transactions after Combining IHE XDS and IHE ATNA Profiles

```
<BusinessCollaboration name="Collaboration of IHE
ITI 14 and IHE ITI 15" nameID="IHE-ITI-14-15">
   <Role name="Document Source" nameID=
   "ITI14-15-XDSSource"/>
   <Role name="Document Repository" nameID=
   "ITI14-15-XDSRepository"/>
   <Role name="Document Registry" nameID=
   "ITI14-15-XDSRegistry"/>
   <TimeToPerform/>
   <Start>
      <ToLink toBusinessStateRef="CA-IHE-ITI-15"/>
   </Start>
   <CollaborationActivity name="IHE ITI 15" nameID
   ="CA-IHE-ITI-15" collaborationRef="XDS-IHE-ITI-15">
      <Performs currentRoleRef="ITI14-15-XDSSource"
      performsRoleRef="PRDSXDSSource"/>
      <Performs currentRoleRef="ITI14-15-XDSRepository"
      performsRoleRef="PRDSXDSRepository"/>
   </CollaborationActivity>
   <Transition>
      <FromLink fromBusinessStateRef="CA-IHE-ITI-15"/>
      <ToLink toBusinessStateRef="CA-IHE-ITI-14"/>
   </Transition>
   <CollaborationActivity name="IHE ITI 14" nameID=
   "CA-IHE-ITI-14" collaborationRef="XDS-IHE-ITI-14">
      <Performs currentRoleRef="ITI14-15-XDSRepository"
      performsRoleRef="RDSXDSRepository"/>
      <Performs currentRoleRef="ITI14-15-XDSRegistry"
      performsRoleRef="RDSXDSRegistry"/>
   </CollaborationActivity>
   <Transition>
      <FromLink fromBusinessStateRef="CA-IHE-ITI-14"/>
      <ToLink toBusinessStateRef="ITI14-15-Success"/>
   </Transition>
   <Success name="Success" nameID="ITI14-15-Success"/>
</BusinessCollaboration>
```

Fig. 4. The ebBP Code Defining IHE-XDS Profile

The code listings of referenced Business Collaborations are available at [10].

1. {⟨XDSDocumentSource, XDSDocumentRepository, XDS-ITI-15⟩ precedes ⟨XDSDocumentRepository, XDSDocumentRegistry, XDS-ITI-14⟩}
2. {⟨XDSDocumentConsumer, XDSDocumentRegistry, XDS-ITI-16⟩ precedes ⟨XDSDocumentConsumer, XDSDocumentRepository, XDS-ITI-17⟩}

Fig. 6. Some Example Profile Precedence Rules for Ordering IHE XDS Transactions

IV. Collaborative IHE Profiles by Grouping IHE Actors

There is a need to automatically generate "custom" design business processes, integrating any possible existing or forthcoming IHE profiles according to user needs.

We developed a software tool for ordering transactions automatically for the grouped IHE Actors, called IHE Actor Grouping Tool (IHE-AGT). The software is based on the algorithm described in this section. The notations and the data structures used in the algorithm are given in Table III and Table IV, respectively.

TABLE III
Notations used in the Algorithm

Notation	Description
$A=\{a_1, a_2, .. , a_n\}$	List of actors in the profile
$T=\{t_1, t_2, .. , t_m\}$	List of transactions in the profile
$G=\{g_1, g_2, .. , g_l\}$	List of graphs in the profile
$PR=\{pr_1, pr_2, .. , pr_k\}$	List of profile rules
$GR=\{gr_1, gr_2, .. , gr_j\}$	List of grouping rules

The first requirement for automation is to have a machine processable definition of IHE Profiles which is described in Section III by defining IHE Profiles through ebBP. The next step is to be able to decide on the execution sequence of transactions coming from different profiles.

TABLE IV
DATA STRUCTURES USED IN THE ALGORITHM

Structure	Description
t_i	Each transaction is a quadraple <name, initiator, responder, seqnum> - *name* is the name of the xAnct, - *initiator* is the initiator actor, - *responder* is the responder actor, - *seqnum* shows the sequence number of the transaction in a collaboration. It is initially set to 1.
g_i	Each graph is a two-dimensional NxN array (i.e. matrix) whose elements point to a transaction list. The indices (i.e. vertices) of the graph are actors. The entries in the graph are a list of transactions (i.e. edges).
pr_i	Each profile rule is a list of transactions. The sequence of the transactions shows their expected order in the profile.
gr_i	Each grouping rule is a tuple of the form <left, right>. - *left:* is a tuple of actors - *right:* is a tuple of xAncts of the form <first precedes next>. This means, if actors in *left* are grouped, then *first* must precede *next*.

1. [⟨XDSDocumentSource, ATNASecureNode⟩,
   ⟨⟨XDSDocumentSource, XDSDocumentRepository, ATNA-ITI-19⟩ precedes ⟨XDSDocumentSource, XDSDocumentRepository, XDS-ITI-15⟩⟩]
2. [⟨XDSDocumentRepository, ATNASecureNode⟩,
   ⟨⟨XDSDocumentSource, XDSDocumentRepository, XDS-ITI-15⟩ precedes ⟨XDSDocumentRepository, AuditRepository, ATNA-ITI-20⟩⟩]
3. [⟨XDSDocumentRepository, ATNASecureNode⟩,
   ⟨⟨XDSDocumentRepository, XDSDocumentRegistry, ATNA-ITI-19⟩ precedes ⟨XDSDocumentRepository, XDSDocumentRegistry, XDS-ITI-14⟩⟩]
4. [⟨XDSDocumentRegistry, ATNASecureNode⟩,
   ⟨⟨XDSDocumentRepository, XDSDocumentRegistry, XDS-ITI-14⟩ precedes ⟨XDSDocumentRegistry, AuditRepository, ATNA-ITI-20⟩⟩]
5. [⟨XDSDocumentRegistry, ATNASecureNode⟩,
   ⟨⟨XDSDocumentConsumer, XDSDocumentRegistry, XDS-ITI-16⟩ precedes ⟨XDSDocumentRegistry, AuditRepository, ATNA-ITI-20⟩⟩]
6. [⟨XDSDocumentConsumer, ATNASecureNode⟩,
   ⟨⟨XDSDocumentConsumer, XDSDocumentRegistry, ATNA-ITI-19⟩ precedes ⟨XDSDocumentConsumer, XDSDocumentRegistry, XDS-ITI-16⟩⟩]

Fig. 7. Some Example Grouping Precedence Rules for Grouping Actors across IHE Profiles

To automate this process, we make use of the precedence rules in the IHE Profiles. There are two types of precedence rules: first, there is a precedence among transactions within an IHE Profile, which we call "Profile Precedence" rules. For example, in IHE XDS, the transaction XDS-ITI-15 "XDSProvide/RegisterDocumentSet" must precede the transaction XDS-ITI-14 "XDSRegisterDocumentSet". Secondly, when two Actors from different profiles are grouped, there is a precedence among their transactions, which we call "Grouping Precedence" rules. As an example, in IHE XDS, there is a need for node authentication and audit trail to provide patient information confidentiality and user accountability. To achieve this, IHE Audit Trail and Node Authentication (ATNA) Integration Profile is used together with IHE XDS. To make these two profiles work together, ATNA Secure Node Actor is grouped with each of the following IHE-XDS Actors: XDS Document Source, XDS Document Repository, XDS Document Registry and XDS Document Consumer.

When these Actors are grouped, the resulting sequence of Transactions is depicted in Figure 5. For example, "Record Audit Event" Transaction (IHE-ITI-20) of ATNA appears between "Provide and Register Document Set" Transaction (IHE-ITI-15) and "Register Document Set" (IHE-ITI-14) of XDS in the figure.

For expressing these rules, we have developed the following notation:
<InitiatingActor1, RespondingActor1, "IHE-Transaction1"> precedes
<InitiatingActor2, RespondingActor2, "IHE-Transaction2"> which means that the "IHE-Transaction1" between InitiatingActor1, and RespondingActor1 precedes "IHE-Transaction2" between InitiatingActor2 and RespondingActor2.

In Figure 6, some example precedence rules among the transactions within the IHE XDS Profile are given. Some example rules for grouping Actors across IHE Profiles are given in Figure 7. To automate processing, these rules are expressed in the Rule Markup Language (RuleML) [12].

The actor grouping algorithm is a graph based algorithm where each profile is represented as a directed graph. In the graph, the vertices are the IHE actors and the edges are the IHE transactions. A directed edge between two vertices shows a transaction whose initiator actor is the source vertex and responder actor is target vertex. In the algorithm, the graphs are represented with two-dimensional NxN arrays whose elements point to a list of transactions. The vertices in the graph (actors) are represented by the indices of the array and the edges (transactions) are represented by the entries of the array. For example, g[a_1][a_2] returns a transaction list (from actor a_1 to a_2) which shows the edges from vertex a_1 to vertex a_2.

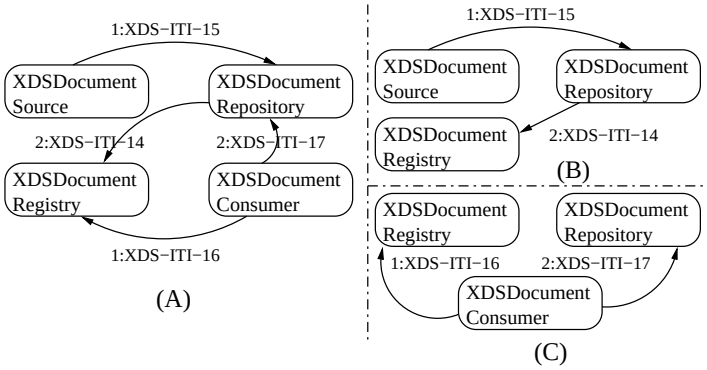


Fig. 8. A. Graph Representation of IHE-XDS, B. A subgrapgh of IHE-XDS, C. Another Subgraph of IHE-XDS

The algorithm gets three types of input: (1) the list of actors, (2) the list of transactions and (3) the precedence rules in the IHE Profiles.

The algorithm has mainly three phases: Construction phase, Partitioning phase and Merging Phase. In the Construction phase, the transaction list is iterated and each transaction is added to the graph by assigning its label indicating its execution sequence (termed as "sequence number" in the algorithm). For assigning the sequence numbers, the "Profile Precedence" rules are used. At the end of this phase the overall profile graph is produced. For example, considering the IHE XDS Profile, the directed graph in Figure 8 A is the output of the construction phase where XDS-ITI-15 has the sequence number "1" and XDS-ITI-14 has the sequence number "2"

```
1.forall g ∈ G do begin
2.   forall gr ∈ GR do begin
3.      firstXanct = gr.right.first;
4.      nextXanct = gr.right.next;
5.      // if the graph contains the first transaction
6.      if(g.contains(firstXanct)) do begin
7.      // indexOf function returns the seqnum of the specified transaction in the graph
8.        index = g.indexOf(firstXanct);
9.        if(firstXanct.initiator == nextXanct.initiator and
          firstXanct.responder == nextXanct.responder)
10.       do begin
11.         edgeList = g[firstXanct.initiator][firstXanct.responder];
12.         edgeList.add(nextXanct);
13.       else
14.         g.addColumn(nextXanct.responder);
15.         g.addRow(nextXanct.responder);
16.         g[nextXanct.initiator][nextXanct.responder].add(nextXanct);
17.       end
18.       // updateIndicesInGraph function increases the seqnum of all transactions
19.       // whose seqnums are greater than or equal to second argument by one
20.       updateIndicesInGraph(g, index+1);
21.       nextXanct.seqnum = index+1;
22.     else
23.       index = g.indexOf(nextXanct);
24.       if(nextXanct.initiator == firstXanct.initiator and
          nextXanct.responder == firstXanct.responder)
25.       do begin
26.         edgeList = g[nextXanct.initiator][nextXanct.responder];
27.         edgeList.add(firstXanct);
28.       else
29.         // addColumn function adds a new column to the graph if there is not.
30.         g.addColumn(firstXanct.responder);
31.         // addRow function adds a new row to the graph if there is not such a row.
32.         g.addRow(firstXanct.responder);
33.         g[firstXanct.initiator][firstXanct.responder].add(firstXanct);
34.       end
35.       updateIndicesInGraph(g, index);
36.       firstXanct.seqnum = index;
37.     end
38.  end
39.end
```

Fig. 9. The Merging Phase

indicating their execution sequence.

Once the profile graph is constructed, connected subcomponents of the graph are formed in the partitioning phase. In other words, the graph is split into subgraphs each of which is a distinct business collaboration among IHE actors. The logic behind this step is to find out the transactions whose sequences depend on each other. For example, if you take "XDSDocumentConsumer" as a starting node, you obtain the graph shown in Figure 8 C. The pseudocodes of construction and partitioning phases are available at [13].

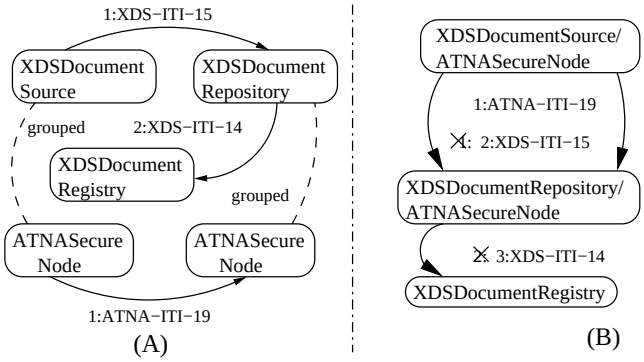


Fig. 10. A. An Example of Grouping IHE Actors, B. An Example of IHE Actor Merging Process

The partitioning phase outputs a set of subgraphs, called *G*.

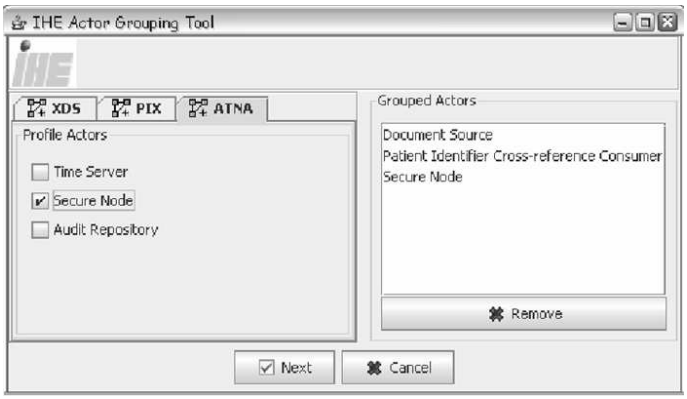


Fig. 11. IHE Actor Grouping Tool

In the last phase, the "Grouping Precedence" rules are applied to each subgraph in *G*. The algorithm applied in the merging phase is depicted in Figure 9. It incorporates the missing transactions in the rules to the graph. For example, assume the following rule exists in *GR* (Table III): $< <a_1, a_2> \text{ precedes } <t_1, t_2> >$, which means that if a_1 is grouped with a_2 then in the process t_1 should precede t_2. Assume further that a_1 already exists in the graph. Then, the algorithm adds a_2 to the graph and adjust the "sequence numbers" of the transactions accordingly. The transaction existency check is performed at Line 6 of the algorithm and the sequence numbers are adjusted at Lines 20-21 and 35-36. Continuing with the previous example, in Figure 10 A, "XDSDocumentSource" is grouped with "ATNASecureNode" and "XDSDocumentRepository" is grouped with "ATNASecureNode". Figure 10 B shows the result of this grouping where "XDSDocumentSource" and "ATNASecureNode" are merged and "XDSDocumentRepository" is merged with "ATNASecureNode". From the rule: 'When "XDSDocumentSource" is grouped with "ATNASecureNode", <XDSDocumentSource, XDSDocumentRepository, "ATNA-ITI-19"> precedes <XDSDocumentSource, XDSDocumentRepository, "XDS-ITI-15">'; it follows that ATNA-ITI-19 must precede XDS-ITI-15 and hence the label on the graph are adjusted as shown in Figure 10 B.

V. Implementation Status

The IHE Actor Grouping Tool (IHE-AGT) is developed using Java Programming Language (version 1.5.0) and the installation of the tool is realized with Jakarta Ant [14] project build tool.

Figure 11 shows a snapshot from the tool where IHE-XDS Document Source, IHE-PIX Patient Identifier Cross-reference Consumer and IHE-ATNA Secure Node are grouped. As shown in Figure 11, the user selects the actors to be grouped by clicking the checkboxes. After pressing to the "Next" button, the IHE transactions in the profiles are displayed as shown in Figure 12 and the user enters the URLs of the corresponding ebBP Business Collaboration definitions. Note that the definitions of most of the IHE Business Collaborations are available from our Web site [10]. Finally, the user presses to the "Group" button (Figure 12) in order to generate the overall business

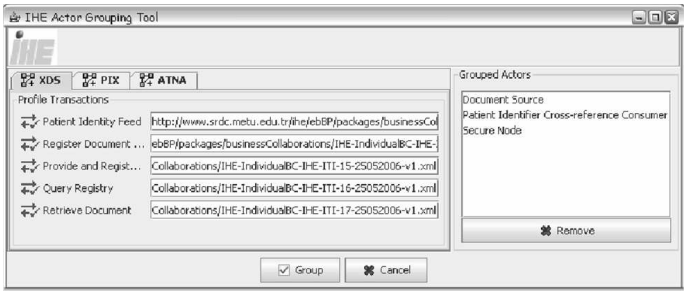


Fig. 12. Specification of the corresponding ebBP Business Collaborations

process definition. The business process definition generated is given at [15].

The IHE-AGT tool has an extensible architecture by considering the fact that IHE profiles are being specified continuously. As new profiles emerge, they can be incorporated into the tool easily by editing the configuration files. The IHE-AGT keeps the following information in its configuration files:

- The names of the IHE Actors
- The names of the IHE Transactions along with the URL references to the corresponding Business Collaboration definitions in ebBP
- The rules for grouping Actors within an IHE Profile
- The rules for grouping Actors across IHE Profiles

VI. Related Work

In the healthcare domain, there is a wide array of shared care delivery collaborative processes such as the placing and tracking of orders (e.g. drug prescriptions, radiology orders, etc.). Parts of these processes are executed by workflows running in different departments of a healthcare institute. These intra-departmental workflows are usually implemented either by proprietary applications or more recently by the composition of medical e-services [17] through the use of Business Process Execution Language (BPEL) [18]. For example, [17] addresses how IHE Scheduled Workflow actors and transactions are defined as BPEL processes. In order to define each departmental workflow in BPEL, first the UML activity diagrams are used as a representation language of the overall process among the participants. The authors further demonstrate the use of Web service technology for modelling the medical workflows such as encoding HL7 and DICOM messages, data and service identification, coordination, transaction, security and composition.

However, a BPEL process provides only the view point of a single participant. For collaborative, inter-departmental or inter-enterprise healthcare business processes, the choreography of multi-party interactions is necessary. To express such multi-party collaboration among departments in an enterprise or between enterprises, a collaborative business process language like ebXML Business Process Specification Language (ebBP) [6] is needed.

UN/CEFACT's modelling methodology (UMM) can also be used to develop global choreographies of inter-organizational business processes. In [19], how UMM models can be mapped to the ebXML Registry information model is described. Furthermore, various aspects of business process management are addressed in [20], [21] and an in depth discussion of IHE Profiles is available in [22].

In [23], we present our initial ideas on how IHE Workflow Profiles can be integrated to IHE XDS by using ebXML Business Processes (ebBP) through an example scenario. The work described in [23] is complemented in this paper by describing the methodology to express IHE Profiles through ebBP Business Transaction Patterns and more importantly by describing how to automate the overall collaborative process through IHE Actor Grouping Tool.

Finally, in [24], we describe a normative specification of how IHE Profiles should be represented through ebXML Business Processes (ebBP). The work described in this paper complements [24] by providing scientific and technical details of this standard proposal.

VII. Conclusions

IHE is an important initiative strongly supported by the healthcare IT industry: more than 160 companies, including most of the market leaders have developed IHE compliant systems between 1999 and 2007 and participated in the cross-vendor testing events organized by IHE. This means that standards recommended by IHE have a high probability of a quick uptake in the medical IT market.

IHE has defined a number of profiles to provide interoperability among healthcare applications. In realizing a real world scenario, there is a need to combine several profiles by "grouping" the related IHE Actors to achieve the required functionality. Grouping IHE Actors involves determining the multiparty collaboration choreography of the resulting business process.

In this paper, we describe a graphical tool to define collaborative business processes by combining IHE Profiles through ebBP. ebBP is designed to specify business processes in a generic way and in this work we specialize it to the healthcare IT domain by expressing IHE Profiles in ebBP. In this way, IHE defined business processes become machine processable. Then, by using the graphical tool that we developed, entitled "IHE Actor Grouping Tool", it becomes possible to group IHE Actors to generate the overall business process automatically. Clearly this helps to automate the business processes in actual healthcare settings as well as for "workflow" testing in IHE Connectathons.

As a future work, it is necessary to define the capabilities of the IHE Actors to be able to establish electronic relationship among them automatically. OASIS ebXML Collaboration-Protocol Profile (CPP) and Agreement (CPA) [25] can be used for this purpose. Through a Collaboration-Protocol Profile (CPP), it is possible to define an Actor's message-exchange capabilities and the Business Collaborations that it supports. Then a Collaboration Protocol Agreement (CPA) can be established between two IHE Actors describing the messaging and the business process requirements that are agreed upon by both of the Actors.

Finally, although we have described this work based on IHE Integration Profiles, it is also applicable to HL7 to define national or local profiles.

REFERENCES

[1] Integrating the Healthcare Enterprise (IHE), http://www.ihe.net
[2] Radiological Society of North America, http://www.rsna.org/
[3] Healthcare Information and Management Systems Society, http://www.himss.org
[4] IHE IT Infrastructure Technical Framework, Cross-Enterprise Clinical Documents Sharing (XDS), by the IHE ITI Technical Committee, http://www.ihe.net/Technical_Framework/index.cfm#IT
[5] IHE IT Infrastructure Technical Framework, Patient Identifier Cross-referencing (PIX), by the IHE ITI Technical Committee, http://www.ihe.net/Technical_Framework/index.cfm#IT
[6] ebXML Business Process Specification Schema Technical Specification v2.0.1, http://docs.oasis-open.org/ebxml-bp/ebbp-2.0/2.0.3/
[7] Health Level 7 (HL7), http://www.hl7.org
[8] DICOM, NEMA Standards Publication PS 3.x, Digital Imaging and Communications in Medicine (DICOM), National Electrical Manufacturers Association, 2004, http://dicom.nema.org.
[9] IHE IT Infrastructure Technical Framework, Audit trail and Node Authenticaton (ATNA), by the IHE ITI Technical Committee, http://www.ihe.net/Technical_Framework/index.cfm#IT
[10] http://www.srdc.metu.edu.tr/ihe/ebBP/packages/
[11] ebBP Business Collaboration Definition for ITI-8 Transaction of IHE PIX (Online). Available: http://www.srdc.metu.edu.tr/ihe/links/IHE-ITI-8-BC.xml
[12] Rule Markup Language (RuleML) 0.91, http://www.ruleml.org/0.91/
[13] http://www.srdc.metu.edu.tr/ihe/ebBP/pseudo.JPG
[14] Jakarta Ant Build Tool, http://ant.apache.org/
[15] IHE Actor Grouping Instance (Across Profiles), http://-www.srdc.metu.edu.tr/ihe/ebBP/IHE-XDS-ATNA-PIX-ebBP.xml
[16] IHE Actor Grouping Instance (Within a Profile), http://-www.srdc.metu.edu.tr/ihe/ebBP/IHE-XDS-DRep-DReg.xml
[17] Anzböck, R., Dustdar, S. "Modeling Medical e-services", 2nd International Conference on Business Process Management (BPM 2004), June 2004, Germany, Springer LNCS.
[18] BEA Systems, IBM, Microsoft, SAP AG, Siebel Systems, "Business Process Execution Language for Web Services 1.1", http://www-128.ibm.com/developerworks/library/specification/ws-bpel/, 2003.
[19] Hofreiter, B., Huemer, C., Zapletal, M., "Registering UMM Business Collaboration Models in an ebXML Registry", Proceedings of the 8th IEEE Conference on E-Commerce Technology (CEC'06), San Francisco, California, June 2006.
[20] Aalst, W., Benatallah, B., Casati, F., Curbera, F. (Eds.), "Business Process Management", 3rd International Conference, BPM 2005, Nancy, France, September, 2005.
[21] Dustdar, S., Fiadeiro, J., Sheth, A., (Eds.), "Business Process Management", 4th International Conference, BPM 2006, Vienna,Austria, September, 2005.
[22] Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A., Laleci, G., "A Survey and Analysis of Electronic Healthcare Record Standards", ACM Computing Surveys, Vol. 37, No:4, December 2005.
[23] Dogac A., Bicer V., Okcan A., "Collaborative Business Process Support in IHE XDS through ebXML Business Processes", In proc. of International Conference on Data Engineering (ICDE2006), Atlanta, USA, April 2006.
[24] Dogac A., Kabak Y., Gulderen O., Namli T., Okcan A., Kilic O., Gurcan Y., Orhan U., Laleci G., "ebBP Profile for Integrating Healthcare Enterprise (IHE)", Submitted to OASIS ebXML Business Process Technical Committee, http://www.oasis-open.org/apps/org/workgroup/ebxml-bp/download.php/19233/ebBPprofile-forIHE-1.0-draft 1-July19-2006.pdf (for non-OASIS members http://www.srdc.metu.edu.tr/webpage/publications/2006/ebBPprofile-forIHE-1.0-draft1-July19-2006.pdf).
[25] ebXML, Collaboration-Protocol Profile and Agreement Specification v1.0, May 2001, http://www.ebxml.org/specs/ebCCP.pdf.
[26] Anzböck, R., Dustdar, S., "Modeling and Implementing Medical Web services", Data and Knowledge Engineering, 55 (2), pp. 203-236, Elsevier.
[27] ebXML, http://www.ebxml.org/

Asuman Dogac is a full professor at the Computer Engineering Dept., Middle East Technical University (METU), Ankara, Turkey and the CEO of the Software Research, Development and Consultancy Company (SRDC). Her research interests include health care informatics, Web Services, and semantic interoperability.

Yildiray Kabak is a PhD candidate at METU, Computer Engineering Dept., and a senior researcher at SRDC. His research interests include Semantic Web, Web Service technology, and health care informatics.

Tuncay Namli is a PhD candidate at METU, Computer Engineering Dept., and a senior researcher at SRDC. His research interests include semantic interoperability, Web Service technology, and health care informatics.

Alper Okcan is a PhD candidate at METU, Computer Engineering Dept., and a senior researcher at SRDC. His research interests include semantic interoperability, Web Service technology, and health care informatics.

A Survey and Analysis of Electronic Business Document Standards

Yildiray Kabak and Asuman Dogac
Dept. of Computer Eng.
Middle East Technical University (METU)

The development of electronic document interoperability standards has been evolutionary based on the traditional EDI technology and affected both by the technological developments such as the Internet and XML and also by the dynamic interoperability needs of the current eBusiness applications.

No document standard is sufficient for all purposes because the requirements significantly differ amongst businesses, industries and geo-political regions. On the other hand, the ultimate aim of business document interoperability is to exchange business data among partners without any prior agreements related with the document syntax and semantics. Therefore, an important characteristic of a document standard is its ability to adapt to different contexts, its extensibility and customization. UN/CEFACT Core Component Technical Specification (CCTS) is an important landmark in this direction.

In this article, we present a survey and an analysis of some of the prominent UN/CEFACT CCTS based electronic document standards. We describe their document design principles and discuss how they handle customization and extensibility. We address their industry relevance and the recent efforts for their harmonization and convergence. We conclude by mentioning some emerging efforts for the semantic interoperability of different document standards.

Categories and Subject Descriptors: J.1 [**Computer Applications**]: Administrative Data Processing—*Business*; J.1 [**Computer Applications**]: Administrative Data Processing—*Government*; H.2.4 [**Database Management**]: Systems—*Distributed databases*

General Terms: Standardization, Design

Additional Key Words and Phrases: eBusiness, Document Interoperability Standards, UN/-CEFACT Core Component Technical Specification (CCTS), OASIS Universal Business Language (UBL), OAGIS Business Object Documents (BODs), Global Standards One (GS1) XML

Authors' addresses: Y. Kabak and A. Dogac, Dept. of Computer Eng., Software R&D Center, Middle East Technical University (METU), 06531, Ankara, Turkey, e-mail: {yildiray, asuman}@srdc.metu.edu.tr.

This work is supported by the European Commission through the IST-213031 iSURF project and in part by the Scientific and Technical Research Council of Turkey (TÜBİTAK), Project No: EEEAG 105E068

1. INTRODUCTION

Interoperability of business applications can be investigated at three broad layers: communication layer, business processes layer and the document layer. In this article we focus on the document layer which addresses the interoperability of the document content exchanged.

Business Document interoperability initiatives started in the 1970s before the invent of the Internet. The first standard developed was the Electronic Data Interchange (EDI) framework [EDI] where document exchange was realised through dialup connections using proprietary networks.

Starting with the late 1990s *eXtensible Markup Language* [XML] became popular for describing data exchanged on the Internet. The relative human readability and the amount of XML tools available made XML a popular basis for a number of new document standards such as *Common Business Library* (CBL) [CBL] and *Commerce XML* [cXML]. This progress has been evolutionary because the later standards used the EDI experience. For example, CBL became *XML Common Business Library* [xCBL] after including EDI experience in CBL.

EDI, CBL and xCBL are horizontal industry standards addressing several industry domains. In the mean time, there are several vertical industry specific standard initiatives such as the ones from the *North American Automotive Industry Action Group* [AiAG], *Health Level 7 (HL7) Standards Development Organization* [HL7], the *Petroleum Industry Data Exchange (PIDX) committee* [PIDX] the *Chemical Industry Data Exchange (CIDX) organization* [CIDX], Open *Travel Alliance* [OTA], and *RosettaNet Consortium* [RosettaNet] to name but a few.

The earlier standards have focused on static message/document definitions which were inflexible to adapt to different requirements that arise according to a given context which could be a vertical industry, a country or a specific business process.

The leading effort for defining flexible and adaptable business documents came from the UN/CEFACT *Core Components Technical Specification* [CCTS] in the early 2000s. UN/CEFACT CCTS provides a methodology to identify a set of reusable building blocks, called *Core Components* to create electronic documents. *Core Components* represent the common data elements of everyday business documents such as "Address", "Amount", or "Line Item". These reusable building blocks are then assembled into business documents such as "Order" or "Invoice" by using the CCTS methodology. Core components are defined to be context-independent so that they can later be restricted to different contexts. Many core components defined by UN/CEFACT are available to users from UN/CEFACT *Core Component Library* [UN/CCL].

This concept of defining context-free reusable building blocks, which are available from a single common repository, is an important innovation in business document interoperability for the following reasons:

—The incompatibility in electronic documents is incremental rather than wholesale. The users are expected to model their business documents by using the existing core components and by restricting them to their context with well defined rules.

—Dynamic creation of interoperable documents becomes possible because if users cannot find proper components to model their documents, they can create and publish new core components.

—The horizontal interoperability among different industries is greatly facilitated by using a single common repository and by customizing the components to different industry contexts.

CCTS is gaining widespread adoption by both the horizontal and the vertical standard groups. Universal Business Language (UBL) [UBL] is one of the first implementation of the CCTS methodology. Some earlier horizontal standards such as *Global Standard One* (GS1) XML [GS1 XML] and *Open Applications Group Integration Specification* (OAGIS), and some vertical industry standards such as CIDX and RosettaNet have also taken up CCTS.

In this article, we survey some of the prominent horizontal business document standards, namely, EDI, UN/CEFACT CCL, UBL 2.0, OAGIS BOD 9.0 and GS1 XML. EDI is not only the earliest standard but the experience and the knowledge gained in its development also affected the other standards development efforts. UN/CEFACT CCL, which is based on UN/CEFACT CCTS, is a promising standard initiative to support dynamic electronic business requirements. The rest of the standards we cover are UBL 2.0, OAGIS BOD 9.0 and GS1 XML which are horizontal standards all based on UN/CEFACT CCTS.

The surveyed standards are first analyzed based on their document design principles: the document design principles involve the document artifacts used in composing the documents, the code lists used to convey the meaning of the values in the elements and the use of XML namespaces. Furthermore, since all the document standards surveyed are based on UN/CEFACT CCTS, how this methodology is used in the design of the documents is also discussed.

We then discuss how the standards handle extensibility and customization. The standards basically handle the customization and extensibility in two ways: either by introducing an "extension" element into the document schema or by allowing users to change the document schema. When an "extension" element is used, the document schema remains unchanged and the user can put any extra information in this element. When the document schemas are modified to accommodate extensions, the document interoperability is reduced.

Another important issue is whether the standards address the other layers in the interoperability stack, namely the communication layer and the business process layer. The communication layer addresses the transport protocol and the message header. The business processes layer involves the sequencing of the messages, and the business processes.

We also point out the industry relevance of these standards by providing some major usage examples. Most of the standards covered have very wide industry take up. Finally we conclude by mentioning the harmonization efforts and an emerging trend for the semantic interoperability of document standards.

Before we proceed any further, we clarify the use of the terms *message* and *document*. Some standards call a *document* what other standards call a *message*. We use these terms to mean the following: the data that is exchanged between parties is called a *message*, which contains a *transport header* and a *payload*. The *payload* may consist of one or more *documents*. It is the *document* that contains the actual business data although most of the time, the document standards also provide transport configuration information to be passed to the *transport header*.

The paper is organized as follows: Section 2 summarizes the EDI initiative. Section 3 describes the UN/CEFACT Core Component Technical Specification. Section 4 introduces the Universal Business Language (UBL) 2.0 standard. In Section 5, Open Applications Group Integration Specification (OAGIS) 9.0 is presented. Global Standard One (GS1) XML standard is covered in Section 6 after briefly introducing the set of standards proposed by GS1. Section 7 contains an analysis of the presented standards with respect to document design principles, customization and extensibility, coverage of other layers of interoperability and the industry relevance. Finally, Section 8 concludes the paper by describing harmonization efforts and the emerging semantic approach to the document standards interoperability. Since a large number of acronyms is introduced throughout the paper, a list of all acronyms and their meaning is provided in Table III.

2. ELECTRONIC DATA INTERCHANGE (EDI)

EDI is developed through two main branches: ANSI X12 and UN/EDIFACT. In the USA, the American National Standards Institute (ANSI) developed ANSI X12 [X12] and internationally EDI is standardised as UN/EDIFACT (United Nations/Electronic Data Interchange For Administration, Commerce, and Transport) [UN/EDIFACT]. Through both of these initiatives, a large number of standard electronic documents in plain-text, quote-delimited formats have been specified for domains like procurement, logistics and finance. EDIFACT has also been standardised by the International Standards Organisation as ISO 9735 [UN/EDIFACT].

Fig. 1. The Basic EDI Architecture

The basic EDI architecture is shown in Figure 1. The communications are through the Value Added Networks (VANs) which are responsible for routing, storing and delivering EDI messages. Special EDI adapters are implemented to interface the internal system of a partner to the value added network. The particulars of the message syntax and interaction process are negotiated between partners in advance. Sometimes a dominant partner imposes its standards on smaller partners.

An EDI "interchange" document, as shown in Figure 2 (a) consists of "messages" which are in turn composed of "data segments". The segments themselves consist of "data elements". Figure 2 (b) shows an example EDI message.

When the Internet became an established networking environment starting with mid 1990s, there were several updates to the EDI architecture. First, the Internet protocol for email, Simple Mail Transfer Protocol (SMTP), and the File Transfer Protocol (FTP) came to be used to transfer EDI documents directly between parties connected to the Internet. Later, once the World Wide Web and its transfer protocol, the Hyper-Text Transfer Protocol (HTTP), was popularised, this became another mechanism for EDI document transfer.

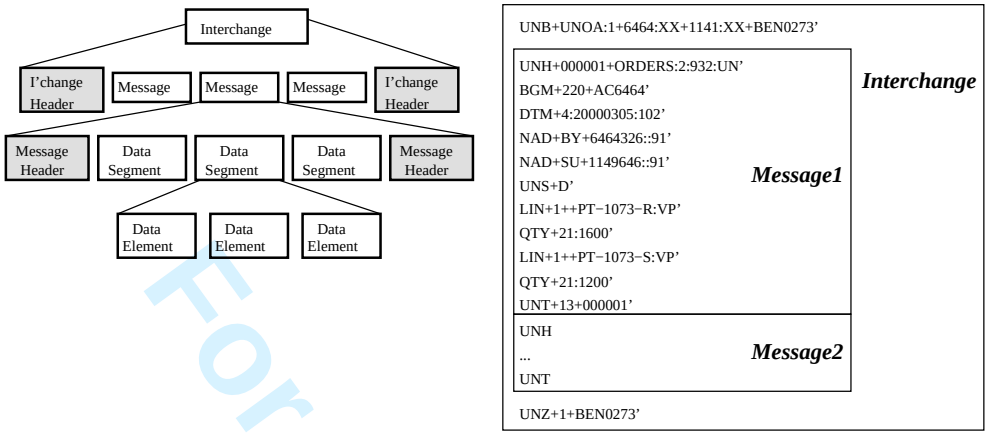


Fig. 2. (a) The Basic EDI Message Structure, (b) An Example EDI Message

3. UN/CEFACT CORE COMPONENT TECHNICAL SPECIFICATION (CCTS)

UN/CEFACT Core Components Technical Specification (CCTS) is defined as Part 8 of the ebXML (electronic business XML) Framework and is approved as ISO 15000-5 [CCTS].

The essence of UN/CEFACT CCTS is to design documents from standard, reusable building blocks, called *Core Components*. Considerable number of *Core Components* are available from the *UN/CEFACT Core Component Library (CCL)* for discovery and reuse and more will be available as the work progresses.

The first step to provide interoperability based on core components is to represent values in the components consistently. Hence the starting point for the design of *Core Components* is the *Core Component Types* and *Data Types*.

3.1 Core Component Types and Data Types

Core Component Types (CCT) constitute the leaf-level type space of UN/CEFACT *Core Components*. They specify the basic information types, such as amount, binary object, code and date time, and they are built from primitive data types (e.g. binary, decimal, integer and string). A CCT is composed of a *Content Component*, where the actual primitive content resides, and one or more *Supplementary Components*, which further describe the *Core Component Types*. In other words, *Supplementary Components* help to interpret a value in the *Content Component*.

For example, the "Code" CCT's *Content Component* is of type string and has a set of *Supplementary Components* such as *Code List Agency Identifier* which is the identifier of the Agency that maintains the code list and *Code List Agency Name* which is the name of the Agency that maintains the code list.

On the other hand, *Data Types* are based on one of the *Core Component Types* and further restrict them. In this respect, CCT's can be thought of as abstract types from which more specialized *Data Types* are produced. For example, in the

current version of the UN/CEFACT *Data Types*, there is a *Data Type*, called the "CurrencyCode". This data type is based on the "Code" CCT and restricts it as follows:

—*Content Component*: The value in the *Content Component* should be a three-letter code.

—*Code List Identifier*: The identifier of the code list is ISO 4217.

—*Code List Version Identifier*: The version of the code list is 2006-11-21.

The relationship among *Core Component Types*, *Data Types* and other types of core components are shown in Figure 3 [CCTS]. Up to now, UN/CEFACT has approved 10 *Core Component Types* and defined 35 permissible *Data Types*, and undertook their maintenance. Furthermore, the *Data Types* provided by UN/CEFACT can be used without restrictions (*Unqualified Data Types (UDT)*) or further restricted (*Qualified Data Types (QDT)*) to accommodate specific business needs. UN/CEFACT also provides the rules to restrict the *Data Types* to *Qualified Data Types*.

3.2 Naming Convention Used

A naming convention is necessary to consistently name the defined components to facilitate the comparison during the discovery and analysis process. Furthermore, ambiguities can be prevented such as developing multiple *Core Components* with different names that have the same semantic meaning. The Naming Convention used in CCTS is derived from ISO 11179 Part 5 [ISO11179]. This naming convention has three major parts: *Object Class*, *Property Term* and *Representation Term*. For example, when the *Core Component* "Invoice. Tax. Amount" is expressed according to the CCTS naming convention, "Invoice" is the *Object Class*, "Tax" is the *Property Term* and "Amount" is the *Representation Term*.

3.3 Types of Core Components

A *Core Component* is a reusable building block for creating electronic business documents. There are three types of *Core Components*:

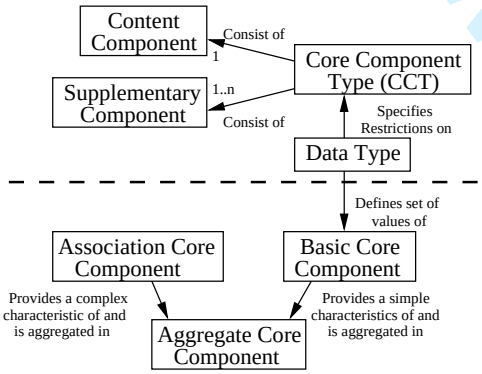


Fig. 3. Core Component Overview [CCTS]

—*Aggregate Core Component (ACC)*: A distinct real world object with a specific business meaning such as "Address" or "Purchase Order" is termed as an *Aggregate Core Component* (ACC). An *Aggregate Core Component* has at least one and possibly more *Basic Core Components* (BCCs). For example, as shown in Figure 4 "Address. Details" is an *Aggregate Core Component (ACC)* containing several *Basic Core Components (BCCs)*.

—A *Basic Core Component* describes a property of an ACC by using a *Data Type*. For example, as shown in Figure 4, "Address. Details. Street" is a *Basic Core Component (BCC)* and is of "Text" Data Type. In other words, the *Data Types* are used as *Representation Terms* of *Basic Core Components*.

—Sometimes it is necessary to define an association between Aggregate Core Components. This is realized through *Association Core Components*. As shown in Figure 4, "Person. Details. Residence" is an *Association Core Component* (ASCC) referencing the "Address. Details" ACC.

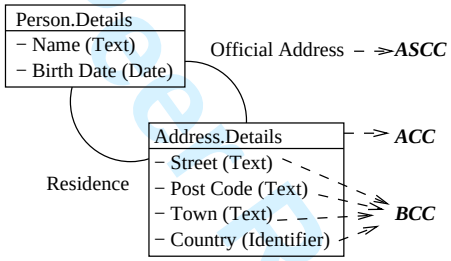


Fig. 4. Examples of *Basic Core Component* (BCC), *Aggregate Core Component* (ACC) and *Association Core Components* (ASCC) [CCTS]

Dictionary Entry Name	Type	Object Class Term	Property Term	Representation Term	Cardinality
Payment.Details	ACC	Payment			
Payment.Paid.Amount	BCC		Paid	Amount	0..unbounded
Payment.Received.Date Time	BCC		Received	Date Time	0..unbounded
...	...				
Payment.Tax.Amount	BCC		Tax	Amount	0..unbounded
Payment.Agent.Party	ASCC		Agent	Party	0..1

Business Process Context

Trade

Dictionary Entry Name	Type	Object Class Term Qualifier	Object Class Term	Property Term	Representation Term	Cardinality
Advance_Payment.Details	ABIE	Advance	Payment			
Advance_Payment.Paid.Amount	BBIE			Paid	Amount	1..1
Advance_Payment.Received.Date Time	BBIE			Received	Date Time	0..1

Fig. 5. Customizing an *Aggregate Core Component* to the *Business Process Context* "Trade"

3.4 Business Information Entity (BIE)

A *Core Component* is designed to be context-independent so that it can later be adapted to different contexts and reused. When a *Core Component* is restricted to be used in a specific business context, it becomes a *Business Information Entity (BIE)* and given its own unique name.

The possible business contexts that can be used are defined to be: *Business Process Context; Product Classification Context; Industry Classification Context; Geopolitical Context; Business Process Role Context; Supporting Role Context; System Capabilities Context* and *Official Constraints Context.*

For example, when the *Business Process Context* is specialized to "Purchasing", and the *Geopolitical Context* is set to be "EU", the "Invoice. Tax. Amount" BCC becomes the "Invoice. VAT_ Tax. Amount" *Basic Business Information Entity (BBIE).*

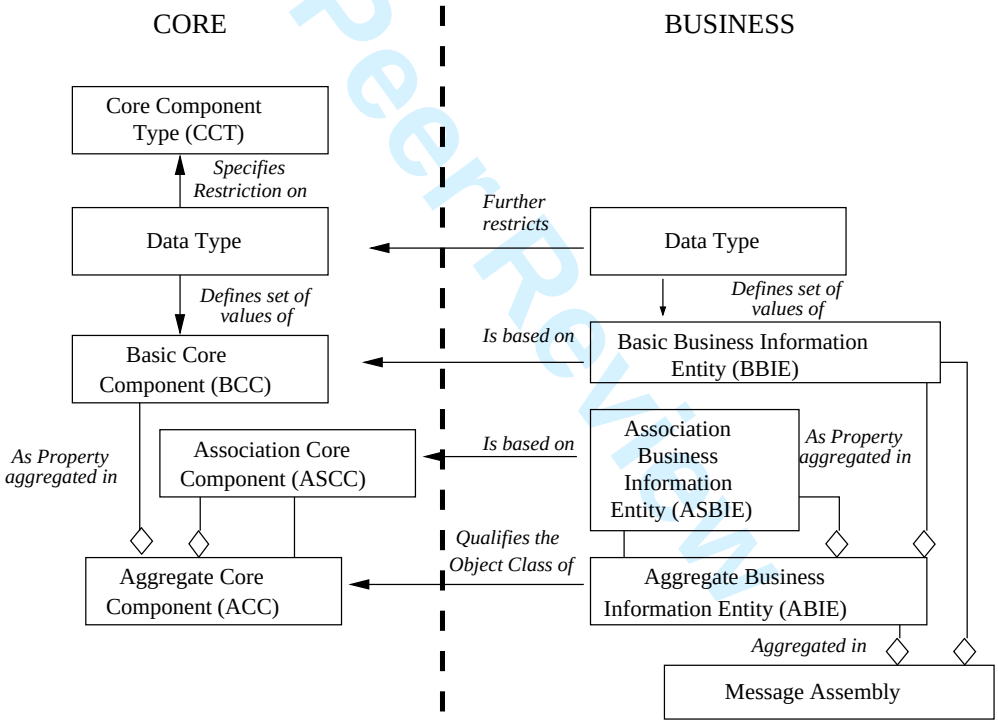


Fig. 6. Relationship between Core Components and Business Information Entities [CCTS]

Similarly, when an *Association Core Component* is used in a context, it becomes *Association Business Information Entity (ASBIE)* and *Aggregate Core Component* becomes *Aggregate Business Information Entity (ABIE).* For example, in Figure 5 an "Advance. Payment. Details" ABIE is created by customizing the "Payment.

Details" ACC to the *Business Process Context* "Trade" as follows: An *Object Class Term Qualifier* is added as an additional property and the related BCCs are customized to create the BBIEs by restricting their cardinality.

Figure 6 [CCTS] gives the relationship between the types of core components and the corresponding business information entities.

3.5 UN/CEFACT Core Component Library

The Core Component Library [UN/CCL] is the repository for UN/CEFACT CCTS artifacts. Currently there are quite a number of UN/CEFACT artifacts in the Core Component Library.

4. UNIVERSAL BUSINESS LANGUAGE 2.0 (UBL)

The Universal Business Language [UBL] initiative from OASIS adopts the UN/-CEFACT Core Component Technical Specification (CCTS) approach and develops a set of standard XML business document definitions.

Currently, the approved version of UBL is 2.0 [UBL] and there are thirty one XML schemas for common business documents such as "Order", "Despatch Advice" and "Invoice". In addition to the document definitions, UBL 2.0 provides a library of XML schemas (XSDs) [UBLSchemas] for reusable common data components like "Address", "Item", and "Payment" from which the documents are constructed. UBL 2.0 reuses *Core Component Type* and *Data Type* definitions from

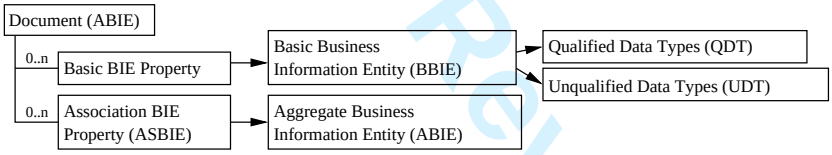


Fig. 7. The UBL Components

UN/CEFACT CCTS such as "AmountType", "CodeType" and "DateTimeType". When UN/CEFACT CCTS *Data Types* are imported to UBL type space, they are termed as the *Unqualified Data Types (UDT)*. Additionally, UBL defines *Qualified Data Types (QDT)* which are primarily for code lists such as *CurrencyCodeType* or *CountryIdentificationCodeType* defined for use within UBL.

At the time UBL initiative has started, UN/CEFACT CCTS has not yet specified core components. Therefore UBL created its own BIEs based on CommerceOne's xCBL (XML Common Business Library) 3.0 [xCBL] and the UN/EDIFACT (EDI for Administration Commerce and Trade) dictionary [UN/EDIFACT]. Hence the UBL vocabulary consists primarily of *Aggregate Business Information Entities* - (ABIEs).

Figure 7 shows the structure of the UBL Documents. It should be noted that in addition to identifying conceptual *Business Information Entities* (BIEs), UBL uses the CCTS artifacts such as ABIE, ASBIE and BBIE to compose its document schemas. This is in contrast to some other standards which use CCTS components in different document artifacts of their own and also name them differently.

In UBL, there are two types of ABIEs: (1) The document ABIEs which represent UBL Documents such as "Order" and "Invoice" and (2) More fine-grained reusable ABIEs such as "Address" and "Party". As shown in Figure 6, an ABIE is composed of BBIEs and ASBIEs as in UN/CEFACT CCTS. In UBL 2.0, according to the UBL 2.0 Naming and Design Rules, this composition is realized through *BIE Properties*. There are two types of *BIE Properties*:

(1) The *Basic BIE Property*, which is used for relating the ABIE with a BBIE, represents an intrinsic property of an ABIE. Note that there is no corresponding concept for *Basic BIE Property* in the UN/CEFACT CCTS. With this artifact, UBL diverges from UN/CEFACT CCTS Methodology where BBIEs are specialized from *Basic Core Components*. However as already mentioned UBL started creating its BIEs before UN/CEFACT *Core Components* were available.

(2) The *Association BIE Property*, which establishes an association from one ABIE to another ABIE, represents an extrinsic property. In other words, it is the *Association BIE Properties* that express the relationship between ABIEs. The *Association BIE Properties* correspond to the *Association Business Information Entities* (ASBIEs) in the UN/CEFACT CCTS.

A BBIE has a single content whose type is specified either with *Qualified Data Types* (QDT) or *Unqualified Data Types* (UDT). Figure 8 shows an example UBL 2.0 "Order" document.

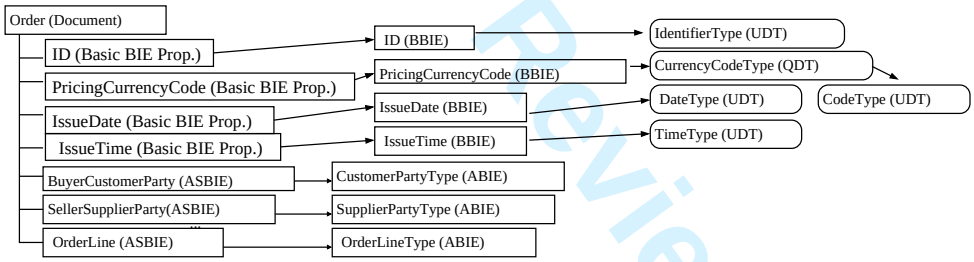


Fig. 8. An Example UBL Document Schema

4.1 UBL Customization and Extensibility

There are two types of customizations specified in UBL 2.0: Conformant customization and Compatible customization.

Before going into details of customization, it is worth mentioning about the validation of UBL documents. UBL 2.0 adopts a two-phase validation technique as shown in Figure 9. In the first phase, an incoming UBL document is validated against UBL 2.0 XSD schemas (or customized versions of them). If the instance passes the first phase, in the second phase it is checked against the rules, which specify additional constraints on the values of the elements in the instance. Generally, the rules are specified through XSL [XSL] or Schematron languages [Schematron]. If the instance passes both of the phases successfully, it is delivered to the processing business application.

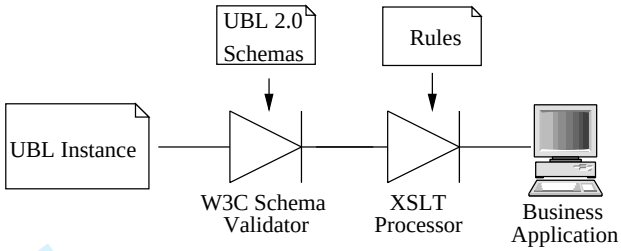


Fig. 9. Two-phase validation of UBL Messages

4.1.1 *Conformant Customization of UBL 2.0.* The key idea behind the conformant customization is that the XML instances in the customized implementation must also conform to the original UBL 2.0 schemas.

There are four ways of performing conformant customizations:

(1) *Inserting additional elements through the use of "UBLExtensions" element*: An optional *UBLExtensions* element appears as the first child of all UBL 2.0 documents which is used to include non-UBL data elements. For example, there could be elements containing data whose inclusion is mandated by law for certain business documents in certain regulatory environments. *UBLExtensions* element is composed of multiple *UBLExtension* elements, each containing a single element *ExtensionContent* of type "xsd:any" to accommodate the widest possible range of extensions. This means that any well-formed XML element from any vocabulary can be inserted into *ExtensionContent* element without modifying the schema.

(2) *Subsetting original UBL 2.0 schemas*: There are very many elements in a UBL document. For example, there are about 50,000 elements in a UBL Order Document. Some applications may not need all this data. Therefore, UBL 2.0 allows the users to create subsets of its documents. Subsets remove any optional information entities that are not necessary to the specific implementation. UBL 2.0 Small Business Subset [UBL-SBS] is an example of this subsetting mechanism.

(3) *Placing constraints on the value space of information entities and/or putting constraints among these values*: In a specific implementation of UBL 2.0, there may be additional constraints on the value space of information entities. For example, "The Total Value of an Order cannot be more than 50,000 USD". There may also be rules about dependencies between values of the elements, such as "The Shipping Address must be the same as the Billing Address" or "The Start Date must be earlier than the End Date". The former type of requirements can be reflected to the UBL schemas by type restriction; however, it requires schema modification. On the other hand, the latter type of requirements cannot be represented through XSD schemas. Therefore, the users can describe these constraints through Schematron [Schematron] or XSL rules [XSL] and feed these rules into the second phase of validation as already described.

(4) *Customizing the code lists*: Code list customization is described in Section 4.1.3.

4.1.2 *Compatible Customization of UBL 2.0.* Sometimes conformant customization may not be sufficient for a specific implementation. The users may need to perform more complex modifications such as extending an ABIE, creating a new ABIE or creating a new document. To handle these cases, compatible customization approach can be used. In compatible customization, the users modify an existing UBL 2.0 schema or create a new one by re-using the "largest suitable" aggregation from the UBL library. When performing compatible customization, the users need to follow the UBL Naming and Design Rules [UBL NDR].

4.1.3 *The Use of Code Lists.* In UBL 1.0, the standard and the default code list values are specified directly in the UBL schemas as XSD enumeration constraints. This allows all UBL 1.0 instances to be validated in a single pass using generic XSD processors. However, the specification of the default values directly in the schemas also makes it difficult to modify the code lists to meet customization requirements.

In UBL 2.0, only three code lists are enumerated in the schemas: (1) The *CurrencyCodeContentType* for internationally standardized currency codes, (2) The *BinaryObjectMimeCodeContentType* for MIME encoding identifiers and (3) The *UnitCodeContentType* for unit codes. In fact, these enumerations are specified in *Unqualified Data Types* from UN/CEFACT and UBL 2.0 includes them as they are.

The other code lists used in UBL are not enumerated in the schema expressions. Instead of enumerating the codes in the XSD schemas, UBL uses a common base type called *CodeType*, which is an extension of "xsd:normalizedString" for all elements expressing values from the code lists. The UBL 2.0 package includes files for every code list. These files are separate from the provided XSD schemas and they are in a custom format. Trading partners can modify or replace any of these files to meet their business requirements. After this step, they can convert these files in proprietary format to Schematron or XSL rules. UBL 2.0 provides tools for this purpose. Later these rules can be fed into the second phase of validation as already described.

5. OPEN APPLICATIONS GROUP INTEGRATION SPECIFICATION (OAGIS) BUSINESS OBJECT DOCUMENTS (BOD) VERSION 9.0

The Open Applications Group, Inc. (OAGi) [OAGi] is a not-for-profit open standards organization that defines electronic document standards called *Business Objects Documents* (BODs). Since its first release in 1995, several versions of Open Applications Group Integration Specification (OAGIS) BODs have been produced, the latest one being the OAGIS BOD version 9.0 [OAGIS]. This version is redesigned to be based on the UN/CEFACT Core Components Technical Specification.

The *Business Object Document* (BOD) is based on a pair of concepts called the *Noun* and the *Verb*. The *Verb* identifies the action to be applied to the *Noun*. *Noun* is the object or document such as "PurchaseOrder", "RequestForQuote", and "Invoice" that is being acted upon. Examples of *Verbs* include "Cancel", "Get", "Process", and "Synchronize". The *Verb* and *Noun* combination provides the name of the BOD. For example, when the *Verb* is "Process" and the *Noun* is "PurchaseOrder", the name of the BOD is "ProcessPurchaseOrder". There are 77 nouns and 12 verbs defined in OAGIS 9.0.

The separation of *Verb* and *Noun* components increases the reusability of data.

For example, the *Noun* "PurchaseOrder" contains all of the information that might be present in a "PurchaseOrder". The instantiation of each of the possible *Verb* and *Noun* combinations then further restricts the document to a context. For example, in a "ProcessPurchaseOrder" transaction, business partners and line item data must be provided, whereas in a "CancelPurchaseOrder" only the order identifier is enough to carry out the transaction. Note that, these constraints do not change the schema of a document. Rather, they provide the constraint rules to be applied in the validation of a BOD. Like UBL, OAGIS recommends a two-phase validation. When an OAGIS document is received, it is first validated against the corresponding XML Schema and afterwards against the corresponding Schematron/XSL rules. Only after the OAGIS instance document passes this two-phase validation, it is delivered to the business application that processes the document content.

OAGIS provides some recommendations on the usage of *Verbs*. *Verbs* may come in pairs meaning that the response to a *Verb* should be another specific *Verb*. For example, the response *Verb* of "Process" is "Acknowledge".

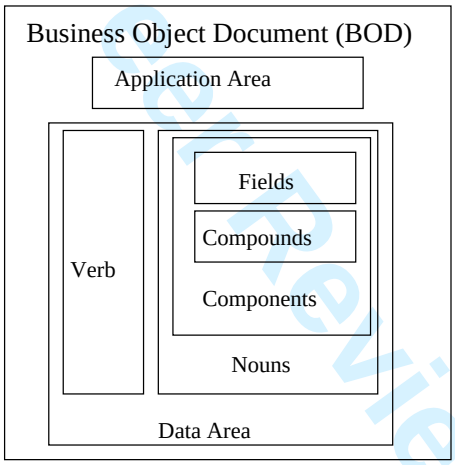


Fig. 10. The Structure of OAGIS Business Object Document (BOD)

As shown in Figure 10, BOD is a message structure composed of an *ApplicationArea* and a *DataArea*. The *ApplicationArea* carries necessary information for a transport software to send the message to the destination such as the sender, the signature of the sender and the unique identifier of the BOD. The need for the *ApplicationArea* stems from the following: the application software that creates a BOD may be separate from the transport software that sends the BOD to the destination. Therefore the application software creating the BOD should provide the transport software with the necessary configuration information to send the BOD. In other words, the *ApplicationArea* contains the configuration information created by the application software and conveyed to the transport software.

The *DataArea* contains a single *Verb* and multiple *Nouns*. A *Noun* may be assembled from *Component*, *Compound*, and *Field* document artifacts. *Components* are

large-grained building blocks and may in turn consist of other *Components*, *Compounds*, and *Fields*. Examples of *Components* include: "PurchaseOrder Header", "Party", and "Address". *Compounds*, which are used across all BODs, are a logical grouping of *Fields* (low level elements). Examples include "Amount", "Quantity", "DateTime", and "Temperature". *Fields* are the lowest level elements used in OAGIS *Components* and *Compounds*. Figure 11 shows an example BOD assembly with OAGIS artifacts.

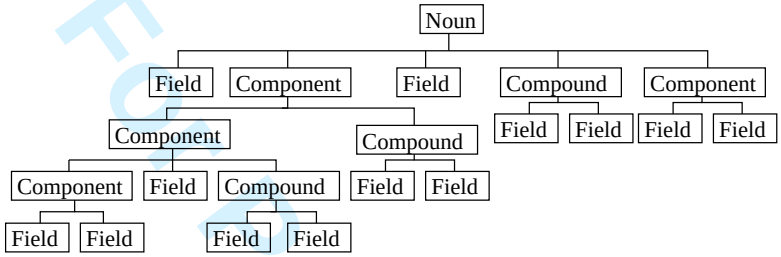


Fig. 11. OAGIS Business Object Document (BOD) assembly example

OAGIS implementation of the Core Component Technical Specification (CCTS) is shown in Figure 12. In OAGIS 9.0, the *Core Component Types* and *Unqualified Data Types* are directly used in the OAGIS Schemas. In other words, all OAGIS *Field* types are based on UN/CEFACT *Core Component Types*. Furthermore, the code lists, such as ISO 54217 Currency Codes and ISO 5639 Language Codes, recommended by UN/CEFACT are also used as described in Section 5.1.3.

As shown in Figure 12 OAGIS, rather than using the UN/CEFACT ABIEs directly, it incorporates them into OAGIS *Components*. When using these ABIEs in their *Components*, OAGIS appends "ABIEType" suffix to the name of the ABIE in order to identify that it is an ABIE from UN/CEFACT.

OAGIS Naming and Design Rules (NDR) are based on the UN/CEFACT ATG2 Naming and Design Rules [ATG2-NDR].

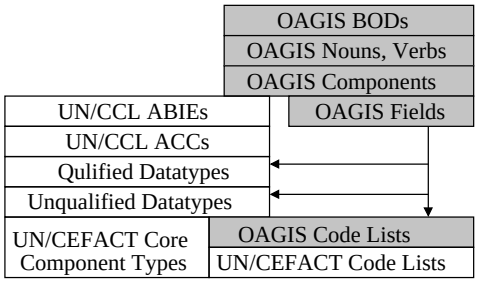


Fig. 12. OAGIS usage of UN/CEFACT CCT

5.1 OAGIS Extensibility

OAGIS provides two mechanisms to extend its specifications: *UserArea* Extensions and *Overlay* Extensions.

5.1.1 *UserArea Extensions.* The *UserArea* extensibility provides a means of adding implementation specific content to an existing OAGIS *Component* in an existing OAGIS BOD. When a few simple fields are needed to complete the information for the exchange, *UserArea* extensions are used. There is a *UserArea* element of type "xsd:any" at the end of each OAGIS *Component* where the users can insert any valid XML instance without changing the original OAGIS schema.

For example, in Turkey, the addresses contain "Mahalle" information, which basically specify a district in a city. In OAGIS, "Address" *Component* does not have such a *Field* to carry "Mahalle" information. This "Mahalle" information can be inserted in the *UserArea* part of "Address" *Component* in a BOD instance when it is used in Turkey, as shown in Figure 13.

```
<Address>
....
        <UserArea xmlns:myTrImpl="http://www.myTrImpl.org">
                <myTrImpl:Mahalle>EsatOglu<myTrImpl:Mahalle>
        </UserArea>
</Address>
```

Fig. 13. UserArea Example

5.1.2 *Overlay Extensions.* When the users need more complex changes such as creation of a new BOD or creation of new a *Component*, *Overlay* extension mechanism is used. The *Overlay* extensions result in the creation of new XML Schemas for the BOD in their own separate namespaces. It should be noted that only *Nouns* and *Components* are overlay extensible.

The *Overlay* extension mechanism adopts a layering approach. New layers, called overlays, are defined in their own respective namespaces on top of core OAGIS Schemas. Specialized BODs and *Components* are defined by extending BODs from lower layers and/or by composing new BODs from a combination of existing components, extended components, and new components. In Figure 14, an example for overlays is shown where "Automotive" overlay is created from core OAGIS schemas, whereas "Auto Parts" that is a subdomain of "Automotive", is built on "Automotive" and OAGIS core.

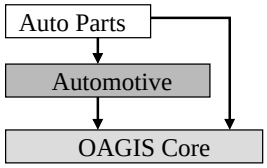


Fig. 14. OAGIS Overlay Layering Example

```
<xs:complexType name="MyInvoiceType">
 <xs:complexContent>
  <xs:extension base="oa:Invoice">
   <xs:sequence>
    <xs:element ref="ia:TotalDiscounts" minOccurs="0"/>
    <xs:element name="GrandTotal" type="oa:Amount" minOccurs="0"/>
    <xs:element name="MyInfo" type="xs:string" minOccurs="0"/>
   </xs:sequence>
  </xs:extension>
 </xs:complexContent>
</xs:complexType>

<xs:element name="MyInvoice" type="my:MyInvoiceType"
  substitutionGroup="oa:Invoice"/>
```

Fig. 15. Overlay Extension Example

With *Overlay* extensions, the users are allowed to create a new BOD, a *Noun*, a *Component*, a *Compound* or a *Field*, or extend any of the previously defined OAGIS artifacts. For example, a user may extend the "Invoice" *Noun* of OAGIS by adding the following: a new *Component* for representing total discounts; an existing *Compound* for grand total and a new *Field* for a special purpose. Figure 15 shows how these extensions are realized. The user first creates a new *Noun* called "MyInvoiceType" by extending the "Invoice" provided by OAGIS. Afterwards, the user inserts the elements mentioned. Finally, the user defines the "MyInvoice" element of type "MyInvoiceType". Note that "MyInvoice" element is in the same "xsd:substitutionGroup" as OAGIS "Invoice", which means that anywhere the OAGIS "Invoice" element is included in a model, the "MyInvoice" element can be inserted as well. In order to preserve interoperability among different *Overlay Extensions*, XSLT transformations are defined to convert an instance document conforming to an overlay into another.

```
In the CodeLists.xsd:

        <xsd:simpleType name="PaymentMethodCodeEnumerationType">
                <xsd:restriction base="xsd:normalizedString">
                        <xsd:enumeration value="Cash"/>
                        <xsd:enumeration value="Cheque"/>
                        <xsd:enumeration value="CreditCard"/>
                        <xsd:enumeration value="DebitCard"/>
                        <xsd:enumeration value="ElectronicFundsTransfer"/>
                        <xsd:enumeration value="ProcurementCard"/>
                        <xsd:enumeration value="BankDraft"/>
                        <xsd:enumeration value="PurchaseOrder"/>
                        <xsd:enumeration value="CreditTransfer"/>
                </xsd:restriction>
        </xsd:simpleType>

        <xsd:simpleType name="PaymentMethodCodeContentType">
                <xsd:union memberTypes="PaymentMethodCodeEnumerationType xsd:normalizedString"/>
        </xsd:simpleType>

In the Fields.xsd:

        <xsd:simpleType name="PaymentMethodCodeContentType">
                <xsd:restriction base="oacl:PaymentMethodCodeContentType"/>
        </xsd:simpleType>
```

Fig. 16. Code List Example

UserArea extensions are faster to apply than *Overlay* extensions. However, they do not provide the same level of control on the schemas as the *Overlay* extensions do. This is because the *UserArea* extensions are applied to the OAGIS BOD XML instance documents and not to the OAGIS BOD schema itself.

5.1.3 *Code List Extensions.* OAGIS uses and recommends the code lists from UN/CEFACT, and allows additional values to be present. This is accomplished as follows: OAGIS defines two "xsd:simpleType" for each coded *Field*: (1) an enumeration type, which lists the codes to be used and (2) a "xsd:simpleType" which is a union of that enumeration type and the "xsd:normalizedString". In other words, with the specification of "xsd:normalizedString" any code can be inserted to a BOD XML instance without affecting the validity against the BOD Schema. For example, as presented in Figure 16, the "PaymentMethodCodeContentType" *Field* is associated with "oacl:PaymentMethodCodeContentType" which is the union of "PaymentMethodCodeEnumerationType" and "xsd:normalizedString". The use of "xsd:normalizedString" allows the users to send codes that are not listed in "PaymentMethodCodeEnumerationType".

6. GLOBAL STANDARDS ONE (GS1)

Global Standards One (GS1) [GS1] is a family of standards focusing on different aspects of supply chain integration such as electronic products codes, product information synchronization and the electronic document standards. GS1 is formed in the early 2005 by the European Article Number [EAN] and the Uniform Commercial Code [UCC] organizations when they joined together. EAN and UCC were two organizations that heavily contributed to the adoption and proliferation of barcodes.

The part addressing the electronic document interoperability in this family of standards is GS1 eCom. In GS1 eCom, there are two distinct categories: the earlier eCom standards that are based on Electronic Document Interchange (EDI), called EANcom [EANCOM] and the newer generation GS1 XML [GS1 XML] which is defined using XML Schema.

The other standards in GS1 family include the Global Data Synchronization Network [GDSN] and EPCglobal [EPCglobal]. The Global Data Synchronization Network (GDSN) enables product data and location information synchronization so that trading partners have consistent item data in their respective systems.

EPCglobal drives the development of the Electronic Product Code (EPC) related with RFID standards. The specifications are based on the Radio Frequency Identification (RFID) research performed at the MIT AutoID Labs [MIT-AutoID].

6.1 GS1 XML

As shown in Figure 17, a GS1 XML document is represented with a *StandardBusinessDocument*, which contains a *StandardBusinessDocumentHeader* (SBDH) and a *Message*. *StandardBusinessDocumentHeader* is based on the SBDH defined by UN/CEFACT [UN/SBDH] and provides information about the routing and processing of the XML instance document contained in the GS1 XML *Message*. The SBDH is used for the same purpose as OAGIS's *ApplicationArea* element; that is, it contains the configuration information for the transport software to send the

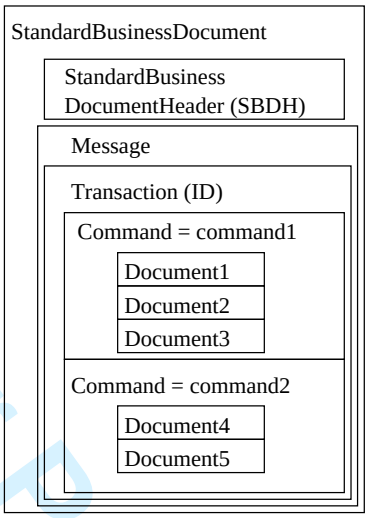


Fig. 17. The Structure of GS1 XML Structure

message to its destination.

A GS1 XML document includes either a set of *Commands* or a set of *Transactions* which in turn contain *Commands*:

—*Command*: A *Command* instructs the recipient to perform a particular action, such as "Add", "Delete" and "Refresh", related to the documents within the command. The use of these commands decreases the number of documents needed. The same document can be used with different commands. Hence, no separate documents like "Add Order", "Change Order" or "Delete Order" are needed; the same "Order" document can be sent with a relevant command. In a similar way, several documents can reuse the same command.

—*Transaction*: A *Transaction* provides the functionality of executing multiple commands atomically as in relational databases. If one command in a transaction fails, the transaction fails causing all other commands in the transaction to be discarded applying the principle of "all or nothing".

As an example, assume that a sender needs to send a message about two products and the first product is related to the second one. Instead of sending two distinct transmissions, the sender can transmit them together in one *Transaction* that contains one *Command*, which holds two *Documents* each of which is for a product. If the products are not related then, the sender can send them without using the *Transaction* element. In other words, the user sends only one *Command* containing two *Documents*.

GS1 XML is compliant with UN/CEFACT CCTS methodology in that GS1 XML uses the same modelling, design and technical principles. However, unlike UBL or OAGIS, which use UN/CEFACT artifacts (such as *Core Component Types*, *Data Types* and *Business Information Entities*), GS1 XML does not use UN/CEFACT CCTS artifacts in their XML Schemas. Yet, the GS1 core components are submit-

ted as an input to UN/CEFACT CCTS development.

While developing their e-business standards, GS1 uses its Global Data Dictionary [GS1 GDD] to store, reuse and share common components and business definitions, and their corresponding representations in XML. In other words, the GDD is the repository of:

—Data components, used to create the GS1 XML standards, developed according to the UN/CEFACT Core Components Technical Specification (CCTS).

—Business terms and their representation in GS1 XML.

Through GDD, the search of previously defined components is facilitated.

In the GS1 XML documents, some of the components such as *Measurement*, *DocumentStatus* and *MontetaryAmount* are common to more than one business document and more than one context. Therefore, these components are included in a common library as a part of the GDD. This approach allows reusing the same information constructs in all business messages.

6.1.1 *Customization and Extensibility.* In GS1 XML, the following context categories are defined for customization:

—*Business Process* context in which collaboration takes place such as ordering or delivery.

—*Industry Sector* context in which the business partners are involved such as automotive.

—*Geopolitical* context reflecting the geographical factors that influence the business semantics. This can be either country-specific, for example, only for France or Sweden, or limited to certain economic regions, for example, NAFTA or European Union, and finally, it can be applicable everywhere in the world in this case the context is defined as "Global".

The context information is reflected to the documents through their namespaces. In other words, the GS1 information components are assigned to a namespace that reflects the context it is defined in. For example, the namespace for the documents that are used in Global Data Synchronization Network (GDSN) is "gdsn=urn:ean.ucc:gdsn:2". As another example, the documents for alignment of trade items in Sweden use "sw=urn:ean.ucc:align:sweden:2" as their namespace. On the other hand, the schemas in the common library have "eanucc=urn:ean.ucc:2" as their namespace, because they do not belong to any specific context.

GS1 XML supports extensibility of its document schemas. Starting from release 2.0, there is an element called "extension" at the end of each business document XML schema where additional context-specific information that are not defined by GS1 XML can be inserted. This element is of type "xsd:any", which allows the users to insert any XML data to the exchanged instance documents without changing the standard GS1 XML schema.

Before starting to exchange GS1 XML instances with other parties, each organization that requires additional elements in their documents publishes their extensions to the "Extended Attributes" section of the Global Data Dictionary Web site. When a sender wishes to send a message to a receiver, the sender first checks whether the receiver has an extension by consulting the GDD Web site.

If there is an extension, the sender sends the message using Attribute/Value Pair mechanism. Attribute/Value Pair mechanism is a way to populate the "extension" area of a document. As an example, assuming that the receiver requires two additional elements: "packagingWeightValue" and "packagingWeightUnitOfMeasure", the sender populates the "extension" area as shown in Figure 18.

```
<extension>
<gdsn:attributeValuePairExtension
  xsi:schemaLocation="urn:ean.ucc:2
  ../Schemas/AttributeValuePairExtensionProxy.xsd">
    <value name="packagingWeightValue">15</value>
    <value name="packagingWeightUnitOfMeasure">kg</value>
</gdsn:attributeValuePairExtension>
</extension>
```

Fig. 18. Attribute/Value Pair Mechanism to populate extension area

```
<xsd:complexType name="ISO3166_1CodeType">
  <xsd:sequence>
    <xsd:element name="countryISOCode">
      <xsd:simpleType>
        <xsd:restriction base="xsd:string">
          <xsd:maxLength value="3"/>
          <xsd:minLength value="1"/>
        </xsd:restriction>
      </xsd:simpleType>
    </xsd:element>
  </xsd:sequence>
</xsd:complexType>
```

Fig. 19. Example Country Code Element

6.2 The Use of Code Lists

In GS1 XML, there are two types of code lists, external and internal. External code lists are defined and maintained by other standard bodies outside GS1 XML. The example external code lists include the following:

—Country Codes - ISO 3166-1:1997

—Country Subdivision Codes - ISO 3166-2:1998

—Currency Codes - ISO 4217:2001

The external code lists are defined as "xsd:string" and restricted to an appropriate number of characters. Figure 19 shows an example for "countryISOCode" element which is defined of type "xsd:string", whose length is three characters. However, GS1 XML does not import the code list values to the GS1 XML Schemas because of the copyright and maintenance issues. In other words, they are not enumerated in the GS1 XML Schemas.

The internal code lists are those developed and maintained within the GS1 System. They are defined as "xsd:enumeration" and imported into the business document schema that uses it. Figure 20 provides an example internal coding list for payment method types used in GS1 XML. It should be noted that all of the possible values are enumerated in the provided XML Schemas.

```
<xsd:simpleType name="PaymentMethodListType">
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="BANK_CHEQUE">
    </xsd:enumeration>
    <xsd:enumeration value="CASH">
    </xsd:enumeration>
    <xsd:enumeration value="CERTIFIED_CHEQUE">
    </xsd:enumeration>
    <xsd:enumeration value="CHEQUE">
    </xsd:enumeration>
    <xsd:enumeration value="CREDIT_CARD">
    </xsd:enumeration>
    <xsd:enumeration value="LETTER_OF_CREDIT">
    </xsd:enumeration>
    ...
  </xsd:restriction>
</xsd:simpleType>
```

Fig. 20. Example Payment Method List Element

7. ANALYSIS OF THE ELECTRONIC BUSINESS DOCUMENT STANDARDS

In this section, the surveyed electronic document standards are analyzed with respect to their document design principles, how they handle customization and extensibility, their coverage of the other layers of interoperability and their industry relevance.

7.1 The Document Design Principles

The document design principles involve the document artifacts used in composing the documents, the code lists used to convey the meaning of the values in the elements and the use of XML namespaces. Furthermore, since all the document standards surveyed are based on UN/CEFACT CCTS, how this methodology is used in the design of the document schemas is also discussed. Table I summarizes the document design principles.

7.1.1 *Document Artifacts and the Use of UN/CEFACT CCTS Methodology.* The document artifacts used in EDI are "Interchange", "Message", "Segment" and "Element" (Section 2). Note that EDI is not based on UN/CEFACT CCTS Methodology. UBL 2.0 uses the CCTS methodology to generate the document artifacts. UBL 2.0 currently considers only the "Business Process" context and identifies the *Business Information Entities* (BIEs) and bases the type of their artifacts to UN/CEFACT *Unqualified Datatypes* and *Core Component Types*. The UN/CEFACT develops its own BIEs, *Core Components* and *Datatypes* and stores them at the UN Core Component Library (UN/CCL). OAGIS 9.0 uses some of the UN/CEFACT ABIEs in their *Components* and bases the types of its *Fields* to UN/CEFACT *Unqualified Datatypes* (UDT) and *Core Component Types* (CCTs). GS1 XML uses the UN/CEFACT CCTS methodology to generate its own artifacts by using its Global Data Dictionary.

7.1.2 *The Use of Code Lists.* Code lists are important to uniquely convey the semantics of elements in electronic documents such as the country codes, currency codes, and the payment units. All of the surveyed document standards provide default code lists and allow them to be modified and/or extended to support local codes.

Table I. Document Design Principles

	Document Artifacts	**Use of CCTS Methodology**	**Use of Codelists**	**Use of namespaces**	**Naming and Design Rules**
EDI	Interchange, Message, Segment, Element	Not used	UN/EDIFACT recommends a number of code lists. Local and external codes are also allowed	Not used	UN/EDIFACT Syntax Rules (ISO 9735) or X12.5 and X12.6 Syntax rules
UN/-CCL	Uses CCTS based document artifacts such as Core Component Types and BIEs	Fully based on CCTS Methodology	Defines five code lists: *Country Codes*, *Subdivision Codes*, *Currency Codes*, *BinaryObject Mime Codes* and *Unit Codes*	It is syntax-independent	ISO 11179-5
UBL 2.0	Uses CCTS Artifacts	Fully based on CCTS Methodology	Through a common base type called *CodeType* "xsd:-normalizedString"	Mostly for document categorization	UBL 2.0 Naming and Design Rules
OAGIS 9.0	BODs, Application Areas, Nouns, Verbs, Components, Compounds, Fields	Fields are UDT and CCT based. Some Components are UN/CEFACT ABIE based	Defines two "xsd:simpleType" for each coded *Field*	To identify the *Overlay* extension elements	UN/CEFACT ATG2 Naming and Design Rules
GS1 XML	SBDH, Transactions, Commands, Documents	Use the CCTS methodology to generate its own document artifacts	External Code Lists; Internal Code Lists defined through "xsd:-enumeration"	The namespaces indicate the document context	GS1 XML's UML to XSD conversion rules

As shown in Table I, EDI provides codes for structuring of the message artifacts (e.g. segment codes). Furthermore, UN/EDIFACT recommends *ISO Country Code, Currency Code, Numerical Representation of Dates, Times, Periods of Time* and *UN/LOCODE* [ISO Codes]. EDI also allows implementers to convey their own local or external codes through the use of two data elements, 1131 [UN/EDIFACT 1131] and 3055 [UN/EDIFACT 3055].

UN/CEFACT defines five code lists: *Country Codes*, *Subdivision Codes*, *Currency Codes*, *BinaryObject Mime Codes* and *Unit Codes*.

UBL 2.0 uses *Currency Codes*, *BinaryObject Mime Codes* and *Unit Codes* from

UN/CEFACT and enumerates them in its schemas. The other code lists used in UBL are not enumerated in the schema expressions. Instead of enumerating the codes in the XSD schemas, UBL uses a common base type called *CodeType*, which is an extension of "xsd:normalizedString", for all elements expressing values from code lists. As described in Section 4.1.3, UBL allows the users to implement their own local/external codes.

For use of code lists, OAGIS defines two "xsd:simpleType" for each coded *Field*: (1) an enumeration type, which lists the codes to be used and (2) a "xsd:simpleType" which is a union of that enumeration type and the "xsd:normalizedString" as explained in Section 5.1.3. With this mechanism, the implementers can use their own local/external code lists.

In GS1 XML, there are two types of code lists, external and internal. External code lists are defined and maintained by other standard bodies outside GS1 XML. The internal code lists are those developed and maintained within the GS1 System. They are defined as "xsd:enumeration" and imported into the business document schema that uses it as described in Section 6.2.

7.1.3 *The Use of Namespaces.* Generally, the namespaces in XML are used for avoiding name conflicts. The document standards make additional use of the namespace mechanism as follows: UBL achieves categorization of documents through namespaces, OAGIS identifies the extensions through namespaces (Section 5.1) and GS1 XML gives context to the both original documents and extended documents through the namespaces as described in Section 6.1.

7.1.4 *Naming and Design Rules.* The naming and design rules specify how to name and structure the artifacts, how to put relations between the artifacts and how to use data types for the artifacts. UN/CCL uses ISO 11179 naming rules, which identify the artifacts in *Object Class*, *Property Term* and *Representation Term* format as described in Section 3. UBL 2.0 uses UBL 2.0 Naming and Design Rules, which are based on the CCTS terms such as ABIE, ASBIE and BBIE. OAGIS 9.0, on the other hand, applies naming and design rules based on Applied Technology Group XML Syntax (ATG2) Naming and Design Rules (NDR) [ATG2-NDR]. UN/CEFACT ATG2 NDR basically specifies how to represent the artifacts such as ABIEs, ASBIEs and BBIEs in XML schemas. For example, for every ABIE, a "xsd:complexType" must be defined and the name of this complexType must be in upper camel case (UCC) format (UCC capitalizes the first character of each word and compounds the name such as "AccountType").

GS1 XML first designs its information model in UML, before creating the corresponding XML schemas. GS1 XML uses its own UML to XSD conversion rules to generate their XML schemas and to name them.

7.1.5 *Analysis of Document Design Principles.* The differences with respect to document design principles as analyzed in this section results in considerable differences in document instances from different standards. As an example, in Figure 21, the OAGIS 9.0 "AddressBaseType" Component and GS1 XML "NameAndAddressType" document elements are compared. As it is clear from this figure, there are differences in the element names, the element positions and structures as well as in the use of code lists.

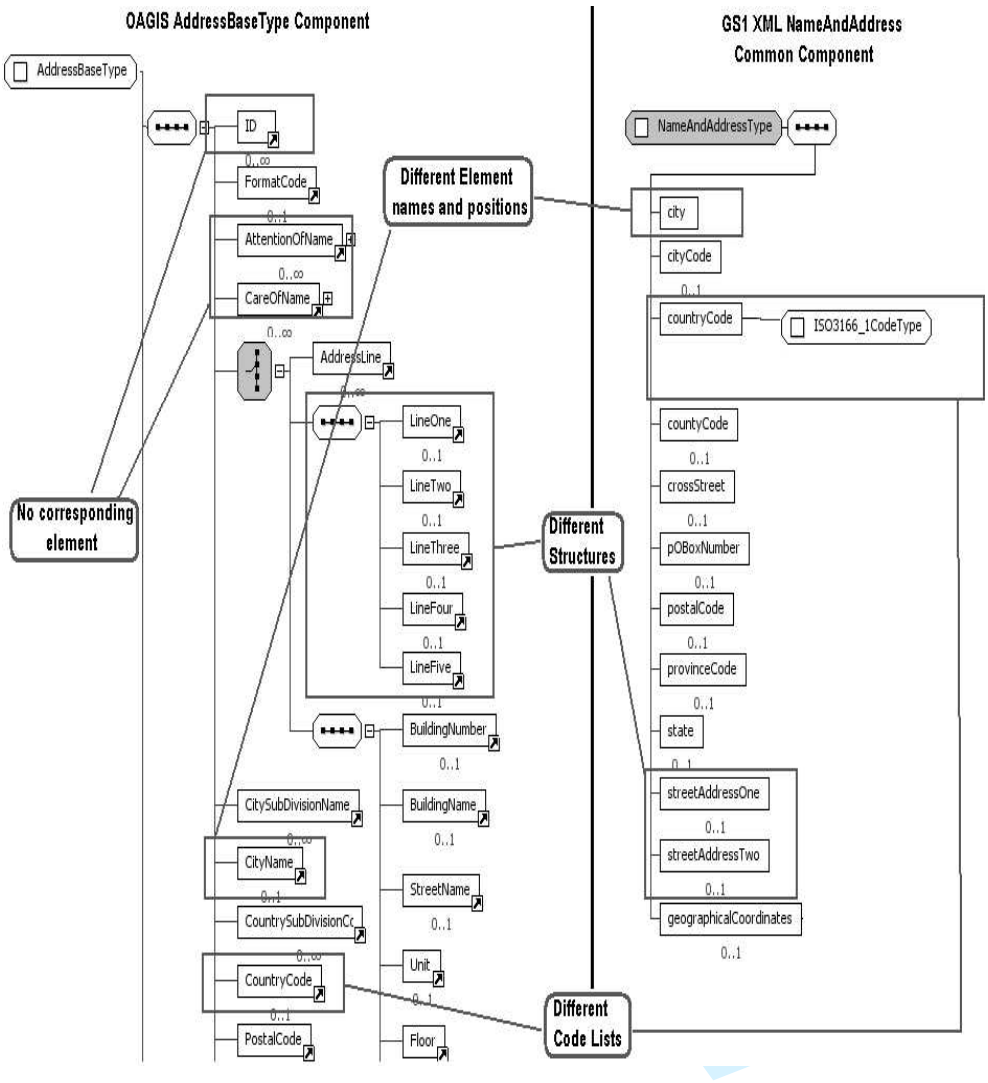


Fig. 21. An Example Comparing Related Parts of OAGIS BOD 9.0 and GS1 XML Documents

7.2 Customization and Extensibility

Any document interoperability standard faces two challenges. First, the standard needs to be extensible to allow definition of information that is not contained in the standard's artifacts because no standard can contain all of the data needed in every environment. Secondly, to be able to address a particular constraint in a specific context, it should be possible to customize the standard's artifacts according to a context.

Customization and extensibility can be handled in two ways: by allowing changes in the standard document schema or without changing the document schema. Doc-

ument schema changes weaken interoperability.

Table II present a summary of how the standards addressed in this article handle customization and extensibility.

Table II. Customization and Extensibility

	Customization	**Extensibility**
EDI	Subsetting EDI documents through context specific Implementation Guidelines. No schema changes.	Introduction of new types of business documents which has to be validated through related EDI Committees. Schema changes
UN/CCL	Core Components are customized according to eight contexts to create BIEs. No schema changes.	New components can be published to the Core Component Library. Schema changes.
UBL 2.0	Conformant customization through "UBLExtensions" element, or subsetting or placing constraints on the value space. No schema changes.	Compatible customization by reusing the largest suitable aggregation from the UBL Library. Schema changes.
OAGIS 9.0	No formal methodology for defining user specific customizations	Through *User Area* and *Overlay* extensions. Schema changes in *Overlay* extensions.
GS1 XML	Through the following three contexts: Business Process, Industry sector, Geopolitical	Through the "extension" element at the end of each document schema. No schema changes.

EDI addresses the customization through a subsetting mechanism to cover the requirements of a specific context. The EDI messages are subsetted first through industry Implementation Guides (IG), which are then subsetted into trading partner IGs, and into departmental IGs.

The extensibility in EDI is difficult because the EDI system is highly static and inflexible: the introduction of a new type or changing an existing type of business document is a complex process. Such changes require the modification of translation software and must be validated in the related EDI committees.

In UN/CEFACT CCTS, a *Core Component* is designed to be context-independent and is customized to one of the eight contexts defined by UN/CEFACT to become a *Business Information Entity (BIE)*. The possible business contexts that can be used are defined to be: *Business Process Context; Product Classification Context; Industry Classification Context; Geopolitical Context; Business Process Role Context; Supporting Role Context; System Capabilities Context* and *Official Constraints Context*.

UN/CEFACT CCTS supports extensibility as follows: if users cannot find proper components in the *Core Component Library* to model their documents, they can create and publish new core components. In other words, UN/CEFACT CCTS thrives on extensibility by allowing users to define core components with possible future harmonizations and removal of redundancies.

UBL 2.0 allows customization through (1) *UBLExtensions* element, (2) subsetting by removing optional information entities that are not needed, and (3) putting constraints to the elements as described in Section 4.1.1. On the other hand, the

users can extend the UBL 2.0 schemas through the mechanisms described in Section 4.1.2.

In OAGIS BODs, there is no formal mechanism to handle user specific constraints. However, the users are free to restrict an already existing BOD as they wish and sharing it with other partners.

OAGIS provides two mechanisms to extend its specifications as detailed in Section 5.1:

—*UserArea* Extensions: *UserArea* extensions provide an optional element within each OAGIS defined *Component* that may be used by an implementer to carry any necessary additional information. This area is of type "xsd:any", which means any valid XML instance can be inserted to this area without modifying the OAGIS standard XML Schemas (XSDs).

—*Overlay* Extensions: *Overlay* extensions allow users to extend an OAGIS BOD, *Noun* and *Component* to meet their own needs, even adding new BODs, *Verbs*, *Nouns* and *Components* where necessary. It is also possible for users to provide additional constraints in their own XSL constraints, which may then be applied to OAGIS document instances. The *Overlay* extension mechanism is used when the implementers have more complex customization requirements than a few additional elements.

Every document in GS1 XML is used in a business context and in GS1 XML, there are three context categories: *Business Process*, *Industry Sector* and *Geopolitical contexts* as described in Section 6.1.1.

GS1 XML supports extensibility of its document schemas. Starting from release 2.0, there is an element called "extension" at the end of each business document XML schema where additional context-specific information that are not defined by GS1 XML can be inserted. This element is of type "xsd:any", which allows the users to insert any XML data to the exchanged instance documents without changing their standard XML Schema.

Before starting to exchange GS1 XML instances with other parties, each organization that requires additional elements in their documents publishes their extensions to the "Extended Attributes" section of Global Data Dictionary (GDD) Web site. When a sender wishes to send a message to a receiver, the sender first checks whether the receiver has an extension by consulting the GDD Web site.

7.2.1 *Analysis of Customization and Extensibility.* The customization and extensibility affect how the documents are processed. There are two cases to be considered:

—In the first case, if the parties use the same document schema with the same extensions and customizations, a two-phase validation at the receiving end is applied: In the first phase, the incoming document instance is validated against the common XSD schema. If the document instance passes the first phase, in the second phase it is checked against the rules, which specify additional domain specific constraints on the values of the elements in the instance. Generally, the rules are specified through XSL [XSL] or Schematron languages [Schematron]. If the instance passes both of the phases successfully, it is delivered to the processing business application.

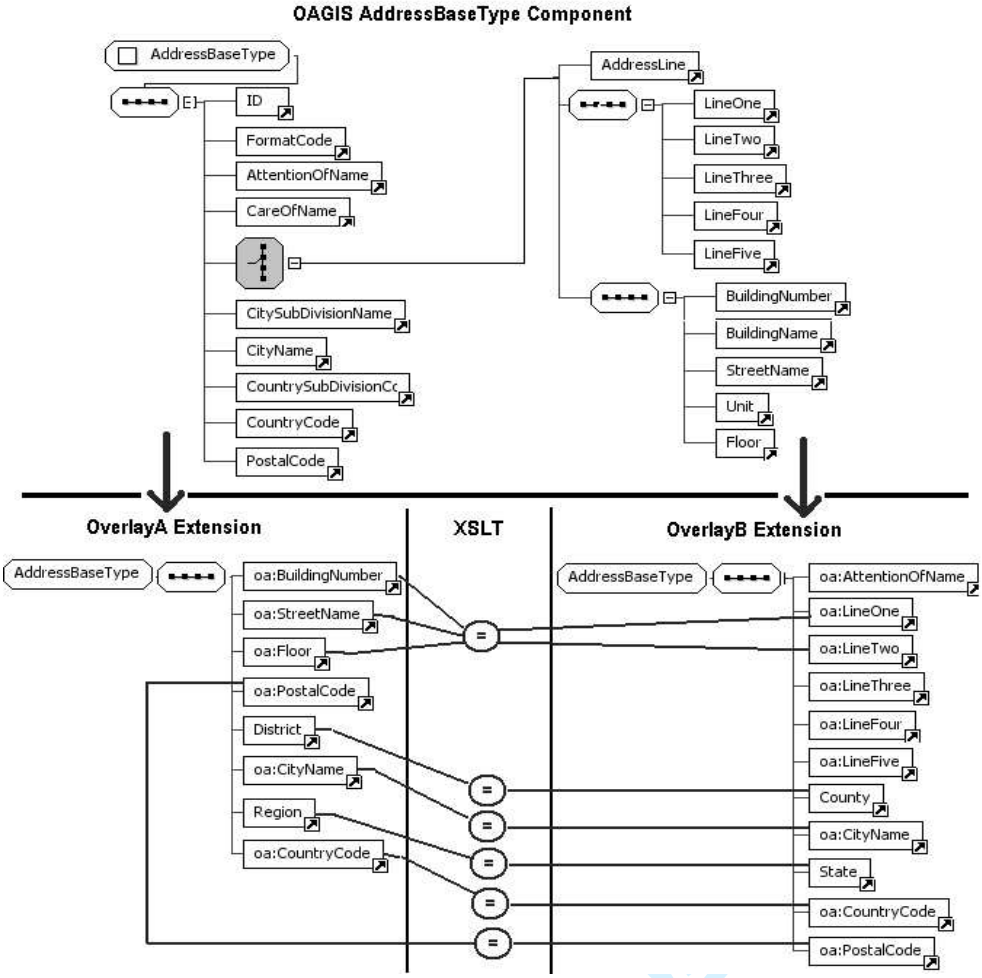


Fig. 22. Example XSL Transformations necessary to map between two different *Overlay* extensions in OAGIS BODs

—In the second case, when two enterprises use different customizations or extensions of the same document schema, the schema changes need to be mapped to each other through manually provided XSL Transformations. For instance, Figure 22 shows the XSL Transformations necessary to map between two different example *Overlay* extensions in OAGIS BODs. A classification of problems and solutions using XSL transformations to convert business documents is given in [Wüstner].

Once the transformations are applied, the document instance goes through the two-phase validation as described for the first case.

7.3 Coverage of Other Layers of Interoperability

Document interoperability is only one of the layers in the interoperability stack. The other layers of interoperability include the transport protocol, the message header and the business processes. A detailed survey of Business-to-Business interactions in general is given in [Medjahed, et. al.].

The standards covered in this survey do not enforce any specific transport protocol. However, some of them recommend certain transport protocols: GS1 XML recommends the use of EDIINT AS1 [EDIINT-AS1] and AS2 [EDIINT-AS2] transport protocols, which define a minimum set of parameters and options to enable secure/reliable transport for the exchange of EDI or XML data. EDIINT-AS1 is based upon SMTP and EDIINT-AS2 is based on HTTP. Among them, AS2 is the transport protocol of choice. However, the exchange of GS1 XML documents is not limited to these standards. OAGIS is currently moving in the Web Service technology direction although any technology can be used to transport BODs.

The document standards first analyze the relevant business processes or scenarios before deciding on the document components. For example, through the analysis of an invoicing business process, it may be revealed that a component is necessary to represent the "tax amount" in the invoice. Hence, "Tax Amount" is defined as a component that can be discovered and reused in any business document. However, no formal business process specification is provided by the standards surveyed in this article. Yet, it worths mentioning that there is work, called Universal Business Process [UBP], for defining UBL 1.0 processes through ebBP 2.0 [ebBP]; however, currently it is only informative.

All of the standards (except for UN/CCL which is syntax independent) provide message header information to be conveyed to the transport protocol header. The EDIFACT message headers are the Interchange Control Header Segment, UNB [ICHS] and the X12 Interchange Control Header, ISA [ICH]. In UBL, the message header information to be conveyed to the transport protocol header is dispersed through out the document instances. The *Application Area* in an OAGIS BOD is used to convey configuration information from application software to transport software. GS1 XML *StandardBusinessDocumentHeader* (SBDH) carries transport related information from application software to the transport software just as in the case of OAGIS *Application Area*.

7.3.1 *Analysis of Layers of Interoperability Addressed.* The surveyed standards do not specify a transport protocol but provide configuration information for the transport protocol message header.

Refraining from specifying other levels of interoperability has the advantage that it allows a wide variety of implementation techniques to be used and hence provides ease in implementation. However, the differences in the implementation techniques may cause interoperability problems.

7.4 Industry Relevance

EDI, being an early horizontal standard, is being used in several industry domains. For example, the financial and monetary systems like *Society for Worldwide Interbank Financial Telecommunication* [SWIFT] and *Electronic Funds Transfer* [EFT] use EDI. Furthermore, all airplane booking and ticketing operations are done over

EDIFACT through the *International Air Transport Association* system [IATA].

Contrary to the popular belief, electronic business interoperability is still achieved heavily through EDI based messages and EDI use is growing 3 to 5 percent every year [Vollmer]. It seems large organizations will continue to use EDI for the foreseeable future mostly due to the existing infrastructure investments.

UN/CEFACT CCTS is gaining widespread adoption by the standards organizations. As already mentioned a number of standardization efforts have taken up CCTS Methodology including UBL, GS1 XML, OAGIS, CIDX and PIDX in addition to UN/CEFACT's own Core Component Library (CCL).

The merits of CCTS for improving interoperability have also been noticed by the industry and the governments. For example, the U.S. Department of the Navy (DON) designed their XML Naming and Design Rules around CCTS. The German Government has made a formal announcement identifying CCTS as the future data standard for domestic affairs [Crawford].

One of the first companies to support UN/CEFACT CCTS methodology and core components in their products is SAP [SAP]. SAP Global Data Types (GDTs) form the basis of Business Objects and Enterprise Services. All leaf elements of these SAP GDTs are based on *Core Component Types* and *Data Types* [Stuhec; Stuhec2].

UBL is being adopted by several communities around the world especially in electronic government applications. The first government to use UBL Invoice is Denmark. The use of UBL Invoice is realized through the "Offentlig Information Online UBL (OIOUBL)" Project and has been mandated by law for all public-sector businesses [OIOUBL] in Denmark. Also in Sweden, the National Financial Management Authority recommended UBL Invoice customized to Sweden, namely, Svefaktura for all government use [Svefaktura].

Following the success of Danish and Swedish examples, representatives from Denmark, Norway, Sweden, UK, Finland and Iceland have set up a working group to develop a Northern European Subset (NES) [NES] for UBL to ensure interoperability among these countries.

In the USA, the Department of Transportation is developing a UBL based pilot project for a demonstration of state-of-the-art electronic commerce in a real-world setting [US/DOT].

OAGIS BODs are being used in more than forty countries in more than thirty eight industries [OAGIS-Usage]. The fact that OAGIS allows BODs to be extended by a vertical industry helps with its extensive use. The vertical standards based on OAGIS BODs include AiAG [AiAG], Odette [ODETTE], STAR [STAR], and Aftermarket [AAIA] in the automotive industry. Other standard bodies focused on human resources, chemical, and aerospace industries also use OAGIS BODs.

There are products based on OAGIS BODs such as Oracle E-Business Suite [Oracle] where OAGIS BODs are implemented as Web Services. As another example, IBM WebSphere Commerce service interfaces are defined using the OAGIS message structure [Rowell].

GS1 XML is being used in more than twenty countries in more than twenty industries all over the world. GS1 is a business solution partner of many companies including Oracle, Siemens and Philips. The GS1 standards are also leveraged in

SAP business solutions packages [SAP].

8. CONCLUSIONS

Today, an enterprise's competitiveness is to a large extent determined by its ability to seamlessly interoperate with others and electronic document standards play an important role in this.

Although all the document standards surveyed in this article (with the exception of EDI) are based on UN/CEFACT CCTS Methodology, their analysis reveal that there are considerable differences in the resulting document schemas. This is mostly because the standards like OAGIS BODs and GS1 XML existed long before UN/CEFACT CCTS Methodology is proposed and therefore these standards adapted their existing document schemas rather than starting from fresh. However, all of these standards are still developing and their future versions may become more harmonized.

In fact, by observing that the divergent and competing approaches to electronic document standardization threatens intersectoral coherence in the field of electronic business, four major standard bodies, namely, the International Electrotechnical Commission (IEC), the International Organization for Standardization (ISO), the International Telecommunication Union (ITU) and the United Nations Economic Commission for Europe (UNECE) signed a "Memorandum of Understanding" to specify a framework of cooperation [MoU]. In the year 2000, they established a Memorandum of Understanding Meeting Group for eBusiness standards harmonization. Up to now, OAGIS 9.0 and UBL 2.0 have achieved a level of harmonization in that they are based on the same UN/CEFACT *Unqualified Datatypes* and *Core Component Types*. However, the harmonization needs to be extended to the upper level artifacts such as the BBIEs and the ABIEs.

An alternative emerging approach to document interoperability is the semantic mediation of the electronic document schemas.

[Yarimagan1] argues that providing syntactic interoperability among document schemas based on XSL transformations or Schematrons alone is not enough. Syntactic interoperability needs to be supported by semantic interoperability, that is, it must be possible for automated processes to discover and reuse customizations provided by other users. For this purpose, the authors describe how the semantic representations of the context domains are provided and how this semantics is utilized by automated processes for component discovery and schema customization in UBL.

In [Yarimagan2], a Component Ontology for UBL is developed by using the Web Ontology Language (OWL) to represent the semantics of individual components and their relationships within customized schemas. Then this ontology is processed through description logic reasoners for the discovery of similar components and the automation of the translation process among different UBL customizations.

In [Anicic], Semantic Web technologies are used to transform documents between two vertical industry standards both based on OAGIS: one conforming to the *Standards in Automotive Retail* [STAR] and the other conforming to the *Automotive Industry Action Group* [AiAG]. First, the STAR and AiAG XML Schemas are converted to Web Ontology Language [OWL]. Then these independently devel-

oped ontologies are merged. By using the merged ontology, the STAR document instances are converted to the corresponding AiAG documents and vice versa.

In [Ye], a supply chain management ontology, called Onto-SCM, is developed which represents a common semantic model of supply chain management. The authors then show how Onto-SCM can be used converting document schemas of different standards.

As a final word, although the electronic document standards developed so far proved to be very useful for the industry and government applications, further efforts are needed for their harmonization and semantic interoperability.

Table III. List of Acronyms and Abbreviations

ABIE	Aggregate Business Information Entity
ACC	Aggregate Core Component
AiAG	Automotive Industry Action Group
ANSI	American National Standards Institute
ASBIE	Aggregate Business Information Entity
ASCC	Association Core Component
ATG	UN/CEFACT Applied Technology Group
BBIE	Basic Business Information Entity
BBC	Basic Core Component
BIE	Business Information Entity
BOD	Business Object Document
CBL	Common Business Library
CCL	Core Component Library
CCT	Core Component Types
CCTS	Core Components Technical Specification
CEFACT	Centre for Trade Facilitation and Electronic Business
CIDX	Chemical Industry Data Exchange
cXML	Commerce XML
EAN	European Article Number
ebXML	Electronic Business eXtensible Markup Language
ebBP	ebXML Business Process
EDI	Electronic Data Interchange
EDIFACT	Electronic Data Interchange For Administration, Commerce and Transport
EFT	Electronic Funds Transfer
EPC	Electronic Product Code
GDD	Global Data Dictionary
GDSN	Global Data Synchronization Network
GDT	Global Data Types
GS1	Global Standards One
HL7	Health Level Seven
HTTP	HyperText Transfer Protocol
HTTPS	Secured HyperText Transfer Protocol
IATA	International Air Transport Association
IEC	International Electrotechnical Commission
IG	Implementation Guides
ISO	International Organization for Standardization
IT	Information Technology
ITU	International Telecommunication Union

continued on next page

continued from previous page	
MIME	Multipurpose Internet Mail Extensions
NDR	Naming and Design Rules
OAGI	Open Applications Group, Inc.
OAGIS	Open Applications Group Integration Specification
OASIS	Organization for the Advancement of Structured Information Standards
OTA	Open Travel Alliance
OWL	Web Ontology Language
PIDX	Petroleum Industry Data Exchange
QDT	Qualified Data Types
RFID	Radio Frequency Identification
SBDH	Standard Business Document Header
SMTP	Simple Mail Transfer Protocol
STAR	Standards in Automotive Retail
SWIFT	Society for Worldwide Interbank Financial Telecommunication
UBL	Universal Business Language
UBP	Universal Business Process
UCC	Uniform Commercial Code
UDT	Unqualified Data Types
UMM	UN/CEFACT Modelling Methodology
UN	United Nations
UNECE	United Nations Economic Commission for Europe
xCBL	XML Common Business Library
XML	eXtensible Markup Language
XSD	XML Schema
XSL	Extensible Stylesheet Language
XSLT	Extensible Stylesheet Language Transformations

REFERENCES

AAIA. Automotive Aftermarket Industry Association. http://www.aftermarket.org/Home.asp.

AiAG. Automotive Industry Action Group. http://www.aiag.org/.

Anicic. N. Anicic, N. Ivezic, Semantic Web Technologies for Enterprise Application Integration. ComSIS, Computer Science and Information Systems, International Journal, Volume 02 , Issue 01, June 2005.

ATG2-NDR. UN/CEFACT Applied Technology Group (ATG) XML Syntax, XML Naming and Design Rules. http://www.uncefactforum.org/ATG/Documents/ATG/Downloads/XMLNamingAndDe%signRulesV2.0.pdf.

CBL. Common Business Library. http://xml.coverpages.org/cbl.html.

CCTS. UN/CEFACT Core Components Technical Specification. http://www.unece.org/cefact/ebxml/CCTS_V2-01_Final.pdf.

CIDX. Chemical Industry Data Exchange. http://www.cidx.org/.

Crawford. M. Crawford, Core Components Adoption On The Rise. https://www.sdn.sap.com/irj/sdn/weblogs?blog=/pub/wlg/5395.

cXML. Commerce XML. http://cxml.org/.

EAN. European Article Number. http://en.wikipedia.org/wiki/European_Article_Number/.

EANCOM. European Article Number Communication. http://www.gs1.org/productssolutions/ecom/eancom/.

ebBP. ebXML Business Process. http://docs.oasis-open.org/ebxml-bp/2.0.4/OS/.

EDI. Electronic Data Interchange. http://en.wikipedia.org/wiki/Electronic_Data_Interchange.

EDIINT-AS1. T. Harding, R. Drummond, C. Shih, MIME-based Secure Peer-to-Peer Business Data Interchange over the Internet, RFC 3335, Sept 2002. http://www.ietf.org/rfc/rfc3335.txt.

EDIINT-AS2. D. Moberg, R. Drummond, MIME-Based Secure Peer-to-Peer Business Data Interchange Using HTTP, Applicability Statement 2 (AS2), RFC 4130, July 2005. http://www.ietf.org/rfc/rfc4130.txt.

EFT. Electronic Funds Transfer. http://en.wikipedia.org/wiki/Electronic_funds_transfer.

EPCglobal. Electronic Product Code Global. http://www.gs1.org/productssolutions/epcglobal/.

GDSN. GS1 Global Data Synchronisation Network. http://www.gs1.org/productssolutions/gdsn/.

GS1. Global Standard One. http://www.gs1.org/.

GS1 GDD. Global Standard One, Global Data Dictionary. http://gdd.gs1.org/.

GS1 XML. Global Standard One XML. http://www.gs1.org/productssolutions/ecom/xml/.

HL7. Health Level 7. http://www.hl7.org/.

IATA. International Air Transport Association. http://www.iata.org/index.htm.

ICH. ANSI ASC X12 ISA Interchange Control Header Segment. http://www.rawlinsecconsulting.com/x12tutorial/x12syn.html.

ICHS. UN/EDIFACT UNB Interchange Header Segment. http://www.unece.org/trade/edifact/untdid/d422_s.htm.

ISO Codes. International Standards Organization Codes. http://www.unece.org/cefact/codesfortrade/codes_index.htm.

ISO11179. ISO/IEC 11179-5: Naming and identification principles. http://standards.iso.org/ittf/PubliclyAvailableStandards/c035347_ISO_IE%C_11179-5_2005(E).zip.

Medjahed, et. al. B. Medjahed, B. Benatallah, A. Bouguettaya, A.H.H. Ngu, A.K. Elmagarmid, Business-to-Business Interactions: Issues and Enabling Technologies. The International Journal on Very Large Data Bases, Vol. 12, No. 1, May, 2003.

MIT-AutoID. Auto-ID Labs at MIT. http://autoid.mit.edu/cs/.

MoU. Memorandum of Understanding on electronic business between IEC, ISO, ITU, and UN/ECE. http://www.itu.int/ITU-T/e-business/files/mou.pdf.

NES. UBL Northern European Subset. http://www.nesubl.eu/.

OAGi. Open Applications Group. http://www.openapplications.org/.

OAGIS. Open Applications Group Integration Specification 9.0. http://www.openapplications.org/downloads/oagis/loadfrm9.htm.

OAGIS-Usage. Open Applications Group (OAGi) at 10 Years: A Look Back and Forward. http://webservices.sys-con.com/read/47282.htm.

ODETTE. Organisation for Data Exchange by Tele Transmission in Europe. http://www.odette.org/html/home.htm.

OIOUBL. Offentlig Information Online UBL. http://www.oio.dk/dataudveksling/ehandel/hoeringer/oioubl.

Oracle. Oracle Corporation. http://www.oracle.com/products/middleware/docs/oracle_ebs_and_soa.pdf http://www.oracle.com/technology/products/applications/integration/1147%_EBS_and_SOA.ppt.

OTA. OpenTravel Alliance. http://www.opentravel.org/.

OWL. Web Ontology Language. http://www.w3.org/2004/OWL/.

PIDX. Petroleum Industry Data Exchange. http://www.pidx.org/.

RosettaNet. http://www.rosettanet.org/.

Rowell. M. Rowell, The Open Applications Group Integration Specification. http://www.ibm.com/developerworks/xml/library/x-oagis/.

SAP. SAP - Systemanalyse und Programmentwicklung. http://www.sap.com/index.epx.

Schematron. http://www.ldodds.com/papers/schematron_xsltuk.html.

STAR. Standards for Technology in Automotive Retail). http://www.starstandard.org/.

Stuhec. Gunther Stuhec, How to Solve the Business Standards Dilemma - The CCTS based Core Data Types. https://www.sdn.sap.com/irj/sdn/go/portal/prtroot/docs/library/uuid/500%db5c9-0e01-0010-81aa-d73cdd30df9a.

Stuhec2. Gunther Stuhec, How to Solve the Business Standards Dilemma: The Context Driven Business Exchange. https://www.sdn.sap.com/irj/servlet/prt/portal/prtroot/docs/library/uui%d/a6c5dce6-0701-0010-45b9-f6ca8c0c6474.

Svefaktura. Swedish Invoice. http://www.svefaktura.se/SFTI_Basic_Invoice20051130_EN/SFTI\%20Basic\%2%0Invoice_1.0/index.html.

SWIFT. Society for Worldwide Interbank Financial Telecommunication. http://www.swift.com/.

UBL. Universal Business Language. http://www.oasis-open.org/committees/ubl/.

UBL NDR. Universal Business Language Naming and Design Rules. http://docs.oasis-open.org/ubl/os-UBL-2.0/doc/ndr/NDR-checklist.pdf.

UBL-SBS. Universal Business Language Small Business Subcommittee. http://www.oasis-open.org/committees/sc_home.php?wg_abbrev=ubl-sbsc.

UBLSchemas. Universal Business Language 2.0 Schemas. http://docs.oasis-open.org/ubl/os-UBL-2.0/.

UBP. Universal Business Process. http://docs.oasis-open.org/ubl/cs-UBL-1.0-SBS-1.0/universal-business-pr%ocess-1.0-ebBP/.

UCC. Uniform Code Council. http://www.uc-council.org/.

UN/CCL. United Nations Core Component Library. http://www.unece.org/cefact/codesfortrade/unccl/CCL07A.xls.

UN/EDIFACT. United Nations Directories for Electronic Data Interchange for Administration, Commerce and Transport. http://www.unece.org/trade/untdid/welcome.htm.

UN/EDIFACT 1131. UN/EDIFACT 1131 Data Element, Code list identification code. http://www.unece.org/trade/untdid/d00a/tred/tred1131.htm.

UN/EDIFACT 3055. UN/EDIFACT 3055 Data Element, Code list responsible agency code, note = http://www.unece.org/trade/untdid/d00a/tred/tred3055.htm.

UN/SBDH. UN/CEFACT Standard Business Document Header Technical Specification. http://www.gs1.org/docs/gsmp/xml/sbdh/CEFACT_SBDH_TS_version1.3.pdf.

US/DOT. US Department of Transportation UBL Implementation. http://www.oasis-open.org/committees/ubl/faq.php.

Vollmer. Ken Vollmer, B2B Integration Trends. Forrester, http://www.forrester.com/Research/Document/Excerpt/0,7211,42735,00.html%.

Wüstner. E. Wustner, T. Hotzel, P. Buxmann, Converting Business Documents: A Classification of Problems and Solutions using XML/XSLT. Proc. of the 4th IEEE Intl Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS), 2002.

X12. EDI ANSI X12. http://www.x12.org/.

xCBL. XML Common Business Library. http://www.xcbl.org/.

XML. Extensible Markup Language. http://www.w3.org/XML/.

XSL. Extensible Stylesheet Language. http://www.w3.org/Style/XSL/.

Yarimagan1. Y. Yarimagan, A. Dogac - Semantics Based Customization of UBL Document Schemas. Journal of Distributed and Parallel Databases, Springer-Verlag, Volume 22, Numbers 2-3 / December 2007, pp. 107-131.

Yarimagan2. Y. Yarimagan, A. Dogac - A Semantic based Solution for the Interoperability of UBL Schemas. http://www.srdc.metu.edu.tr/webpage/publications/2007/YarimaganDogac-UB%LTranslationPaper.pdf.

Ye. Y. Ye, D. Yang, Z. Jiang, L. Tong - Ontology-based semantic models for supply chain management. The International Journal of Advanced Manufacturing Technology, Published online, Springer London, May 2007.

Received ; revised

Multi-Event Adaptive Clustering (MEAC) Protocol for Heterogeneous Wireless Sensor Networks

Ghalib A. Shah, Muslim Bozyiğit
Department of Computer Engineering
Middle East Technical University, Ankara
{asad, bozyigit}@ceng.metu.edu.tr

Özgür B. Akan
Department of Electrical Engineering
Middle East Technical University, Ankara
akan@eee.metu.edu.tr

***Abstract*— Clustering of nodes is one of the most effective approach for conserving energy in wireless sensor networks (WSNs). Cluster formation protocols generally consider the heterogeneity of sensor nodes in terms of energy difference of nodes but ignore the different sensing mechanisms (multiple events detection) of them. Observing different type of phenomenas and reporting them at different rates is an important factor effecting the homogeneity. It is, therefore, imperative to consider the multi-event sources in the design of clustering protocols. In this paper, a multi-event adaptive cluster ($MEAC$) formation protocol is proposed that aims to conserve the energy of sensor nodes in the presence of heterogeneity. It is achieved by considering three design factors; (1) electing an appropriate node to function as cluster-head, (2) limiting the number of clusters in the network and (3) reducing the frequency of clusters reformation. Performance evaluation results show that MEAC improves the stability and energy conservation of the heterogeneous wireless sensor networks.**

***Index Terms*— Clustering, Heterogeneous WSN, Multi-Events Adaptive**

I. Introduction

RECENT advances in the field of wireless sensor networks (WSN) have lead to the revolution of Ambient systems. Ambient systems are networked embedded systems intimately integrated with the everyday environment and are supporting people in their activities. Smart environment is a key example of such a system. Smart environment relies purely on the sensory data observed from the real world. The information needed by such an application is provided by hundreds or thousands of low-power nodes of same or different types which are densely deployed in the environment. These nodes are responsible for sensing as well as relaying the data to a central node called sink. It is, therefore, desirable to make these nodes as energy-efficient as possible to increase the lifetime of the individual sensor nodes as well as the network.

Sensor networks are characterized by a highly dynamic topology, due to a significant level of node failures (e.g. because of energy depletion) or re-energizing caused by deploying new nodes. Therefore, the network must be able to periodically reconfigure itself so that it can continue to function. The implementation of self-configuration then become a requirement in order to guarantee efficient network operation.

The importance of self-configuring clustering protocols for heterogeneous wireless sensor networks has been highlighted in [1]. It emphasizes that any clustering protocol should consider the node energy and traffic rate as key elements. A number of protocols [3], [5], [7], [8], [10], [11], [12], [13] have been proposed for WSN. In most of these studies, sensor nodes are assumed to be homogeneous. However, depending on the application, sensor nodes can have different role or capability making the network heterogeneous. These special sensors can be either deployed independently or the different functionalities can be included in the same sensor nodes. For example, some applications might require a diverse mixture of sensors for monitoring temperature, pressure and humidity of the surrounding environment and capturing the image or video tracking of objects. Even data reading and reporting can be generated from these sensors at different rates and can also follow multiple data reporting models.

In this paper, we present Multi-Event Adaptive Clustering (MEAC) protocol for heterogeneous wireless sensor networks. MEAC constructs clusters to cope with uniform as well as non-uniform deployment of nodes in heterogeneous wireless sensor network. It uses an application-oriented weight-based clustering algorithm to select optimal number of cluster-heads.

Generally, the clustering protocols [5], [7] focus on the currently available energy of the nodes and periodically reorganize clusters to do energy balancing. But this strategy is not practical when the nodes are sending traffic at different rates due to different events characteristics. If all the nodes have the same probability to become cluster heads then the nodes reporting events at higher rate will eventually loose their energy earlier than the others. Therefore, the load on a node, that is, its data

rate is used as a key factor for cluster-head selection. Hence, MEAC protocol distributes the energy usage of nodes by adapting to the multiple events in the field, in order to increase the stable period of the network. Simulation results have shown that, MEAC is more energy efficient than existing clustering protocols and is capable of handling the network dynamics.

The remainder of the paper is organized as follows. In Section II, we highlight the basic assumptions and a model of WSN. We derive the parameters for optimal configuration of nodes in Section III. The protocol operations are described in Section IV. Finally, performance evaluation and results obtained by a simulation model are considered in Section V.

II. System Model

This section describes the model of heterogeneous wireless sensor network where the heterogeneity of nodes is considered regarding their different initial energy levels and observation of multiple events. Multiple events can be detected either by a single node or nodes have different sensing mechanisms to detect different events. Multi-events observation generate different reporting rates due to the different characteristics and requirement. Hence, it greatly effects the energy consumption of nodes.

The network is stable as long as all the nodes are alive and they have enough energy to detect and relay packets. This period is known as *stable period* of the network and the throughput is maximum during this time. In heterogeneous wireless sensor networks, sensor nodes may have different energy levels and might report events at different data rate. The nodes having initial energy E_o are termed as energy-constrained (EC) devices and all other nodes having energy higher than E_o i.e. $E_o + \delta$ are energy-rich (ER) devices. The degree of heterogeneity is also affected by multiple data rates in the network. Other factors like computational or memory capabilities also contribute to network's heterogeneity, but they are not considered in this work.

The degree of the heterogeneity (λ) is due to change in energy level and data rate that can be measured as:

$$\lambda = \lambda_{energy} + \lambda_{rate} \tag{1}$$

where λ_{energy} is the contribution of energy to λ and λ_{rate} is contribution in λ due to different data rates.

In next subsections, we show how individually λ_{energy} and λ_{rate} contribute to the overall degree of heterogeneity.

1) Contribution of λ_{energy} to λ: Let us assume that there exist *m* number of ER nodes among total of *n* nodes in the network. Let $\delta_1, \delta_2, ..., \delta_i, \delta_j, ..., \delta_m$, be the extra energies of *m* nodes. When δ is constant for all *m* nodes, then $\delta_i = \delta_j$; where $i \neq j$. In other words, there are only two different kind of nodes having energy levels E_o or $E_o + \delta$. In this case, λ is large when $m \approx \frac{n}{2}$, but λ is small when $m < \frac{n}{2}$ or $m > \frac{n}{2}$ for $m \leq n$. Then the fraction of nodes (m_λ) making the network heterogeneous due to energy is:

$$m_\lambda = 1 - |\frac{n - 2m}{n}| \tag{2}$$

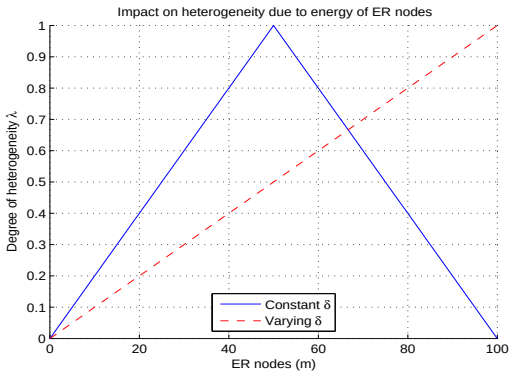


Fig. 1. Heterogeneity due to number of *m* nodes.

On the other hand, when $\delta_i \neq \delta_j$, where $i \neq j$; ER nodes have different energy levels above E_o. Hence, energy levels of ER nodes will be $E_o + \delta_1, E_o + \delta_2, ..., E_o + \delta_i, E_o + \delta_j, ..., E_o + \delta_m$. This behavior of δ is depicted in Fig 1, λ increases continuously by increasing *m*. Therefore, when δ is variable, the fraction m_λ is simply m/n. Let α be the energy factor that ER nodes have higher than EC nodes. We can calculate α as:

$$\alpha = \frac{1}{mE_o} \sum_{i=1}^{m} (E_i - E_o) \tag{3}$$

The above equation can be simplified for constant δ as:

$$\alpha = \frac{m(E_i - E_o)}{mE_o} = \frac{E_i}{E_o} - 1 \tag{4}$$

Therefore, heterogeneity due to energy λ_{energy} or the energy gain in the network due to ER nodes is α times m_λ i.e. $\lambda_{energy} = \alpha \times m_\lambda$

2) Contribution of λ_{rate} to λ: Let ρ_o be the lowest initial data rate of a node(s) in the network and *q* be the number of nodes among the total of *n* nodes which have data rate higher than ρ_o in the network. Similar to Eq. 2 for constant δ, the fraction of nodes (q_λ) making the

network heterogeneous due to data rate difference can be calculated as:

$$q_\lambda = 1 - |\frac{n-2q}{n}| \tag{5}$$

Let $\rho_i = \rho_o + \delta$ be the data rate of node i, then the data rate fraction φ that *k* nodes produce more than ρ_o can be defined similar to α as:

$$\varphi = \frac{1}{q\rho_o}\sum_{i=1}^{q}\rho_i - \rho_o \tag{6}$$

The simplified equation of φ for constant δ is:

$$\varphi = \frac{q(\rho_i - \rho_o)}{q\rho_o} = \frac{\rho_i}{\rho_o} - 1 \tag{7}$$

Hence, the fraction of heterogeneity due to different data rates (λ_{rate}) can be given as $q_\lambda \times \varphi$.

For example, assume that the network contains 30% ER nodes having $E_o = 1.5j$ and EC nodes with $E_o = 0.5j$. 10% nodes report readings at 40 packets/sec while the other nodes at 10 packets/sec. We find out the value of λ. We get $\alpha = 0.2$ by using Eq 3 and $m_\lambda = 0.6$ due to constant change δ in all the *m* nodes. Similarly, for $\rho_o = 10$, $q = 10$, the higher data rate factor $\varphi = 3.0$ making $q_\lambda = 0.2$. Hence, λ is 1.8 by using Eq 1.

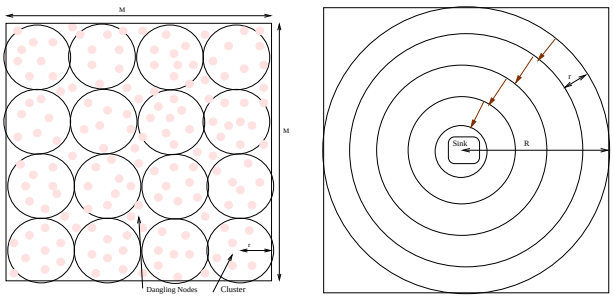
III. Design Parameters of MEAC Protocol

In this section, the design parameters for clustering protocol are derived. We call the node responsible for collecting the data locally as *cluster-head* and the other nodes in a group as *members* of the cluster. Under some circumstances, there may exist some nodes which have not joined any group or cluster are referred as *dangling nodes* (DN). All the member nodes transmit their data packets to their cluster-heads. The basic design of MEAC consists of calculating the optimal number of clusters (k_{opt}) and the optimal members of a cluster (N_{opt}).

A. Optimal Number of Clusters (k_{opt})

MEAC computes optimal number of clusters (k_{opt}) such that it decreases the energy consumption, while providing high degree of connectivity. We devise a formula to find k_{opt} that is directly proportional to total number of nodes and the area of the network, while it is inversely proportional to the transmission range.

Let r be the transmission radius of each node regardless of its functioning. In the clustering process, there is some probability that a number of DN nodes may exist due to the deployment of nodes or coverage of the elected cluster-head. To find out the probability of such nodes, we map the sensor field ($M \times M$) to non-overlapping circles of radius r as shown in Fig 2(a) and assume that the nodes lying outside the boundary of the circle are DN nodes and the others are member nodes. These DN nodes affiliate to cluster-heads through the nodes insides the circle (member nodes) and become the multi-hop members of cluster-heads. The square field M^2 can be packed by $M^2/(2r)^2$ non-overlapping circles of radius r. Thus, the probability p_{DN} of a multi-hop member is



(a) (b)

Fig. 2. Network model to formulate the optimal clusters. Fig 2(a) represents the model to find the probability of DN nodes. Fig 2(b) illustrates the model of routing packets from cluster-heads to the sink node.

$$p_{DN} = \frac{M^2}{(2r)^2} \times \frac{(2r)^2 - \pi r^2}{M^2} \approx 0.214$$

Let E_{elec} be the energy consumed by the electronic circuitry in coding, modulation, filtration and spreading of the signal. Whereas, $\epsilon_{amp}r^2$ is the energy consumed for signal amplification over a short distance r. Thus, the energy consumed by each member node is

$$E_{Member} = l(E_{elec} + \epsilon_{amp}r^2(1 + p_{DN}))$$

where l is the size of data packet. The above equation can be simplified by taking the area as circle given in Eq. 16 of [6].

$$E_{Member} = l(E_{elec} + \epsilon_{amp}\frac{M^2(1+p_{DN})}{2\pi k})$$

Let us assume that the sensory field is covered by a circle of radius R, where the sink node lies at the center of this circle as shown in Fig 2(b). This assumption is made for sending packets from cluster-heads to the sink. Cluster-heads do not extend their transmission range and, therefore, has the same radius r as member nodes. This adapts the multi-hop model proposed by [4] to route packets from cluster-head to the sink.

In the model, a circle is divided into concentric rings with the distance of r. The energy spent to relay the packet from outside ring towards inside ring is $l(2E_{elec} + \epsilon_{amp}r^2)$. The number of hops H_{CH-S} require to route

packet from cluster-head to sink node can be calculated by $\frac{R}{r}(1-p_{hops})$, where p_{hops} is the probability indicating the distance in terms of hops to the sink. This probability can be calculated by using the nodes distribution in the rings given in [4].

$$p_{hops} = \frac{r}{R}\sum_{i=1}^{R/r}\frac{R^2-(ir)^2}{M^2}$$

Packets from cluster-heads that are far from the sink are relayed through intermediate nodes. Therefore, if N_s is the number of neighbors of the source node s then $N_s \times E_{elec}$ is the energy consumed by the electronic circuitry of the neighbors during the transmission of a data packet by s. The number of neighbors N_s of a node can be computed by $n\frac{\pi r^2}{M^2}$. Hence, the energy consumed in forwarding data from cluster-head to sink is measured as:

$$\begin{aligned} E_{CH-S} &= l(N_s E_{elec} + E_{elec} \\ &+ (2E_{elec} + \epsilon_{amp} r^2 + N_s E_{elec}) H_{CH-S}) \end{aligned}$$

The total energy dissipated by the network is

$$\begin{aligned} E_{total} &= l((n + nN_s)E_{elec} \\ &+ k(2E_{elec} + \epsilon_{amp} r^2 + N_s E_{elec}) H_{CH-S} \\ &+ n\epsilon_{amp}\frac{M^2(1+p_{DN})}{2\pi k}) \end{aligned}$$

For $r < R$, the optimal value of k can be found by taking the derivative of above equation with respect to k and equating to zero

$$k_{opt} \approx \sqrt{\frac{n(1+p_{DN})}{(2\pi(1+\frac{2E_{elec}}{\epsilon_{amp} r^2}+\frac{N_s E_{elec}}{\epsilon_{amp} r^2}))H_{CH-S}}} \times \frac{M}{r} \quad (8)$$

The optimal value depends on the transmission range r. For long range transmission, the value of optimal clusters k_{opt} is small. For example, Let $n = 100$, $M = 100$ and the sink is at the center of the field $(x = 50, y = 50)$. Then the value of radius R is obtained by drawing a circle at $x = 50, y = 50$ to cover the field. The estimated value is $R = 60$ and let set the range r of individual nodes to 25. In this scenario, we obtain the value of $k_{opt} \approx 10$. By increasing the range of nodes to 40 meters, we obtain $k_{opt} \approx 7$. Whereas, the value of k_{opt} in SEP [7] is 10 regardless of the transmission coverage of individual nodes.

B. Optimal Cluster Size (N_{opt})

When the deployment is uniform, the optimal value of member nodes N_{opt} can be easily found by n/k_{opt}. However, for non-uniform deployment, the number of member nodes depends on the density in a particular zone of the sensor field. Therefore, we put the maximum and minimum limits N_{Min} and N_{Max} respectively on the size of cluster, such that, we still achieve k_{opt} clusters in non-uniform deployment. Let N_i be the number of neighboring nodes of any ith node. $Max(N_i)$ is the maximum number of neighboring nodes that any of the ith neighbor node have. We measure density of nodes in a particular zone by comparing the neighbor nodes N_i with N_{opt}. It can be concluded that the deployment is:

$$\begin{aligned} N_i/N_{opt} &> 1, \quad dense \\ N_i/N_{opt} &\approx 1, \quad uniform \\ N_i/N_{opt} &< 1, \quad sparse \end{aligned}$$

Therefore, the number of nodes in a cluster can be constrained by setting the lower bound N_{Min} and upper bound N_{Max} according to the deployment as:

$$N_{Max} = Max(N_{opt}, Max(N_i))$$

That is, the maximum of N_{opt} and maximum number of neighbors of any cluster-head at the time of cluster formation.

$$N_{Min} = N_{opt} \times Min(N_{opt}, Max(N_i))/N_{Max}$$

These limits allow the configuration to manage the dense as well as sparse deployment of nodes.

IV. MEAC Operation

When nodes are initially deployed in the field, every node broadcasts hi beacons. A receiving node updates its neighborhood table {*ID, Weight, Energy, Neighbors, Hops, Expiry*}. If a node does not hear any hi beacon from a neighbor during the duration of $Expiry$ value, it is considered unreachable and is deleted. These beacons are exchanged periodically to deal with the network dynamics. After the exchange of hi beacons initially, every node calculates K_{opt} and N_{opt} values. The clustering process starts by electing a cluster-head and then linking the clusters together to form a hierarchical clustered network.

A. Cluster-head Election Procedure

Cluster-heads are elected by effectively combining the required system parameters with certain weighting factors. Every node calculates its weight based on its available power, data rate and the density of nodes. Values of these factors can be chosen according to the application needs. For example, power control is very

Algorithm 1 Elect Cluster-head

```
Pseudo-code executed by each node N in each round
W_max = 0
for all neighbor N_i do
    if W_max < W_n then
        W_max = W_n
    end if
end for
W_i = my-weight()
if status = NONE then
    if W_i > W_max then
        announce-head()
        W_th = W_i × threshold_factor
    else if status = HEAD then
        if W_i < W_th then
            if W_i < W_max then
                withdraw-head()
            else
                W_th = W_i × threshold_factor
            end if
        end if
    end if
end if
```

important in CDMA-based networks. Thus, weight of the power factor can be made larger. In order to achieve the goal of energy saving, it minimizes the frequency of cluster reformations. It is achieved by encouraging the current cluster-heads to remain cluster-heads as long as possible. On the other hand, the distance weighting factor can be made larger if the density of nodes is high or the deployment is made in hostile environment. This ensures that a node is elected as a cluster-head that can receive the transmission from farther nodes and the number of clusters formed remain close to the optimal value.

Let D_i be the average distance of node i to its neighbors, N_i be the total number of its neighbors, E_i be its available energy and ρ_i be its reporting rate. Node i computes its weight W_i as:

$$W_i = c_1 \frac{\rho_o}{\rho_i} \times \frac{E_i}{E_o} + c_2 \frac{D_i}{r} \times \frac{N_i}{N_{opt}} \tag{9}$$

where the coefficients c_1, c_2 are the weighting factors for the energy and data rate parameters. Node i announces itself a cluster-head if its weight is high among all its neighbors and sets its threshold $W_{Th} = cW_i$, where c is the threshold adjusting factor that can be set relative to λ. The pseudo-code of the operations executed by a sensor node in each round of cluster formation is reported in Algorithm 1.

1) Adaptivity to Multi-Events: It can be seen from Eq. 9 that a node having higher energy level than its neighbors is the potential candidate of becoming cluster-head. However, a node reporting events at higher rate is less likely to be elected as cluster-head. The weighting equation includes the ratio of data rate (ρ_o/ρ_i) to consider the multiple data rates due to different events. If a node i has $\rho_i > \rho_o$ then its weight is reduced. Therefore, it has lesser chances to become cluster-head than the other nodes. It is due to the fact that nodes sending packets at higher rates exhaust their energy soon and, thereby, the probability of becoming cluster-heads is reduced.

2) Reducing Clustering Reformation: MEAC reduces the frequency of clusters reformation by setting weight threshold at the time of cluster formation. In each round, each cluster-head recomputes its weight and compares with its threshold value. If W_i of cluster-head i is higher than its W_{Th} value then it keeps functioning as head. if $W_i < W_{Th}$ then it checks whether its W_i is also lower than any of its member node weight. If so, it withdraws itself from being cluster-head and cluster election procedure is initiated.

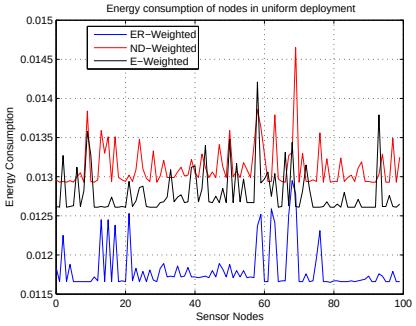
B. Inter-Clusters Connectivity

Energy of nodes is conserved by using minimum transmission energy (MTE) scheme. Obviously, MTE scheme requires multi-hop routing when the sink is not in the transmission range of cluster-head. Thus cluster-heads are linked allowing the packets forwarded through the clusters on the path toward the sink node. Once the clusters are formed after the first round of cluster formation, member nodes keep updating their heads about any adjacent cluster found. The member nodes also forward the membership request of any DN node to heads. It works as follows:

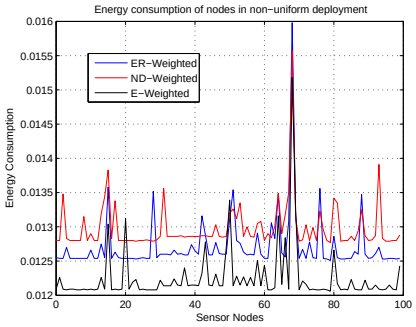
Each node keeps broadcasting hi beacon that contains its cluster-head ID or empty for DN nodes. If a receiving node N_i has different cluster-had ID, a neighbor cluster is found. N_i sends the NEIGHBOR-CLUSTER message to its cluster-head that updates its neighboring cluster table. When the cluster-head ID field is empty in the periodic beacon, member node considers it as membership request for DN and forwards it to its cluster-head. Cluster-head checks whether its member nodes does not exceed the limit N_{Max}. If the limit is not reached yet then it replies with the ACCEPTED message otherwise REFUSED message. N_i ignores the REFUSED message from its head but forwards the ACCEPTED message to the DN node.

C. Adaptivity to Nodes Deployment

One of the key issues in WSN is the deployment of mobile sensor nodes in the region of interest (ROI) [15]. Before a sensor can report observation to the monitoring system, it must be deployed in a location that is contextually appropriate. Optimum placement of sensors results in the maximum utilization of the energy of nodes. However, The deployment can not be determined *a priori* when the environment is unknown or hostile in which case the sensors may be air-dropped from an aircraft [16] or deployed by other means. The proper choice for sensor locations based on application requirements is difficult.



(a)



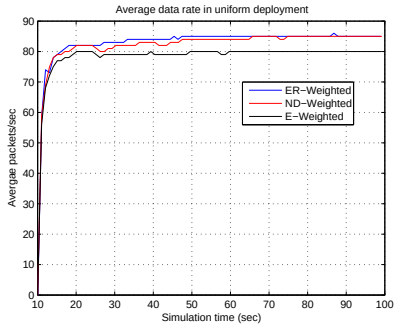
(b)

Fig. 3. Energy consumption in heterogeneous WSN ($k = 30\%, \varphi = 1, m = 20\%, \alpha = 1$) for uniform 3(a), and non-uniform 3(b) deployment of nodes.

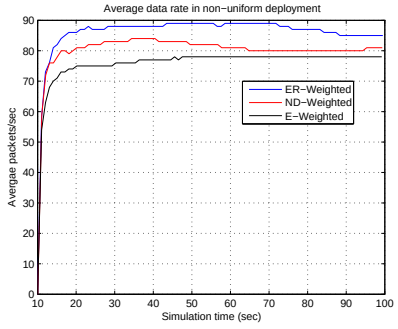
The deployment pattern of sensor nodes greatly affect the performance of the self-configuring clustering protocols. Due to the unpredictable distribution of nodes, MEAC takes into consideration the different system parameters as described in Section IV-A to adapt to the deployment. The performance is evaluated by weighting the parameters according to uniform deployment of nodes as well as non-uniform deployment. We create two different scenarios of deployment; first, 100 nodes are uniformly deployed in 100×100 meters area and second, 50 nodes are deployed in 100×100 meters area at first and then 50 more are dropped in 50×50 meters area of the same region to make it non-uniform.

The performance is measured in both scenarios by adjusting the weighting factors. The experiments are run by; (1) keeping the weighting factors of energy and reporting rate (ER-Weighted) high, (2) considering only the energy parameter (E-Weighted) and (3) setting the factor of neighbor nodes and distance (ND-Weighted). Fig 3 illustrates the energy consumption for both scenarios.

Clearly, the energy consumption is small when the weighting factor of ER parameters is set large in uniform case as shown in Fig 3(a). The energy gain in considering R parameter along with E is about 12% as compared to just E-Weighted clustering that the heterogeneity-aware clustering protocol exploit [5], [7]. Fig 4(a) shows that the reporting rate (packets/sec) is also 7% higher in ER-Weighted approach than the E-Weighted clustering approach. Hence, by including the data rate due to multiple events detection in the clustering of sensor nodes, it not only achieves the gain in energy but also high data rate.



(a)



(b)

Fig. 4. Average reporting rate in heterogeneous WSN ($q = 30\%, \varphi = 1, m = 20\%, \alpha = 1$) for uniform 4(a), and non-uniform 4(b) deployment of nodes.

For non-uniform case, Fig 3(b) illustrates that although the energy consumption is 4% lower in E-Weighted than ER-Weighted, the data rate is also lower by 15%. Therefore, the 4% gain is not actually due to the effeciency of

E-Weighted approach but due to the fact of low data rate. Even if all the sensor nodes have same data rate, ND-Weighted approach delivers events at higher rate than E-Weighted as shown in Fig 4(b), with some extra cost of energy. Hence, there is a tradeoff between high data rate and energy consumption.

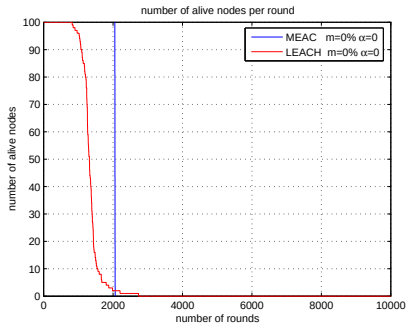
V. Performance Evaluation

We evaluate the performance of protocol in terms of energy consumption, network stability with multiple events and throughput metrics. The heterogeneous WSN is composed of nodes of different energy levels and sensing modules for multiple events detection. The example scenario of wireless sensor networks consists of 100 sensors deployed randomly in a field of 100×100. The sink node is placed at the center of field i.e. $x = 50, y = 50$. The initial energy E_o of EC nodes is set to 0.5 joules. The transmission and reception power is set to 50 nJ/bit and sources produce traffic at 4 kbps.

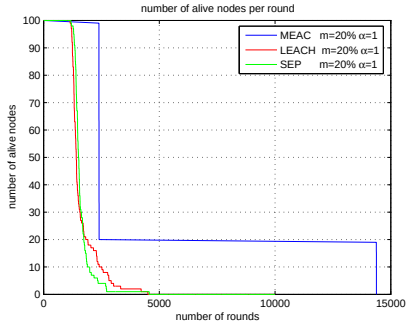
1) Energy: The energy efficiency of MEAC is compared with SEP and LEACH. Both SEP and LEACH periodically elect cluster-heads to balance the energy of nodes. Fig 5 illustrates a detailed view of the behavior of MEAC, LEACH and SEP for different values of the parameters. The number of alive nodes are plotted for the scenarios ($m = 0\%$, $\alpha = 0$), ($m = 20\%$, $\alpha = 1$) and ($m = 20\%$, $\alpha = 3$) in Fig 5(a), 5(b) and 5(c) respectively. Unlike SEP and LEACH, MEAC considers the available energy to elect cluster-heads and a node keep working as head as long as its available energy is higher than its theshold value. This approach reduces the frequency of cluster-head election.

It is obvious in Fig 5(a) that MEAC extends the stable region compared to LEACH by 55% for homogeneous network. The behavior of SEP is the same for $m = 0$ and, therefore, the gain in stability is similar to LEACH. Fig 5(b) shows the results for $m = 20\%$ and $\alpha = 1$ parameters. The stable period is 41% and 33% more than LEACH and SEP respectively. Besides the stable period, the unstable period is also quite large which keep the network alive for 250% more than LEACH and SEP. Fig 5(c) illustrates the stability gain of MEAC for $m = 20\%$ and $\alpha = 3$. MEAC achieves the gain of 58% in comparison with LEACH and 35% from SEP. The unstable region is remarkably larger than both these candidate protocols.

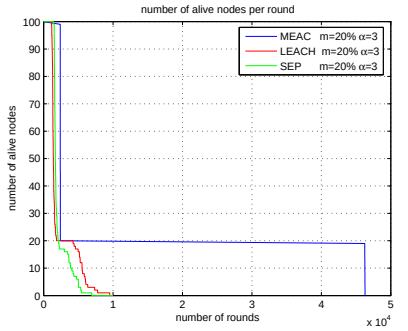
2) Stability with Multiple Events: Fig 6 shows the impact of multiple data rates with and without deployment of ER nodes. The stability of network increases by deploying more and more number of ER nodes. Whereas, it decreases when sensor nodes are reporting events at different rates to sink. It is obvious from Fig



(a)



(b)



(c)

Fig. 5. Energy consumption comparison among MEAC, LEACH and SEP in the presence of heterogeneity due to energy for $\alpha = 0$ in 5(a), $m = 20\%, \alpha = 1$ in 5(b) and $m = 20\%, \alpha = 3$ in 5(c).

6 that the stability is high in the presence of ER nodes ($\lambda_{energy} > 0$). The extra energy of ER nodes is utilized to accommodate the high data rate. If we keep increasing λ_{rate} (by increasing φ) then the loss in stability is very small as compared to increase in φ.

3) Throughput: MEAC does not imply any aggregation technique at cluster-heads because it does not suit for the reliability measure in terms of packet delay for delay-sensitive applications. Fig 7 shows the throughput comparison of MEAC with LEACH and SEP. MEAC aims to provide in-time packet delivery and sacrifices some throughput at cost of packet delay that increases due to aggregating data [17]. Although the throughput in MEAC is less than SEP but it continues for the

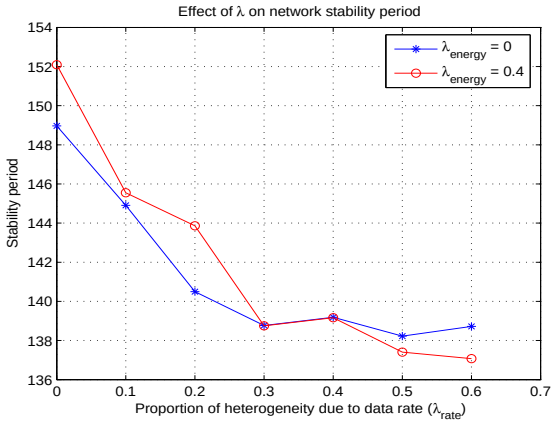


Fig. 6. MEAC stable period in heterogeneous WSNs for different values of λ_{rate}

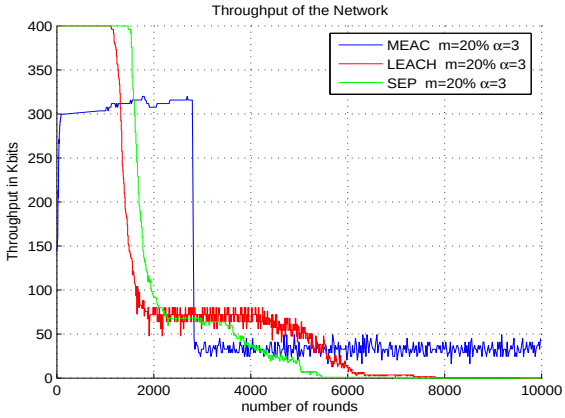


Fig. 7. Throughput comparison among MEAC, LEACH and SEP.

longer time than in LEACH and SEP. Therefore, the low throughput is compensated by longer period. It is observed that when the ER nodes are close to sink then the throughput is high in unstable period but the period is short. When ER nodes are placed far from the sink node then some ER nodes might not reach sink directly or indirectly and, therefore, reduces the throughput but keeps the network alive for longer period. Thus the deployment of ER nodes greatly effect the network performance during unstable period.

VI. Conclusion

A number of clustering protocols have been proposed for heterogeneous wireless sensor networks. However, they do not consider the presence of multiple phenomenon in the sensor field. When a sensor node detects either multiple events or an event whose required reporting rate is higher than the other nodes, it consumes relatively higher energy. $MEAC$ is a cluster-based routing protocol that considers the heterogeneity of nodes due to energy as well as multiple events. MEAC makes use of heterogeneity factors in such a way that energy consumption is reduced and stability period is extended compensating for reduced throughput of non-aggregated data. This conclusion was verified by the simulation experiments compared with SEP and LEACH.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam anf E. Cayirci, "A Survey On Sensor Networks," *IEEE Communications Magazine, pp. 102-114, Aug 2002.*

[2] I. Chatzigiannakis, A. Kinalis and S. Nikoletseas, "An Adaptive Power Conservation Scheme for Heterogeneous Wireless Sensor Networks with Node Redeployment," *Proc. of the 17th annual ACM SPAA'05, Jul 2005.*

[3] E. J. Duarte-Melo and M. Liu, "Analysis of Energy Consumption and Lifetime of Heterogeneous Wireless Sensor Networks," *IEEE GLOBECOM'02. vol. 1 pp. 21-25, Nov 2002.*

[4] V. Mhatre and C. Rosenberg, "Homogeneous vs heterogeneous clustered sensor networks: A comparative study," *IEEE ICC'04 , vol. 27, no. 1, pp. 3646-3651, Jun 2004.*

[5] M. J. Handy, M. Haase and D. Timmermann, "Low Energy Adaptive Clustering Hierarchy with Deterministic Cluster-Head Selection," *IEEE Intl. Conference on Mobile and Wireless, 2002.*

[6] W. R. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless microsensor networks," *Proc. of the 33rd Hawaii Intl. Conference on System Sciences (HICSS-33), Jan 2000.*

[7] G. Smaragdakis, I. Matta and A. Bestavros "SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks," *2nd Intl. Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004).*

[8] W. R. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "Minimizing communication costs in hierarchically-clustered networks of wireless sensors," *Computer Networks, vol. 44, no. 1, pp. 1-16, Jan 2004.*

[9] M. Chatterjee, S.K. Das and D. Turgut, "A Weight Based Distributed Clustering Algorithm for Mobile Ad hoc Networks," *High Performance Computing (HiPC), pp. 511-521, 2000.*

[10] G. Gupta and M. Younis , "Performance Evaluation of Load-Balanced Clustering of Wireless Sensor Networks," *10th Intl. Conference on Telecommunications (ICT'03), vol. 2, Mar 2003.*

[11] A.D. Amis, and R. Prakash, "Load-Balancing Clusters in Wireless Ad Hoc Networks", *Proc. of ASSET, Mar 2000.*

[12] A.D. Amis, R. Prakash, T.H.P. Vuong and D.T. Huynh, "Max-Min-D-Cluster Formation in Wireless Ad hoc Networks," *IEEE INFOCOM, March 2000.*

[13] S. Bandyopadhyay and E.J. Coyle, "An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks," *IEEE INFOCOM 2003.*

[14] S. Basagni, "Distributed Clustering for Ad Hoc Networks," *Proc. of the Intl. Symposium on parallel architectures, Algorithms and Networks, 1999.*

[15] N. Heo and P.K. Varshney, "Energy-efficient deployment of Intelligent Mobile sensor networks," *Systems, Man and Cybernetics, Part A, IEEE Transactions on vol. 35, issue 1, pp. 78-92, Jan 2005.*

[16] S. S. Dhillon and K. Chakrabarty, "Sensor placement for effective coverage and surveillance in distributed sensor networks," *Proc. of the IEEE Wireless Communication Networking Conference, pp. 1609-1614, 2003.*

[17] T. He, B. M. Blum, J. A. Stankovic and T. Abdelzaher, "AIDA: Adaptive Application-Independent Data Aggregation in Wireless Sensor Networks," *ACM Transactions on Embedded Computing Systems, vol. 3, no. 2, pp. 426-457, May 2004.*

RAT: Routing by Adaptive Targeting in Wireless Sensor/Actor Networks

Ghalib A. Shah, Muslim Bozyiğit
Department of Computer Engineering
Middle East Technical University, Ankara
{e135333, bozyigit}@metu.edu.tr

Demet Aksoy
Department of Computer Science
University of California, Davis
aksoy@cs.ucdavis.edu

***Abstract*— Wireless sensor networks have attracted significant interest for various scientific, military, and e-health applications. Recently a new class of sensor networks "sensor/actuator" has been introducing new research challenges due to the coordination of data push and pull messages. In particular, actuators that have the capability to move and respond to emergencies necessitate adaptive destination targeting in comparison to static node-to-sink communication. In this regard, little progress has been made in terms of the co-existing push and pull data flows in the network and also the positioning of the actuators in response to observations. In this paper we propose an adaptive targeting (RAT) protocol that allows low-delay sensor-to-actor communication and dynamic coordination of actors in response to emergencies. RAT exploits the mobility of actor nodes to form dynamic responsibility clusters that ensures a specific response time to emergencies. Our simulation results suggest significant performance improvements.**

***Index Terms*— Adaptive Targeting, Routing, Integrated Pull/Push Coordination, Heterogeneous WSAN**

I. Introduction

RECENT advances in wireless technology and the development of small, low-cost, low-energy electronics has led to highly distributed and intelligent sensor networks. Increasing onboard processing capabilities of sensors allow for spontaneous self-organizing sensor networks, dynamic adaptation to failures, and quick reaction to changes in task and network requirements. Such intelligence allows scientists and engineers to deploy autonomous and untethered networks. Various defense, scientific, and engineering applications require such intelligent sensor networks [1] in reconnaissance, surveillance, environmental and habitat monitoring, and wildlife tracking. In practice it is not only important to monitor the environment, but also to react to it. Recently, for certain applications, sensor/actuator configurations have been proposed [2], [4], [5]. These wireless sensor and actor networks (WSANs) consist of nodes that will react to changes in the environment, such as distributed mobile robots, light switches or climate regulators.

Generally, when sensors detect an event (a change in the environment), they communicate with each other to confirm the event or generate new data based on the event, which is reported to the information collector (sink/actuator). There are two ways to disseminate the event data, namely push-based and pull-based or querying approach. Using push policy, event information is sent out to the sink without explicit requests. In contrast, with pull based approach events are reported in response to explicit requests received from sink/actors. In our study we refer to the communication initiated by the actors to disseminate their location and their acting capabilities as pull communication.

It has been long argued that push is more beneficial than pull since it provides scalability and well suits to monitoring application, which demands continuous event reports. On the other hand, it is unnecessary to report a particular event scenario which is not of interest for the applications. One focus of this paper is to minimize data communications through pull/push coordination. There have been a number of studies for developing push (data collection) [8], [9], [11], [13], [14] based routing in WSN as well as pull based (querying) approach [15]. Some hybrid push-pull techniques [16], [17] are suggested to overcome the drawbacks of pure push or pure pull based approaches in WSN. However, there exist no integrated push/pull [7] solution for distributed processing of query in wireless sensor and actuator networks.

In this paper, we propose a multi-level communication solution where the network can respond to emergencies without coordinating with a centralized control, when possible. In particular, we propose a routing by adaptive targeting (RAT) protocol that allows low-delay sensor-to-actor communication and dynamic coordination of actors in response to emergencies. RAT comprises of two component; Resilient Delay-constrained geographical-based routing (RED-GEO) and Integrated Pull/Push (IPP) coordination. RED-GEO relays the packets in greedy mode such that delay constraint can be meet as well as energy consumption of forwarding nodes is balanced. In IPP, actor nodes subscribe specific events of their interest in the sensor field and sensor nodes disseminate the event readings to subscribed actor for a time period of subscription life. As a result, actor nodes does not require to send a query every time they need event readings. Similarly, sensor nodes pushes the data as long as there is a subscribed actor interested for the observed event.

Additionally, RAT exploits the mobility of actor nodes to form dynamic responsibility clusters that ensures a specific response time to emergencies. In this regard, we are able to alter the actor positions to keep them close to emergency areas and reduce the response time significantly. Our simulation results suggest significant performance improvements.

The remainder of the paper is organized as follows. A delay constrained geographical-based routing mechanism is proposed in Section II that provides resilient path in a highly

non-uniform deployment of sensor nodes. Section III describes the proposed architecture for adaptive targeting in WSAN. Performance evaluation and results are considered in Section IV. Finally, the paper is concluded in Section V.

II. RESILIENT DELAY-CONSTRAINED GEOGRAPHICAL (RED-GEO) ROUTING MECHANISM

Geographic forwarding is a stateless routing algorithm based on the location information of the neighboring nodes. More specifically, a forwarding node relays an incoming data packet to the neighbor who is the nearest to the destination among its one-hop neighbors. The forwarding set (FS) of a given node can be defined as the neighbor nodes, which are more close to the destination. Generally, the forwarding set is a subset of neighboring nodes, according to the given location of the single destination (sink) in WSN. However, in WSAN, multiple actors, placed or moved at different locations of the field, are the destinations of the sensors readings and, therefore, a relaying node keeps all the neighbors in FS initially. Geographic routing is more attractive in resource constrained nodes since it requires that nodes only maintain the location of their one-hop neighbors and do not need any energy consuming route requests. Routing decision can be made locally and dynamically.

A. Delay Measurement

When nodes are initially deployed in the field, every node i broadcasts *presence beacon*, which is added in the neighbors list by all the nodes that receive this broadcast. Consequently, a neighborhood table or FS is maintained by all the nodes. A node that receives this broadcast, computes the delay T of the packet and records the value in FS along with the entry of sending node.

The value of T in transmitting a packet from one node to a node in its neighbor is measured by the following factors; queue, MAC, propagation and receiving delay represented by T_q, T_{Mac}, T_{Prop} and T_{Rec} respectively. The wireless channel is asymmetric that does not imply any synchronization mechanism. Therefore, the delay is measured partially at both the sender and receiver. Sender measures the delay T_s until the start of transmission that includes the queue delay as well as the MAC contention delay. Whereas, the receiver adds the factor T_r as sum of propagation delay and receiving delay to get the total packet delay T.

$$T_s = T_q + T_{Mac}, \quad T_r = T_{Prop} + T_{Rec}$$

The hop latency T can be computed as sum of these factors:

$$T = T_s + T_r$$

The delay is measured by exchanging the *presence beacon* that represents the load on that node. Each node i maintains this value in its FS that contains the fields $\{ID, T, energy, distance\}$, where the values of $energy$ and $location$ of sender are reported in *presence beacon*. However, this delay value for a particular neighbor node j does not remain constant and varies under different traffic load. The large delay value reflects the high volume of traffic a node is experiencing. As a result, this periodic beacon keeps the neighbors informed about the current state of the node that includes its energy, location and traffic load by computing its delay.

B. Delay-constrained Forwarding Set DFS

RAT focuses on applications that require time critical actions. Such as monitoring a strategic area where emergencies should be responded as quickly as possible. In this regard it is more efficient to report the abnormal event readings in WSAN to nearby actors, rather than a stationary sink node that further needs to communicate with the actor. We define the delay tolerance of events as τ and try to minimize $\tau - \epsilon$ where ϵ represents the time that is required by the actor to respond to sensor input after its arrival. The value of ϵ depends on two factors; first, the time required by an actor to get ready and take certain action and second, how many action requests are in the waiting queue for response. The process of forwarding the data packet towards destination actor consists of two phases:

- Construct a delay-constrained forwarding subset (DFS) of nodes from the set FS such that the given delay constraint τ can be meet. The source node set the time to live (TTL) field in the data packet to $\tau - \epsilon$ and each forwarding node updates the TTL field by deducting the traversed hop delay. The packet is considered obsolete and dropped whenever it reduces to 0.
- Balance the load of the nodes in the subset DFS in terms of energy consumption by selecting the forwarding node from the set DFS that has the higher energy level.

All the nodes in the neighbors list form the forwarding set FS due to different targets at varying locations. More specifically, multiple mobile actors in the field are the possible destinations of the packets generated by the sensor nodes. However, the greedy approach is modified in order to consider the response time demanded by the application and efficient energy consumption of nodes. Hence, a forwarding node is selected such that the response time can be meet as well as energy consumption is balanced. For each possible destination actor, a node i decomposes its set FS into subsets DFS_i^a for each destination actor a as:

$$DFS_i^a = \{j |\ j \in FS_i \wedge D(i,a) > D(j,a) \wedge \frac{TTL}{T(j)} > \frac{D(i,a)}{D(i,j)}\}$$

That is, a node i builds a subset $DFS_i^a \subseteq FS_i$ such that a node $j \in FS_i$ is closer than i to the destination actor a and the expected number of hops with respect to delay from i to j should be larger than the expected number of hops according to the distance between i and j. Where, $T(j)$ is the time required to relay the packet to node j, $D(i,a)$ is the distance between the node i and actor a and $D(i,j)$ is the distance between the node i and node j. From this subset, it selects the node which has the highest residual energy.

When the distance between the node i and j is large, it is high likely that the delay is also large. It is due to the long propagation delay and the intermediate nodes interfering the transmission between them. At larger distance between

sender and receiver, more nodes share the same channel and thereby delaying the transmission. However it is not always true when the traffic is low between two distant nodes. In such case, packets mostly experience the propagation delay which is not dominant at different distance. Therefore, RED-GEO modifies the greedy routing to adapt to different circumstances in the network as well as select the forwarding node which is more close to the destination and guarantee the in-time packet delivery. Intuitively, it is rationale to relate the delay with distance in terms of packet relaying speed. Fig. 1 justifies our delay-constrained routing. Although the hops in greedy routing are smaller than RED-GEO but the delay in RED-GEO is lower than the delay in pure greedy routing.

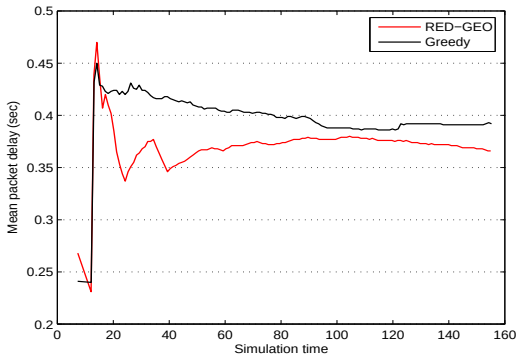


Fig. 1. Delay in RED-GEO vs pure greedy routing.

A node i forwards the request progressively to its neighbor node $j \in DFS_i^a$, which has the highest energy level among the other elements in DFS_i^a. However, if $DFS_i^a = \phi$ then it cannot forward the incoming data packet further; the packet is stuck in a local minimum where FS contains no element for the destination target as shown in Fig. 2 that node 1 and 6 are stuck and find no element in FS. In such a case, there exists a void region [1] and greedy geographical approach fails.

C. Resilient Recovery Mode

The recovery mode for the void region has been proposed in [6], [11], [12]. In [6], exploiting virtual coordinates (VC) in recovery mode is the innovative solution. However this approach seems cost inefficient that requires additional anchor nodes as reference stations to compute the virtual coordinates apart from the localization devices. On the other hand [11], [12] follows the single route along the perimeter of the void region. For a highly non-uniform or terrain deployment of nodes, it is high likely that successive void regions exist as shown in Fig. 2, which introduce significant delay for delay intolerant applications in WSAN in recovering the path.

We overcome the existence of void region by using resilient geographical approach. The term resilient is used in the context that, on detection of void region, a packet is relayed by sending two copies along the left and right side of the region to increase the reliability of packet delivery. In this scenario, node i forwards the request to its left node j such that j is the closest to destination among the nodes on the left of i and similarly, a copy of the packet towards the node k on its right side. Moreover, i is marked as a void forwarding node by the source node for the destination actor. The two nodes j and k, receiving the request from i, transmits the packets individually and selects the forwarding node other than i in a greedy mode. These individual transmissions might reach to a single node x after crossing the void region or they may follow disjoint paths. But it is high likely that they converge to single route after some transmissions and select the same forwarding node x. If a node receives the same packet, it drops the duplicate.

We maintain another set VS of nodes, which have detected void regions for relaying packets to the destination actors. When a packet is transmitted back to left and right of void region, the transmitting node puts its ID as void node ID in the packet. The source of the packet also hears the transmission in addition to the left and right nodes and consequently, it makes an entry of the this ID in the set VS. However, it is possible that a void node for a particular destination actor might be a forwarding node for some other actor. Therefore, the element of set VS have two attributes; the void node ID (NID) and destination actor ID (AID). Before picking an element i from the set DFS as the candidate forwarding node in greedy mode, the relaying node lookups the set VS for i and if it is found then the node i is not selected for relaying the packet and the next closest node is picked from the set DFS.

To terminate the packet forwarding in case of a void region that can never reaches to destination or stuck in some region, source node puts two kinds of information in each data packet; time to live (TTL) and maximum number of hops (H_{max}). As described earlier that TTL is set to $\tau - \epsilon$ and is decremented at each intermediate hop. The value of H_{max} is computed on the basis of distance between source node and destination and the assumption that each forwarding node is half of the transmission radius r closer to the destination a, on average, than the current node i. Therefore, we can compute the limit on number of hops; $H_{max} = 2 \times D(i,a)/r$. The packet is no more relayed and dropped whenever one of the two values TTL or H_{max} becomes 0.

Algorithm 1 Select Forwarding Node

Ensure: forwarding node $i \notin VS$

1: Relaying node executes this procedure to select a forwarding node from the set FS for a destination actor a.
2: $f = NULL$
3: $E_{max} = 0$ {Maximum energy}
4: **for all** $i \in DFS$ **do**
5: E_i represents the energy of node i
6: **if** $E_{max} < E_i$ AND $FindVN(i,a,VS) = NO$ **then**
7: $f = i$
8: $E_{max} = E_i$
9: **end if**
10: **end for**
11: **return** f

[1]The region is called void if a node fails to find the forwarding node in that region, which is more close to the target than itself by using the greedy approach.

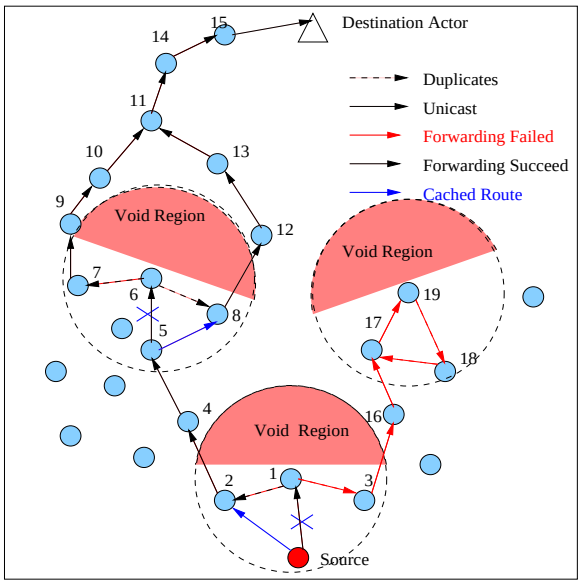


Fig. 2. Packet forwarding from source to destination actor in the presence of void region in a highly non-uniform deployment of nodes.

Fig. 2 demonstrates the recovery mode of geographical based data pushing approach in void region. Source node selects node 1 as the forwarding node initially when the set VS is empty. However node 1 detects a void region due to non-existence of any forwarding node in set DFS, which is close to target. Node 1 transmits the request to the node 2 on its left and node 3 on its right so that a path must be established to make it resilient and in-time delivery. Node 2 and 3 receives the request and forwards the request to 4 and 16 respectively, which are closer to the destination actor D. However, the path is established through node 2 but the path through node 3 can not be established as shown in Fig. 2. The source node adds the the node ID 1 to the set VS and future requests are relayed through the node 2. Similarly, whenever the request is forwarded to node 3, it also encounters the path failure due to void region around node 19 and is also added in the set VS.

D. RED-GEO Overheads

Each node i broadcasts *presence beacon* periodically to maintain the neighbors list that contains the location information (x_i, y_i) of node i and its residual energy E_i. The size of this beacon is 16 bytes that is considered as overhead. The frequency ϕ of the beacon is small because we assume that the sensor nodes are stationary and the neighbors list remains valid as long as the nodes are alive. Therefore, the overheads in terms of bandwidth for maintaining the list of n neighboring nodes sharing the same channel is $(128 \times \phi \times n \ bits/sec)$ of the total bandwidth β. At $n = 20$, $\phi = 0.2$ i.e., a beacon is transmitted after 5 sec, the overhead is only 512 bps which is very trivial cost over 1 Mbps channel capacity. Similarly, the overhead in terms of energy consumption is $(512 \times E_{elec}) + (128 \times 0.2) \times E_{amp}$ which is negligible. where E_{elec} is the energy consumed per bit by the transceiver circuitry of the nodes and E_{amp} is the energy consumed per bit by the sender for signal amplification. If there is no incremental deployment then we further decrease the frequency ϕ to minimize overheads.

III. Location-aware Integrated Pull/Push (IPP) Coordination

In WSAN, appropriate actions corresponding to the sensed phenomenon cannot be performed unless event information is transmitted from sensors to the actor capable of taking action in the event area. Unlike sensors with limited resources, however, actors have abundant resources and can move in the area covered by the sensor network. Actors can therefore use long range transmission to communicate pull requests as opposed to sensors that rely on hoping to report events. Sensors and actors should locally coordinate with each other to trigger the right action. It is assumed that each actor has limited range of action. If an actor does not cover the event area, it should coordinate with the other actors so that the uncovered field is targeted by some actor. Hence, the communication paradigm in RAT is based on the effective sensor-actor and actor-actor coordination.

A. Our Approach

To enable a quick response time, as actors move within a geographical delta they report their location, *event type*, *interest duration* (γ), value of ϵ and action response capabilities using a long-range in-network *subscription message*. where ϵ is the time required by an actor to response to sensor input after its arrival. In case of multiple events, actors explicitly mention the type of event they are interested in and also the *interest duration* during which an actor is interested in event readings. For small value of γ, this refers to a pull-based dissemination to get sensors to report directly to the actors. The frequency of such pull messages is very low at large value of γ when the actors are moving infrequently. However, the frequency of the interest message increases as an actor moves because its coverage or targeting area also changes and there needs an updated interest *subscription message*. When sensor nodes detect some event, they push the information to the interested actor for a time period of γ. Consequently, the event readings are not pulled by the actors every time they occur but are pushed by the sensor nodes once they know about the actors capable of taking action. Hence, we propose an integrated Pull/Push coordination mechanism to efficiently operate actors in WSANs.

B. Sensor-Actor (S-A) Coordination

In WSANs, multiple destinations (actors) of sensor readings are possible as opposed to single destination (sink) in WSNs. In other words, all sensor nodes may not transmit data to the single actor selected before the transmission starts. Hence, the problems in S-A coordination are [1]; which sensors communicate with which actors and how to establish a communication path between sensors and actors. Additionally, the given response time must be considered to achieve the right action upon the inputs received from sensors. In the

next section, we propose a solution for S-A coordination by employing an integrated Pull/Push (IPP) approach using RED-GEO routing described in Section II.

1) Pull/Push in S-A Coordination: Applications in WSAN involve dissemination of observed event readings to the interested clients i.e. sink or actors and thus require an efficient information dissemination mechanism for resource-constrained sensor nodes. There are two ways to disseminate the information. The first way is to reactively send the queries in the field and pull the relevant information out of the field. The second way is to proactively push all the relevant information out of the network regardless of the current queries. Both the approaches have some benefits as well as drawbacks.

Sensor networks are proposed to play role in detecting and characterizing Chemical, Biological, Radiological, Nuclear, and Explosive (CBRNE) attacks and material to prevent disastrous conditions. Obviously, an immediate action is imperative to encounter the threats in such applications. In such circumstances, pure-pull policy is not applicable because the applications need to pay immediate attention to certain events, which is not possible unless the queries are sent frequently. However, frequent queries introduce significant overheads in terms of resource utilization and delay. On the other hand, pure-push policy is cost inefficient when the events are received from the regions where actions can not be executed due to the restricted movement of actors or lack of resources to execute action or, even more specifically, the events are out of the region of interest.

We deal with the above mentioned problems by using integrated Pull/Push Coordination. Hybrid Push/Pull coordination is studied in [3], [16], [17]. These studies are proposed for monitoring and control purposes in WSN that has a single destination sink. However, there are multiple mobile destinations actors in WSAN that requires an efficient coordination mechanism to overcome the mobility of multiple destination actors. Unlike these studies, we exploit the mobility of actors to keep the efficient pull/push coordination. We assume that all the nodes are location aware because we exploit the location information to decide sensor-actor coordination as well as actor movement for targeting uncovered sources. Actor nodes show their interest about a particular event by broadcasting the *subscribe* message request. The request is announced to help in deciding the suitable actor that can react when the event is reported. Sensor nodes receive the request from a subset of actor nodes. However, we assume that the event is reported to only a single actor for a time period of γ. Furthermore, we assume that the subscription is made on the basis of geographical locations of the actors and sensors accept the request of the nearest actor. This criterion can be extended to include the load factor and energy of the actor to decide the subscription in an overlapping region. There is an associated life γ of the actor subscription and it expires if no further *subscribe* message is received during that time or the subscribed actor has moved.

An actor broadcasts a *subscribe* message at two occasions; first, when it changes its location by some threshold (M_{th}) value or leaves the event area it is currently targeting. If the movement of actor is higher than M_{th}, it broadcasts the *subscribe* message. This is done to handle the actor mobility and keep the coordination effective. Second, the interest is refreshed after each subscription period. Hence, the control message cost is directly related to the mobility of actors and integrated pull mechanism. Short subscription life leads to more frequent queries and intuitively more overheads. It is choice of application to select these parameters appropriately. The mobility can be restrained by deploying sufficient number of actors to cover the whole region. The communication path between sensor and target actor is established by RED-GEO approach described in Section II.

2) Subscribe (Pull) Message Frequency (f): The frequency of *subscribe* message (f) is computed by $1/\gamma$ initially which is the minimum bound on f. However, the mobility of actors causes to dynamically alter this frequency. In such case, we define maximum bound on f in order to avoid the frequent queries sent as in pure pull-based approach. This is achieved by setting the value of M_{th} according to the velocity of actors. The interest message is also propagated in the network for incremental deployment of nodes in addition to the mobility. For such topological changes, we restrict f not to go beyond 1 message per unit time. As a result, the value of M_{th} is set according to the movement speed and the density of sensor nodes. It is calculated such that the coordination remains effective. By effectiveness we mean that whenever an event occurs, it should be propagated to some actor. If the movement speed is high then the large value of f (resulted from low threshold value relative to the speed) would result in wrong subscription of actors and loss of event reports. Clearly, it seems that the threshold value should be relative to the speed of actors somehow. That is, if the speed is high, M_{th} should be relatively large to reduce f. However, this does not apply to the slow movement of actors. At lower velocity of actors, the value of M_{th} will also be lower by relating threshold with velocity. Intuitively, there would be unnecessary *subscribe* messages that limit the benefits of integrated Pull/Push coordination. Hence, we define a relation between M_{th} and the movement of actors to obtain the suitable value of threshold.

Let s, n and M, respectively, be the mobility speed of an actor, total number of nodes and dimension of the field. Then the unit θ occupied per node can be computed by $\sqrt{M^2/n}$. By using the node occupancy value θ, we can find the value of M_{th} as:

$$M_{th} = \begin{cases} s & \text{if } s > \theta \\ \theta & \text{if } s \leq \theta \end{cases}$$

Hence, the value of f is computed dynamically to minimize the overheads experienced due to subscription requests.

$$f = \begin{cases} \frac{1}{\gamma} & \text{if } s \approx 0 \\ \frac{s}{M_{th}} & \text{if } s > 0 \end{cases}$$

C. Actor-Actor (A-A) Coordination

Actors are assumed to attend an emergency anywhere in the field. We move an actor towards emergency zone if the source node is not within the action range of any actor. This is a controlled targeting of the emergency zone i.e. actor moves towards the requesting source such that the source nodes is

within the boundary of the action range of actor. In S-A coordination, sensor nodes send their readings to the nearest actor. However, the actor receiving sensor data may not act on the event area due to small action range or low energy. It requires exploiting the coordination between actor nodes to track the target area. Each actor node maintains neighborhood list of nearby actors in order to trigger coordination. However, it does not require to broadcast any specific beacon for neighborhood rather use the *subscribe message* received from other actors. This message also contains the information about the number of requests (sensors) currently the sender node is serving. Thus every actor maintains a list of actors that records the location information as well as the load on that actor. We assume that there are sufficient number of actors to cover the whole sensor field. Sensor nodes send event readings with their location information. As an actor receives a request from a sensor node that lies outside the coverage area of the actor, it starts the event targeting procedure.

1) Event Targeting: We assume that all the actors have the same circular action range of radius R, which is limited and different from the transmission range. When an actor a_i, residing at the point $I(x_i, y_i)$, receives a request from the source s at position $J(x_j, y_j)$, it calculates the distance $D(I,J) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ from the source. if $D(I,J) \leq R$, a_i is able to respond to the request. However, if $D(I,J) > R$, there are two possible scenarios. First, actor a_i needs to move $D(I,J) - R$ units of distance towards the target location at some new point $\bar{I}(\bar{x_i}, \bar{y_i})$ so that the coverage is provided to s as shown in Fig. 3. The new location $\bar{I}$ of the actor can be computed as follows

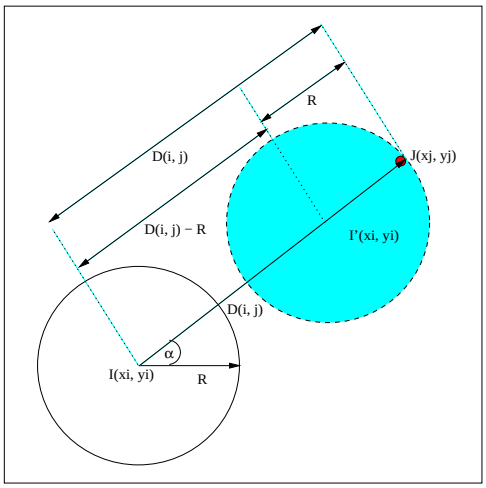


Fig. 3. Positioning actor towards uncovered source

Taking $I(x_i, y_i)$ as the center of axis on the XY plane, we can find the coordinates $\bar{x_i}$ and $\bar{y_i}$ as

$$\bar{x_i} = x_i + (D(I,J) - R) \times cos(\alpha)$$
$$\bar{y_i} = y_i + (D(I,J) - R) \times sin(\alpha)$$

where α is the angle between the vector $X(R, 0)$ along x-axis and the vector $A(x_j - x_i, y_j - y_i)$.

To find out the value of α, first compute the length $|X|$ and $|R|$ of two vectors X and R.

$$|X| = R,\ |A| = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Normalize each vector (unit vector)

$$X_u = (R, 0)/|X|,\ A_u = (x_j - x_i, y_j - y_i)/|A|$$

Hence, the angle between X and A is:

$$\alpha = cos^{-1}(X_u.A_u) \quad (1)$$

However, the displacement is only possible if the movement of a_i does not leave the presently attended sources uncovered. In the second scenario, if the displacement of a_i is not valid then it initiates the actor-actor coordination algorithm.

Algorithm 2 Action Coverage of a Target

1: Pseudo-code executed by the actor a to determine the coverage of source i.
2: $d(i,a) = \sqrt{(x_a - x_i)^2 + (y_a - y_i)^2}$
3: $C_a = coverSource((d(i,a) - R), (x_i, y_i))$ {New center position of a if move $d(i,a) - R$ towards (x_i, y_i)}.
4: $MoveOK = YES$
5: **for all** $s \in sources_a$ **do**
6: $MoveOK = canCover(C_a, R, (x_s, y_s))$
7: **if** $MoveOK = NO$ **then**
8: $Exit$
9: **end if**
10: **end for**
11: $moveTo(C_a)$

The coordination among actors is triggered by broadcasting the *relocate* message that contains the location of source s. Each actor receiving this message runs the Algorithm 2 to check for the possibility of attending the source s. An actor a_j sends back the *relocation-ok* message to a_i if it is able to move towards the source s while attending its own sources at the same time. This message contains the residual energy of actor a_j as well as the number of sources being attended. If the field is being operated by sufficient number of actors such that all the sources can get response from the actors then the actor a_i receives *relocation-ok* message from some of the neighboring actors. Consequently, the event occurred near the source s can be responded and a_i forwards the request of s to the actor which is covering lesser number of sources and has the highest residual energy.

When the actors have either limited action range or they are small in number to cover the whole field, it is quite possible that the action requests remain unsatisfied. This situation arises when the intensity of event is too high to affect the entire field. We can compute the number of unattended sources μ in $M \times M$ field as

$$\mu = \frac{M^2 - m \times \pi R^2}{M^2} \times n \quad (2)$$

where m is the number of mobile actors, n the total number of deployed sensors. However, the value of μ depends on the density of nodes in the field. In case of non-uniform deployment when the action range is limited, actors might

be stuck in sparse zone of the field that do not let the actors to attend the nodes in dense part as described above, which greatly effects the value of μ.

2) Example Scenario: Fig. 4 illustrates the target tracking procedure in two different scenarios. Sensor X subscribes the actor A because its the nearest one and sensor Y subscribes the actor B due to the shorter distance than A and C. Although node X reports its event to A but its out of the action range of A. Actor A runs the target tracking procedure and finds that node X can be covered with its existing sources. Therefore, actor A moves $(D(A \rightarrow X) - R)$ units towards the node X and is able to fulfill the future requests of node X without any further movement. This is depicted in case 1 of Fig. 4.

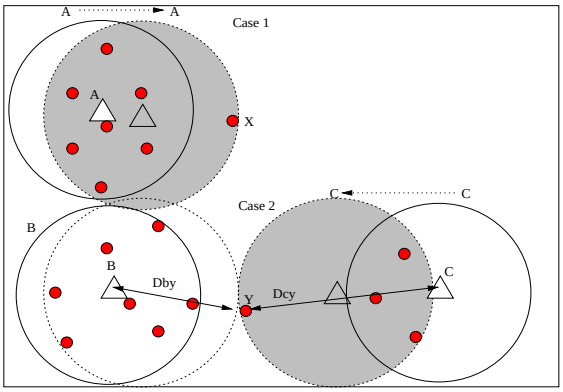


Fig. 4. A-A coordination to target the event area.

The case 2 requires the A-A coordination because the actor B is not able to serve node Y. If it moves to cover the event area of Y then it leaves the region of current sources uncovered. Therefore it coordinates with A and C and finds C as an appropriate actor to cover the event area of Y. Although $D_{by} < D_{cy}$, C is the right actor because the position of its existing sources permit it to attend the new source.

IV. Performance Evaluation

To evaluate the performance of RAT, the experiment model consists of 150 sensor nodes, having 20 meters of transmission radius, deployed randomly in $100 \times 100\ m^2$ area. Additionally, 5 actors are also placed at appropriate locations in the field to cover the equal regions (possibly non-overlapping to provide maximum coverage). Experiments are run with different event radius and action ranges. Table I illustrates the number of nodes detecting event at different event radii. In the example runs, sources generate 3 packets per second, each 100 bytes long, for 100 seconds.

TABLE I
Number of Nodes vs Event Radius

Event Radius	20	22.5	25	30	32.5	35	40	50
Nodes (≈)	19	24	30	43	50	56	76	118

Fig. 5 illustrates the average response time, in terms of delay per hop, in RAT as a function of event radius. It is obvious from the graph that the delay per hop increases as we increase the number of sources by expanding the event radius. By increasing the event radius, the number of sensors detecting the events also increase producing more traffic. Whereas, the large value of action range of actors covers wide area and results in attending requests from farther nodes increasing the average number of hops. Intuitively, the impact of large event radius and action range causes more traffic effecting the performance in terms of delivery ratio, average hops and packet delay. Obviously, the requests per actor are increased as well and the average number of hops is approximately 2.2 per request at $R = 50$ as shown in Fig. 6 as compared to 1.4 at $R = 30$. The increase in hops introduces more delay and poor response time. We can conclude from the results that the average hops are approximately R/r.

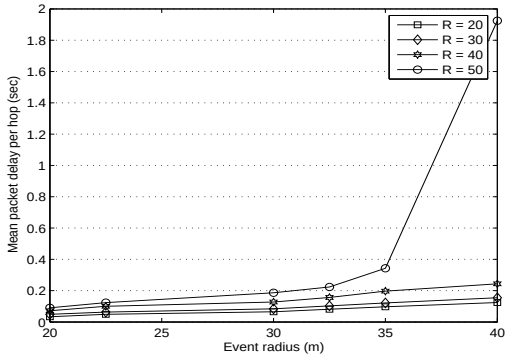


Fig. 5. Average response time in RAT.

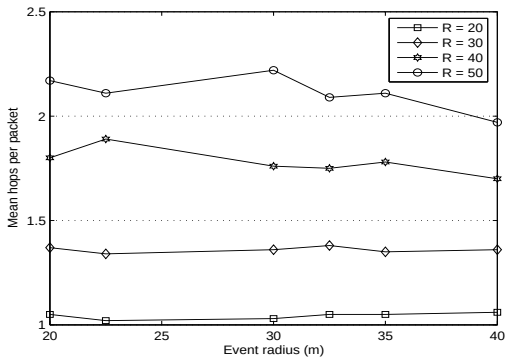


Fig. 6. Average number of hops for different action range of actors in RAT.

By employing movement model in A-A coordination, we can improve the response time and the packet delivery ratio as shown in Fig. 7. By keeping the action range ($R = 20$) too small results in lower delay. However, it reduces the delivery ratio. However, a large fraction of requests remain unsatisfied as shown in Fig. 8. Hence, for $R/r \approx 1.5$ (where R = 30 and r = 20), the response time is improved as well as the overall packet delivery ratio as reflected in the above figures.

The theoretical value of unattended requests computed by using Eq 2 for $R = 20$ with 5 actors is also plotted in Fig. 8. It can be seen that when event radius goes beyond 30, the actual unsatisfied request ratio found by simulation becomes worst. This is due to the random distribution of nodes that leaves some area sparse and some dense. The effect seems

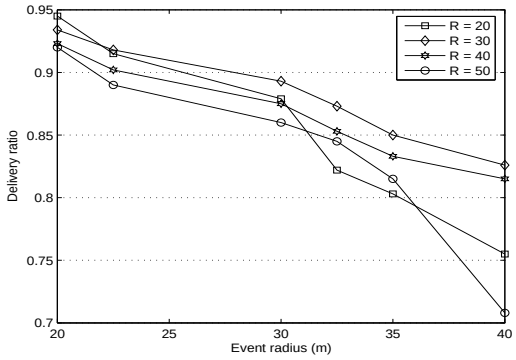


Fig. 7. Delivery ratio in RAT.

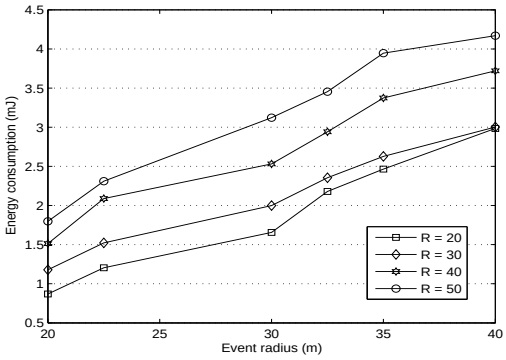


Fig. 9. Average energy consumption in RAT.

more significant as we increase the event radius with small action range. However, for large action range $R \geq 30$, the field is mostly covered by the actors, hence actors can handle more requests.

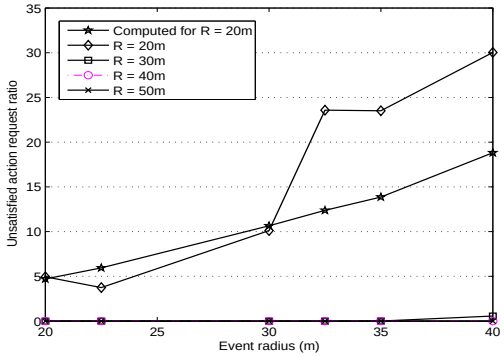
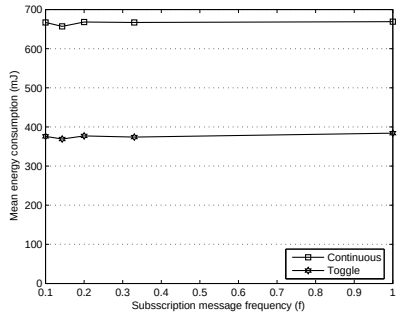


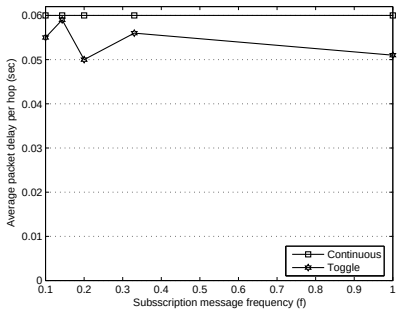
Fig. 8. Unattended requests in RAT.

Fig. 9 refers to overall energy consumption with different values of R. Although the energy consumption for $R = 20$ is very low but this value offers lower delivery rate that might not be acceptable for reliable measurements. The energy consumption is slightly higher at $R = 30$ but it provides high delivery rate with good response time. Hence, the performance results suggest that response time and delivery ratio can be improved with efficient energy consumption when the route length is approximately 1.4 that can be achieved by taking $R/r \approx 1.5$.

Due to long range transmission of actor nodes, all the nodes receive the *subscribe message* directly. As a result, the large value of f does not make any significant difference in energy consumption of sensor nodes as shown in Fig. 10(a). The energy consumption is almost same for all the values of f. However, the energy is saved by pushing the event observations only when there is an interested actor. The results plotted in Fig. 10(a) illustrates the energy consumption of two scenarios. In first scenario, nodes continuously push data to the subscribed actors. In second case, actors are interested in event readings only for the time period of γ and remain silent for an equal time period i.e., they toggle their interest. It is apparent that the energy can be saved in *toggle* mode for approximately 60%. Eventualy, pushing data only when it is required by the interested clients saves energy significantly. Fig. 10(b) shows the average packet delay in aforementioned scenarios. It is obvious that the delay is not reduced much in *toggle* scenario because the data rate remains same whenever data is reported. However, if γ is small i.e. f is large then event flows for shorter time that results in small delay.



(a)



(b)

Fig. 10. Energy and delay comparison at different subscription period.

V. Conclusion

We have presented a unique integrated push/pull coordination in order to solve the communication paradigm in pull/push based wireless sensor and actor networks. The information dissemination is triggered by a *subscribe* message from actors showing their interest and range of action as a pull-based strategy. Pure pull-based approach is not applicable, especially

when the frequency of this message is small as compared to the events reported by the sensor nodes. Once the subscription of an actor is performed by the sensor nodes, we adapt to push policy and continue to push the event data as long as the subscription is valid. The subscription is refreshed whenever it expires or actor moves for continuous operations of the network.

Moreover, it is not possible to predetermine the exact location of a limited number of actors to provide full coverage in the field. We exploit the mobility of actors in order to target the uncovered event area. Simulation results reveal that with the proper choice of action range of actors and A-A coordination, we achieve high delivery rate, better response time and better energy consumption. The results prove that IPP shows better performance under varying traffic load.

REFERENCES

[1] I. F. Akyildiz and I. H. Kasimoglu, "Wireless Sensor and Actor Networks: Research Challenges," *Ad Hoc Networks, Vol. 2, Issue 4, pp. 351-367, October 2004*.

[2] G. A. Shah, M. Bozyiğit, Ö. B. Akan and B. Baykal, "Real-time Coordination and Routing in Wireless Sensor and Actor Networks," *in proc. of 6th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN), May 2006*.

[3] S. Kapadia1, B. Krishnamachari, "Comparative Analysis of Push-Pull Query Strategies for Wireless Sensor Networks," *in proc. of International Conference on Distributed Computing in Sensor Systems (DCOSS), June 2006*.

[4] T. Melodia, D. Pompili, V. C. Gungor and I. F. Akyildiz, "A Distributed Coordination Framework for Wireless Sensor and Actor Networks," *ACM MobiHoc'05, May 2005*.

[5] Arjan Durresi, Vamsi Paruchuri and Leonard Barolli, "Delay-Energy Aware Routing Protocol for Sensor and Actor Networks," *Proc. of the 2005 11th International Conference on Parallel and Distributed Systems (ICPADS'05)*.

[6] K. Liu and N. Abu-Ghazaleh, "HGR: Hybrid Geographic and Virtual Coordinate Routing for WSN," *http://arxiv.org/abs/cs.NI/0511059 November 2005*

[7] A. Aksoy, M. Leung, "Pull vs. Push: A Quantitative Comparison," *IEEE GLOBECOM, Dec 2004*.

[8] J. Kulik, W. Heinzelman, and H. Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks," *Wireless Networks, Vol. 8, 2002, pp. 169-185*.

[9] K. Akkaya, M. Younis, "An Energy-Aware QoS Routing Protocol for Wireless Sensor Networks," *23rd ICDCSW, pp. 710-715, 2003*.

[10] D. Tian and N.D. Georganas, "Energy Efficient Routing with Guaranteed Delivery in Wireless Sensor Networks," *Mobile Computing and Communications Review, 2001*.

[11] Y. Yu, R. Govindan and D. Estrin, "Geographical and Energy Aware Routing (GEAR): a recursive dissemination protocol for wireless sensor networks," *UCLA/CSD-TR-01-0023, Tech. Rep., 2001*.

[12] B. Karp and H. T. Kung, "GPSR: Greedy Perimeter Stateless Routing for Wireless Networks," *MobiCom 2000*

[13] M. J. Handy, M. Haase and D. Timmermann, "Low Energy Adaptive Clustering Hierarchy with Deterministic Cluster-Head Selection," *IEEE International Conference on Mobile and Wireless, 2002*.

[14] T. He, J.A. Stankovic, C. Lu and T. Abdelzaher, "A Real-Time Routing Protocol for Sensor Networks," *In Proc. of IEEE International Conference on Distributed Computing Systems, May 2003*.

[15] C. Intanagonwiwat, R. Govindan and D. Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," *In Proc. of ACM MobiCom, March 2000*.

[16] X. Liu, Q. Huang and Y. Zhang, "Combs, Needles, Haystacks: Balancing Push and Pull for Discovery in LargeScale Sensor Networks," emphACM Sensys 2004, November 3-5, 2004

[17] N. Trigoni, Y. Yao, A. Demers and J. Gehrke, "Hybrid Push-Pull Querying Processing for Sensor Networks," emphIn Proc. of the Workshop on Sensor Networks (WSN) 2004

[18] UC Berkeley, LBL, USC/ISI and Xerox PARC, "The Network Simulator ns-2," *The VINT Project, http://www.isi.edu/nsnam/ns/*.

Exploiting Energy-aware Spatial Correlation in Wireless Sensor Networks

Ghalib A. Shah and Muslim Bozyiğit
Department of Computer Engineering
Middle East Technical University,
Ankara Turkey, 06531
{asad, bozyigit}@ceng.metu.edu.tr

***Abstract*— Wireless sensor networks (WSNs) promise fine-grain monitoring in a wide variety of applications, which require dense sensor nodes deployment. Due to high density of nodes, spatially redundant or correlated data is generated. Redundancy increases the reliability level of information delivery but increases the energy consumption of the nodes too. Since energy conservation is a key issue for WSNs, therefore, spatial correlation can be exploited to deactivate some of the nodes generating redundant information. In this paper, we present an energy-aware spatial correlation based on a clustering protocol. In this approach, only the cluster-heads are responsible of exploiting spatial correlation of their member nodes and selecting the appropriate member nodes to remain active. The correlation is based on the distortion tolerance and the residual energy of member nodes. Each cluster-head divides its clustered region into correlation regions and selects a representative node in each correlation region which is closer to the center of correlation region and has the higher residual energy. Hence, the whole field is represented by a subset of active nodes which perform the task well. Simulation results prove that the required reporting rate can be achieved with lesser number of nodes by exploiting spatial correlation and eventually conserving the nodes energy.**

***Index Terms*— Wireless Sensor Networks (WSNs), Clustering, Spatial Correlation, Energy-efficient**

I. INTRODUCTION

In wireless sensor networks (WSNs), small efficient sensors in terms of sensing, processing, storage and energy are placed or scattered at different locations for the purpose of sensing or more precisely information gathering. The basic characteristics of sensor networks include large scale random deployment, dense nature, unattended mode of operation, varying nature of radio links, and of course limited energy resource. However, energy is the basic and critical resource that separates sensor networks from all other forms of wireless networks. The main objective of the WSN is to reliably detect/extract event features from the collective information provided by sensor nodes [1]. Therefore, the energy and processing constraints of small wireless sensor nodes can be overcome by their collaboration.

Wireless sensor networks promise fine-grain monitoring in a wide variety of applications. The environments in some of these applications e.g. indoor environments, battlefield or habitats, can be harsh for wireless communication. Due to hostile environments and limited energy or transmission range, sensor nodes are highly prone to failures. Sensor nodes may fail or be blocked due to lack of power, physical damage or environmental interference. The failure of sensor nodes should not affect the overall task performed by the network. This is the reliability or fault tolerance issue [20].

The node densities in the deploymeny field may vary from application to application. For example, it may be as high as 20 nodes/m3 [19] or more. As a result of high node density, spatially redundant or correlated information is very likely in the network. Redundancy increases the reliability level of information delivery while increasing energy consumption of nodes as well. Since energy conservation is a key issue for WSNs, hence, spatial correlation can be exploited to deactivate some of the redundant nodes. Aggregation is another energy conserving approach. However, it provides estimated results that may not be appropriate for applications demanding intolerable distortion in event readings. Furthermore, all the nodes in network remain active unnecessarily participating in observing the phenomenon that may be otherwise avoided by exploiting spatial correlation.

Spatial correlation has been explored in the literature to some extent [5]- [12]. However, none of these studies focus on selecting appropriate representative nodes [1] in terms of their residual energy. In this paper, we present an energy-aware spatial correlation mechanism based on the cluster-based [2] configuration in WSNs. The contributions of this approach are to balance the residual energy of sensor nodes and dynamically adjust the number of representative nodes according to the distortion tolerance.

In this approach, only the cluster-heads are responsible of applying spatial correlation to their member nodes and selecting the appropriate member nodes to remain active for observing the phenomenon. Sink on behalf of application defines the correlation region [2] in which only one reading is sufficient for the event reading precision. However, the correlation factor can be changed dynamically in order to achieve the distortion tolerance. Each cluster-head divides its clustered region into correlation regions and selects a representative node in each correlation region which is closer to the center of correlation region and has the higher residual energy. The non-representative nodes remain passive until the energy of active nodes go down to some threshold value. Consequently, the whole field is efficiently represented by a

[1] A node which reports event on behalf of a group of nodes detecting similar readings.

[2] The region in which the sensor nodes report almost the similar readings is called correlation region.

subset of active nodes which perform the task well, equal to that of the original deployment in terms of sensing coverage.

The remainder of the paper is organized as follows. Section II provides an overview of the existing approaches exploiting spatial correlation. The cluster-based gridiron spatial correlation mechanism is presented in Section III. Detailed performance evaluation and simulation results are presented in Section IV. Finally, in Section V we draw the main conclusion.

II. Related Work

In this section, we investigate some of the existing approaches exploiting spatial correlation in WSNs. The information theoretical aspects of the correlation are explored in depth in [11],[12]. In other words, these studies aim to find the optimum rate in the sensor observation taking into account the compression of the redundant information. However, no correlation (spatial or temporal) between sensor observations is considered in these studies. The work in [7] exploits the spatial correlation to measure the link quality of wireless sensor nodes. The intuition behind spatial correlation is that sensor nodes geographically close to each other may have correlated link quality. It shows that the spatial correlation in link quality of neighbor sensor nodes can be captured to estimate the link quality with substantially less transmission cost than the link quality estimators based on temporal correlation. The history information of link quality for one node may be used for estimating not only its own link quality but also that of other neighbor sensor nodes geographically close. These approaches do not exploit spatial correlation in eliminating the redundant event reports.

In [5], the relation between spatial locations of the sensor nodes in the event area and the event detection reliability has been formulated. It estimates the minimum number of representative nodes in the field, each representing a group of spatially correlated sensor nodes. The correlation region is determined by applying Vector Quantization (VQ) statistical approach [21]. However in this approach, the energy has not been considered in the selection of a representative node criterion. It achieves the overall energy gain in the network regardless of the energy of individual node, which is not a load-balancing approach. Moreover, if a representative node, which is the center of correlation region, moves or its energy depletes, the selection of another representative node for that region effects the tolerable distortion. Frequent change in representative nodes produces more distortion and hence, the whole network is required to reconfigure to achieve the distortion limit. On the other hand, nodes working as representatives in their correlation regions for longer period results in unfair load distribution that reduces the network life. CC-MAC [3] uses the above formulation of spatial correlation to work with the medium access control (MAC).

Spatial correlation is also considered as a quality of service (QoS) parameter in some studies [8], [9], [10]. They use the idea of allowing the base station to communicate QoS information to each of the sensors using a broadcast channel and use the mathematical paradigm of the Gur Game to dynamically adjust to the optimum number of sensors. The sink node dynamically adjusts the number of sensors being activated to achieve the required reliability, thereby controlling the resolution of QoS it receives from the sensors. However, this approach controls just the number of active nodes, ignoring their locations. In short, all of the above approaches do not consider the energy of nodes in selecting the representative nodes and therefore do not efficiently exploit the spatial correlation.

One may also think the sensors coverage algorithm [13], [14], [15], [16] as one of the energy saving approach which eliminates the redundancy. The coverage algorithms ensure that the original sensing coverage is maintained after turning off redundant nodes. The primary goal of coverage algorithm is to cover the field with the given sensing range of the sensor nodes regardless of the characteristics of the event features and environment conditions. In noisy environments, it is highly likely that the event signal is faded, until certain distance, though it is observed within the sensing range. Intuitively, applications may require the observations at closer points and a number of such points may lie within the coverage of single sensor. Therefore, the coverage approach which is inflexible of such requirements may not be applied to achieve the desired accuracy in observation. Hence, an adaptive spatial correlation of sensor nodes can be exploited to overcome such problem.

III. Cluster-based Spatial Correlation of Sensor Nodes

In this section, we present a grid-based spatial correlation of sensor nodes that is deterministic and more independent of network dynamics. Generally, the nodes coverage control mechanism is required when the number of sensor nodes deployed in the field are more than the sufficient number of nodes to provide the coverage in the whole field. Intuitively, the high density of nodes produce more traffic and results in congestion. We control the spatial resolution of nodes by deactivating the redundant nodes. By redundant nodes, we mean the nodes which are so close to each other that identical event information is reported. However, we divert the drawback of redundancy into the network favor. An active node is deactivated when its energy goes down below the energy of its nearby inactive node by a certain threshold value and, hence, inactive node is activated.

A. Correlation Model

We assume that the sensor nodes report almost similar values when they are close to each other. However, this closeness (θ) depends on the application requirements and event characteristics. Some applications are more critical and are less tolerant to discrepencies in the event readings requiring closer nodes to report event readings. While the others can be more toleratant requiring the farther nodes to report events. *We define the region as correlation region in which the readings reported by the sensor nodes are considered similar by the applications and therefore a single report is sufficient to represent that region.* It varies from application to application and event to event. For intolerable applications, the region is very small and large for tolerable. Hence the region is directly

related to the application defined distortion tolerance. Similarly for certain events, the event features vary significantly as the event signal propagates in the field. Such an event requires to be reported by relatively nearby nodes as opposed to the events whose detected features do not change at short distance.

The distortion in readings is observed when the sensor nodes fail to report the event from within the defined correlation region. The correlation region can then be changed dynamically according to the observed reliability. For example, an application may require to get readings at locations 10 meters apart. In that case the dimensions of correlation region is considered $10 \times 10\ m^2$. If a sink node receives readings from the nodes which are more than 10 meters far from each other then the distortion is observed and readings are considered unreliable relative to their closeness. *Hence, the distortion is observed if the reporting nodes are apart more than θ and therefore it is measured in terms of closeness rather than the event features. It is upto the application to evaluate the event features and set the value of θ accordingly*. Hence, θ controls the number of active nodes (K).

In order to investigate the distortion achieved when smaller number of nodes sending information, we assume that only K out of n packets are received by the sink, where n is the total number of sensor nodes in the event area. The distortion function $D(K)$ to find out the relation between the spatial locations of the sensor nodes in the event area and the event estimation precision has been formulated in [3] as

$$\begin{aligned} D(K) &= \sigma_S^2 - \frac{\sigma_S^4}{K(\sigma_S^2+\sigma_N^2)}(2\sum_{i=1} K\rho(s,i) - 1) \\ &+ \frac{\sigma_S^6}{K^2(\sigma_S^2+\sigma_N^2)^2}\sum_{i=1}^{K}\sum_{j\neq i}^{K}\rho(i,j) \end{aligned}$$

where σ_S^2 and σ_N^2 are the variances of the i^{th} event reading S_i and the observation noise N_i, respectively. $D(K)$ shows the event distortion achieved at the sink as a function of the number of sensor nodes K that send information to the sink and correlation coefficients $\rho(i,j)$ and $\rho(s,i)$ between nodes n_i and n_j, and the event S from the sensor field and the sensor node n_i, respectively. The correlation coefficient between nodes n_i and n_j can be computed by $\frac{E[S_iS_j]}{\sigma_S^2}$.

B. Gridiron Spatial Correlation (GSC)

The gridiron spatial correlation mechanism is adaptive to achieve the required reliability by dynamically changing the correlation region. The correlation regions are formed as squared rectangles and nodes lying in the rectangle are assumed to be spatially correlated. However, the correlation is fine tuned by resizing the rectangle according to the feedback provided by the sink computed on the basis of redundancy and reliability. The rectangles are considered independent of each other and, hence, switching the representation among nodes in a rectangle does not require the reconfiguration of the entire network. The switching is performed to balance the energy usage of correlated nodes in a heterogeneous network. Hence, the network is highly adaptive to the events of different intensity and heterogeneity.

Cluster-head identifies the redundant and close sources in its vicinity and turns off the activity of nodes by considering their energy level and closeness as criterion, where the closeness θ is either defined by the application or evaluated according to the density of nodes in the zone.

$$\theta = \sqrt{R^2/n} \tag{1}$$

where, R is the dimension of the squared field and n represents the total number of nodes deployed in $R \times R$ field.

Furthermore, the number of active sources can also be adjusted according to the required reliability level announced by some central node. However, this is applicable in continuous data thirst applications that would like to monitor the environment continuously. By controlling the spatial resolution, we implicitly avoid the congestion in network to happen.

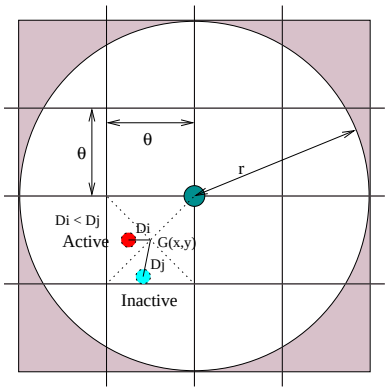


Fig. 1. Spatial resolution control mechanism employed by the cluster-head.

Fig 1 illustrates the mechanism to control the spatial resolution. Cluster zone is decomposed into grid of $2r/\theta \times 2r/\theta$ rectangles, where the cluster-head lies at the center of grid. The aim of our spatial resolution control mechanism is to select a single node in each rectangle of $\theta \times \theta$ dimensions. Let $G(x_i, y_i)$ be the center of an i^{th} rectangle. We measure the distance of each node lying in the i^{th} rectangle from the point $G(x_i, y_i)$. The closest node is selected as an active member of the cluster and other nodes in that rectangle are deactivated by the cluster-head. Similarly, the inactive nodes do not participate in relaying packets i.e. their transceivers are also turned off during the current cycle. Fig 2 shows the active members in a cluster selected by applying our resolution control procedure. It is important to note that the multi-hop members do not lie in the direct coverage of cluster-head and, therefore, control mechanism is not applied to multi-hop members.

Cluster formation procedure in [2] is followed by exploiting gridiron spatial correlation. Cluster-head applies the gridiron procedure given in Algorithm 1 over its member nodes and decides the active nodes. It then announces the list of active nodes in a single *AList* message. The size of *AList* message is dynamic relative to the number of active nodes. Cluster-heads run this procedure periodically and whenever there is a change in active nodes list, they broadcast the updated list.

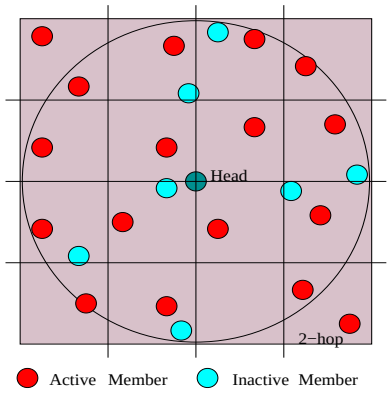


Fig. 2. Spatial resolution in cluster zone.

Algorithm 1 Gridiron Spatial Correlation

Cluster-head runs the procedure to decide the active nodes.
$nRects = r/\theta$
for $x = -nRects; x++; x < nRects$ **do**
  for $y = -nRects; y++; y < nRects$ **do**
    $rect_x = head_x + x \times \theta$
    $rect_y = head_y + y \times \theta$
    $C_x = rect_x + \theta/2$
    $C_y = rect_y + \theta/2$
    $near_{dist} = \infty$
    $active_{member} = null$
    for all $m \in members$ **do**
      $d = \sqrt{((C_x - m.nbr_x)^2 + (C_y - m.nbr_y)^2)}$;
      if $d < near_{dist}$&&$m.energy + threshold > MaxMembersEnergy()$ **then**
        $near_{dist} = d$
        $active_{member} = m$;
      end if
    end for
    if $active_{member} \neq null$ **then**
      $active_{member}.is\_active = true$
    end if
  end for
end for

This change might be either due to changed θ or the energy of active nodes drop to threshold.

C. Scheduling Nodes

GSC does not apply duty cycle to alternate the activity of nodes. Cluster-heads rather use threshold to decide about the activeness of nodes. However, it follows the duty cycle so that all the member nodes (active or inactive) hear the transmission of their cluster-heads at some time. For instance, the cluster-head announces the initial list (*AList message*) of active nodes at time $t = t_1$. All the member nodes obey the decision of cluster-head for time period of Δ seconds. During this time i.e. $t_1 \rightarrow t_1 + \Delta$, inactive remains inactive and do not overhear any transmission. However, as this time period passes, all the nodes become active for a shorter time period of ϵ seconds. Cluster-heads, during this time period, broadcast an updated list of active nodes. Henceforth, the member nodes obey the new decision and unnamed node in this list go into sleep mode for Δ seconds. Fig. 3 illustrates the state transition of the member nodes.

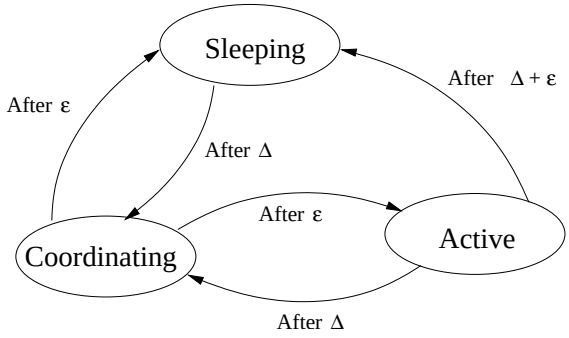


Fig. 3. State transition of member nodes in exploiting GSC.

The list of active nodes is modified if the energy of current active nodes goes down to some threshold value as compared their correlated nodes. Hence, GSC maintains equilibrium of the energy of correlated nodes in order to extend the life of network. Cluster-head also keep information about their neighboring cluster-heads as described in [2] in order to establish route with the sink node. In order to exploit GSC, it makes sure that deactivating the member nodes does not break its connection to the cluster-head through which it forwards the packet. If it suffers in such a problem then it does not deactivate the node working as a gateway and let it to remain active. Hence, a steiner tree is maintained in order to ensure the minimum connectivity of the nodes to the sink.

IV. Performance Evaluation

The performance of the spatial correlation mechanism is evaluated by using the network simulator *ns-2* [22]. The example scenario consists of various number of nodes randomly deployed in the field of 100×100 m^2 . An event source is also included in the scenario to trigger events by using the NRL phenomenon node extensions [23] for *ns-2*. The experiments are run at different density of nodes ($\mu = n\pi r^2/M$), where the density is changed by varying the number of nodes (n) while keeping the area of the field (M) and transmission radius of nodes (r) constant. The performance metric consists of the event readings distortion, number of active nodes, average packet delay and, of course, energy consumption.

By exploiting spatial correlation, we control the number of active nodes. For larger value of θ, lesser number of nodes are active and eventually lower reporting rate is observed. Fig. 4 shows the reporting rate of active nodes at different number of deployed nodes. Applications can define the correlation region by setting the value of θ according to their required reporting rate for different number of deployed nodes. Although the nodes are densely deployed but GSC keeps the number of nodes as active as required and eventually the reporting report is almost same at higher densities ($\mu = 22.5$ and $\mu = 30$). However, this is applicable only when the number of deployed nodes are higher than the sufficient number of nodes to cover the field. At $\theta = 10$ $meters$, 40% nodes remain passive for $n = 200$, which is 20% for $n = 100$ and just 5% for $n =$

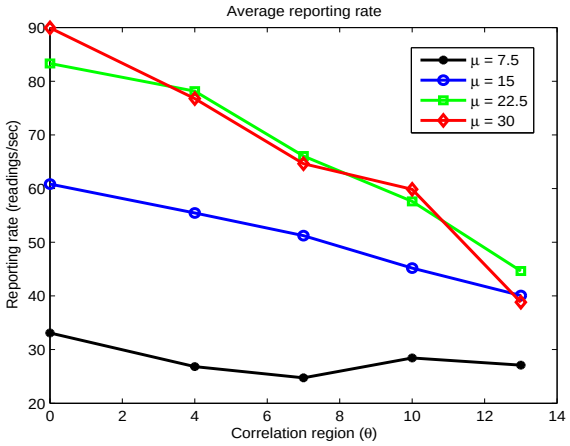


Fig. 4. Event reporting rate of active nodes.

50. Ideally, there should not be any inactive nodes at lower density. But it is due to the random deployment in which some nodes may be deployed very close to each other and are made inactive as given for $n = 50$. Hence, the results in Fig. 4 reveal that the number of nodes are activated relative to the ratio of their deployment or density. Similarly, the number of active nodes are directly related to the value of θ at higher density ($n = 200$), which shows a linear relation.

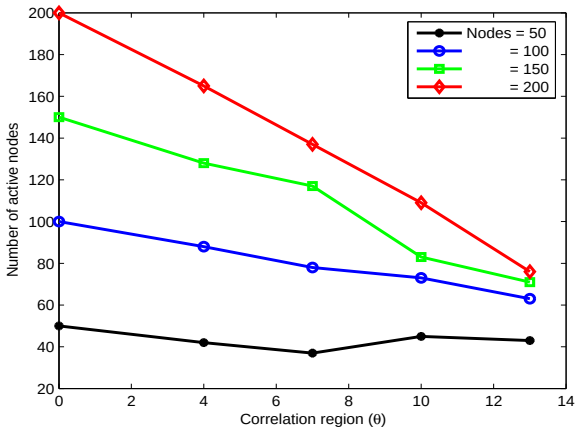


Fig. 5. Active nodes for different values of correlation region θ.

Fig. 6 shows the distortion at different correlation regions. It is clear that the distortion is lesser at higher density and larger θ. The distortion in readings occurs not only due to low density but also due to the non-uniform deployment. Generally the distortion at lower density should be reduced by the factor of additional deployment of nodes. However this reduction is until certain limit and do not improve further because the nodes are randomly deployed and are placed non-uniformly in some part of the field. At lower density ($\mu = 7.5$), the reduction is linear and is reduced 100% by expanding the correlation region while an increase in θ at higher density ($\mu = 30$) reduce distortion upto 60%.

When the density of nodes is high, the wireless channel is contended by large number of nodes that result in large

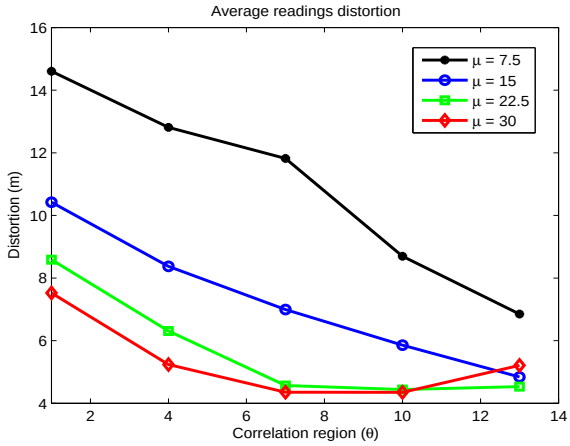


Fig. 6. Distortion in sensor readings at different values of correlation region θ

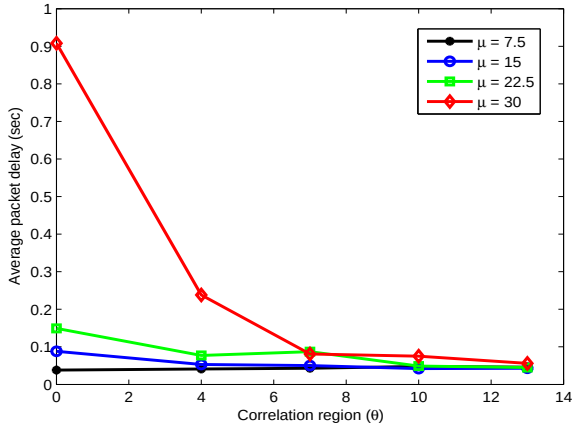


Fig. 7. Average packet delay observed at different values of correlation region.

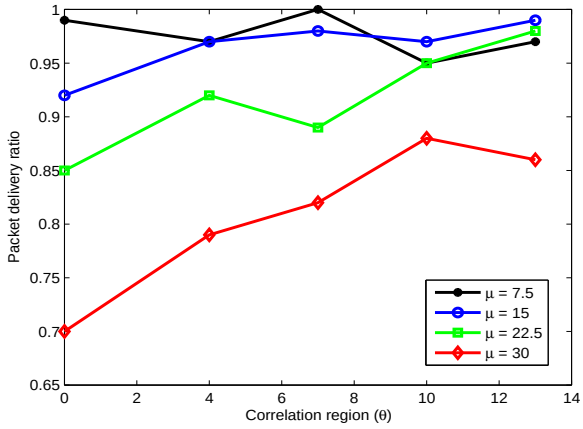


Fig. 8. Packet delivery ratio of active nodes at different values of correlation region.

communication delay. Fig. 7 shows that the average packet delay at $\mu = 30$ and $\theta = 0$ is 9 times larger than at lower densities and $\theta > 0$. By exploiting spatial correlation at

higher density ($\mu = 30$), we reduce the delay significantly by increasing the value of θ and hence it approaches to the same value as for lower densities i.e. it is reduced by 9 times. Similarly the delivery ratio is improved by increasing the value of θ as shown in Fig. 8.

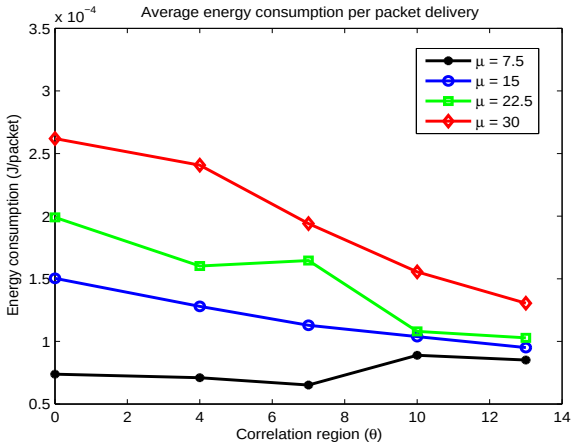


Fig. 9. Energy consumed per packet delivery at different values of correlation region.

Likewise, the energy consumed per packet delivery is also conserved at higher density by exploiting spatial correlation which is not true for lower density. It is obvious from Fig. 9 that for $\mu = 30$ the energy is saved approximately two times at $\theta = 13$. However, this improvement can not be experienced at lower density and eventually there is no energy conservation at $\mu = 7.5$.

V. Conclusion

Spatial correlation has been exploited in the literature to conserve the energy of sensor nodes. The existing approaches do not consider the residual energy of nodes in selecting the representative nodes and therefore do not efficiently exploit the spatial correlation. In this paper, we present an energy-aware spatial correlation based on the clustering protocol. Cluster-heads apply spatial correlation in their regions independently to keep the member nodes active subject to the information reliability. Each cluster-head divides its clustered region into correlation regions and selects a representative nodes per correlation region which is closer to the center of correlation region and has the higher residual energy. However the correlation regions can be resized according to distortion tolerance which is measured and announced by the sink node. Hence, the whole field is efficiently represented by a subset of active nodes which perform the task well equal to that of all the deployed nodes. Simulation results prove that the required reporting rate can be achieved with lesser number of nodes by exploiting spatial correlation and eventually conserves the nodes energy.

References

[1] I. F. Akyildiz, M. C. Vuran, O. B. Akan, and W. Su, "Wireless Sensor Networks: A Survey REVISITED," *to appear in Computer Networks Journal (Elsevier), 2006.*

[2] G. A. Shah, O. B. Akan and M. Bozyigit, "Multi-Event Adaptive Clustering (MEAC) Protocol for Heterogeneous Wireless Sensor Networks" *in Proc. Fifth Annual Mediterranean Ad Hoc Networking Workshop (MedHoc-Net 2006), June 2006.*

[3] M. C. Vuran and I. F. Akyildiz, "Spatial Correlation-based Collaborative Medium Access Control in Wireless Sensor Networks," *IEEE/ACM Transactions on Networking, vol. 14, no. 2, pp. 316 -329, April 2006.*

[4] E. Felemban, C.G. Lee, E. Ekici, R. Boder, and S. Vural, "Probabilistic QoS Guarantee in Reliability and Timeliness Domains in Wireless Sensor Networks," *in Proc. of IEEE INFOCOM 2005, Miami, FL, USA, March 2005.*

[5] I. F. Akyildiz, M. C. Vuran and O. B. Akan, "On Exploiting Spatial and Temporal Correlation in Wireless Sensor Networks," *in Proc. IEEE WiOpt 2004, University of Cambridge, UK, March 2004.*

[6] SunHee Yoon and Cyrus Shahabi, ""Exploiting Spatial Correlation Towards an Energy Efficient Clustered AGgregation Technique (CAG)," *IEEE International Conference on Communications (ICC), pp. 16-20 May 2005.*

[7] Yingqi Xu and Wang-Chien Lee, "Exploring Spatial Correlation for Link Quality Estimation in Wireless Sensor Networks," *in Proc. Fourth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2006) March 2006.*

[8] J. Kay and J. Frolik, "Quality of Service Analysis and Control for Wireless Sensor Networks," *1st IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS), October 2004.*

[9] I., Ranjit and L. Kleinrock, "QoS Control for Sensor Networks," *IEEE International Communications Conference (ICC 2003), Anchorage, AK, May 2003.*

[10] Frolik, J. "QoS Control for wireless sensor networks," *Wireless Communication and Networking Conference (WCNC), Atlanta, GA, March 2003.*

[11] M. Gastpar and M. Vetterli, "Source-Channel Communication in Sensor Networks," *in Proc. of 2nd International Workshop on Information Processing in Sensor Networks (IPSN'03), vol. 219, pp. 162-177, 2003.*

[12] S. S. Pradhan, J. Kusuma and K. Ramchandran, "Distributed Compression in a Dense Microsensor Network," *IEEE Signal Processing Magazine, vol. 19, no. 2, pp. 51-60, Mar 2002.*

[13] Y. Xu, J. Heidemann and D. Estrin, "Geography-informed energy conservation for Ad Hoc routing," *In Proc. of the 7th annual international conference on Mobile computing and networking (MobiCom), pp. 70-84, 2001.*

[14] D. Tian and N. D. Georganas, "A coverage-preserving node scheduling scheme for large wireless sensor networks," *In Proc. of the 1st ACM international workshop on Wireless sensor networks and applications, pp. 32-41, 2002.*

[15] X. Wang, G. Xing, Y. Zhang, C. Lu, R. Pless and C. Gill , "Integrated coverage and connectivity configuration in wireless sensor networks," *In Proc. of the 1st international conference on Embedded networked sensor systems, pp. 28-39, 2003.*

[16] M. Lu, J. Wu, M. Cardei, and M. Li, "Energy-Efficient Connected Coverage of Discrete Targets in Wireless Sensor Networks," *In Proc. of the international conference on networking and mobile computing 9ICCNMC), pp. 43-52, 2005.*

[17] G. A. Shah, M. Bozyigit, Ö. B. Akan and B. Baykal, "Real-time Coordination and Routing in Wireless Sensor and Actor Networks," *in Proc. 6th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN), May 2006.*

[18] C. Intanagonwiwat, R. Govindan, D. Estrin, "Directed diffusion: a scalable and robust communication paradigm for sensor networks," *In Proc. of the ACM MobiCom'00, Boston, MA, 2000, pp. 56-67.*

[19] E. Shih, S. Cho, N. Ickes, R. Min, A. Sinha, A. Wang, A. Chandrakasan, "Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks," *In Proc. of ACM MobiCom'01, Rome, Italy, July 2001, pp. 272-286.*

[20] G. Hoblos, M. Staroswiecki, and A. Aitouche, "Optimal Design of Fault Tolerant Sensor Networks," *IEEE International Conference on Control Applications, pp. 467-472, Anchorage, AK, September 2000.*

[21] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. Comm., vol. COM-28, no.1, pp. 84-95, Jan. 1980.*

[22] UC Berkeley, LBL, USC/ISI and Xerox PARC, "The Network Simulator ns-2," *The VINT Project, http://www.isi.edu/nsnam/ns/.*

[23] NRL Sensor Network Extension for ns-2, *http://downloads.pf.itd.nrl.navy.mil/nrlsensorsim/.*

A Mashup-Based Strategy for Migration to Service-Oriented Computing

Semih Cetin[1,2], N. Ilker Altintas[1,2], Halit Oguztuzun[2], Ali H. Dogru[2],
Ozgur Tufekci[1] and Selma Suloglu[2]
[1] *Cybersoft Information Technologies*
[2] *Department of Computer Engineering, Middle East Technical University*
{semih.cetin, ilker.altintas, ozgur.tufekci}@cs.com.tr
{oguztuzn, dogru, selma.suloglu}@ceng.metu.edu.tr

Abstract

Service-Oriented Computing holds great promises to realize the vision of on-demand services available anytime and anywhere. It is still not clear, however, when and how the existing systems will benefit from this new wave. The problem is particularly acute for the software embedded in myriad devices. This work charts a roadmap for migration of legacy software to pervasive service-oriented computing. A key idea is to achieve integration even at the presentation layer, not only at backend layers like application or data. This requires re-inventing the popular mashup technology for the enterprise level. Domain-Specific-Kits, which has been originally introduced within the context of software factory automation, has been reshaped as another enabling technology towards migrating to the service harmonization platform.

1. Introduction

Today's computing necessitates emerging services available as quickly as possible but this is not so easy to achieve since enormous amount of functionality is somehow buried in billions of dollars worth of existing code. Service-Oriented Computing (SOC) may help tackle this problem by decomposing monolithic legacy systems into loosely coupled parts wrapped by service adapters. This is quite challenging since such legacy systems have mostly single tier architectural designs, which complicates service partitioning and assuring the Quality of Service (QoS) attributes such as durability, maintainability, security, and efficiency.

Another difficulty stems from the fact that service pervasiveness essentially requires the content-oriented harmonization rather than the data or behavior-oriented one, which is common today. Known as the "mashup" technology, this new trend adds more value to existing services but such services are expected to be REST services rather than the classical Web services [17]. This makes even today's state of the art Web services legacy as well.

All these challenges reveal that a generic migration strategy to SOC may not scale well for miscellaneous business requirements. Alternatively, domain specific approaches will tackle the low-level details more effectively but with a certain surcharge. Therefore, a compromise is needed for the effectiveness of domain specificity at the cost of generality. This paper puts forth such a strategy for migrating legacy solutions to SOC by mashing them up in a harmonization platform: "MigrAtion to Service Harmonization compUting Platform (MASHUP)". MASHUP provides a six-step roadmap to mash existing binaries in a generic Mashup Server where domain specific wrappers are plugged in to wrap existing logic either as B2B or B2C services.

2. Legacy software, Web 2.0 and Mashup

Services are actually engineered with the purpose of composing them with different applications. They are "components" whose coupling with many different systems is inherently facilitated. Legacy software, on the other hand, constitutes frequently monolithic, big sized applications. The philosophy had been to write code for the given system specifications, but now it turned out to "compose" existing services – even "on the fly", in the age of ubiquitous / pervasive services.

The Web 2.0 designation was originally coined by O'Reilly's Dale Dougherty to describe the Web experiences that fundamentally engage users by: (a) allowing them to participate in sharing information and

enriching data freely; (b) readily offering their core functionality as open services to be composed or "mashed up" into new services and sites; (c) placing the Web at the center of software experience both in terms of data location and where the software is [5].

Web 2.0 takes us to a deeply service-oriented world where we can exploit everyday services such as news, instant messaging, and blogging. These services seem re-writing of the existing services but they are smart compositions (mashups) of the existing services with new ones. Also known as Web application hybrid, this smart way of combining the content from more than one source into an integrated experience is called "mashup" technology. Further convergence of Web 2.0 and SOA creates a new trend of applications so-called Enterprise Web 2.0, which embraces the union of Web 2.0 technologies such as Ajax with legacy systems, and Web services to enable deploying robust, reliable, and secure business applications over the Web.

3. Existing migration approaches to SOA

Migrating a legacy system to SOA by wrapping with Web Services may be relatively straightforward. However, certain characteristics of legacy systems like platform, language, architecture, and the target SOA may complicate the task. Table 1 classifies the existing SOA migration strategies.

Table 1. SOA migration strategies

	SMART [6]	SWA * [3]	EMDMA ** [7]	IBM [8]	ORACLE [9]	MICROSOFT [10]	SAP [11]
Provides a Roadmap	√	√	?	√	√	√	√
Platform/Vendor/Tool Independent	√	√	√				√
Scales to Different Domains		√	√	√	√	?	
Process/Service Integration	√	√	√	√	√	√	√
Content Integration		?	?	√	√	√	√
Web 2.0 Enablement				?	?	√	?
Mashup Enablement				?	?		
Quality of Services				√	?	?	?

√ Fully match
? Partly match
* Salvaging & Wrapping Approach
** Enterprise Model Driven Migration Approach

Existing migration approaches propose merely the back-end integration with reverse engineering and architectural reconstruction of source codes [1, 2, 3, 4, 7]. Such context sensitive attempts, however, are not practical enough for enterprise legacy applications and they cannot be generalized easily.

4. The MASHUP migration strategy

The proposed MASHUP migration strategy has six steps to address both the behavioral and architectural aspects of the migration as shown in Fig. 1:

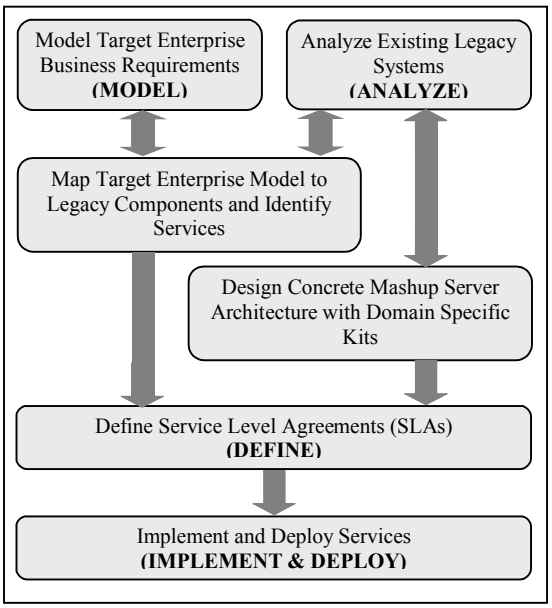


Figure 1. Migration activities

Step 1 (MODEL) is modeling the target enterprise business requirements. The business requirements are modeled for understanding the functional requirements of the target system. The business domain model might be expressed using a semantic notation similar to Business Process Modeling Notation (BPMN) [12], which is a methodology independent and unambiguous notation to express any business process.

Step 2 (ANALYZE) is analyzing the existing legacy systems. The analysis of existing systems reveals the important data for designing domain specific kits and reference architecture and for investigating reuse potential of existing legacy components. The results of this step include information on system architecture, system components, infrastructure details, interfaces, presentation characteristics, QoS attributes, level of maintainability, level of complexity and coupling, etc.

Step 3 (MAP & IDENTIFY) is mapping business requirements to system components and identifying services. System components might be reused legacy components or freshly developed ones. Thus, this step uses the results of business modeling and legacy analysis activities.

The iterative mapping process is as follows: (a) if a business requirement can be satisfied by one of the existing legacy components, then simply wrap it by considering the QoS attributes; (b) if there is a gap with the existing legacy component and requirement, and the gap can be filled during service wrapping, then accustom the legacy component into a new service; (c) if the gap cannot be fulfilled, then develop a new service for the requirement. The new service need not be implemented in any specific language as Mashup Server reference architecture can utilize any component (implemented in any language) using appropriate Domain Specific Kits (DSKs).

Continue this step as the services can be broken down into further sub-services and those can be mapped using the above mappings (either into legacy components or new ones). A successful mapping might use Commercial-Off-The-Shelf (COTS) components as service sources since those COTS components can be plugged into reference architecture with proper DSKs that can utilize component orientation techniques [13]. As part of this activity, investigate any implicit dependencies. The implicit dependencies can occur at data level, platform level, service quality level, etc. For instance, if there are two services from two different systems but they need a common data set, then mapping process should introduce suitable data feeder services. At the end of this step, the business requirement to service (BR2S) and service-to-service (S2S) breakdown has been determined at the conceptual level.

Step 4 (DESIGN) is for designing a concrete Mashup Server Architecture with DSKs. The design of Mashup Server Architecture (MSA) has been adapted from the architecture modeling approach already employed for software factory automation [15]. Architecture modeling relates architectural aspects and quality targets to running DSKs and choreography rules. The "symmetric alignment" technique [14] is assisted by a methodical approach to identify components (Domain Specific Engines – DSEs), and connectors (composition of DSEs) in the solution domain. Identification of architectural properties facilitates the definition of contextual information, which contains the stateful/stateless information to connect individual DSEs, and is needed for independent design and implementation of these individual DSEs through a standard communication schema across Domain Specific Languages (DSLs). Defining DSLs under the governance of a service meta-model enables the exposition of service from different sources with varying attributes. This step results in MSA with a set of specific pluggable DSEs, contextual information, internal requirements of DSEs and a runtime model.

Step 5 (DEFINE) is defining the Service Level Agreements (SLAs). At the end of Step 3, the service mappings are finalized; and the MSA (from Step 4) is ready to execute those services. SLA is defined in a common service repository and contains information on: a) contextual information; b) QoS characteristics of the service. QoS characteristics are basically inherited from the provider system and managed by the DSE serving the service, but the service implementation may introduce additional quality concerns. The Mashup Server manages both the contextual information and QoS definitions.

Step 6 (IMPLEMENT & DEPLOY) is implementing and deploying the services. This includes wrapping of existing legacy components, customization of existing legacy components with some additions, development of new services, and introduction of COTS components as services. The development of new services does not require that those new services need to be implemented using a common language on a common platform since they are abstracted by appropriate DSEs at the Mashup Server level.

The reference architecture model for the Mashup Server, depicted in Fig. 2, highlights the generic reference architecture that enables plugging specific DSEs for different systems (service sources). Mashup reference architecture depends on a meta-model for specifying DSLs for different service sources, a common service repository, and a policy for managing contextual information. The reference architecture enables the integration of another Mashup Server using a particular DSE (see DSE_5 in Fig. 2).

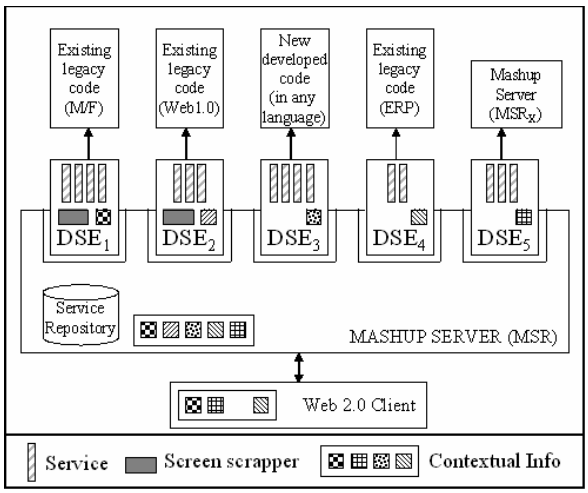


Figure 2. MASHUP reference architecture

The reference architecture employs a choreography engine that manages the DSEs. Dynamic plugging and context-awareness of DSEs are crucial for the runtime execution model. The choreography engine enables

this communication and coordination among DSEs. It ensures state coordination, communication, messaging, nested processes, context management, and service-oriented exception handling. The corresponding DSLs are required to comply with a certain meta-model to be plugged into the Choreography Engine. DSLs are declarative, context-free and loosely coupled to each other; hence we can readily apply the golden principle of separation of concerns for different domains.

The DSEs can perform specific tasks for different systems. Consider, for instance, a Web 1.0 system without an API. It heavily mixes presentation with content and makes it hard to sift out meaningful data from the rest of the elements used for formatting, spacing, decoration or site navigation. In such a situation the DSE can employ "Screen Scrapping/Web Reading" techniques by analyzing the page structure and wrapping out the relevant records. In some cases the task is even more complex than that: The data can be scattered over more pages. Then, triggering of a GET/POST request may be needed to get the input page for the extraction or authorization that might be required to navigate to the page of interest. The situation may be even more complex if there are workflows running in the provider system and if tight security policies are applied.

In case the legacy code is provided by a mainframe, screen-scrapping techniques are essential to simulate the working of the mainframe terminals. Using screen scrapping, the data would be retrieved from the host and also posted onto the host. The mashed up services are accessed through Web 2.0 clients supported with AJAX, Flash, Flex, JavaScript, and other XML-based rendering [16] technologies. Even smarter clients can be used to access several Mashup Servers from a single client with core Mashup Choreography abilities.

4. Conclusions

This paper introduced a migration strategy to SOC, based on one of the Web 2.0 challenges, namely the mashup technology. The six-step MASHUP roadmap propounded here aims to overcome many difficulties; especially QoS related ones, in the migration process by means of architecture-driven designs with DSEs. The MASHUP approach is already being tried to create a Service-Oriented Business Application through the mashup of financial gateway services; almost all of them are legacy and out-of-organization services.

In order to have such a pervasive service providing a single interface to customers but managing a variety of resources at the backend, the approach should be capable of mashing up in a harmonized platform. This harmonization can be achieved due to the generality of Mashup Choreography engine, which enables plugging multiple legacy service adapters implemented within a domain specific perspective. Having enormous amount of legacy systems without the source code, MASHUP approach will facilitate their migration to SOC as well.

5. Acknowledgement

This work has been partly supported by TUBITAK within the context of METU-ISTEC project.

6. References

[1] Sneed H.M., "Encapsulation of Legacy Software: A Technique for Reusing Legacy Software Components", Annals of Software Engineering, 2000, Vol. 9, 293-313.

[2] Papazoglou M.P. and Georgakopoulos D., "Service-Oriented Computing", Communications of the ACM, Vol. 46 No: 10, October 2003, 25-28.

[3] Sneed H.M., "Integrating Legacy Software into a Service Oriented Architecture", CSMR'06, IEEE CSP, 2006.

[4] Benatallah B. and Nezhad H.R.M., "Service Oriented Computing: Opportunities and Challenges", SWDB, 2005.

[5] Hinchcliffe, D., "i-Technology Viewpoint: Is Web 2.0 the Global SOA?", SOA World Magazine, 2006.

[6] SEI, "SMART: The Service-Oriented Migration and Reuse Technique, (CMU/SEI-2005-TN-029), 2005.

[7] Ziemann, J., Leyking, K., Kahl, T. and Dirk, W.: "Enterprise Model driven Migration from Legacy to SOA", Software Reengineering and Services Workshop, 2006.

[8] Channabasavaiah, K and Holley, K., "Migrating to a service-oriented architecture", IBM White Paper, 2004.

[9] Oracle, "Oracle IT Modernization Series: The Types of Modernization", Oracle White Paper, 2006.

[10] Microsoft Corporation, "Enabling Real World SOA through the Microsoft Platform", 2006.

[11] SAP, "NetWeaver Open Integration Platform", https://www.sdn.sap.com/irj/sdn/developerareas/netweaver

[12] BPMI, "Business Process Management Initiative", http://www.bpmi.org/.

[13] Dogru, A. H., and Tanik, M. M., "A Process Model for Component-Oriented Software Engineering", IEEE Software, Volume 20(2), 2003.

[14] Cetin, S., Altintas, N. I., and Sener C., "An Architectural Modeling Approach with Symmetric Alignment of Multiple Concern Spaces", ICSEA 2006, IEEE CSP, 2006.

[15] Altintas N. I., Cetin S., and Dogru H. A., "Industrializing Software Development: The 'Factory Automation' Way", TEAA-2006, LNCS, 2006.

[16] Altintas, N. I. and Cetin, S., "Integrating a Software Product Line with Rule-Based Business Process Modeling", TEAA 2005, LNCS 3888, pp. 15-28, 2005.

[17] Vinoski S., "REST Eye for the SOA Guy", IEEE Internet Computing, 2007.

A Hybrid Method For The Identification of Expression Patterns From Microarray Data

Zerrin Sokmen[1], Mehmet Ozturk[2], Volkan Atalay[1], Rengul Cetin Atalay[2]

[1] Computer Engineering Department, Middle East Technical University, 06531 Ankara, Turkey

[2] Department of Molecular Biology and Genetics, Bilkent University, 06800 Ankara, Turkey

Expression microarrays provide a platform for the investigation of thousands of genes simultaneously. There are various methods which aim to identify differentially expressed significant gene list over a microarray experiment [1]. Many of those analyses focus in pairwise comparison of the selected subset data of the experiment rather than considering the entire data. The aim of this study is to extract gene expression patterns with computational clustering techniques using not only a subset but the entire experimental microarray dataset. Therefore we applied a hybrid computational method to identify expression patterns on an affymetrix microarray dataset containing 54000 probe sets of 6 different cell fates with 3 replicates for each stage.

The raw probe data were normalized by RMA pre-processing method. Subsequently K-means clustering algorithm [2] was applied to explore a group of genes which were in similar behavior at each experimental cell fate. We set the total number of clusters to 100 in K-means algorithm. In order to identify significant clusters, the variance in cell fates for each cluster was computed and the most variant 13 clusters were selected as candidate sets. Among 13 candidate clusters, we selected an expressing pattern cluster; called as promising cluster which seemed biologically significant based the experimental design. The promising cluster then was used as the profile to train a hidden Markov model (HMM) [3]. Before training of HMM, outlier genes were removed according to their mean and standard deviations. In other words, the expression value of a gene should be in a valid range higher or lower than the standard deviation, otherwise that gene is removed from promising cluster. Our HMM comprised of 6 states and each state generates one of the 3 observation symbols: up regulated, down regulated and no-change. HMM was trained with Baum-Welch method by using expression value of 83 genes in the promising cluster.

In order to explore genes that showed expression patterns similar to that of promising cluster profile, we tested all genes in the data set over the trained HMM. The genes which gave higher probability in HMM test process were labeled as significant. There were 620 significant genes. We have observed that the profiles of these significant genes were very similar to that of the profile of the promising cluster. Therefore, our hybrid method could achieve to identify a group of genes sharing a similar expression pattern.

Acknowledgement: This study is supported under TUBITAK EEEAG-105E068.

References:

1. J. Claverie. “Computational methods for the identification of differential and coordinated gene expression” *Human Molecular Genetics* 8: 1821-1832, 1999.

2. J. B. MacQueen. "Some Methods for classification and Analysis of Multivariate Observations” *Proceedings of 5-th Berkeley Symp. on Math. Statistics and Probability"*, 1:281-297, 1967.

3. L. R. Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition” *Proceedings of the IEEE*, 77(2):257-286, 1989.

Legacy Migration to Service-Oriented Computing with Mashups

Semih Cetin[1,2], N. Ilker Altintas[1,2], Halit Oguztuzun[2], Ali H. Dogru[2],
Ozgur Tufekci[1] and Selma Suloglu[2]

[1] *Cybersoft Information Technologies*

[2] *Department of Computer Engineering, Middle East Technical University*

{semih.cetin, ilker.altintas, ozgur.tufekci}@cs.com.tr

{oguztuzn, dogru, selma.suloglu}@ceng.metu.edu.tr

Abstract

Although Service-Oriented Computing holds great promises, it is still not clear when and how the existing systems will exploit this new computational model. The problem is particularly severe for the software having several years of use. This work provides a roadmap for the migration of legacy software to Service-Oriented Computing by means of right levels of abstraction. The key idea is having integration even at the presentation layer, not only at backend layers such as application or data. This requires re-inventing the popular mashup technology of Web 2.0 at the enterprise level. Domain-Specific-Kits and Choreography Engine concepts that were originally introduced by the Software Factory Automation approach have been reshaped as another enabling technology towards migrating to the service harmonization platform. The paper also exemplifies the proposed approach on a simple case problem.

1. Introduction

Today's computing necessitates emerging services available as quickly as possible but this is not so easy to achieve since enormous amount of functionality is somehow buried in billions of dollars worth of existing code. Service-Oriented Computing (SOC) may help tackle with this problem by decomposing monolithic legacy systems into loosely coupled parts wrapped by service adapters. This is also quite challenging since these legacy codes mainly have single tier architectural designs, which complicates service partitioning as well as assurance of the Quality of Service (QoS) attributes like durability, maintainability and efficiency.

Another difficulty stems from the fact that service pervasiveness essentially requires the content-oriented harmonization rather than the data or behavior-oriented one, which is common today. Known as "mashup" technology, this new trend adds more value to existing services but such services are expected to be REST services (explained in Section 3.1) rather than classical Web services. This prerequisite makes even today's state of the art Web services [9] legacy as well.

All these challenges reveal that a generic migration strategy to SOC may not scale well for miscellaneous business requirements. Alternatively, domain specific approaches will tackle with the itty-bitty details more effectively but with a certain surcharge. Therefore, a compromise is needed for the effectiveness of domain specificity at the cost of generality. This paper puts forth such a strategy for migrating legacy solutions to SOC by mashing them up in a harmonization platform: "MigrAtion to Service Harmonization compUting Platform (MASHUP)". MASHUP provides a six-step roadmap to mash existing binaries in a generic mashup server where domain specific wrappers are plugged in to wrap existing logic either as B2B or B2C services.

2. What is legacy software?

Developed using better practices or not, existing software artifacts, very expensive by nature, dominate everyday life. They cannot be discarded simply or converted to new technologies easily, and worse, for some, re-engineered due to their unstructured design.

The shift from early-unstructured development to "structured/traditional", then to Object-Oriented (OO), and currently to Service-Oriented Architecture (SOA) is well known. Besides, several other technologies and approaches were utilized in the supplying industries of software. Aspect Orientation, Component Orientation (CO), Client/server Architecture, 3-tiered Architecture, Rich Client, Thin Client, and so on, can be listed as examples to the recent trends in industry. Some of

them will not even complete their evolution before the expected domination of SOA on the market. Among them, CO is a concept that is still in need for the wide adaptation of techniques. We regard CO in support of the realization of Service Oriented Computing (SOC).

We intend to facilitate software development in an effort to respond to the sophisticated demand in the business world. While trying to leverage Web services, we also intend to reuse the wealth of previously built software. In recent literature, legacy software was mostly related to the earliest development that could not carry any design documentation to the engineer that was in position to change or incorporate the old assets. Legacy software components have been defined as "*jobs, programs, modules, or procedures within existing application systems which are more than five years old*" in [1]. Another definition paying attention to the outdated technology for the development of the programs can be found in a later work of Sneed [3]. MASHUP has a perspective that any software artifact that was built using pre-SOA techniques is legacy. The challenge is to adopt such software in the new development that will be service-oriented.

Services are actually engineered with the purpose of composing them with different applications. They are "components" whose coupling with many different systems is inherently facilitated. Legacy software, on the other hand, constitutes frequently monolithic, big sized applications. The philosophy had been to write code for the given system specifications, but now it turned out to "compose" existing services – even "on the fly", in the age of ubiquitous / pervasive services.

3. Service-Oriented Computing (SOC)

Service-Oriented Computing is a new computing paradigm that takes services as basic elements. SOC relies on SOA when constituting the service model. Basic tenets of SOC are loosely coupled asynchronous interactions on the basis of open standards to support complex business processes and transactions as reusable and accessible services, in contrast to tightly integrated monolithic applications [2].

The constituents of SOA can be providers (basic service providers and aggregators), consumers (service aggregators and end users) and brokers (middleware and registries). Providers do advertise their services to registries and consumers query registries in order to discover required services that satisfy their goals.

Adapted from [2], the crosscutting concerns of SOA can be described at three different service levels:

- *Basic level* includes service description portions (capability, interface, behavior, QoS) and basic operations on services (publication, discovery, selection, binding and invocation) for offering reusable, adaptable and context-aware services to conform a constructible model.
- *Composite level* includes coordination activity (orchestration of services), conformance (integrity insurance of interfaces), monitoring, QoS for offering static/dynamically composable, verified with regard to quality concerns and seamlessly integrated services to conform a composable model.
- *Managed level* includes operations (providing control and feedback) and market considerations for offering satisfied regarding market needs, correlated and controllable services to conform a canonical model.

The fundamental challenges of SOC are finding the effective and efficient ways of service description, discovery, selection, composition, monitoring and integration while focusing on semantics those point out intelligent, dynamically adaptive and context sensitive services. Thanks to the developments in networks and microelectronics, in the foreseeable future, ordinary devices will be equipped with embedded processors, each with wired or wireless communication facilities and embedded software. Networks of such devices, mobile or stationary, will form parts of larger systems, providing and requiring services. These services will be ubiquitous in the sense that they will be part of our everyday environment, and they will be pervasive in the sense that they will emerge in the context of interactions experienced among people and devices.

In the context of service composition challenge, the definition of service component adds an abstraction layer to facilitate the representation of modularized service based applications to overcome complexity. Service Component Architecture (SCA) emerges with a set of specifications describing a model for service component as a cohesive and conceptual module which includes services assembled by wiring of service oriented interfaces and orchestrated according to stated business logic [5]. SCA can be coupled with Service Data Objects (SDO) to provide uniform representation of business data for accessing the messages that arrive at or are sent from components.

3.1. Existing technologies for services

Internet standards assist realizing SOA with Web services through exposing them as services that can be described, advertised, discovered and interoperated with [4, 6]. Web services can be described using Web Service Description Language (WSDL), which defines operations along with input/output messages and data residing in messages [7]. The interaction between

services can be achieved by an XML document whose schema is specified by Simple Object Access Protocol (SOAP) using HTTP at transportation layer [8].

Representational State Transfer (REST) technology mainly differs from SOAP in representing the service definition and invocation design distinct from RPC. The architectural constraints of REST consists of uniform interface constraint meaning that all clients relate to a certain unique interface for each resource, statelessness signifying that state information is carried by all requests and cashing that denotes letting clients or intermediaries cache responses [9, 10].

The universally accepted standard to facilitate the discovery of Web services is Universal Description, Discovery, and Integration (UDDI) [11]. However, UDDI supports only keyword-based matches, which often lead to poor performance. To overcome this challenge Service Aggregation Matchmaking (SAM), an algorithm for composition-oriented discovery of Web services features more flexible matching queries with service registries using OWL-S ontologies [12].

Composite services can be built from existing basic or composite services following the composition rules that determine the ordering of service invocations. Composite services have two views complementing each other, namely, orchestration and choreography. Orchestration of Web services enables coordination of services by assigning an orchestrator, which is a central manager responsible for invoking and combining sub-activities. However, Web service choreography defines inter and intra collaboration of each service to realize the system target goal without a central mechanism. Business Process Execution Language (BPEL) can be used to realize orchestration and choreography in order to achieve interoperability between applications. But, any of the service composition language is sufficient to represent business agreement support, which defines the contract between two parties on QoS [6].

Also, the lack of semantic representation of services restricts the capability to support service composition. W3C suggests a semantic-based description of Web services using OWL-S ontology [13].

3.2. Business process management

Business Process Management (BPM) involves a control mechanism for defining, altering, orchestrating, executing and monitoring business processes taking into account business rules. BPM defines behavioral roles of business processes, which are seen as an assembly of activities realized regarding workflow and business rules with the human intervention. Business Process Modeling Notation (BPMN), Business Process Modeling Language (BPML) and BPEL are the key notations for visualizing, altering, modeling and executing. Inter and intra coordination and interaction of business processes is inevitable on the notion of business transactions [14].

Within SOC, business processes act as a conceptual player, whereas services spread over logical layer of the picture. When appropriately represented and put into development, business processes can provide the application-wide glue in composing Web services. Without such processes, the SOA cannot account for the sequencing of the service activations. The new trend currently points out networks of orchestration for collaborating different enterprise applications within and across organizational boundaries; context adaptive, ambient intelligence type services are spread over the network and are accessed potentially from any device and any location [15].

4. Web 2.0 and Mashup technologies

Adoption of SOA holds the promise of enabling businesses to more effectively adapt to change and to add new offerings to their existing products in a more efficient manner. These offerings have first started to be realized on the technology stack as pure behavioral requirements. However, it did not take so long for service community to realize that moving to a service-oriented world cannot happen without having service enablement in its widest range covering the satisfaction of quality of services requirements - requirements that address aspects of the system that do not directly affect the business functionality. Such an understanding has been dramatically changing the path to SOA realization mainly from the viewpoint of Web content rendering.

The Web 2.0 designation was originally used to describe Web experiences that fundamentally engage users by: (a) allowing them to participate in sharing information and enriching data freely, (b) readily offering their core functionality as open services to be composed or "mashed up" into new services and sites, and (c) placing the Web at the center of software experience both in terms of data location and where the software is [16].

One of the key concepts for convergence of Web 2.0 and SOA is the idea of turning applications into platforms. This concept and the focus on software via services seem to be more readily accessible through both B2C and B2B channels. The convergence and overlap between SOA and Web 2.0 organizing principles in software appears obvious from this perspective as depicted in Fig. 1. SOA initiatives tend to have application integration as one of their primary goals, while Web 2.0 software usually has communities, user participation, and unintended uses

as primary goals that are within reach, thanks to the extremely simple use and integration patterns. Thus, Web 2.0 targets at making the complete picture of "Software as a Service (SaaS)" a reality.

Figure 1. Convergence of SOA and Web 2.0

Web 2.0 takes us to a deeply service-oriented world where we can exploit everyday services such as news, instant messaging, and blogging via our desktops, mobile phones, PDAs, etc. Most of the time, we even don't know what is behind the scenes in terms of the massive network of service interactions. Such services seem re-writing of the existing services. In fact, they are smart compositions (mashups) of the existing services with new ones. Known as "Web application hybrid", this smart way of combining the content from more than one source into an integrated experience is called "mashup" technology.

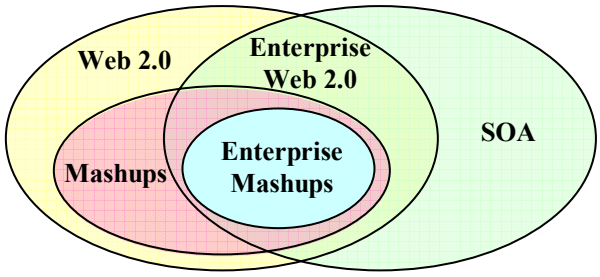


Figure 2. Enterprise service mashups and Web 2.0

Mashup technology leverages the Service-Oriented Business Application (SOBA), which is known as a composite application built out of the services that realize a business process. Why mashup technology is different can be clearer by considering that it enables SOBA not only at middle tiers (application and data tiers), but massively at presentation tier as well.

As shown in Fig. 2, further convergence of Web 2.0 and SOA creates a new trend of applications so-called Enterprise Web 2.0, which embraces the union of Web 2.0 technologies such as Ajax with legacy systems, and Web services to enable organizations to deploy robust, reliable, and secure business applications over the Web. Besides, as mashup technologies take the logical next step, Enterprise Web 2.0 will turn out to be an Enterprise Mashup.

For a mashup to be an enterprise mashup in that it addresses a particular business problem, tight coupling between provider and consumer software would be a serious concern. Most of today's mashups, however, care little about loose coupling. Mashups that meet business needs, therefore, will require SOA, and the SOA infrastructure necessary to guarantee loose coupling. Most importantly, however, SOBAs require governance as shown in Fig. 1. Clearly, no business would risk allowing any of its employees to assemble and reassemble business processes with no controls in place. It must be ensured that the resulting SOBAs followed corporate policies. The weakness of today's mashups is their inherently ungoverned nature. In any case, the true promise of SOBAs depends upon user interfaces being sophisticated enough for a broader business audience to use. Few such tools exist today, but the writing is on the wall: the enterprise mashup is the future of the SOBA consumer [32].

5. Existing SOA migration strategies

Converting legacy applications to services allows systems to remain largely unchanged while exposing functionality to a large number of clients through well-defined service interfaces. Migrating a legacy system to SOA, e.g. wrapped as Web Services, may be relatively straightforward. However, characteristics of legacy systems like platform, language, architecture, and the target SOA may unexpectedly complicate the task. This is particularly the case during migration to highly demanding SOA enforcing the rich content rendering, service composition and mashups.

There exist some proposals for SOA migration and each approaches to the problem mainly from different viewpoints though they have commonalities. Table 1 summarizes well-known SOA migration strategies and how they look at the SOA migration problem.

Service-Oriented Migration and Reuse Technique (SMART), proposed by SEI at CMU, is an initial approach to the identification and analysis of issues in migration to services [17]. It helps organizations analyze legacy systems to determine whether their functionality can be reasonably exposed as services in an SOA. Driven from the Options Analysis for Reengineering (OAR) method [18], SMART gathers a wide range of information about legacy components,

the target SOA, and potential services to produce a service migration strategy for the organization as its primary product. Demanding source code engineering, SMART offers a quite clear roadmap. It is apparently platform/vendor/tool independent and it targets the process/service integration only at middle tier. But, it does not address QoS issues, content integration and other Web 2.0 technologies.

Table 1. SOA migration strategies

	SMART [6]	SWA * [3]	EMDMA ** [7]	IBM [8]	ORACLE [9]	MICROSOFT [10]	SAP [11]
Provides a Roadmap	√	√	?	√	√	√	√
Platform/Vendor/Tool Independent	√	√	√				√
Scales to Different Domains		√	√	√	√	?	
Process/Service Integration	√	√	√	√	√	√	√
Content Integration		?	?	√	√	√	√
Web 2.0 Enablement				?	?	√	?
Mashup Enablement				?	?		
Quality of Services				√	?	?	?

√ Fully match
? Partly match
* Salvaging & Wrapping Approach
** Enterprise Model Driven Migration Approach

Salvaging & Wrapping Approach (SWA) proposed by Sneed [3] involves a tree-step procedure for creating Web services from legacy code. These are salvaging the legacy code, wrapping the salvaged code and making the code available as a Web service. Here, the business rules are discovered by identifying the variables and hence the returned functions. Once the code is reengineered a tool is utilized for automatic code extraction, and then the legacy code is wrapped behind an XML shell to offer Web services to external users. SWA can be effective at process and service integration, but it has limited support for content integration by means of wrapping second-level Web Services and no support for rich content integration. It does not address the QoS as a first class concern.

Ziemann et al [19] describe a business driven legacy-to-SOA migration approach (EMDMA) based on enterprise modeling by introducing an elementary process model between the business function tree and the tree related to the legacy system, which is later aligned to function tree of the legacy system. Later, it applies a transformation from the legacy business process model to the SOA process model. EMDMA draws attention to the fact that aspects of functional granularity, security, reliability, etc. are not taken into account sufficiently and to the need for investigating how special characteristics of legacy systems (reliability, scalability, policies) can be matched to SOA systems. It doesn't either address rich content integration issues like Web 2.0 and mashups.

Table 1 reveals that IBM has a clear picture for a contemporary SOA migration strategy. IBM SOA strategists think that it is not enough to consider only process/service integration, but also the integration at the end-user interface. Application connectivity, process integration, information integration and a build-to-integrate development model should be considered [20] as well. The roadmap of IBM aims to help identification of appropriate best practices, merge with existing IT processes, and support governance of the capabilities introduced with SOA. Even new content integration techniques of Web 2.0 and mashup can be partly addressed in the SOA migration activities.

Oracle is a very close runner to IBM with its clear migration strategy, known as Oracle Modernization Framework (OMF) [21]. The first step an organization must take is to carry out an Application Portfolio Analysis (APA) of the current applications and their environment. Subsequently, organization will have the information necessary to choose among various modernization techniques such as re-architecting, enablement, re-hosting, automated migration or COTS replacement. This migration roadmap has been strongly supported by a tool suite, known as Oracle SOA Suite [22], which guides users from service choreography and orchestration to even Web services wrapping of legacy resources. New content generation techniques (Web 2.0 and mashup) have been partly supported so far.

Microsoft advocates what it calls a "middle out" approach in which SOA efforts are driven by strategic vision and business needs, and are met through iterative and incremental manner. Microsoft builds SOA through its technology stack, from the developer tools that build Web services, such as .NET to server side products, i.e. BizTalk and Microsoft Office SharePoint Servers, which further Web service construction by connecting and orchestrating services and composite applications [23]. Thus, SOA migration strategy of Microsoft is primarily leveraged by BizTalk Server concept and it heavily uses XML technology, which has some limitations in scaling to other domains. Though Web 2.0 is available today, mashup and QoS issues are partly supported.

SAP introduced the concept of Enterprise Services Architecture (ESA) or Enterprise SOA in 2003. The company has developed NetWeaver, a Web-based,

open integration and application platform and foundation for enterprise SOA [24]. NetWeaver allows the third parties offering solutions such as Seagull Software's LegaSuite or Software AG's Legacy Integrator that use an SOA approach to create reusable assets from core legacy applications to complement SAP NetWeaver customers seeking to integrate legacy. SAP provides the ESA Adoption Program to steer its customers on a smoother path to a SOA and ease the learning process of its NetWeaver development and integration platform.

Although vendors already offer complex proprietary software systems to enact an SOA, they are lack of a holistic method to systematically prepare a SOA implementation, taking into account requirements of existing (legacy) system and of current business needs as well. Almost all of the existing approaches are implicitly bottom up neglecting business requirements. Focusing on the IT perspective yields brittle Web Services and case-by-case, experimental systems that fall short of implementing SOA principles [25].

Moreover, existing approaches mainly demand context sensitive solutions for migrating legacy applications to service orientation. In fact, they propose merely back-end integration based on the reverse engineering and architectural reconstruction of source codes. Such context sensitive attempts, however, are not practical enough for enterprise scale legacy applications and they cannot be generalized easily.

Consequently, the migration of legacy systems into SOA environments is a major research challenge with a significant economical impact. Despite major research efforts in the field of legacy reengineering and SOA, current approaches only focus on the technical migration itself and do not sufficiently reflect business needs especially QoS.

6. The MASHUP migration approach

The proposed mashup migration strategy has six steps shown in Fig. 3. The strategy addresses both behavioral and architectural aspects of the migration. The first two activities are the modeling of the target enterprise business (MODEL) and the analysis of legacy systems and infrastructure (ANALYZE). These activities lead to two main steps: (MAP & IDENTIFY) maps model requirements to legacy components and services identifications; and (DESIGN) models mashup server architecture with Domain Specific Kits (DSKs), which abstracts legacy components.

Both the mapping and architecture design activities might cause a loopback to MODEL and ANALYZE activities to re-reconsider some of the decisions and improvements. As a result of these major activities, target system service dependency graph has been constructed and mashup server concrete architecture has been designed. Defining the Service Level Agreements (SLAs), including non-functional and contextual properties, is the next step (DEFINE) that will obviously be followed by implementation and deployment activities (IMPLEMENT & DEPLOY).

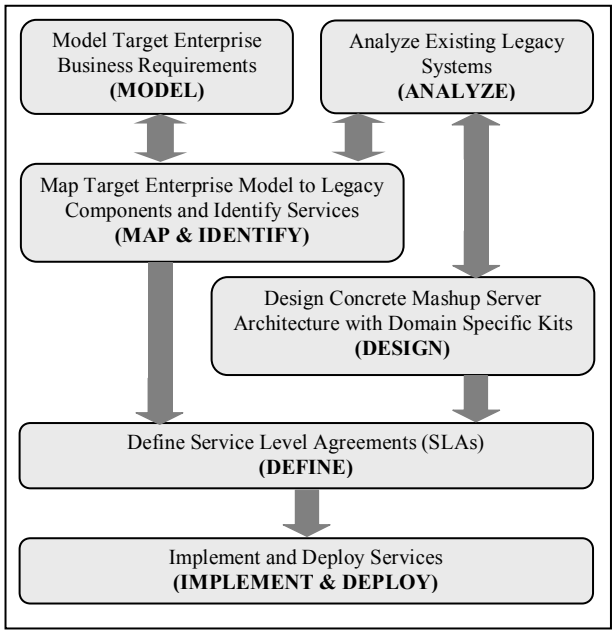


Figure 3. Migration activities

Our migration approach is based on target enterprise business model, similar to the approach proposed by [19]. However our model differs by being architecture-driven, by being extendable using domain specific languages and by considering the quality of service and contextual information as key concerns. The details of migration steps are as follows:

Step 1 (MODEL) Model business requirements of target enterprise: The business requirements are modeled for understanding the functional requirements of the target system. The business domain model might be expressed using a semantic notation similar to Business Process Modeling Notation (BPMN) [12], which is a methodology independent and unambiguous notation to express any business process.

Step 2 (ANALYZE) Analyze legacy systems: The analysis of existing systems reveals the important data for designing domain specific kits and reference architecture and for investigating reuse potential of existing legacy components. The results of this step include information on system architecture, system components, infrastructure details, interfaces, the

presentation characteristics, QoS attributes, level of maintainability, level of complexity and coupling, etc.

Step 3 (MAP & IDENTIFY) Map business requirements to system components and identify services: System components might be reused legacy components or freshly developed ones. Thus, this step uses the results of business modeling and legacy analysis activities. The iterative mapping process is as follows: (a) if a business requirement can be satisfied by one of the existing legacy components, then simply wrap it by considering the QoS attributes; (b) if there is a gap with the existing legacy component and requirement, and the gap can be filled during service wrapping, then accustom the legacy component into a new service; (c) if the gap cannot be fulfilled, then develop a new service for the requirement. The new service need not be implemented in any specific language as Mashup Server reference architecture can utilize any component (implemented in any language) using appropriate Domain Specific Kits (DSKs).

Continue this step as the services can be broken down into further sub-services and those can be mapped using the above mappings (either into legacy components or new ones). A successful mapping might use Commercial-Off-The-Shelf (COTS) components as service sources since those COTS components can be plugged into reference architecture with proper DSKs that can utilize component orientation techniques [13]. As part of this activity, investigate any implicit dependencies. The implicit dependencies can occur at data level, platform level, service quality level, etc. For instance, if there are two services from two different systems but they need a common data set, then mapping process should introduce suitable data feeder services. At the end of this step, the business requirement to service (BR2S) and service-to-service (S2S) breakdown has been determined at the conceptual level.

Step 4 (DESIGN) Design Concrete Mashup Server Architecture with Domain Specific Kits (DSKs): The design of Mashup Server Architecture (MSA) has been adapted from the architecture modeling approach already employed for software factory automation [15]. Architecture modeling relates architectural aspects and quality targets to running DSKs and choreography rules. The "symmetric alignment" technique [14] is assisted by a methodical approach to identify components (Domain Specific Engines – DSEs), and connectors (composition of DSEs) in the solution domain. Identification of architectural properties facilitates the definition of contextual information, which contains the stateful/stateless information to connect individual DSEs, and is needed for independent design and implementation of these individual DSEs through a standard communication schema across Domain Specific Languages (DSLs). Defining DSLs under the governance of a service meta-model enables the exposition of service from different sources with varying attributes. This step results in MSA with a set of specific pluggable DSEs, contextual information, internal requirements of DSEs and a runtime model.

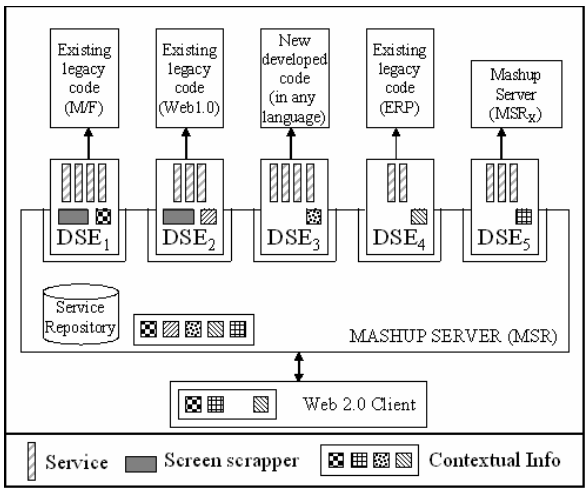


Figure 4. MASHUP reference architecture

Step 5 (DEFINE) Define service level agreements (SLAs): At the end of Step 3, the service mappings are finalized; and the MSA (from Step 4) is ready to execute those services. SLA is defined in a common service repository and contains information on: a) contextual information; b) QoS characteristics of the service. QoS characteristics are basically inherited from the provider system and managed by the DSE serving the service, but the service implementation may introduce additional quality concerns. The Mashup Server manages both the contextual information and QoS definitions.

Step 6 (IMPLEMENT & DEPLOY) Implement and Deploy Services: This includes wrapping of existing legacy components, customization of existing legacy components with some additions, development of new services, and introduction of COTS components as services. The development of new services does not require that those new services need to be implemented using a common language on a common platform since they are abstracted by appropriate DSEs at the Mashup Server level.

The reference architecture model for the mashup server has been depicted in Fig. 4, which highlights the generic reference architecture that enables plugging the specific DSEs for different systems (service sources). The reference architecture depends on a meta-model for specifying DSLs for different service sources, a common service repository, and a policy for managing

contextual information. The reference architecture enables the integration of another mashup server using a particular DSE (see DSE_5 in Fig. 4).

The reference architecture employs a choreography engine that manages the DSEs. Dynamic plugging and context-awareness of DSEs are crucial for the runtime execution model. The choreography engine, as a central authority, enables this communication and coordination among DSEs, which ensures the context management, state coordination, communication, produce/consume messaging, nested processes, and service-oriented exception handling. Corresponding DSLs are required to comply with a certain meta-model to be plugged into the Choreography Engine. DSLs are declarative, context-free and loosely coupled to each other; hence we can readily apply the golden principle of separation of concerns for different domains.

The DSEs can perform specific tasks for different systems. Consider, for instance, a Web 1.0 system without an API. It heavily mixes presentation with content and makes it hard to sift out meaningful data from the rest of the elements used for formatting, spacing, decoration or site navigation. In such a situation the DSE can employ "Screen/Web scraping" techniques by analyzing the page structure and wrapping out the relevant records. In some cases the task is even more complex than that: The data can be scattered over more pages. Then, triggering of a GET/POST request may be needed to get the input page for the extraction or authorization that might be required to navigate to the page of interest. The situation may be even more complex if there are workflows running in the provider system and if tight security policies are applied.

In case the legacy code is provided by a mainframe, the screen scrapping techniques are essential to simulate the working of the mainframe terminals. Using screen scrapping, the data would be retrieved from the host and also posted onto the host. Mashed up services are accessed through Web 2.0 clients supported with AJAX, Flash, Flex, JavaScript, and other XML-based rendering [30] technologies. Even smarter clients can be used to access several mashup servers from a single client empowered with mashup choreography abilities.

7. MASHUP migration case study

The example to demonstrate the mashup reference model and migration strategy comes from our financial gateway product line: Mashing up financial gateways and black list management (BLM) sources. The BLM sources include Central Bank lists, credit history lists, capital market black lists, internal black lists of the bank, etc. These services are accessed via different ways: (a) through a standalone legacy program accessed via the user screens, (b) through gateway accessing mainframe, (c) through MQ-based access, (d) through existing AS/400 based core banking solution.

The mashup problem is as follows:

i. "Provide a third party service to the banks so that they can access these varying sources via a single interface."

ii. "Enrich this service with value-added services such as combining the black list records with Ministry of Finance records accessed through a web service."

iii. "Enable bank to associate their customer information while querying black lists on-demand."

Initial modeling of the problem revealed that the domain-specific approach is effective in absorbing variations in service sources. The design of domain specific engines can hide all the dirty tricks and bits-and-bytes of external accesses. The modeled DSEs are as follows:

- DSE-CB: Modeled to access Central Bank list with the old legacy program. This engine must do screen scrapping to initiate a query and extract information from the display screen.
- DSE-MF: Modeled to access mainframe to retrieve/post data from/onto the host.
- DSE-MQ: Modeled to access message-based system with message queues.
- DSE-AS400: Modeled to access AS/400 system to retrieve/post data from/onto the AS/400. This engine must also screen scrapping and manage complex screen flow logic.
- DSE-WS: Modeled to encapsulate web service access.

The approach can utilize the existing legacy application, such as Central Bank lists, mainframe lists, and AS/400 screens, without any change at the code level. Furthermore, it is relatively easy to integrate Ministry of Finance querying with a particular DSE. "Querying black list" service can support advanced functionality such as record matching among all black lists and data linkage algorithms. Those new services can be implemented using any language of choice during implementation.

Our initial investigation showed that the model is quite tolerant to variations in service sources and can help to reuse existing legacy components compared to our earlier implementations where such external links are managed at the code level.

8. Conclusions

As a programming model, SOC motivates a loosely coupled way of thinking about building applications. The SOC platform is full of promises like enabling you to design software systems that provide services to other applications through published and discoverable interfaces, and where the services can be invoked anytime and anywhere.

However, achieving SOC platforms is not that much easy since service pervasiveness mainly needs content-oriented harmonization of heterogeneous resources having usual legacy implementations. Migrating legacy binaries to SOC is a challenging research area and likely to be so in the near future since the dynamics of Web 2.0 keep pushing software engineers with new content rendering tools, techniques and protocols.

This paper introduced a migration strategy to SOC, based on one of the Web 2.0 challenges, namely the mashup technology. The six-step MASHUP roadmap propounded here aims to overcome many difficulties; especially QoS related ones, in the migration process by means of architecture-driven designs with DSEs. The MASHUP approach is already being tried to create a Service-Oriented Business Application through the mashup of financial gateway services; almost all of them are legacy and out-of-organization services.

In order to have such a pervasive service providing a single interface to customers but managing a variety of resources at the backend, the approach should be capable of mashing up in a harmonized platform. This harmonization can be achieved due to the generality of the mashup choreography engine, enabling the pluggability of multiple legacy service adapters implemented with domain specific perspective.

9. Acknowledgement

This work has been partly supported by TUBITAK within the context of METU-ISTEC project.

10. References

[1] Sneed H.M., “Encapsulation of Legacy Software: A Technique for Reusing Legacy Software Components”, Annals of Software Engineering, 2000, Vol. 9, 293-313.

[2] Papazoglou M.P. and Georgakopoulos D., “Service-Oriented Computing”, Communications of the ACM, Vol. 46 No: 10, October 2003, 25-28.

[3] Sneed H.M., “Integrating Legacy Software into a Service Oriented Architecture”, CSMR’06, IEEE CSP, 2006, 3-14.

[4] Benatallah B. and Nezhad H.R.M., “Service Oriented Computing: Opportunities and Challenges”, SWDB, 2004, 1-8.

[5] “Service Component Architecture”, OSOA, http://www.osoa.org/display/Main/Home.

[6] Bucchiarone A., Gnesi S., “A Survey on Services Composition Languages and Models”, International Workshop on Web Services Modeling and Testing, 2006, 37-49.

[7] Christensen E., Curbera F., Meredith G. and Weerawarana S., “Web Services Description Language (WSDL) ver. 1.1”, http://www.w3.org/TR/2001/NOTE-wsdl-20010315, 2001.

[8] W3C. SOAP 1.2 Working draft, http://www.w3.org/TR/2001/WD-soap12-part0-20011217, 2001.

[9] Vinoski S., “REST Eye for the SOA Guy”, IEEE Internet Computing, 2007.

[10] Fielding R. T., “Architectural Styles and the Design of Network-based Software Architectures”, PhD Thesis, University of California, Irvine, 2000.

[11] UDDI, “The UDDI Technical White Paper”, http://www.uddi.org/, 2001.

[12] Brogi A., Corfini S. and Popescu R., “Composition Oriented Service Discovery”, Springer Verlag 2005.

[13] W3C, “Semantic Web Services Interest Group Charter”, http://w3c.org/2003/10/swsig-charter.

[14] Tufekci O., Cetin S., Altintas N. I., “How to Process [Business] Processes”, IDPT-2006, 2006.

[15] Software Technologies Unit Information Society and Media DG European Commission, “Software, Services and Complexity Research in the IST Programme”, 2006.

[16] Hinchcliffe, D., “i-Technology Viewpoint: Is Web 2.0 the Global SOA?”, SOA World Magazine, 2006.

[17] SEI, “SMART: The Service-Oriented Migration and Reuse Technique, (CMU/SEI-2005-TN-029), 2005.

[18] Bergey, J., O'Brien, L. and Smith, D.: “Using the Options Analysis for Reengineering (OAR) Method for Mining Components for a Product Line”, SPLC2, 2002.

[19] Ziemann, J., Leyking, K., Kahl, T. and Dirk, W.: "Enterprise Model driven Migration from Legacy to SOA", Software Reengineering and Services Workshop, 2006.

[20] Channabasavaiah, K and Holley, K., "Migrating to a service-oriented architecture", IBM White Paper, 2004.

[21] Oracle, "Oracle IT Modernization Series: The Types of Modernization", Oracle White Paper, 2006.

[22] Mohammed, F., "Oracle SOA Suite", Sys-Con XML Journal, 2007.

[23] Microsoft Corporation, “Enabling Real World SOA through the Microsoft Platform”, 2006.

[24] SAP, “NetWeaver Open Integration Platform”, https://www.sdn.sap.com/irj/sdn/developerareas/netweaver.

[25] Zimmermann, O, Krogdahl, O. P., and Gee, C., "Elements of Service-Oriented analysis and Design: an interdisciplinary modeling approach for SOA projects", http://www-128.ibm.com/developerworks/library/ws-soad1/, 2006.

[26] BPMI, Business Process Management Initiative, http://www.bpmi.org/.

[27] Dogru, A. H., and Tanik, M. M., “A Process Model for Component-Oriented Software Engineering”, IEEE Software, Volume 20(2), 2003.

[28] Cetin, S., Altintas, N. I., and Sener C., “An Architectural Modeling Approach with Symmetric Alignment of Multiple Concern Spaces”, ICSEA 2006, IEEE CSP, 2006.

[29] Altintas N. I., Cetin S., and Dogru H. A., “Industrializing Software Development: The ‘Factory Automation’ Way”, TEAA-2006, LNCS, 2006.

[30] Altintas, N. I. and Cetin, S., “Integrating a Software Product Line with Rule-Based Business Process Modeling”, TEAA 2005, LNCS 3888, pp. 15-28, 2005.

[31] Gregory, J., “Weave Web 2.0 into Business Strategy”, http://www.kickstartall.com/documents/KS_Articles/Weave_Web20.html, 2005.

[32] This, D., “Mashups and Service-Oriented Business Applications To Feature at SOA World 2007”, http://soa.sys-con.com/read/339613.htm, 2005.

Computing Answer Sets Using Model Generation Theorem Provers

Orkunt Sabuncu and Ferda N. Alpaslan

Department of Computer Engineering
Middle East Technical University,
06531 Ankara, Turkey
{orkunt, alpaslan}@ceng.metu.edu.tr

Abstract. Model generation theorem provers have the capability of producing a model when the first-order input theory is satisfiable. Because grounding step may generate huge propositional instances of the program it hardens the search process of answer set solvers. We propose the use of model generation theorem provers as computational engines for Answer Set Programming (ASP). It can be seen as lifting of SAT-based ASP to the first-order level for tight programs to eliminate the grounding step of ASP or do it more intelligently.

1 Introduction

ASP is a declarative approach for solving search problems. A search problem is represented by a logic program whose models, according to the answer set semantics, correspond to the solutions of the problem. It has originated from logic programming and nonmonotonic reasoning during the late 90s [1–3]. With the emergence of the fast answer set solvers and its highly declarative nature, ASP has become an effective and widely accepted paradigm for knowledge representation and reasoning tasks [4]. Among the areas of application, we can list planning, configuration, and diagnosis.

Current answer set solvers work on grounded logic programs. The grounding step becomes a bottleneck of ASP when size of the program is relatively large, which results in a huge propositional instance of the program. This makes the search process of the answer set solver difficult. This is the main motivation behind lifting SAT-based ASP to the first-order level. We use model generation theorem provers as computational engines for ASP, which can process first-order input.

Refutational theorem provers search for the unsatisfiability of input theory usually given in first-order logic. Since the negated conjecture is added to the input theory, the unsatisfiability result proves the conjecture. However, if the theory is satisfiable many provers say only that it is satisfiable. Usually, there is no additional information which can be used as a clue to understand whether the conjecture is invalid or there is an error in the logical representation of the domain and the theorem. A model of the theory can be a valuable information

in this sense. It may help to pinpoint a possible error in the representation or convince us that the conjecture is actually invalid. Unlike traditional theorem provers, model generation theorem provers can output a model when the input is satisfiable.

The main support which makes model generation provers interesting is the paradigm of solving search problems. A search problem is represented by a first-order theory whose models correspond to the solutions of the problem. This support is especially important for knowledge representation and reasoning applications which, in turn, makes the model generation theorem provers flexible engines [5].

We have concentrated on Darwin[1], FM-Darwin[1] and Paradox[2] as model generation systems.

The ability of completing a logic program with variables and forming a first-order theory in this way is the basis of this work. The first-order theory becomes the input to the model generation prover. When the completion semantics [6] and the answer set semantics [7] coincide [8–10] for a program, we can find the answer sets of it by searching the models of the formed first-order theory. In this way, we plan to eliminate the grounding step of ASP or do it more intelligently by the model generation system.

Answer set computation based on the relationship between the completion and answer set semantics has been studied earlier and has lead to successful answer set solvers like Assat[3] and Cmodels[4]. However, these solvers complete the grounded logic program [11, 12]. Since the completed theory is propositional, they use SAT solvers to find the models of it. They have also managed to find answer sets in a sound way even for non-tight programs where the completion semantics and the answer set semantics differ. Our work does not include non-tight programs. Answer set computation experimented here can be thought as a lifting of the process of solvers, like Assat and Cmodels, to the first-order level for tight programs.

Altough not all the results we have obtained so far are very promising, our work can shed light on ASP on the first-order level. It also may increase the awareness of the ASP community in some recent theorem provers. The calculi and techniques used in the model generation theorem provers can be beneficial to ASP on the first-order level.

The brief descriptions of the model generation theorem provers used are given in the next section. Section 3 describes the lifting of the process of SAT-based answer set solvers to the first-order level. Section 4 includes the experimental results. In the final section, conclusions and future work can be found.

---

[1] http://combination.cs.uiowa.edu/Darwin/
[2] http://www.cs.chalmers.se/~koen/folkung/
[3] http://assat.cs.ust.hk/
[4] http://www.cs.utexas.edu/users/tag/cmodels.html

2 Model Generation Theorem Provers

The following is a brief description of the provers used and the calculi they are based on.

2.1 Darwin

Almost all modern SAT solvers are based on DPLL procedure. Successful techniques that are applicable in DPLL-based SAT, like unit propagation, backjumping and learning, make the modern SAT solvers very efficient. The Model Evolution (ME) calculus lifts the Davis-Putnam-Logemann-Loveland (DPLL) procedure to the first-order level [13]. It also aims that those SAT techniques will be still applicable in the lifted DPLL procedure. Darwin is an implementation of the ME calculus and has the first-order versions of successful SAT techniques [14].

The ME calculus tries to find a Herbrand model of the input set of clauses, if one exists. The derivation rules of the calculus updates an initial context until it becomes a model or the unsatisfiability of the input clauses is detected.

A *context* Λ is a finite representation of a Herbrand interpretation I_Λ. It is a set which is composed of ground and non-ground literals. An initial context represents an interpretation which falsifies all the ground atoms. When there is a clause which is not satisfied by the Herbrand interpretation represented by the current context, the derivation rules repair I_Λ by adding literals to the context. If all the clauses are satisfied, then the context actually represents a Herbrand model of the input theory. The ME calculus can also detect the situation when there are no possible repairs of the context. This implies the unsatisfiability of the input theory.

There are two kinds of variables defined in the ME calculus; *universal variables* (simply variables) and *parametric variables* (simply parameters). Context literals are either *universal* (i.e., has only variables) or *parametric* (i.e., has only parameters). All the literals of a context, whether it is universal or parametric, stand for all of its ground instances, which, as a whole, compose a Herbrand interpretation. While an universal literal stands for all of its ground instances with no exceptions, a parametric literal represents all of its ground instances except the ones that are instances of another context literal of the same predicate with the opposite sign.

Here is an example from a presentation of the ME calculus[5]. Let $\Lambda_1 = \{p(u,v), -p(u,u), p(f(u),f(u))\}$ where u and v are parameters. The Herbrand interpretation I_{Λ_1} satisfies every instance of $p(u,v)$ except that u and v are the same. It satisfies every instance of $-p(u,u)$ except the ones of the positive literal $p(f(u),f(u))$. If we change the negative literal of Λ_1 to a universal one, we get $\Lambda_2 = \{p(u,v), -p(x,x), p(f(u),f(u))\}$ where x is a variable. The universal literal $-p(x,x)$ imposes a strict restriction that when the arguments of p are

[5] http://users.rsise.anu.edu.au/~baumgart/slides/MEGoeteborg.pdf

the same, the context produces it negatively. Since the literal $p(f(u), f(u))$ is in contradiction with this restriction, Λ_2 is contradictory.

There are two important situations which the ME calculus should detect during the derivation. First one is the case when the current context falsifies an input clause. The other is the one when the current context permanently falsifies a clause no matter how the context is repaired [13]. A context Λ *falsifies* a clause $C = L_1 \vee \cdots \vee L_n$ (i.e., $I_\Lambda \not\models C$) if there are fresh variants $K_1, \cdots, K_n$ of literals in Λ and a substitution σ, where σ is a simultaneous most general unifier of literal sets $\{K_1, \overline{L_1}\}, \cdots, \{K_n, \overline{L_n}\}$[6]. The σ is defined as a *context unifier* of the clause C against the context Λ. A context Λ *permanently falsifies* C (i.e., $I_\Lambda \not\Vdash C$) if there is a context unifier σ of C against Λ and for all i, $(Params(K_i))\sigma \subseteq P$ where $Params(K)$ denotes the set of parameters found in the literal K and P denotes the set of all parameters.

The Split rule applies when the context falsifies a clause. The rule repairs the context by adding a literal from the falsified instance of the clause (i.e., one of the literals $L_1\sigma, \cdots, L_n\sigma$) in order to satisfy it. However, some of the literals are permanently falsified by the context[7]. The clause composed of the other literals is called the *remainder*. If the remainder is $L_{m+1}\sigma \vee \cdots \vee L_n\sigma$, the left side of the Split rule tries to add $L\sigma$, a literal selected from the remainder, to Λ. The right side adds the Skolemized version of $\overline{L\sigma}$ when the derivation along the left side has been closed. The addition of the new literals should not make the context contradictory. The Split rule is the corresponding rule for DPLL's 'guess' or 'choice' rule.

The Close rule detects the case when the current context in the derivation permanently falsifies a clause. This case implies that there is no possible repair of the context. After the application of Close rule, the calculus tries to backtrack to the recent application of Split rule and continues the derivation with the right side of that application of Split. If there are no Split applications to backtrack, then all the branches are closed and the input theory is unsatisfiable.

Consider the clause $C \vee L$ where C is a (possibly empty) sub-clause and L is a literal. When the current context permanently falsifies the sub-clause C with a context unifier σ, $L\sigma$ must be valid in order to satisfy the clause. The Assert rule detects this situation and adds the literal $L\sigma$ to the context. Its precondition checks that $L\sigma$ is parameter-free (i.e., universal). This makes the application of Assert nonretractable. In the special case when C is an empty clause, the Assert rule adds the unit clause L to the context[8].

The ME calculus have other rules for a fair lifting of DPLL procedure to the first-order level. The Subsume and Resolve rules simplify the input clause set. The Compact rule simplifies the context.

---

[6] This condition is necessary but not sufficient for the unsatisfiability of C. See [13] for the necessary and sufficient condition.

[7] for all $i = 1, \cdots, m, (Params(K_i))\sigma \subseteq P$.

[8] A context permanently falsifies an empty clause with the context unifier which is an empty substitution.

Although Darwin's proof procedure is refutationally complete, it is not guaranteed to find a finite model, if one exists [14]. However, Darwin is a decider for Bernays-Schönfinkel formulas [13]. The clausal form of these formulas has no functions.

2.2 Paradox

If there is an interpretation which satisfies a first-order theory, one can easily construct another interpretation from a new domain and a bijection from the domain of the first interpretation to the new one. The constructed interpretation is also a model of the theory [15]. This fact implies that the actual members of the domain is not important in forming a model. An arbitrary set with a sufficient number of elements can be used as a domain since only the size of the domain matters.

A Mace-style finite model builder [16] searches for a model with a finite domain by enumerating the finite domains with sizes in increasing order like $\{0\}, \{0,1\}, \{0,1,2\}, \cdots$ and so on. The builder checks for each finite domain if there exists a model. The domain sizes increase in order until a model is found. The actual search for a model is achieved by instantiating the input clauses for each domain and running a SAT solver on the propositional theory.

The builder handles equality while instantiating the clauses. The equality predicate has a fixed interpretation in finite model building. The model should assign the literal $d_1 = d_2$ true only when it assigns the terms d_1 and d_2 the same domain element. For the other cases it should be false. The builder can eliminate or simplify a clause instance having equality literals.

Paradox is a Mace-style finite model builder [15]. It has many optimizations that enhance the original Mace-style finite model building method. In order to reduce the number of instances of a clause, clause splitting technique is used. It reduces the number of variables in a clause. The SAT solver used has the incremental solving capability. It does not start from scratch for every domain size. The conflict clauses that have been learned during an unsuccessful search for a model in previous iterations can be reused for the next domain sizes.

Many symmetric models in finite model building makes the job of SAT solver harder since it considers them as different models. A model assigns each constant a domain element. Any permutation of this assignment will generate symmetric models. Function interpretations can also cause symmetric models in the same way. Paradox adds some extra clauses to the propositional theory in order to reduce symmetries.

Paradox also uses sort inference as another optimization. Multiple sorts form a typed first-order logic. The quantification of the variables in a clause should be over its assigned sort. There may be less instances of the clauses respecting sorts compared to the unsorted case. This will reduce the instantiation time and ease model finding. Same domain elements can be used for interpreting functions and predicates that are of different sorts. This will decrease the size of the model. Paradox tries to find sorts of the problem by analyzing its first-order logic encoding. It uses the fact that same variables in a clause should be of the

same sort. Also, the arguments of the equality symbol should be of the same sort.

Paradox is complete for finding models with finite domains, but not refutationally complete in general. However, it is complete for first-order input in Effectively Propositional (EPR) class [15]. EPR class problems have no functions of arity greater than 0. The number of the elements in the largest sort inferred for an EPR problem is the limit for its satisfiability. There can be no models with a domain of size greater than this limit [15]. If no model can be found of size up to that number, Paradox decides that the input is unsatisfiable.

2.3 FM-Darwin

FM-Darwin[9] is also a Mace-style finite model builder [17]. Paradox transforms the problem to a propositional satisfiability problem, however FM-Darwin transforms it to a satisfiability problem of function-free clause logic without equality. The ME calculus and Darwin can decide that class of first-order logic [13]. FM-Darwin uses Darwin as an engine to solve the transformed satisfiability problem. Note that FM-Darwin implements the Mace-style finite model building without grounding, since Darwin can work on the first-order level. FM-Darwin has optimizations similar to those in Paradox for reducing symmetries.

3 Lifting SAT-based ASP to First-Order Level

The theory behind SAT-based answer set solvers is the relationship between the completion semantics and the answer set semantics. When they coincide the models of the completed program are equal to the answer sets of the program. A condition for the equivalence of these semantics have been set first by Fages' theorem [8]. This syntactic condition, which is called tightness in [18], is extended to tightness on a set of literals [9] and to programs with nested expressions [10]. For non-tight programs since the two semantics do not coincide, there can be some models of the completion that are not answer sets.

A logic program with variables can be completed. We can run a model generation theorem prover on the completed first-order theory to find the models of the program's completion. In this way, for tight programs the flow in SAT-based ASP can be lifted to the first-order level.

Let the logic program $\mathcal{P} = \{q(2)., r(a,1)., r(b,2)., p(X) \leftarrow r(X,Y), not\ q(Y).\}$. $\mathcal{P}$ is tight and has only one answer set which is $\{q(2), r(a,1), r(b,2), p(a)\}$. After completing $\mathcal{P}$, the first-order theory $\mathcal{C}$ is found as:

$$\begin{aligned}
&\forall x \quad (q(x) \Leftrightarrow (x = 2)),\\
&\forall xy \quad (r(x,y) \Leftrightarrow ((x = a \wedge y = 1) \vee (x = b \wedge y = 2))),\\
&\forall x \quad (p(x) \Leftrightarrow \exists y\ (r(x,y) \wedge \neg q(y)))\ .
\end{aligned}$$

[9] FM-Darwin comes embedded in Darwin. It can be used by running Darwin with some command line arguments.

A prover should be capable of handling equality in order to reason on the theory $\mathcal{C}$. Although equality reasoning can be embedded to a prover by adding the necessary axioms of equality predicate, it is efficient to have dedicated equality inference rules. Unfortunately, Darwin does not have efficient equality reasoning yet[10]. However, Darwin has a command line option for enabling axiomatic equality reasoning. The equality reasoning is relatively easy for finite model builders since the equality predicate has a fixed interpretation (see Section 2.2).

The first-order formulas in $\mathcal{C}$ can easily be input to Darwin, Paradox, and FM-Darwin in TPTP format[11]. However, an input with arbitrary formulas should be clausified first. While Darwin and FM-Darwin uses eprover[12] as a clausifier, Paradox can do the clausification itself.

When we run the finite model builders Paradox and FM-Darwin on $\mathcal{C}$, they generate a model but it is counterintuitive. The model found with a finite domain of size 1 assigns every constant to the same and only domain element. This is not intended in ASP since unique names are assumed. Let the unique names axioms $\mathcal{U} = \{a \neq b,\ a \neq 1,\ \cdots\}$. The model of $\mathcal{C} \cup \mathcal{U}$ found by Paradox (and also FM-Darwin) has a finite domain $\{1',\ 2',\ 3',\ 4'\}$. This model coincides with the unique answer set of the program $\mathcal{P}$ and defined as

$$
\begin{array}{l}
1 = 1' \\
2 = 2' \\
a = 3' \\
b = 4'
\end{array}
\qquad
\begin{array}{c|cccc}
p & 1' & 2' & 3' & 4' \\
\hline
 & F & F & T & F
\end{array}
\qquad
\begin{array}{c|cccc}
q & 1' & 2' & 3' & 4' \\
\hline
 & F & T & F & F
\end{array}
\qquad
\begin{array}{c|cccc}
r & 1' & 2' & 3' & 4' \\
\hline
1' & F & F & F & F \\
2' & F & F & F & F \\
3' & T & F & F & F \\
4' & F & T & F & F
\end{array}
\quad .
$$

Darwin does not halt for the theory $\mathcal{C} \cup \mathcal{U}$ because of the Skolem function introduced during clausification and the equality reasoning. The existential variable in the last formula of $\mathcal{C}$ cause the introduction of a Skolem function f during clausification. The substitution axiom $f(x) = f(y) \leftarrow x = y$ is added by Darwin for equality reasoning. When an equality like $f(a) = 1$ is in the Herbrand model of the theory, the substitution axiom endlessly generates new equations that make the Herbrand model infinite. For the theory $\mathcal{C} \cup \mathcal{U}$, Darwin generates the equations $\{f(a) = 1,\ f(f(a)) = f(1),\ f(f(f(a))) = f(f(1)),\ \cdots\}$ endlessly and can not output a model.

An n-ary Skolem function can be eliminated by introducing an $n + 1$-ary predicate. The unary Skolem function f introduced for the existential variable in the last formula of $\mathcal{C}$ can be eliminated by a binary predicate F. Here are the implicative clauses $\mathcal{D}$ generated after clausifying the theory $\mathcal{C}$ and eliminating

[10] Although the ME calculus with equality has been defined in [19], it has not been implemented in Darwin yet.

[11] http://www.cs.miami.edu/~tptp/TPTP/SyntaxBNF.html

[12] http://www.eprover.org/

the Skolem functions.

$$
\begin{array}{ll}
q(2). & r(a,1). \\
x = 2 \leftarrow q(x). & r(b,2). \\
 & r_1(x,y) \vee r_2(x,y) \leftarrow r(x,y). \\
 & x = a \leftarrow r_1(x,y). \\
p(x) \leftarrow r(x,y),\ \neg q(y). & y = 1 \leftarrow r_1(x,y). \\
r(x,y) \leftarrow p(x),\ F(x,y). & x = b \leftarrow r_2(x,y). \\
\neg q(y) \leftarrow p(x),\ F(x,y). & y = 2 \leftarrow r_2(x,y).
\end{array}
$$

In order to define the new predicate F, we need to add the following axioms $\mathcal{T}$ which state its totality over the set of all constants appearing in the theory and its uniqueness in the last argument.

$$
\begin{array}{l}
F(x,a) \vee F(x,b) \vee F(x,1) \vee F(x,2). \\
y = y' \leftarrow F(x,y),\ F(x,y').
\end{array}
$$

The clausification used in $\mathcal{D}$ is similar to the naive clausification process which converts arbitrary formulas to formulas in clausal normal form. However the naive clausification can generate exponential number of clauses when a predicate in the logic program has many generating rules having many literals in the premises. The new literals r_1 and r_2 are introduced during the clausification for keeping the size of the clauses polynomial. Similarly, Cmodels also introduces new propositional symbols for the rule bodies for polynomial-spaced clausification [20].

Darwin can find a model for the theory $\mathcal{D} \cup \mathcal{U} \cup \mathcal{T}$ which corresponds to the unique answer set of the program $\mathcal{P}$. The found model represented as disjunction of implicit generalizations (DIG) is

$$
\begin{array}{ll}
x = x & q(2) \\
F(a,1) & r(a,1) \\
F(x,2) & r(b,2) \quad . \\
\quad \text{with exception: } F(a,2) & r_1(a,1) \\
p(a) & r_2(b,2)
\end{array}
$$

Let the logic program $\mathcal{N} = \{p(a).,\ k \leftarrow not\ p(X).,\ \leftarrow not\ k.\}$. Although $\mathcal{N}$ has no answer sets, the non-clausal theory formed after completion $\mathcal{NC} = \{\forall x\,(p(x) \Leftrightarrow (x = a)),\ k \Leftrightarrow \exists y\,(\neg p(y)),\ false \Leftarrow \neg k\}$ has a non-Herbrand model. $\{k,\ p(1'),\ \neg p(2')\}$ is a model of $\mathcal{NC}$ with a domain $\{1',\ 2'\}$. It is reasonable to expect that the model generation theorem provers find the unsatisfiability of the theory $\mathcal{NC}$, however they can output the mentioned non-Herbrand model. The domain of the Skolem function introduced for the existentially quantified variable in $\mathcal{NC}$ is not restricted to $\{a\}$, which is the set of all constants in the program $\mathcal{N}$. In order to force the provers to search for Herbrand models, we should add the domain closure axiom to the theory. For $\mathcal{N}$, the domain closure axiom $\mathcal{DC} = \{\forall x\ \ (x = a)\}$. The theory $\mathcal{NC} \cup \mathcal{DC}$ has no models.

There is no need to add a domain closure axiom to a theory formed by eliminating Skolem functions as explained in this section. The totality axioms

added for the introduced predicates actually restrict the domains of the Skolem functions to the set of all constants appear in the program. For example, there is no need to add a domain closure axiom to the theory $\mathcal{D} \cup \mathcal{U} \cup \mathcal{T}$.

The theories formed by completing the *domain-restricted programs*[13], whether they are in non-clausal or clausal form, satisfy the domain closure axiom implicitly. All of the rules of a domain-restricted program are domain-restricted (i.e., every variable appearing in a rule also appears in a positive domain literal in the body). Lparse[14] uses domain predicates to restrict the domains of variables and requires domain-restricted programs. Since the domain of every variable in a domain-restricted program is restricted to a subset of all constants appearing in the program, the need for adding the domain closure axiom to the theory formed by completing such a program is eliminated. Note that the second rule of the program $\mathcal{N}$ is not domain-restricted. All the logic programs used in Section 4 are domain-restricted.

4 Experiments

We have carried out the following experiments on the blocks world planning problem and the quasigroup existence problem. All the tests in this section have been performed on an 1.8Ghz Core2Duo machine with 2GB of RAM running Linux 2.6.20.

4.1 Blocks World Problem

The logic program used here is due to Ilkka Niemelä. It has been shown in [9] that the models of its completion correspond to its answer sets.

The encoding of the blocks world problem is suitable for sort inference optimization. Both FM-Darwin and Paradox have sort inference optimization (see Section 2.2). Time values and block names (including table) constitute important concepts of the problem. These concepts can be considered as different sorts. The constants in different sorts can share the same domain elements in the finite model. This will help searching for a model since the size of its finite domain is less than the size of the unsorted one.

In order to take advantage of sort optimization we need to separate the unique names axioms for each sort (i.e., the axioms $\{a \neq table, a \neq b, \cdots\} \cup \{0 \neq 1, 0 \neq 2, \cdots\}$ where inequalities like $a \neq 0$ are eliminated). Otherwise, finite model builders will infer only one sort. The totality axioms of the predicates that are introduced to eliminate the Skolem functions should also respect the sort information. Theses axioms will not be over all the constants appearing in the theory, but over the elements of the sort of the existential variable for which the Skolem function is introduced.

The experiments use 4 blocks world problem instances: an instance encoding of Susman anomaly case (bw1) with 3 blocks, instances that correspond to the

[13] See Section 4.5 of Lparse manual http://www.tcs.hut.fi/Software/smodels/lparse.ps
[14] http://www.tcs.hut.fi/Software/smodels/

problems PLA026+1 (bw2) and PLA025+1 (bw3) from TPTP library with 5 and 9 blocks respectively, and the large.c instance (bw4) from [2] with 15 blocks. They have plans of 3, 3, 4 and 8 steps respectively.

Table 1. The run times of Smodels, Cmodels and Clasp on Blocks World problems. The entries are CPU times in seconds, '–' means that the solver could not output a result within a time limit of 600 seconds. The 'Grounding' column gives the CPU times of grounding the problem instances by Lparse.

Problem	Grounding	Smodels	Cmodels	Clasp
bw1	0.02	0.01	0.01	0.01
bw2	0.02	0.02	0.02	0.01
bw3	0.1	0.1	0.1	0.04
bw4	0.4	1.5	0.5	0.3

Table 1 shows the run times of Smodels (version 2.32), Cmodels (version 3.70) and Clasp (version 1.01) on the blocks world problems. The time values are CPU times in seconds that correspond to the sum of user and system time values given by the shell's time command[15]. All the solvers have solved the problems instantly even with the addition of grounding times of Lparse (version 1.0.17).

Table 2. The run times of FM-Darwin, Paradox and Darwin on Blocks World problems. The entries are CPU times in seconds, '–' means that the prover did not halt within a time limit of 600 seconds. The 'non-clausal' column is for non-clausal first-order theory formed by completing the logic program. Both 'clausal' and 'no Skolem' columns are for the completed and clausified theories, but the Skolem functions introduced during clausification are eliminated in the latter one.

	FM-Darwin			Paradox			Darwin
Prb.	non-clausal	clausal	no Skolem	non-clausal	clausal	no Skolem	no Skolem
bw1	3.1	0.3	0.4	0.1	0.2	0.2	0.6
bw2	22.4	7.7	7.7	0.4	0.5	0.4	7.4
bw3	–	–	–	8.1	9.8	2.9	–
bw4	–	–	–	–	–	109.5	–

Table 2 shows the run times of FM-Darwin (version 1.4), Paradox (version 2.1) and Darwin (version 1.4) on these problems. We have completed the logic program with variables manually. The results of the corresponding non-clausal first-order theories are shown in the *non-clausal* columns of Table 2. The *clausal*

[15] Since the evaluation machine has a multi-core CPU, we used the sum of user and system time values instead of real time value. The real time may be less than the user time since the programs of shell pipes in the test scripts and other programs like eprover called within a prover may run on different cores in parallel.

columns are representing the results of the inputs that are formed by clausification of the completed theory. They have Skolem functions introduced during the clausification. We have also eliminated these Skolem functions (see Section 3 for the elimination method) and obtained another input. The *no Skolem* columns are for results of this type of input. The run times of Darwin are only for the *no Skolem* column. For the others, because of the Skolem functions and the equality reasoning, Darwin may not halt and try to generate an infinite model.

If we take the program $\mathcal{P}$ in Section 3 as an example, here are the theories which the columns in Table 2 represents. The non-clausal column represents the theory $\mathcal{C} \cup \mathcal{U}$. Here is the theory $\mathcal{S}$ which is formed after clausifying $\mathcal{C}$.

$$
\begin{array}{ll}
q(2). & r(a,1). \\
x = 2 \leftarrow q(x). & r(b,2). \\
 & r_1(x,y) \vee r_2(x,y) \leftarrow r(x,y). \\
 & x = a \leftarrow r_1(x,y). \\
p(x) \leftarrow r(x,y),\ \neg q(y). & y = 1 \leftarrow r_1(x,y). \\
r(x, f(x)) \leftarrow p(x). & x = b \leftarrow r_2(x,y). \\
\neg q(f(x)) \leftarrow p(x). & y = 2 \leftarrow r_2(x,y).
\end{array}
$$

The theory $\mathcal{S}$ is same as $\mathcal{D}$ except that the Skolem function f is not eliminated. The clausal column represents the theory $\mathcal{S} \cup \mathcal{U}$. The no Skolem column represents the theory $\mathcal{D} \cup \mathcal{U} \cup \mathcal{T}$ for the program $\mathcal{P}$.

Darwin (FM-Darwin as well) uses eprover as a clausifier for the non-clausal inputs. The results in the non-clausal and clausal columns of FM-Darwin in Table 2 show that the clausification used in our work, explained in Section 3, leads to better results compared to the results obtained from eprover. The corresponding analysis for Paradox show that the clausification done by Paradox leads to slightly better performance compared to the one used in our work. However, all the results in Table 2 show that all the model generation theorem provers tested do not scale up well for the blocks world problems whether it is non-clausal, clausal with Skolem functions, or clausal without Skolem functions.

Both Darwin and FM-Darwin have displayed weak performance in the experiments. [14] states that Darwin is weak for problems with equality. The axiomatic equality reasoning can be the bottleneck for Darwin in the results shown in Table 2. FM-Darwin transforms the problem to a satisfiability problem over function-free clause logic without equality. It uses Darwin as an engine to build a finite model. Since the transformation eliminates equality, only the inefficient equality reasoning can not explain the weak results. The transformation used in FM-Darwin introduces new predicate symbols and add new clauses for the totality constraints needed over the finite domain. These increase the size of the theory and the search space which, in turn, hardens the job of Darwin.

4.2 Quasigroup Existence Problem

We have also done experiments on the quasigroup existence (QG) problem in algebra. We have selected the QG5 variation. QG problems impose constraints

on a relation whose multiplication table is a order n Latin square The QG5 problem's constraints on the relation $\circ$ are $a \circ a = a$ and $a \circ b \circ a \circ a = b$ for all a and b in a order n Latin square defined for $\circ$. The QG problem is selected since the rules in the logic program representation has many instances [21]. The grounding step of ASP can generate a huge propositional program which makes the job of answer set solvers hard. Since the model generation theorem provers accept first-order theories, they may show better performance relatively.

Here is the tight[16] logic program in Lparse format that is used to encode the QG5 problem.

```
range(1..n).

val(X,Y,Z) :- not nval(X,Y,Z), range(X;Y;Z).
:- val(X,Y,Z), val(X,Y,Z1), Z != Z1, range(X;Y;Z;Z1).

hasval(X,Y) :- val(X,Y,Z), range(X;Y;Z).
:- not hasval(X,Y), range(X;Y).

nval(X,Y1,Z) :- val(X,Y,Z), Y != Y1, range(X;Y;Z;Y1).
nval(X1,Y,Z) :- val(X,Y,Z), X != X1, range(X;Y;Z;X1).

:- val(X,X,X1), X != X1, range(X;X1).
:- val(X,Y,Z), val(Z,X,T), val(T,X,Y1), Y != Y1, range(X;Y;Z;T;Y1).
```

Table 3. The run times of Smodels, Cmodels and Clasp on QG5 problem. The entries are CPU times in seconds, '–' means that the solver could not output a result within a time limit of 600 seconds. The structure of the table is analog to Table 1.

Order	Grounding	Smodels	Cmodels	Clasp
5	0.03	0.03	0.02	0.02
6	0.1	0.2	0.1	0.1
7	0.1	0.2	0.2	0.1
8	0.2	4.3	0.3	0.3
9	0.4	–	540.1	274.4
10	0.7	–	–	–
11	1.2	–	24.4	8.8

Similar to the blocks world problem we have 3 types of input for every instance of the QG5 problem: The first-order completion of the logic program, its clausal form with Skolem functions, and its clausal form after eliminating the Skolem functions. We have tested the QG5 problem instances of order 5 to 11. For orders 5,7,8 and 11 there are QGs satisfying the QG5 constraints. The instances of order 6,9 and 10 are unsatisfiable.

[16] The tightness condition [9] is satisfied by the existence of a function λ which assigns every range atom to 0, every val atom to 1 and every nval and hasval atoms to 2.

Table 3 shows the run times of Smodels, Cmodels and Clasp on the QG5 problem instances. Smodels could not solved the instances of order 9,10 and 11 within a time limit of 600 seconds. Clasp has better results than Cmodels.

Table 4. The run times of FM-Darwin, Paradox and Darwin on QG5 problem. The entries are CPU times in seconds, '–' means that the prover did not halt within a time limit of 600 seconds. The structure of the table is analog to Table 2.

	FM-Darwin			Paradox			Darwin
Ord.	non-clausal	clausal	no Skolem	non-clausal	clausal	no Skolem	no Skolem
5	0.3	0.2	0.3	0.1	0.1	0.1	0.4
6	–	–	4.6	–	–	0.1	10.3
7	93.6	78.7	37.6	0.1	0.2	0.3	98.6
8	–	–	–	0.2	0.4	0.5	–
9	–	–	–	–	–	567.4	–
10	–	–	–	–	–	–	–
11	–	–	–	3.0	6.7	6.5	–

The run time results of Paradox in Table 4 are similar to those of Cmodels and Clasp in Table 3 for the small and satisfiable instances of the QG5 problem. However, these results show that Paradox seems to perform better for the satisfiable instances as the order increase, especially with the non-clausal encoding of the problem. The rules of the logic program for this problem has many instances. The results in Table 3 show that the propositional instance of the logic program formed by the grounder becomes harder for the solver as the order increases (especially the results for Smodels). Note that Paradox also performs grounding before calling the SAT solver but it is done incrementally. It calls the SAT solver for every finite domain starting with the one of size 1 and increase incrementally. It is interesting to compare Cmodels and Paradox since Cmodels is a SAT-based answer set solver. The order of grounding and completion steps is interchanged in Paradox and Cmodels. The completed first-order theory is given to Paradox, but in the case of Cmodels the input is the grounded logic program. Paradox performs the grounding itself, while Cmodels performs the completion. The incremental nature of finite model building and the other optimizations of Paradox seem to be efficient for this problem.

The clausal theory formed after eliminating the Skolem functions has a special importance for the finite model builders. When the input is in this form, it is in the EPR problem class. A finite model builder can be refutationally complete for this class [15]. Paradox have found the unsatisfiability of the input for the problem instances of order 6 and 9 (i.e., no quasigroup exists for them). Similarly, FM-Darwin found the unsatisfiability result for the instance of order 6 as depicted in the *no Skolem* columns of the Table 4. Note that in the columns *non-clausal* and *clausal*, where the input is not in the EPR class, both FM-Darwin and Paradox could not halt for the unsatisfiable instances (orders 6,9 and 10).

The results depicted in Tables 4 show that both FM-Darwin and Darwin have weak performances for this problem. The comments that have been stated in the case of blocks world problem also apply to this problem.

5 Conclusions and Related Work

We proposed and experimented a way of lifting SAT-based ASP to the first-order level for tight programs. The completion step in the lifted case produces a first-order theory. The place of the SAT solver should be filled by a system which can process first-order input and output a model when it is satisfiable. The model generation theorem provers can be suitable to play this role. One important outcome of this lifting is that grounding can be either eliminated or done intelligently in the model generation prover. Model generation theorem provers are gaining increasing popularity in the automated reasoning community. We have used Darwin, FM-Darwin and Paradox in the experiments.

The experiments with Paradox showed good results for the quasigroup existence problem. Darwin and FM-Darwin displayed weak performance for our benchmark problems. Axiomatic equality reasoning, which is used by Darwin, is inefficient. The equality literals generated by completion and the unique names axioms increase the importance of equality reasoning.

The ME calculus is a lifting of the DPLL procedure to the first-order level. Note that the search procedures of successful answer set solvers, like smodels and dlv, are based on the DPLL procedure on the propositional level. We think that the ME calculus and its implementation Darwin are interesting for the future answer set solvers working on the first-order level.

Paradox and FM-Darwin are refutationally complete for theories in the EPR class. The clausal theories formed in our work by eliminating the Skolem functions are in the EPR class. The relation between the EPR formulas and the logic programs under the answer set semantics has been examined in [22] for an application area of textual inference. It has a translation from any EPR formula to a logic program such that the formed logic program has an answer set iff the EPR formula is satisfiable.

Recently, a definition of the stable models of an arbitrary first-order formula has been given in [23]. That work also extends the tightness condition to first-order formulas. Using that definition, it can be decided whether a logic program is tight or not without referring to the original tightness condition. The tightness condition based on Fages' theorem refers to grounding, but the new one does not.

One limitation of the provers used is that they can only output one model for an input. The modification of Darwin needed for finding all models should be easy. After a model is found, it can backtrack to the last application of the Split rule and continue with the right side of that rule. The symmetric models that correspond to the same answer set should be eliminated when we are searching for all models using a finite model builder.

We have concentrated on Mace-style finite model building. One future work can be examining the Sem-style finite model builders like Sem[17] and Finder[18]. A Sem-style finite model builder searches for a model by working on the input directly. There is no translation of the problem into a satisfiability problem [15]. During the lifting of SAT-based ASP, we formed the completed theory and added the necessary axioms manually. Automating this process is another subject of future work.

Acknowledgments

This work is supported by the Scientific and Technical Research Council of Turkey, Project No: EEEAG 105E068. We are grateful to Esra Erdem, Vladimir Lifschitz, Ilkka Niemelä, Tomi Janhunen and anonymous referees for useful comments and discussions related to the subject of this paper.

References

1. Lifschitz, V.: Answer set planning. In Proceedings of the 16th International Conference on Logic Programming (1999) 23–37
2. Niemelä, I.: Logic programs with stable model semantics as a constraint programming paradigm. Annals of Mathematics and Artificial Intelligence **25** (1999) 241–273
3. Marek, V., Truszczynski, M.: Stable models and an alternative logic programming paradigm. In the Logic Programming Paradigm: a 25-Year Perspective (1999) 375–398
4. Anger, C., Konczak, K., Linke, T., Schaub, T.: A glimpse of answer set programming. Künstliche Intelligenz **19** (2005) 12–17
5. Baumgartner, P., Furbach, U., Yahya, A.: Automated reasoning, knowledge representation and management. Künstliche Intelligenz **19** (2005) 5–11
6. Clark, K.: Negation as failure. In Logic and Data Bases (1978) 293–322
7. Gelfond, M., Lifschitz, V.: Classical negation in logic programs and disjunctive databases. New Generation Computing **9** (1991) 365–386
8. Fages, F.: Consistency of Clark's completion and existence of stable models. Journal of Methods of Logic in Computer Science **1** (1994) 51–60
9. Babovich, Y., Erdem, E., Lifschitz, V.: Fages' theorem and answer set programming. In Proceedings of 8th International Workshop on Non-Monotonic Reasoning (2000)
10. Erdem, E., Lifschitz, V.: Tight logic programs. Theory and Practice of Logic Programming **3** (2003) 499–518
11. Lin, F., Zhao, Y.: ASSAT: Computing answer sets of a logic program by SAT solvers. Artificial Intelligence **157** (2004) 115–137
12. Lierler, Y., Maratea, M.: Cmodels-2: SAT-based answer set solver enhanced to non-tight programs. In Proceedings of the 7th International Conference on Logic Programming and Non-Monotonic Reasoning (2004) 346–350

[17] http://www.cs.uiowa.edu/~hzhang/sem.html
[18] http://users.rsise.anu.edu.au/~jks/finder.html

13. Baumgartner, P., Tinelli, C.: The Model Evolution Calculus. Fachberichte Informatik 1–2003, Universität Koblenz-Landau, Institut für Informatik (2003)
14. Baumgartner, P., Fuchs, A., Tinelli, C.: Implementing the Model Evolution Calculus. International Journal of Artificial Intelligence Tools **15** (2006) 21–52
15. Claessen, K., Sörensson, N.: New Techniques that Improve MACE-style Finite Model Finding. In Proceedings of Workshop: Model Computation – Principles, Algorithms, Applications at the 19th International Conference on Automated Deduction (2003)
16. McCune, W.: A Davis-Putnam Program and Its Application to Finite First-Order Model Search: Quasigroup Existence Problems. Technical Report ANL/MCS-TM-194, Argonne National Laboratory (1994)
17. Baumgartner, P., Fuchs, A., de Nivelle, H., Tinelli, C.: Computing Finite Models by Reduction to Function-Free Clause Logic. In Proceedings of 3rd Workshop on Disproving - Non-Theorems, Non-Validity, Non-Provability at the 3rd International Joint Conference on Automated Reasoning (2006)
18. Lifschitz, V.: Foundations of logic programming. Principles of Knowledge Representation (1996) 69–127
19. Baumgartner, P., Tinelli, C.: The Model Evolution Calculus with Equality. In Proceedings of the 20th International Conference on Automated Deduction (2005)
20. Babovich, Y., Lifschitz, V.: Computing answer sets using program completion. Unpublished draft (2003)
21. Shirai, Y., Hasegawa, R.: Answer Set Computation Based on a Minimal Model Generation Theorem Prover. In Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence (2004) 43–52
22. Lierler, Y., Lifschitz, V.: Logic Programs vs. First-Order Formulas in Textual Inference. Unpublished draft
23. Ferraris, P., Lee, J., Lifschitz, V.: A New Perspective on Stable Models. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (2007) 372–379

Key Issues of Technical Interoperability Solutions in eHealth and the RIDE Project*

Asuman DOGAC[1], Tuncay NAMLI[1], Alper OKCAN[1], Gokce B. LALECI[1], Yildiray KABAK[1], Marco EICHELBERG[2]

[1]*Software R&D Center, Dept. of Computer Eng., Middle East Technical University, Ankara, 06531, Turkey*

Tel: +90 312 2105598, Fax: +90 312 2105572, Email: asuman@srdc.metu.edu.tr

[2] *OFFIS - Oldenburger F&E-Institut für Informatik-Werkzeuge und -Systeme Escherweg 2, D-26121 Oldenburg, Germany*

Tel: +49 441 9722147, Fax: + 49 441 9722102, Email: eichelberg@offis.de

Abstract: One of the key problems in healthcare informatics is the lack of interoperability among different healthcare information systems. Interoperability can be investigated in different categories in the eHealth domain, such as the interoperability of the messages exchanged between healthcare applications, interoperability of Electronic Healthcare Records (EHRs), interoperability of patient identifiers, coding terms, clinical guidelines and healthcare business processes. Furthermore, all these categories can be investigated in two major layers: syntactic interoperability layer and the semantic interoperability layer. This paper describes the concepts involved in eHealth interoperability; briefly assesses the current state in some of the countries in the world and discusses the technical issues to be addressed for achieving interoperability.

1. Introduction

Interoperability is the ability of different information technology systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged. Making healthcare information systems interoperable will reduce cost of health care and will contribute to more effective and efficient patient care.

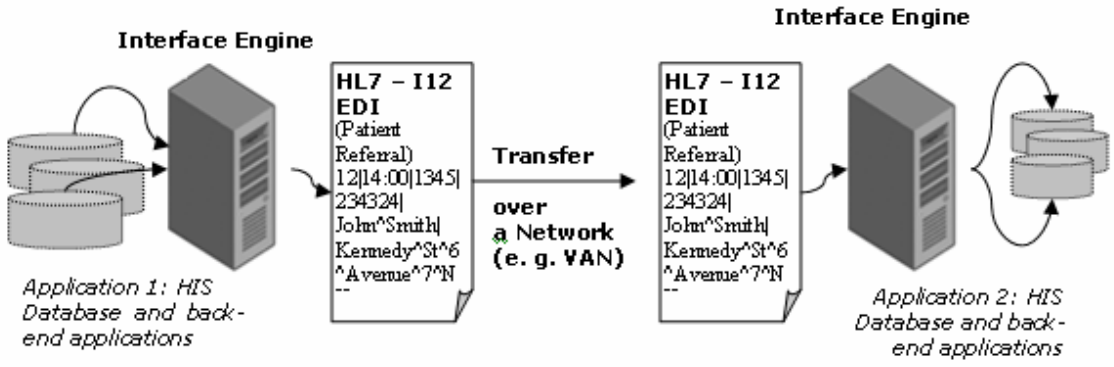


Figure 1. Message exchange between heterogeneous applications

* This work is supported by the European Commission through IST- 027065 RIDE project, and in part by the Scientific and Technical Research Council of Turkey, Project No: EEEAG 105E068

The healthcare interoperability problem can be investigated in the following categories.

1.1 – Interoperability of the Healthcare Messages Exchanged

To be able to exchange information among heterogeneous healthcare information systems, messaging interfaces (also called interface engines) are used. Typically, a messaging interface gathers data from the back-end application systems, encodes the data into a message, and transmits the data over a network such as a Value Added Network (VAN) to another application. On the receiver side, the received messages are decoded, processed and the data which have been received are fed into the receiver's back-end systems to be stored and processed as shown in Figure 1.

When proprietary formats are used in messaging, the number of the interfaces to be developed increases drastically. For example, if there are four applications that need to exchange messages, each of them needs to develop three interfaces. In fact, the total number of interfaces to be developed is $n*(n-1)/2$, i. e. $O(n^2)$, for n applications. To overcome this problem, message standards are preferred since an application can, in principle, talk to any other application conforming to the same message standard by using the same message interface.

Currently, the Health Level 7 (HL7) Version 2 Messaging Standard [HL7, HL7v2.5] is the most widely implemented message interface standard in the healthcare domain. However, being HL7 Version 2 compliant does not imply direct interoperability between healthcare systems. This stems from the fact that Version 2 messages have no explicit information model, rather vague definitions for many data fields and contain many optional fields. This optionality provides great flexibility, but necessitates detailed bilateral agreements among the healthcare systems to achieve interoperability. To remedy this problem, HL7 Version 3 [HL7V3] is developed, which is based on an object-oriented data model, called Reference Information Model (RIM) [HL7RIM]. It should be noted that there is an interoperability problem between HL7 v2.x and HL7 v3 messages – there is no well-defined mapping between HL7 v2.x and v3 messages.

1.2 – Interoperability of Electronic Healthcare Records (EHR)

The Electronic Healthcare Record (EHR) of a patient can be defined as digitally stored health care information about individual's lifetime with the purpose of supporting continuity of care, education and research, and ensuring confidentiality at all times [Iakovidis 1998]. A patient's healthcare information may be spread out over a number of different institutes which do not interoperate. In order to provide continuity of care, clinicians should be able to capture the complete clinical history of a patient. A number of standardization efforts are progressing to provide the interoperability of electronic healthcare records such as CEN/TC 251 EHRcom [EHRcom 2004], openEHR [OpenEHR] and HL7 Clinical Document Architecture (CDA) [HL7CDA]. However, an exchange of well-structured and machine processable electronic healthcare records has not been achieved yet in practice. Also, given the large number of standards for this purpose, conforming to a single standard does not solve the interoperability problem.

There are several other aspects of healthcare domain in need of interoperability such as patient identifiers, coding terms, clinical guidelines and healthcare business processes.

2. Interoperability Layers

The interoperability categories introduced in section 1 can be investigated in two major layers: syntactic interoperability layer and the semantic interoperability layer.

Syntactic interoperability (which we term as messaging layer), involves the ability of two or more systems to exchange information. Syntactic interoperability involves several layers: network and transport layer (such as Internet), application protocol layer (such as HTTP or email), messaging protocol and message format layer (such as ebXML messaging or SOAP), and the sequencing of the messages.

Syntactic interoperability guarantees the message to be delivered but does not guarantee that the content of the message will be machine processable at the receiving end. To guarantee message content interoperability, either the message content should conform to a single machine processable standard or semantic interoperability must be provided. Semantic interoperability is the ability for information shared by systems to be understood at the level of formally defined domain concepts.

2.1 – Layers of Syntactic Interoperability in eHealth

A prerequisite for interoperability is the ability to communicate: that is, the bits running on the wires. In transferring healthcare messages between application systems, network and transport protocols are needed, such as Internet. In fact, today, TCP/IP (Internet) is the de-facto on-line communication standard. On top of this, an application protocol standard is needed such as HTTP or SMTP (email). On top of this layer, standard messaging protocol layer is necessary such as SOAP [SOAP] or ebXML messaging [ebMS]. The sequencing of the messages also needs to be standardized. For example, in HL7, when "I05 RQC Request Clinical Information" message is sent, the expected return message is "I05 RCI Return Clinical Information". There are also different types of messages: each message is either a message with the intent of action or an acknowledgment message indicating the successful transmission of a message or an error message indicating an error situation. Finally, for the message content to be processed correctly by the receiving application, the message content structure and the data items in the message must be standardized, for example as proposed by HL7 Version 3.

As an example of different layers of interoperability in eHealth, consider the IHE Patient Identifier Cross-referencing (PIX) Profile [IHE TF]. In IHE PIX Profile, the network and transport protocol can be Internet; the messaging protocol is EDI and the content in Patient Identity feed message is defined through HL7 ADT messages A01, A04, A05, A08 or A40.

2.2 – Semantic Interoperability in eHealth

Semantics, which is the metadata describing data, is described through ontologies. An ontology can be defined as "a formal, explicit specification of a shared conceptualization" [Gruber 1993]. Formal means that the meaning specification is given in a machine processable language, called the ontology language. An explicit specification means that the concepts and the relationships in the abstract model are given explicit names and definitions. An important feature of ontology languages is that they provide for automated inference to derive new, implicit information from these explicit specifications. Shared means that an ontology describes consensual knowledge, that is, it describes meaning which has been accepted by a group, not by a single individual; in other words, it provides a common vocabulary for those who have agreed to use it. An ontology together with a set of concrete instances constitute a knowledge base. Currently, Web Ontology Language (OWL) [OWL] is a widely accepted ontology language.

A common usage of the term "semantic interoperability in eHealth" can be found in [CEN/ISSS]: "Semantic interoperability implies that the structure of the 'documents' is interpretable, and that their content is understandable. Making this content understandable

sometimes requires that the keys for its correct and safe interpretation, such as the terminological systems used, are identified and easily available."

An overview and assessment of the currently available state-of-the-art ontologies and ontology-like artifacts (controlled vocabularies) in healthcare are given in [Ceusters 2006]. For example, SNOMED CT which is a Description Logics supported, concept based ontology, contains over 366,000 healthcare concepts organized into hierarchies, with approximately 1.46 million semantic relationships between them, and more than 993,420 terms.

Another important use of semantic interoperability in the healthcare domain is the integration of data from heterogeneous sources through semantic mediation. Semantic mediation can be used to convert healthcare messages defined in one standard format into another as realized with the scope of the Artemis project [Artemis, Dogac 2006, Bicer 2005a]. Furthermore, an approach to archetype based semantic interoperability of EHR standards, as realized within the scope of the Artemis project, is described in [Bicer 2005b].

3. Interoperability through Electronic Healthcare Records

Considerable clinical information about a patient is passed around through the messages exchanged among healthcare applications. What differentiates an Electronic Healthcare Record (EHR) from the patient data contained in such messages is that, an EHR as defined in [Iakovidis 1998] is "digitally stored health care information about an individual's lifetime with the purpose of supporting continuity of care, education and research, and ensuring confidentiality at all times". In other words, EHR is the collection of relevant clinical data about an individual's lifetime usually in a document structure.

To address the EHR interoperability problem, there are several standards currently under development such as the Health Level 7 (HL7) Clinical Document Architecture (CDA) [HL7CDA], CEN EN 13606 EHRcom [EHRcom 2004] and openEHR [OpenEHR]. A detailed survey and analysis of EHR standards are given in [Eichelberg 2005].

These standards aim to structure and markup the clinical content for the purpose of exchange. There is also an industry initiative called Integrating the Healthcare Enterprise (IHE) [IHE] which specified the Cross-Enterprise Document Sharing (XDS) integration profile [IHE TF] for this purpose. The basic idea of IHE XDS is to store healthcare documents in an ebXML registry/repository [ebXMLRR] architecture to facilitate their sharing.

In the following subsections, we introduce related technologies identified to realize EHR interoperability.

3.1 – IHE Cross-Enterprise Document Sharing (XDS)

The basic idea of IHE XDS [IHE TF] is to store healthcare documents in an ebXML registry/repository [ebXMLRR] to facilitate their sharing. IHE XDS is not concerned with document content; it only specifies metadata to facilitate the discovery of documents.

In the IHE XDS integration profile, a group of healthcare enterprises that agree to work together for clinical document sharing is called the "Clinical Affinity Domain". Such institutes agree on a common set of policies such as how the patients are identified, the consent is obtained, the access is controlled, and the common set of coding terms to represent the metadata of the documents.

As already mentioned, IHE XDS handles healthcare documents in a content neutral way, that is, a document may include any type of information in any standard format such as simple text, formatted text (e.g., HL7 CDA Release One), images (e.g., DICOM [DICOM]) or structured and vocabulary coded clinical information (e.g., CDA Release Two, CEN ENV 13606 or DICOM SR). Given this, to ensure the interoperability between

the document sources and the document consumers, the clinical affinity domains also agree on the document format, the structure and the content.

3.2 IHE Cross-Enterprise Sharing of Medical Summaries (XDS-MS)

Cross-Enterprise Sharing of Medical Summaries (XDS-MS) is a mechanism to automate sharing of Medical Summaries between care providers. The main characteristics of XDS-MS are as follows:

- XDS-MS Profile uses the Actors and Transactions of IHE XDS; only the Document types used in XDS-MS are more specific Medical Summaries.
- Two types of Medical Summary content are currently specified: one for episodic care, the other for collaborative care.
- A third type of Medical Summary for permanent care is yet to be defined by IHE.
- XDS-MS specifies content of Medical summaries by building on and further constraining the HL7 Clinical Document Architecture (CDA) standard and Care Record Summary (CRS) CDA implementation guides.
- Document Sources provide an XML style sheet to render the content of the Medical Summary document.
- Medical summaries are shared within predefined domains (called XDS Affinity Domains) by storing the medical summaries in Registry/Repositories. Note however that IHE also plans the federated XDS Affinity domains; therefore the exchange of medical documents will not be restricted to XDS Affinity Domains in the near future.
- Registry/Repository architectures facilitate the discovery of the Medical Summaries in an XDS Affinity Domain.

3.3 IHE Retrieve Information for Display (RID)

Retrieve Information for Display (RID) [IHE TF] provides a simple and rapid read-only access to patient-centric clinical information that is located outside the user's current application. It supports access to existing persistent documents in well-known presentation formats such as CDA Level One, PDF and JPEG. It also provides access to specific key patient-centric information such as allergies, current medications, and summary of reports for presentation to a clinician.

IHE defined RID as a Web service by providing its WSDL (Web Service Description Language) [WSDL] description with a binding to HTTP GET.

3.4 Other Issues to be Addressed in EHR Interoperability

For EHR interoperability, the further technical issues that must also be addressed include:

- Mapping the patient identifiers among different healthcare applications: A key issue in accessing the EHR of a patient is his/her patient identifier. Yet different healthcare enterprises or even different departments in a healthcare institute may be using different identifiers for the same patient. Some of the possible mechanisms are as follows:
  - A central database containing all person identification numbers linked to demographic data
  - Smart card containing person identification numbers and demographic data
  - Master Patient Indexes mapping patient identifiers in different systems to each other.
- Authenticating the users across the enterprises: The users must be authenticated not only in their own domain but also across the enterprises.

- Guaranteeing that all the computers involved have consistent time: For distributed applications to work correctly it is essential that the system clocks and time stamps of the many computers in the network are well synchronized.
- Authenticating Nodes and Obtaining Audit Trail: Limiting access control to authorized users is not enough; it is necessary to limit network access between nodes and to limit access to each node in a healthcare setting. Put it differently, an entire host must be secured, not just individual users. Furthermore, audit trail is essential. It is necessary to allow a security officer in a healthcare institution to audit activities to detect improper creation, access, modification and deletion of Protected Health Information (PHI). The audit trail must contain information to answer the following questions:
  - For some user: which patients' PHI was accessed?
  - For some patient PHI: which users accessed it?
  - What user authentication failures were reported?
  - What node authentication failures were reported?

4. Implementation Scenario

Within the scope of the RIDE project, an interoperability scenario is implemented using the technologies described in section 3. The demonstration is implemented as one possible alternative to different routes to interoperability. Basically in this prototype, the integration of IHE XDS, ATNA, PIX and CT Profiles has been performed. The steps of the prototype (Figure 2) are presented in Table 1:

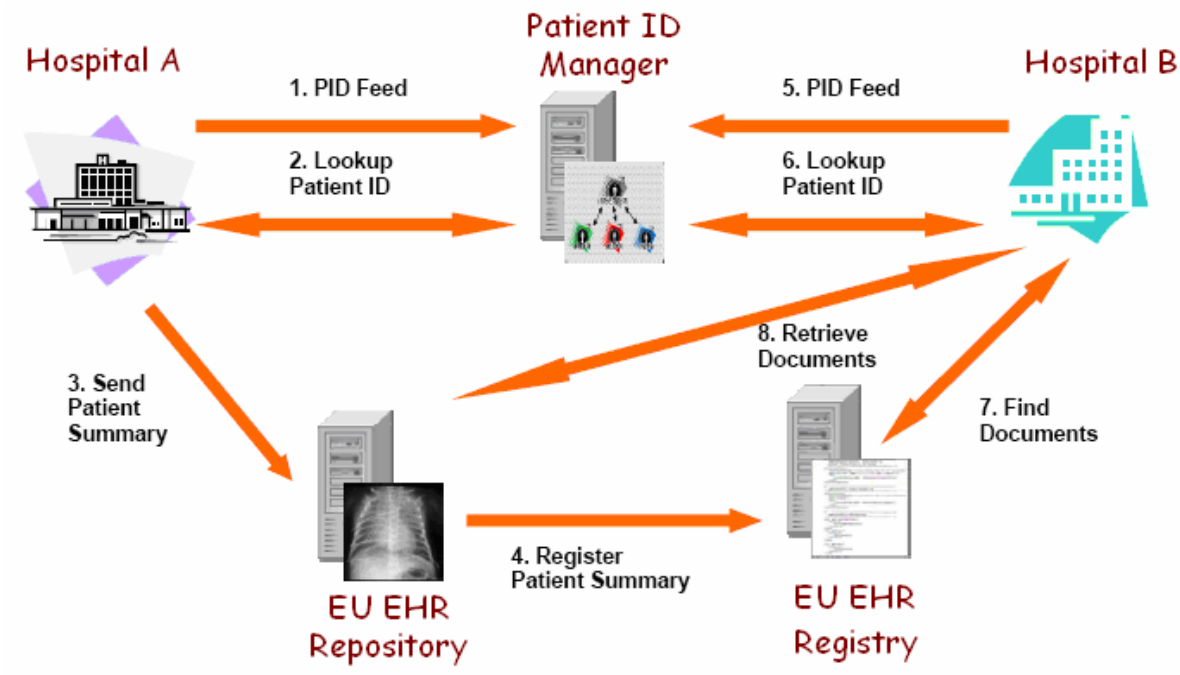


Figure 2 Interoperability Scenario

Table 1. Prototype Steps

Step	Scenario	Interoperability Mechanism
1	A patient with cardiovascular problems had some surgeries and treatment at Hospital A. When the patient is first registered to the hospital the patient identifier used is sent to Patient ID Manager along with the demographics	**IHE PIX**

	information.	
2	Hospital A creates a "Patient Summary" in CEN ENV 13606-1 EHRCom format and and wants to register the document to EU EHR Registry/ Repository. However the Patient ID's used in Hospital A and in the EU EHR Registry/ Repository are different. The Brusells hospital maps the internal Patient ID to the one used in the repository using the Patient ID Manager.	**IHE PIX**
3	Through the Patient Identifier received from the Patient ID Manager, Hospital A stores the "Patient Summary" to the EU EHR Repository	**IHE XDS**
4	The EU EHR Repository registers the "Patient Summary" document to the EU EHR Registry.	**IHE XDS**
5	One day, the same person expriences a heart attack. From the Ambulance, through a mobile device, the patient is admitted to Hospital B. Hospital B sends the Patient ID created to Patient ID Manager along with the demographics information.	**Web services, IHE PIX**
6	The doctor in Hospital B wishes to see the available Electronic Healthcare Records of the Patient from the EU EHR Registry/Repository. However the Patient ID's used in Hospital B and in the EU EHR Registry/ Repository are different. Hospital B maps the internal Patient ID to the one used in the repository using the Patient ID Manager.	**IHE PIX**
7	With the Patient Identifier received from the Patient ID Manager, the Hospital B queries the available EHRs of the patient from the EU EHR Registry.	**IHE XDS**
8	Using the document link provided by the EU EHR Registry, the doctor retrives the "Patient Summary" in CEN ENV 13606-1 EHRCom format from the EU EHR Repository.	**IHE XDS**
9	All transactions are secure and logged at the Audit Record Repository.	**IHE ATNA, IHE CT**

5. Conclusions

It is clear from the discussion that in order to resolve interoperability at the EU level, the issues need to be addressed include:

- Providing the interoperability of the various different messaging infrastructures being used
- Providing the interoperability of various EHR standards being used
- Providing the interoperability of various patient identification mechanisms
- Providing security, privacy and authentication in accessing clinical information.

RIDE (http://www.srdc.metu.edu.tr/webpage/projects/ride/) Project is addressing these issues to propose possible alternatives. It is a roadmap project for interoperability of eHealth systems leading to recommendations for actions and to preparatory actions at the European level. This roadmap will prepare the ground for future actions as envisioned in the action plan of the eHealth Communication COM 356 by coordinating various efforts on eHealth interoperability in member states and the associated states. Since it is not realistic to expect to have a single universally accepted clinical data model that will be adhered to all over the Europe and that the clinical practice, terminology systems and EHR systems are all

a long way from such a complete harmonization; the RIDE project will address the interoperability of eHealth systems with special emphasis on semantic interoperability.

References

[Artemis] The ARTEMIS Project. http://www.srdc.metu.edu.tr/webpage/projects/artemis

[Bicer 2005a] Bicer V., Laleci G., Dogac A., and Kabak Y., "Artemis Message Exchange Framework: Semantic Interoperability of Exchanged Messages in the Healthcare Domain", ACM Sigmod Record 34, 3, September 2005.

[Bicer 2005b]Bicer V., Kilic O., Dogac A., Laleci G., "Archetype-based Semantic Interoperability of Web Service Messages in the Healthcare Domain", Int'l Journal on Semantic Web & Information Systems, Vol. 1, No.4, October 2005, pp. 1-22.

[CEN/ISSS] Report from the CEN/ISSS eHealth Standardization Focus Group: Current and future standardization issues in the e-Health domain: Achieving interoperability, Final version, March 2005.

[Ceusters 2006] Ceusters, W., RIDE Deliverable 2.1.1 – Current Practices in Providing Semantic Interoperability in eHealth Domain: State-of-the-Art in Ontology with Special Emphasis on Biomedical Ontology.

[DICOM] Digital Imaging and Communications in Medicine. NEMA Standards Publication PS 3, National Electrical Manufacturers Association, Rosslyn VA, USA.

[Dogac 2006] Dogac A., Laleci G., Kirbas S., Kabak Y., Sinir S., Yildiz A., Gurcan, Y., "Artemis: Deploying Semantically Enriched Web Services in the Healthcare Domain", Information Systems Journal (Elsevier), Volume 31, Issues 4-5, June-July 2006, pp.321-339

[ebMS] ebXML Messaging Service, http://www.oasis-open.org/committees/documents.php?wg_abbrev=ebxml-msg

[ebXMLRR] Electronic Business XML (ebXML) Registry/Repository, http://www.oasis-open.org/committees/ tc_home.php?wg_abbrev=regrep

[EHRcom 2004] prEN 13606-1, "Health informatics - Electronic health record communication - Part 1: Reference model", Draft for CEN Enquiry, CEN/TC 251 Health Informatics, European Committee for Standardization, Brussels, Belgium.

[Eichelberg 2005] Eichelberg M., Aden T., Riesmeier J., Dogac A., Laleci G., "A Survey and Analysis of Electronic Healthcare Record Standards", ACM Computing Surveys, Vol. 37, No:4, December 2005.

[Gruber 1993] Gruber T., "A translation approach to portable ontology specifications". Knowledge Acquisition 5:199-220; 1993 [HL7] Health Level 7 (HL7), http://www.hl7.org/

[HL7CDA] CDA, HL7 Clinical Document Architecture - Release 2.0, http://xml.coverpages.org/CDA-20040830v3.pdf

[HL7RIM] HL7 Reference Information Model, http://www.hl7.org/library/data-model/RIM/modelpage_mem.htm

[HL7v2.5] HL7, "Application Protocol for Electronic Data Exchange in Healthcare Environments", Version 2.5, ANSI Standard, Ann Arbor MI, USA (2000).

[HL7V3] HL7 v3 Message Development Framework, http://www.hl7.org/library/mdf99/mdf99.pdf.

[Iakovidis 1998] Towards Personal Health Record: Current situation, obstacles and trends in implementation of Electronic Healthcare Records in Europe," Intern. J. of Medical Informatics, vol 52, No 123, pg. 105 –117 (1998).

[IHE] Integrating Healthcare Enterprise, http://www.ihe.net/

[IHE TF] IHE IT Infrastructure Integration Profiles, http://www.ihe.net/Technical_Framework/index.cfm#IT

[LOINC]. Logical Observation Identifiers Names and Codes (LOINC). http://www.loinc.org/.

[OpenEHR] openEHR Community, http://www.openehr.org/

[OpenEHR03] Beale, T. and Heard, S. 2003. The openEHR EHR Service Model, Revision 0.2. openEHR Reference Model, the openEHR foundation.

[OpenEHR05] Beale, T., Heard, S., Kalra, D., and Lloyd, D. 2005f. The openEHR Data Structures Information Model, Revision 1.6rc1 (Release 1.0 draft). openEHR Reference Model, the openEHR foundation.

[OWL] Web Ontology Language (OWL), http://www.w3.org/TR/owl-features/

[RIDE] A Roadmap for Interoperability of eHealth Systems in Support of COM 356 with Special Emphasis on Semantic Interoperability, http://www.srdc.metu.edu.tr/webpage/projects/ride

[SOAP] Simple Object Access Protocol (SOAP), http://www.w3.org/TR/soap/

[SNOMED CT] SNOMED (The Systematized Nomenclature of Medicine) Clinical Terms, http://www.snomed.org/snomedct txt.html

[WSDL]Web Service Description Language (WSDL), http://www.w3.org/TR/wsdl

Ontology on Semantic Web[1]

Semra DOGANDAG, Ferda Nur ALPASLAN[2]
[1] *ASELSAN Inc. MGEO Division P.O.Box: 30 Etlik 06011 Ankara, Turkey*
[2] *Computer Engineering Department, Middle East Technical University, 06531, Ankara, Turkey,*

Abstract: OWL (Web Ontology Language) is a markup ontology language based on Description Logics (DL). It is about to become standard ontology language on the web, therefore different communities try to export their ontologies to OWL. Although OWL is highly expressive, it is not always easy to map a formalism to another one. There are some important conceptual modeling problems that should be revisited by OWL. In this paper we have chosen MPEG-7 as an example meta-language. MPEG-7 is written using XML-Schema. We discussed the problems we came across while mapping MPEG-7 to OWL and gave a conceptual mapping for a part of MPEG-7. OWL standard is yet evolving and we claim that with the employment of new standards a complete mapping from MPEG-7 to OWL will be possible.

1. Introduction

Ontology is a very broad term. A catalog, a thesauri or a logic program can be defined as an ontology. The simplest ontology is a catalog where you have only a finite set of terms and the most expressive ontologies are the general logical programs where you can express disjointness, part-whole relationships, and so on.

Markup languages which can be regarded as ontology languages are SHOE, DAML+OIL, OIL, RDF, RDF Schema, OWL [6, 7]. Among these languages OWL is becoming the most popular one, because it encompasses some important properties an ontology language should have. It has a precise definition, it is based on description logics so you can define a complex structure and it has the reasoning capability which makes it suitable for knowledge representation tasks. OWL has emerged from the idea of Semantic Web [2] and it is the most powerful candidate as a common ontology language.

The promises of the Semantic Web encouraged different communities to put their ontology on the web. Although OWL is highly expressive, it is not always easy to map a formalism to another one. There are some important conceptual modeling problems that should be revisited by OWL. There are already some works done by database and multimedia communities to export the previously designed ontologies to OWL ontologies.

In this paper, we discuss a mapping from MPEG-7 to OWL using a movie database application as an example. Giving a mapping from MPEG-7 to OWL will enable the users to combine the applications that use MPEG-7 with those that use OWL. We believe that in near future there will be a lot of applications which use MPEG-7 to personalize the multimedia content usage. These applications will store the profiles of users and search for the multimedia content which fits best to the users' preferences. It is obvious that MPEG-7 will play an important role here, because it enables to tag a multimedia in very different levels. These applications should also be easy to integrate with web applications. To be able to manage huge multimedia database on the web, one needs to map MPEG-7 to OWL. This will also bring the opportunity of using the tools related with semantic web for the movie database. In this way, an application can collect the user preferences from the MPEG-7 data

of the users' movie repository and then create a corresponding OWL database which can be used for semantic web applications.

MPEG-7 uses XML-Schema. There are several translations of MPEG-7 to an ontology language [9, 3]. Our approach is similar to the one given in [9]. But, we propose a new mapping for the "choice" markup where we see some inefficiencies, and we also discuss how this mapping can be used in semantic web applications. We consider the capabilities of the recent editors and reasoners. We claim that the use of data types in MPEG-7 is an important obstacle for a complete mapping. XML-Schema enables one to define composite data types. However in OWL there is no standard defined for the composite data types. We have used Protégé to create a simple mapping from MPEG-7 to OWL and for reasoning we have used Pellet [8].

The structure of this paper is as follows: First, background information on OWL and MPEG-7 is given. Then, the details of the development are discussed. Finally, a discussion about the current state of the mapping and future work is given.

2. Technology Description

2.1 OWL

An important part of the semantic web is the ontology. XML only gives a surface syntax where RDF adds simple semantics to this syntax. RDF Schema has a small vocabulary to represent class, property, subclass, sub-property, individual, and conjunction. However, these are not enough to build an ontology on the web. Web Ontology Language (OWL) is derived from DAML+OIL and builds upon RDFS. OWL is based on description logic and is very expressive to represent the classes and their properties. There are three layers of OWL, namely, OWL Full, OWL DL, and OWL Lite.

OWL Lite is a subset of OWL DL. It is intended for the users who need to create only taxonomies and simple constraints. It has the most efficient reasoning mechanism. OWL DL is based on $\mathcal{SHOIN}(\mathcal{D})$ DL. It includes whole OWL vocabulary and it is guaranteed to be decidable and computable. The vocabulary of OWL Full is the same as OWL DL. However, it has less restriction on the constructors. OWL Full is an extension of RDF. It has meta-modeling facilities. A class can be an instance of another class. Also one can use arbitrary classes in the constructs. It has maximum expressivity and flexibility but it is neither decidable nor a description logic.

2.2 MPEG-7

MPEG-7 is ISO/IEC International Standard 15938 developed by MPEG (Moving Picture Experts Group). Its difference from its ancestors (MPEG-1, MPEG-2, MPEG-4) is that one can describe the multimedia content using it. Previous standards were for media storage and retrieval, but they can not be used for automatic processing task. The huge number of multimedia data and the need for intelligent search, retrieval and indexing of the multimedia require embedding the content information in the multimedia content. In MPEG-7, the multimedia content can be defined in different abstraction levels. For example for video, one can describe the low level components like colors and shape or more high level components. One can also store the creation information (by whom, when, where) or storage types (tape, dvd etc) of the multimedia data. MPEG-7 has several predefined description schemes and by using DDL (description definition language) one can define new description schemes. A good introduction to MPEG-7 standard is given in [4, 5]. When preparing an MPEG-7 file, one can use a wrapper to wrap the description needed or one can

use one of the MPEG-7 top level elements instead. This top level elements are grouped in 3 parts, as shown in Figure 2.

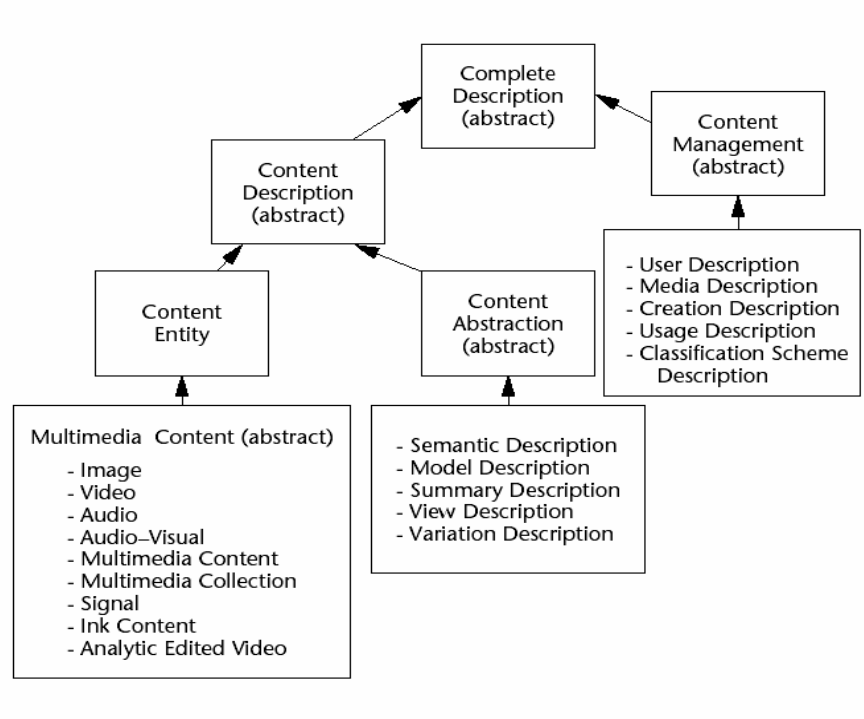


Figure 1: Hierarchy of Description Schemes

3. Development

Our research aims to find the ways of storing, searching, managing, and consuming multimedia content in movie databases to make them suitable for semantic web applications. As stated previously, we map MPEG-7 standard to OWL representation. MPEG-7 is a very detailed standard, describing the multimedia content in different levels. For our application using the "CreationInformation DS" of the Content Management part will be sufficient. "CreationInformation DS" is defined as given in Figure 2. In the movie database we will hold the actors, the director, date, type and other creation informations of the movie. Because we will discuss the mapping for important XML-Schema constructs, mapping other parts of the standard will be similar. Below, we list important modeling primitives of MPEG-7.

- Subclass relationship: Subclass relation in MPEG-7 is shown by <extension> or <restriction> markup. Corresponding construct in OWL is "SubClassOf".
- Sequence markup: Sequence in MPEG-7 represents a list of ordered elements. Representing them in OWL can be done using "object property" or "datatype property". If the element in the sequence is an XML-Schema datatype then "datatype property, otherwise "object property" is used. Order can not be represented using these OWL tags, however when we consider the semantics this is not very crucial.
- Choice markup: Choice in MPEG-7 represents exclusive-or. Therefore there is not an easy mapping for the choice markup. In the following section we will discuss the representation of "choice" in OWL.
- Attributes: Attributes can be represented by "datatype properties" in OWL. The attribute of an element can take only one value, therefore the attributes are defined as "functional" properties.

```
<!-- ############################################## -->
<!--  Definition of CreationInformation DS (9.2.1)  -->
<!-- ############################################## -->
<!-- Definition of CreationInformation DS  -->
<complexType name="CreationInformationType">
  <complexContent>
       <extension base="mpeg7:DSType">
             <sequence>
                   <element name="Creation"/>
                   <element name="Classification" minOccurs=0/>
                   <element name="RelatedMaterial" minOccurs=0
                   maxoccurs=””unbounded”/>
             </sequence>
       </extension>
  </complexContent>
</complexType>
```

Figure 2: Creation Information DS

By considering the above given primitives we have implemented a mapping from MPEG-7 to OWL. There are two properties of XML-Schema, whose mapping to OWL is not straightforward. One of them is the use of user-defined data types in MPEG-7 and the other is the “choice” markup. We discuss our approach for these constructs in the following subsections.

3.1 Use of customised datatypes

XML-Schema has predefined data types, like integer, language, date, and so on. Users can define new data types as well. MPEG-7 uses these composite data types in addition to the other data types, like array and matrix which do not exist in XML-Schema. The use of these data types is very crucial for MPEG-7 standard. The standard defines not only the conceptual properties of a multimedia data but also very low level properties, like audio coding, frames etc. Figure 3 gives an example of composite data type usage in MPEG-7. In this example, “colorDomain” is defined as a union of an enumerated type and “termReferenceType”, which is by itself a composite data type.

```
<attribute name="colorDomain" use="optional" default="color"
  <simpleType>
    <union>
       <simpleType>
          <restriction base="NMTOKEN">
             <enumeration value="binary" />
             <enumeration value="color" />
             <enumeration value="graylevel" />
             <enumeration value="colorized" />
         </restriction>
      </simpleType>
      <simpleType>
         <restriction base="mpeg7:termReferenceType" />
      </simpleType>
   </union>
  </simpleType>
</attribute>
```

Figure 3: Composite datatype example

OWL does not permit the use of user-defined data types, except enumerated data types. The extension of OWL with customised data types is discussed in [10]. OWL 1.1, which is recently proposed, gives more support to data types. However, still the problem with OWL is not resolved completely. Ontology editors and reasoners should also give full support to the use of composite data types. Without the support to customised data types, a full mapping of MPEG-7 to OWL will be problematic, because MPEG-7 metadata depends highly on user-defined data types. Without the support for user-defined data types we can only mimic the definition of composite data types. We explain how to do this by considering the example given in Figure 4.

```
<element name="CaptionLanguage" minOccurs="0" maxOccurs="unbounded">
  <complexType>
    <simpleContent>
      <extension base="language">
       <attribute name="closed" type="boolean"/>
       <attribute name="supplemental" type="boolean"/>
      </extension>
    </simpleContent>
  </complexType>
</element>
```

Figure 4: "Caption Language" element

In this example an extension to "language" data type is defined. As we previously discussed, we represent extension and restriction tags using Subclass. However data types are not defined as class in OWL and it is not possible to define a subclass of a data type. In OWL data types can be represented using "datatype property". Therefore one needs to define another solution. For the example given in Figure 4, the solution is given as in Figure 5. Our solution is similar to the one proposed in [9].

```
<owl:Class rdf:ID="CaptionLanguageType">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="language_redefined"/>
    </rdfs:subClassOf>
  </owl:Class>

<owl:DatatypeProperty rdf:ID="contentLanguage">
    <rdfs:range
rdf:resource="http://www.w3.org/2001/XMLSchema#language"/>
    <rdfs:domain rdf:resource="#language_redefined"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="captionsupplemental">
    <rdfs:domain rdf:resource="#CaptionLanguageType"/>
    <rdfs:range
rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
</owl:DatatypeProperty>

<owl:DatatypeProperty rdf:ID="captionclosed">
  <rdfs:domain rdf:resource="#CaptionLanguageType"/>
    <rdfs:range
rdf:resource="http://www.w3.org/2001/XMLSchema#boolean"/>
</owl:DatatypeProperty>
```

Figure 5: "Caption Language" in OWL

Here we define a new class called “language_redefined” which has “contentlanguage” as datatype property whose range is “language”. In this way, we actually define a class for the data type and use it where the data type would be actually used. Now we can define “CaptionLanguageType” as the subclass of the “language_redefined” class.

3.2 “choice” markup

MPEG-7 “choice” markup gives a list of elements from which only one element should be selected. An example for “choice” markup is given in Figure 6. In this section we will use this example for our discussion.

```
<complexType name="A">
      <choice>
            <element name="B" />
            <element name="C"/>
      </choice>
</complexType>
```

Figure 6: “Choice” markup

We modeled our solution using Protégé. For the example given in Figure 5, we create an object property (role in DL terminology) called “BC” and classes A, Btype and Ctype. We define BC to be a property for which the domain is A and the range is B or C. For the range we use the *unionof* construct of OWL. In this way, the user can choose either an individual from type B or C for the property BC. The following Figure is the output OWL document produced by Protégé.

```
<owl:Class rdf:ID="ClassCType"/>
  <owl:Class rdf:ID="ClassA"/>
  <owl:Class rdf:ID="ClassBType"/>
  <owl:ObjectProperty rdf:ID="ClassBC">
    <rdfs:domain rdf:resource="#ClassA"/>
    <rdfs:range>
      <owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:about="#ClassBType"/>
          <owl:Class rdf:about="#ClassCType"/>
        </owl:unionOf>
      </owl:Class>
    </rdfs:range>
  </owl:ObjectProperty>
  <ClassCType rdf:ID="c3"/>
  <ClassA rdf:ID="a1">
    <ClassBC>
      <ClassCType rdf:ID="c1"/>
    </ClassBC>
  </ClassA>
```

Figure 7:”Choice” markup in OWL

This translation is in OWL DL language. The language subset is $\mathcal{ALC}$, which is the basic attribute language. When the user uses an ontology editor, it restricts the user to choose only one element which prevents incorrect modeling. This solution, however, has a deficiency when a choice of data type elements and object type elements is constructed. In this case, union operator can not be used, because OWL does not accept data types as class.

This problem can be overcome by defining a class having the data type as property as it is suggested in Figure 5.

4. Results

We have created an OWL ontology for a part of MPEG-7 standard and we have applied several queries on this ontology. For creating OWL ontology we have used Protégé. Protégé is an ontology editor, employing a graphical user interface. It has its own reasoner for running SPARQL[11] queries and also communicates with other reasoners with DIG (Description Logic Implementors Group) interface. Protégé enabled us to enter also XML-schema datatypes. However as we discussed above, becuase the current OWL standard does not have support for user-defined datatypes we could only define them by employing the solution given in Figure 5. We omitted more compex datatypes and we gave a conceptual mapping. For reasoning we have used Pellet. Pellet can reason on OWL 1.1 language and it accepts XML-Schema datatypes. Pellet and Protégé can communicate with DIG interface. The version of Protege we use (version 3.2) uses DIG 1.1 interface which does not support XML-Schema datatypes. However DIG 2.0 [1] is compatible with OWL 1.1; hence supports XML-Schema datatypes. The resulting OWL ontology we got is classified as ALCOF(D) language by Pellet. Pellet also checks the consistency of the ontology and runs SPARQL queries.

In this research we gave a mapping for "Creation Information DS". This part of the ontology contains important information that can be used by the movie database applications. We have written several SPARQL queries to query our ontology. SPARQL is a query language for getting information from RDF graphs. We have seen that our ontology can be used for many user applications, like, searching for the films an actor is starring, or the films which are classified in the category "Drama". An example SPARQL query is given in Figure 8. In this example the name of the films which are rated UK:10 for parental guidance field are searched.

```
SELECT ?filmname
WHERE { ?film :Classification ?classification.
        ?film :Creation ?filmcreation.
        ?filmcreation :Title ?title.
        ?title :contentString ?filmname.
        ?classification :ParentalGuidance ?pg.
        ?pg :Region :UK  .
        ?pg :parentalratingminage ?minag.
        FILTER (?minag > 10)
       }
```

Figure 8: SPARQL Query example

5. Conclusions

Semantic Web idea is discussed in several layers. Among them ontology layer takes an important place. OWL is about to become standard language for the Ontology Layer of Semantic Web. However, there are already established ontologies which use different ontology languages. To have a common understanding, these communities will port their ontologies to OWL. An example for this is the multimedia community. Multimedia community, at the time being, uses its own meta-data to create ontologies. MPEG-7 is such a meta-language used to describe multimedia content. It enables us to store information about the multimedia content, usage and creation of it, and so on. in different abstraction levels. In this study we have discussed a mapping from MPEG-7 to OWL. We have

discussed current solutions for mapping from MPEG-7 to OWL and proposed a new solution for the "choice" markup. We pointed out that the most important problem is the use of composite data types in MPEG-7. Current OWL standard does not employ user defined data types (except enumerations). However OWL 1.1 extends OWL with user-defined data types. With the integration of OWL 1.1 standard in all applications of semantic web (ontology editors, reasoners, interfaces), it will be possible to define the composite data types in many applications like the movie database application we used in this paper.

Acknowledgments

This work is supported by the Scientific and Technical Research Council of Turkey, Project No: EEEAG 105E068.

References

[1] Bechhofer S., and Motik B. DIG 2.0 Specification. Available at http://www.cs.man.ac.uk/~bmotik/dig/dig_specification.html, January 2007.
[2] Berners-Lee T., Hendler J.; and Lassila O. The Semantic Web. *Scientific American* 284(5):34-43, 2001.
[3] Hunter J. Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology. *Proceedings of International Semantic Web Working Symposium (SWWS)*, Stanford, 2001.
[4] Martinez J. M., MPEG-7: Overview of MPEG-7 Description Tools, Part 2. In *IEEE MultiMedia*, 9(2):78-87, 2002.
[5] Martinez J. M., Koenen R., Pereira F., MPEG-7: The Generic Multimedia Content Description Standard,Part 1. In *IEEE MultiMedia*, 9(3):83-93, 2002.
[6] OWL Web Ontology Language Guide. W3C Recommendation (10 February 2004). Available at www.w3.org/TR/2004/REC-owl-guide-20040210.
[7] OWL Web Ontology Language Overview. W3C Recommendation (10 February 2004). Available at www.w3.org/TR/2004/REC-owl-features-20040210.
[8] Sirin E., Parsia B., Grau B. C., Kalyanpur A., and Katz Y. Pellet: A practical owl-dl reasoner. Technical report, University of Maryland Institute for Advanced Computer Studies (UMIACS), 2005-68. Available at http://www.mindswap.org/papers/PelletDemo.pdf.
[9] Tsinaraki C., Polydoros P., Moumoutzis N., Christodoulakis S. "Coupling OWL with MPEG-7 and TV-Anytime for Domain-specific Multimedia Information Integration and Retrieval". Proceedings of RIAO 2004, Avignon, France, 2004.
[10] Pan J. Z., and Horrocks I., OWL-Eu Adding Customised Datatypes into OWL. In Proc. Of Second European Semantic Web Conferences. (ESWC 2005), 2005.
[11] Prud'hommeaux E., and Seaborne A., SPARQL Query Language for RDF. Available at http://www.w3.org/TR/2005/WD-rdf-sparql-query-20050217/, Feb 2005.

Semantics Based Customization of UBL Document Schemas *

YALIN YARIMAGAN yalin@srdc.metu.edu.tr
ASUMAN DOGAC asuman@srdc.metu.edu.tr
Software Research and Development Center
Department of Computer Engineering
Middle East Technical University (METU)
06531, Ankara, Turkiye

Abstract. Universal Business Language (UBL) is an OASIS initiative to develop common business document schemas to provide document interoperability in the eBusiness domain. Since the data requirements change according to a context, UBL schemas need to be customized and UBL defines a guideline to be followed for customization of schemas. XSD derivation based customization as proposed by UBL provides syntactic interoperability, that is, an XML parser that can interpret standard UBL documents can also interpret customized UBL documents. We argue that for UBL to become mainstream, syntactic interoperability alone is not enough. It needs to be supported by semantic interoperability, that is, it must be possible for users and even automated processes to discover and reuse customizations provided by other users.

In this paper, we describe how to improve the UBL customization mechanism by providing semantic representations for context domains and describe how these semantics can be utilized by automated processes for component discovery and schema customization. For this purpose, we derive ontologies from taxonomies like the North American Industry Classification System (NAICS), the Universal Standard Products and Services Classification (UNSPSC) and relate corresponding concepts from different ontologies through ontology alignment. Then, we process these aligned ontologies using a reasoner to compute inferred ontologies representing context domains. We show that when custom UBL components are annotated using classes from these ontologies, automated discovery and customization becomes possible.

1. Introduction

Businesses need to exchange data to execute transactions with their trading partners. Reusing well-understood standard patterns in documents makes business processes easier to implement, manage, and improve [12]. Adopting common standards can reduce development and maintenance costs, improve performance, and enhance business relationships.

Universal Business Language (UBL) [33] is an OASIS standard addressing this standardization need for electronic business documents. UBL provides a library of reusable components such as Address, Price and a set of document schemas such as Order, Invoice to be used in electronic business and is rapidly being adopted by both public and private sector organizations around the world.

* This work is supported in part by the European Commission, Project No: IST-027306-STP ABILITIES and the Scientific and Technical Research Council of Turkey (TÜBİTAK), METU-ISTEC Project, No: EEEAG 105E068.

Even though businesses share a lot of common data requirements, they operate in different industry, geopolitical, and regulatory contexts. As a result of these differences, they have different rules and requirements for the information they exchange in their business documents. Hence, it is reasonable to expect that national and industry groups, smaller user communities and individual organizations will need to customize UBL schemas.

In order to address this customization requirement, UBL provides a customization methodology [8] to be followed by implementers. This methodology allows the modification of standard UBL components in response to contextual needs through XSD derivation operations. Schemas can be extended by adding new components or can be restricted by removing components or by limiting cardinality of components to a subset. Eight context categories are introduced: Geopolitical, Industry Classification, Product Classification, Business Process, Official Constraint, Business Process Role, Supporting Role and System Constraint. Context of a document or a component is expressed using a set of name/value pairs (context category, context value).

Current UBL customization methodology focuses on the goal of providing syntactic interoperability, that is, it ensures that an XSD parser that can interpret standard UBL schemas can also interpret customized UBL schemas. On the other hand, it does not address semantic interoperability, that is, when a document schema needs to be customized for a business context, users need to manually discover or provide component versions applicable to that particular context and replace them in place of their standard counterparts to perform the schema customization.

In this paper, we present how to improve the UBL customization mechanism by providing machine processable semantic representations for context domains and describe how these semantics are utilized for automating tasks required for proper customization of UBL schemas.

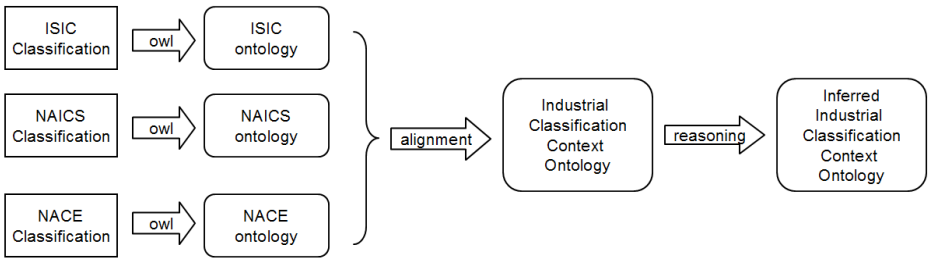


Figure 1. Developing a Semantic Representation for the Industrial Classification Context

Figure 1 displays the steps we take for developing semantic representations for context domains. We start by developing specialized converters to derive Web Ontology Language (OWL) [16] ontologies from classification schemas such as the *International Standard Industrial Classification* (ISIC) [10], the *North American*

Industry Classification System (NAICS) [13]. Then we present how similar concepts from these ontologies are related to each other using ontology alignment techniques and how resulting aligned ontologies are processed using reasoners to infer implicit relationships among their classes. This inference results in a set of ontologies that formally represent the semantics of context domains. It should be noted that our aim here is not to take on the actual alignment task but demonstrate how such correspondences can be specified.

Within the scope of this work, we developed specialized converters to derive individual ontologies based on the taxonomy of the following classifications: ISIC, NAICS, *Universal Standard Product and Service Specification* (UNSPSC) [34], *Classification of the Economic Activities in the European Community* (NACE) [24] and the *Central Product Classification* (CPC) [5]. These ontologies are available in our web site [15]. In addition to those, our approach is flexible enough to support any other ontology derived from classification schemas or code lists relevant to context domains to allow UBL users and component customizers to work with the classification they are accustomed to.

Once context ontologies are developed, customized UBL components are annotated using classes from these ontologies by using graphical tools implemented. As specified in the UBL customization methodology, UBL components can only be customized in response to contextual needs. Annotating customized versions with classes from context ontologies expresses the context they apply to in a machine processable manner. This allows the development of automated processes that intelligently search through custom components by interpreting class-subclass, equivalence and other relationships specified in context ontologies and gather applicable versions of components. These versions then replace their standard counterparts to customize document schemas for particular business context values.

The ability to formally represent context domain values using an ontology language allows us to reuse the work that exists in the general area of interoperability. The work on *Model Management* [3] proposes the development of general-purpose functions in order to reduce the amount of effort required to manage complex models such as UBL components and proposes high-level operations such as matching, merging, function application, selection and composition that work on formal structures representing models. Following that pattern and using context ontologies, we define the following operations to manage UBL components:

- *Create:* Register a new custom component version for the specified context value.
- *Discover:* Find custom component versions applicable for the specified context value.
- *Merge:* Compose multiple custom versions of a component into a single version that represents all other versions.
- *Transform:* Customize a UBL schema by replacing all included components with corresponding custom versions applicable for the specified context value.

As suggested above, the ability to automate component discovery leads to a significant new flexibility, that is, it becomes possible to merge multiple versions of a component to generate additional versions of that particular component. This greatly simplifies the component customization effort as it is no longer necessary to manually provide customized components for every single business context. Instead, it is sufficient to customize components for individual context categories since combinations for multiple categories are automatically generated as needed.

Supported with a global custom component repository and a knowledge base about repository contents, our approach has the potential to streamline the customization process especially for those users who do not have the necessary resources or expertise required for properly tailoring standard UBL schemas for their business needs.

The rest of this paper is organized as follows. Section 2 provides a brief overview of technologies and standards that the UBL is build upon. Section 3 discusses our process for developing UBL context ontologies. Section 4 presents an architecture and a collection of tools and services that utilize context ontologies to implement the operations necessary to manage UBL components. Section 5 describes the system implementation. Section 6 provides a brief overview of the related work on using ontologies for context modeling. Section 7 concludes the paper and proposes future research directions.

2. Enabling Technologies

There are many standards for structuring information to be electronically exchanged between and within businesses, organizations, government entities and other groups. These standards describe structures that represent documents, for example purchase orders to automate purchasing. The motive for such standards is reusing well-understood patterns for reducing development and maintenance costs as it becomes easier to implement, manage and improve business processes.

Electronic Data Interchange (EDI) is a set of such standards for computer-to-computer exchange of business data in standard formats. There are two major sets of EDI standards: the United Nations recommended *United Nations/Electronic Data Interchange For Administration, Commerce, and Transport* (UN/EDIFACT) [32] is the international standard and is predominant outside of North America; and the U.S. standard *American National Standards Institute Accredited Standards Committee X12* (X12) [1] is predominant in North America. Both standards are widely used in electronic commerce transactions around the world.

With the proliferation of XML and Internet based technologies, new standards like the RosettaNet [22], Electronic Business using eXtensible Markup Language (ebXML) [6], XML Common Business Library (xCBL) [35] have emerged adapting EDI ideas to contemporary technologies.

xCBL is one such standard which provides a collection of common business elements that underlie all EDI and Internet commerce protocols. xCBL has been developed and modeled after X12 and EDIFACT to preserve and extend the ex-

isting EDI investments. It is based on a reusable component model to speed up the implementation of standards and facilitate their interoperation by providing a common framework.

ebXML is another set of standards which can be characterized as an XML and Internet realization of the EDI. The ebXML Core Components Technical Specification (CCTS) [31], which is a part of the ebXML standards family, provides a set of reusable concepts, called "Core Components". A "Core Component" is a building block for the creation of a semantically correct and meaningful information exchange package and it contains only the information pieces necessary to describe a specific concept. "Core Components" represent common data elements of everyday business documents such as "Address", "Amount", or "Line Item" in a syntax neutral way. CCTS also defines the "Business Context" concept as a mechanism for qualifying and refining Core Components according to their use under particular business circumstances. Once Business Contexts are identified, Core Components can be differentiated to take into account any necessary qualification and refinement needed to support the use of the Core Component in the given Business Context.

For this purpose, CCTS defines the "Core Components Context Constraints Language", which is used to express the relationship between specific Business Contexts and how rules are applied to Core Components to produce context specific entities that can be assembled into larger business documents such as "Order" or "Invoice".

Constraint Language depends on "Context Categories" to allow users to uniquely identify and distinguish between different Business Contexts. For each context category, one or more standard classifications are specified to provide values for the category. Constraint rules are tied to values from these particular set of classifications for identifying and distinguishing Contexts.

The UBL effort is started as a program to provide specifications for the standard documents corresponding to the EDI transaction sets from X12 and UN/EDIFACT through providing an implementation of the CCTS. Instead of starting from scratch, UBL designers based their work on the xCBL vocabulary, a mature XML standard already being used in a number of ecommerce marketplaces. Since xCBL reuses the semantics from the EDI experience, this ensured the UBL to be build upon the work and experience in the area. Moreover, the component library model of UBL provides much better alignment among document types derived from the library than had been the case with older message standards in which different document types were developed independently.

Following are some real-world implementation examples using UBL:

- UBL Invoice has been mandated by law for all public-sector business in Denmark and an estimated 1.2 million UBL invoices are currently exchanged in Denmark every month [4].

- A subset of the UBL Invoice has been recommended for all government use by the Swedish National Financial Management Authority [29]

- The Electronic Freight Management project of the U.S. Department of Transportation is developing a UBL based pilot project for a complex demonstration of state-of-the-art electronic commerce in a real-world setting [30]

3. Use of Ontologies in Semantic Annotation of UBL Schemas

As we have already mentioned, current UBL customization methodology ensures syntactic interoperability, yet, it does not address semantic interoperability. In order to provide that, we develop ontologies to represent context domains and annotate UBL components using classes from these ontologies. Expressing context information through a formal ontology language, such as OWL [11], provides the following:

- An ontology language is machine processable since it conforms to a formal, well-defined syntax [23]. A description given in an ontology language can be automatically processed to obtain the information it represents. A description in OWL can be parsed into the classes, properties and corresponding values, even when an application knows only the OWL syntax and has no understanding of a particular domain specific ontology. Furthermore, any program having a prior knowledge of the syntax and semantics of the ontology, can parse the description, extract the represented information and interpret it since the syntax and the semantics are already known by the application using it.

- An ontology describes *consensual knowledge*, that is, it describes meaning which has been accepted by a group not by a single individual. In other words, it provides a common vocabulary for those who have agreed to use it. Hence when we annotate a UBL component with an ontology class, it inherits the well-defined, shared meaning attributed to that class.

- Class-subclass relationships expressed through ontology languages are especially useful for the purposes of representing concepts organized as hierarchies. A component customized for a context class is applicable to hierarchically lower level context classes as well. Therefore the ability to traverse class-subclass relations of an ontology enables the development of automated discovery processes for UBL components and schemas.

- An ontology provides the ability to define relationships among classes, properties and instances which can then be used for reasoning. Especially in the case of independently developed ontologies representing similar concepts, different ontologies may include descriptions for corresponding entities. The reasoning capability is the *glue* holding different ontologies together for such cases.

In the following, we present how ontologies can be utilized for expressing context domains by providing examples in the *Industrial Classification* context. Then, we discuss how these ideas can be generalized for other context domains.

3.1. Ontology Development for the Industrial Classification Context

UBL inherits the *business context* concept from the CCTS which defines the Industrial Classification context to provide a description of the industry or sub-industry in which the business takes place [31]. CCTS requires context values for the Industrial Classification context to be used from two code lists: International Standard Industrial Classification (ISIC) [10] and first two digits of Universal Standard Product and Service Specification (UNSPSC) [34]. However, in addition to these two, there are other code lists being used by different organizations to classify industrial activities like the North American Industry Classification System (NAICS) [13] and the Classification of Economic Activities in the European Community (NACE) [24].

All these classifications provide detailed taxonomies for industrial activities and disclose considerable amount of semantic information. Therefore, we propose to use the taxonomy of these classifications to provide context values for UBL components and schemas.

However, there is also a need to relate context values from different classifications to each other so that customizations provided for context values specified using a particular classification can be discovered and reused even when context is specified using values from different classifications. Furthermore, this semantic interoperability needs to be expressed in a machine processable way so that automated processes can interpret it as well as humans.

In order to fulfill these goals, we express taxonomies as ontologies through an ontology language, namely OWL. Taxonomies reveal limited semantics; however this semantics proves to be very useful when transformed to ontologies:

- Once taxonomies are expressed through an ontology language like OWL, they become machine processable. Hence, it becomes possible to process class-subclass relationships.
- It becomes possible to formally specify relationships among classes from different ontologies. This, in turn, allows to define machine processable semantic information among different ontologies and relate them.
- Formally expressed relationships in OWL are processed by reasoners to infer additional relationships between ontology classes. Hence, the users continue to use whichever classification they prefer for specifying context values, yet automated processes relate context values from one classification to others.

Within the scope of this work, we have developed specialized converters to derive individual ontologies based on the taxonomy of various classifications. These ontologies are available in [15].

Figure 2 gives a high level view of class hierarchies for ontologies representing ISIC and NACE classifications. Once such ontologies are developed, we apply ontology reconciliation techniques to specify relationships between classes belonging to different ontologies. This *ontology alignment* [7] task of resolving semantic correspondences between different ontologies can in fact, best be assumed by domain

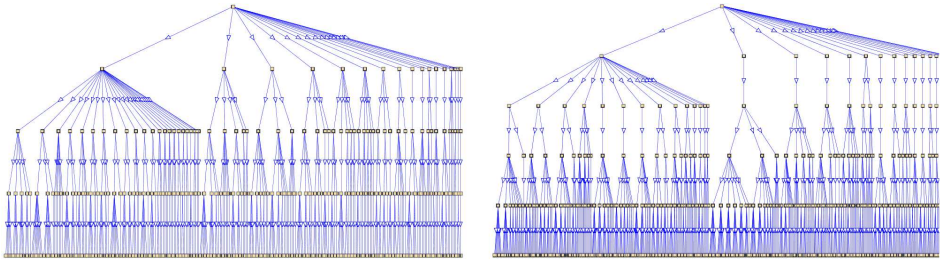


Figure 2. Class hierarchies of ISIC and NACE ontologies

experts and Standard Development Organizations (SDOs). Our aim here is not to take on this role but to demonstrate how such correspondences can be exploited once they are specified. Note that, a significant amount of such relationships are already known and documented, albeit for human use. As an example, [25] provides the correspondences between NACE and ISIC codes.

We define all such correspondences in a machine processable way, by using the following three operations for ontology alignment:

- *Specifying equivalance:* It is very common for multiple classifications to include code for semantically equivalent concepts. As an example, ISIC, NAICS and NACE all specify *Construction* as an industry. For such cases, specification of corresponding classes from different ontologies as equivalent classes is performed by using the `<owl:equivalentClass>` construct.

- *Specifying composition:* There are many cases where a concept represented as a single code by one classification corresponds to multiple codes in other classifications. As an example, the NAICS code *11-Agriculture, Forestry, Fishing and Hunting* correspond to two ISIC codes: *A-Agriculture, Hunting and Forestry* and *B-Fishing*. Such cases are aligned by specifying one class in an ontology as the composition of multiple classes in another ontology by using the `<owl:unionOf>` construct.

- *Specifying subsumption:* Another common pattern among different classifications is that a code from one classification simply subsumes a code from another classification. As an example, NACE code *C-Mining and Quarrying* express a concept that is a superset of the NAICS code *21-Mining*. Such cases are aligned by specifying subsumption relations between classes in different ontologies by using the `<owl:subClassOf>` construct.

Once we align different ontologies by specifying all such correspondences, we use a descriptive logic reasoner to compute the inferred class hierarchy for the *Industrial Classification* domain. This inferred class hierarchy provides a single ontology that represents all individual ontologies and the complete set of relationships among

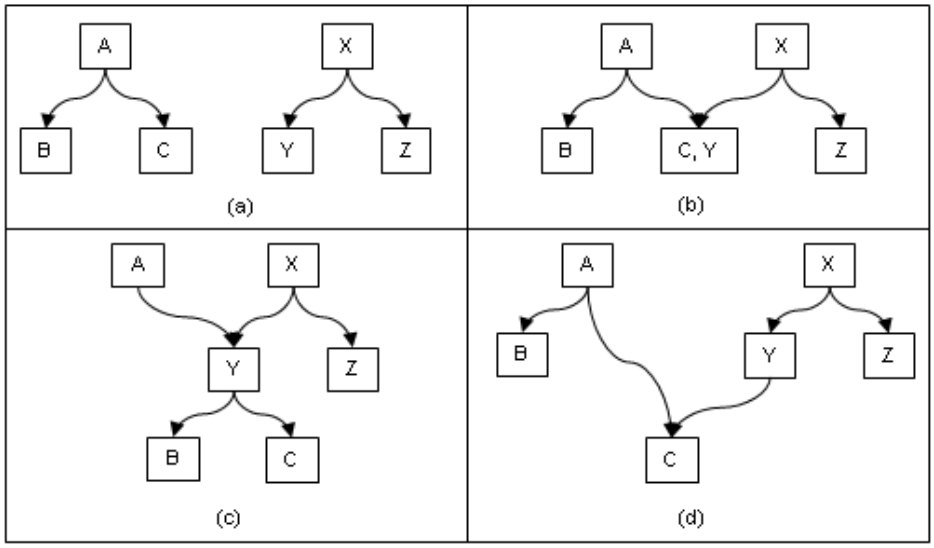


Figure 3. Ontology Alignment Operations

their classes. Figure 3 displays the effects of alignment operations after reasoning: (a) shows two example ontologies, (b) shows the inferred ontology after $C \equiv Y$ is expressed through `<owl:equivalentClass>`. Notice that the axiom provided for alignment cause the reasoner to add additional parents to `C` and `Y`. (c) gives the inferred ontology after $(C \cup B) \equiv Y$ is expressed through `<owl:unionOf>`. For this case, alignment operation changes direct parents of `B` and `C` and adds a new parent to the `Y`. (d) depicts the inferred ontology after $C \subseteq Y$ is expressed through `<owl:subClassOf>`. This alignment operation results in an additional parent for `C`.

It should be noted that all alignment operations result in classes with multiple inheritance. Since a UBL component customized for a context value is also applicable for context values at hierarchically lower levels, multiple inheritance has significant consequences for the purposes of component discovery. While searching for component versions applicable for a context value, component discovery needs to consider versions applicable to all parents and merge them. Details of how this is achieved are given in Sections 4.3 and 4.5.

Once the inferred ontology representing the Industrial Classification domain is computed, classes from that ontology are used for annotating customized UBL components to express their context, and it becomes possible to develop automated processes that can intelligently discover UBL components by utilizing context ontologies and corresponding annotations.

Consider, for example, a fragment from the inferred ontology in Figure 4 representing the Industrial Classification context, where solid lines represent asserted relations and dashed lines represent inferred relations and the following two cases:

- Assume standard UBL `PartyType` is customized for the *naics:Manufacturing* context with the addition of `Manufacturing License`. When an aircraft parts manufacturing company requires `PartyType` component, an automated process, through interpreting class-subclass relations defined between *naics:Manufacturing*, *naics:Transport Equipment Manufacturing* and *naics:Aerospace Product and Parts Manufacturing* can discover that the version of `PartyType` customized for *naics:Manufacturing* context is the one to use rather than the standard UBL version since the applicable context *naics:Aerospace Product and Parts Manufacturing* is a subcontext of *naics:Manufacturing*.

- Assume International Civil Aviation Organization (ICAO) approval is required for all aerospace component manufacturers hence standard UBL `PartyType` is extended by adding `ICAOApproval` and annotated with *isic:Manufacture of Air and Spacecraft and Related Machinery* context. When users accustomed to NAICS classification request components applicable to *naics:Aircraft Manufacturing* context, through class-subclass relation between *naics:Aircraft Manufacturing* and *isic:Manufacture of Air and Spacecraft and Related Machinery* (which is inferred as a result of the asserted equivalence between *naics:Aerospace Products and Parts Manufacturing* and *isic:Manufacture of Air and Spacecraft and Related Machinery*), an automated process can discover that the version of `PartyType` customized for *isic:Manufacture of Air and Spacecraft and Related Machinery* context is also applicable to the *naics:Aircraft Manufacturing* context.

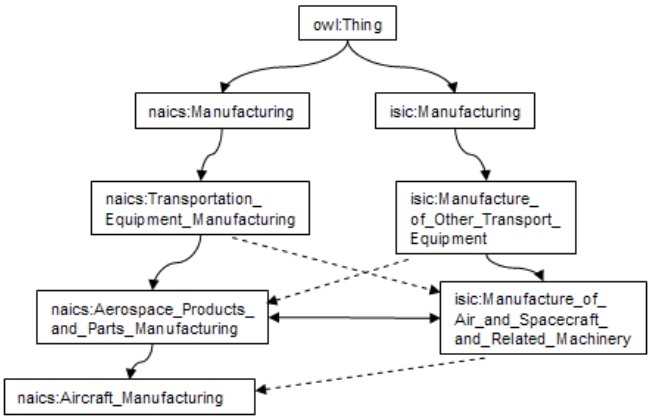


Figure 4. Context Ontology fragments

3.2. Generalization to Other Context Domains

There are existing classifications related to other context domains as well. For the *Product Classification* context, examples include Universal Standard Products and Services Classification (UNSPSC) [34], North American Product Classification System (NAPCS) [14], Classification of Products by Activity in EU (CPA) [26], Central Product Classification (CPC) [5]. For the *Geopolitical* context, CCTS suggests a four leveled hierarchy consisting of Continent, Economic Region, Country (using ISO 3166.1 codes [9]), Region (using ISO 3166.2 codes [9]). For the *Business Process* context, CCTS suggests the use of UN/CEFACT Catalogue of Common Business Processes.

The essence of our approach is to provide an open mechanism that is flexible enough to support all classifications applicable for a context, so that users can continue to work with the classifications they are accustomed to. In the meantime, these classifications need to be interpretable by automated processes, therefore we use individual ontologies to reflect the taxonomy of classifications and relate them to one another through ontology alignment operations described above. The `<owl:import>` construct of OWL makes it possible to develop global ontologies for each context domain that imports individual ontologies representing classifications related to that context.

It is reasonable to assume the introduction of new classification schemes with time. Similarly, user groups and organizations may suggest different structures for representing domain values. With our approach, it is possible to develop corresponding ontologies and relate them to the existing set of ontologies.

4. System Architecture and Operation

Development of ontologies that represent context domains lays the necessary foundation for the implementation of automated processes that can intelligently discover components and customize document schemas. In order to support the development of such processes, we provide a system architecture whose overall view is given in Figure 5.

This architecture provides UBL users a number of tools for component discovery and document schema customization. These tools are supported by internal services which interact with a knowledge base through a reasoning layer. The knowledge base stores metadata about system artifacts such as context ontologies and customized components.

Context ontology metadata shown in Figure 5 gives information about context ontologies such as the domain they represent and their URI on the Internet. The context ontologies are processed by a reasoner and inferred ontologies representing context domains are computed. All requests from internal services concerning context ontologies are served using these inferred ontologies.

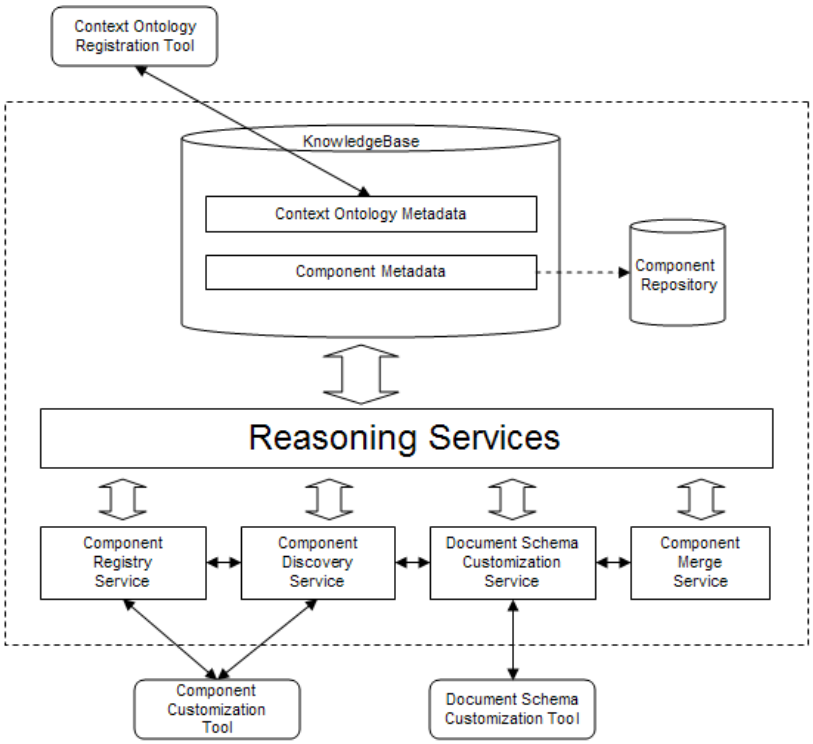


Figure 5. System Architecture

Component metadata gives information about standard and customized components and the relationships among them. This metadata is used for the following purposes:

- When a component is customized for a context, its metadata is created to express the standard UBL component it is derived from and the context it is applicable to by specifying references to classes from inferred ontologies.

- When a custom version of a component is required for a specific context, component metadata is queried to gather applicable versions with the help of inferred context ontologies.

- When a document schema needs to be customized for a specific context, component metadata is queried to gather custom versions of components included in that schema and those versions are used to replace the original components in the customized document schema.

A user accesses the system through three external interfaces which allow him to register context ontologies, to customize components and to customize document schemas as described in the following:

- *Context Ontology Registration Tool* provides a GUI for registering new context ontologies and maintaining existing ones. As described in Section 3.2, we have developed individual ontologies representing classifications related to context domains. And there may be additional ontologies developed by user groups and organizations. Assuming all such ontologies are published on the Internet and maintained by their creators, Context Ontology Registration Tool helps to register those ontologies so that their metadata is maintained by the system.

- *Component Customization Tool* provides *Domain Experts* a GUI for creating and maintaining custom versions of UBL components. In a typical component customization scenario, domain expert specifies the component to be customized and the target context. The Component Discovery Service provides a baseline version of the component for the specified context. Domain expert customizes baseline version locally through XSD extension and restriction operations and submits the resulting component for registration. The Component Registration Service verifies that the submitted version is a valid specialization of the baseline, creates necessary metadata and ensures proper registration of the component.

- *Document Schema Customization Tool* provides *UBL users* a GUI for generating customized versions of UBL document schemas based on their business context. In a typical document schema customization scenario, the user specifies a UBL document schema to be customized and its business context as a combination of classes from context ontologies. The Document Schema Customization Service, by collaborating with the Component Discovery and Component Merge services, generates a customized version of the UBL document schema for the specified context.

Following subsections explain the internal services (Figure 5) and how they work together for supporting the user tools.

4.1. Reasoning Services

This layer provides reasoning services to the rest of the system. It consists of an OWL Reasoner, namely Pellet [18], together with supporting service and utility classes. Internal services do not access the knowledge base directly. Instead, they query the knowledge base through the reasoner, which loads all aligned ontologies and computes inferred ontologies before answering any queries. This ensures all knowledge base queries are executed on inferred ontologies rather than individual asserted ontologies.

The ontologies representing context domains consist of classes representing the taxonomy of corresponding classifications and do not include any instances. Since there are no instances, the time consuming reasoning operations necessary to process instances such as consistency checking (which ensures the axioms about instances and classes are consistent) and realization (which computes a direct class for each instance) are not required. This greatly improves the time it takes for reasoning over ontologies.

4.2. Component Registration Service

Given the XSD definition of a customized UBL component and the applicable context information, Component Registration Service ensures proper registration of components.

Before registering a new customized version, Component Registration Service verifies that the new version is a valid UBL specialization. In order to perform that, Component Discovery Service is requested to provide the baseline version of the same component for the specified context. Then, the new version is compared against the baseline version to ensure that it is generated only through UBL conformant XSD derivation operations, namely extension and restriction.

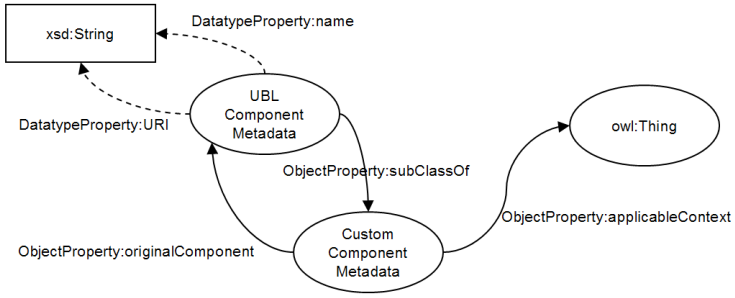


Figure 6. OWL Classes representing the metadata of standard and custom UBL components

In order to allow component discovery, knowledge base stores metadata about standard and customized UBL components as instances of `UBLComponentMetadata` and `CustomComponentMetadata` classes shown in Figure 6. Range of `applicableContext` object property can be any class from context ontologies therefore it is defined as `owl:Thing`. XSD definitions of components are stored in component repository locations pointed by these metadata. UBL specification provides XSD definitions of standard components and the corresponding `UBLComponentMetadata` instances are created and stored in the knowledge base. XSD definitions of customized components provided by users are stored in the component repository locations and the corresponding `CustomComponentMetadata` instances are created

by the Component Registration Service. Figure 7 displays the metadata for the standard `ItemType` and Figure 8 displays the metadata for a customized `ItemType` version.

```
<UBLComponentMetadata rdf:ID="UBLItemType">
  <componentName rdf:datatype="string">ubl:ItemType</componentName>
  <componentURI rdf:datatype="string">
    http://docs.oasis-open.org/ubl/prd2-UBL-2.0/xsdrt/common/
      UBL-CommonAggregateComponents-2.xsd#ItemType
  </componentURI>
</UBLComponentMetadata>
```

Figure 7. Metadata for the Standard ItemType

```
<CustomComponentMetadata rdf:ID="AerospaceItemType">
  <componentName rdf:datatype="string">industry:aerospace:ItemType</componentName>
  <componentURI rdf:datatype="string">
     http://www.srdc.metu.edu.tr/componentRepository/industry/aerospace_ItemType.xsd
  </componentURI>
  <originalComponent rdf:resource="UBLItemType"/>
  <applicableContext
     ref:resource="isic:Manufacture_of_Air_and_Space_and_Related_Machinery"/>
</CustomizedComponentMetadata>
```

Figure 8. Metadata for a Custom ItemType

It should be noted that for a particular context value, there can only be one custom version of a component. Context ontologies can have multiple sibling (semantically equivalent) classes all representing the same context value, therefore, any custom version registered for a particular context class is considered to be applicable to all sibling classes of that class as well.

After a component is stored in the repository, it becomes available for discovery and is included in document schema customizations regarding the context it is applicable to.

4.3. Component Discovery Service

In order to gather the applicable version of a component for a particular context value, the Component Discovery Service queries the knowledge base for a `CustomComponentMetadata` instance with the `applicableContext` property set to any of the semantically equivalent classes representing the specified context value (as described in Section 4.2, there can be only one custom version for a particular context value).

As an example, consider the ontology fragment in Figure 4 and metadata in Figure 8. When Component Discovery is called to gather the `ItemType` version applicable to the *naics:Aerospace Products and Parts Manufacturing* context, the reasoner is queried for equivalent classes, revealing the *isic:Manufacture of Air and Spacecraft and Related Machinery* class. Then, knowledge base is queried for component metadata with `ApplicableContext` property equal to any of those classes, matching the `AerospaceItemType` instance.

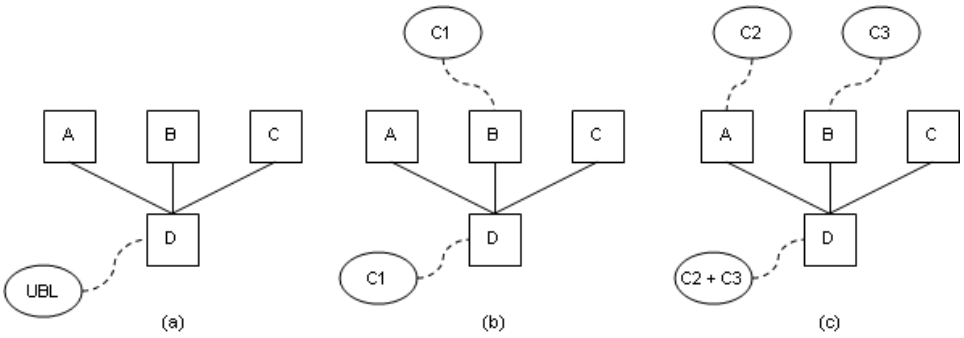


Figure 9. Context class with multiple inheritance

For cases which knowledge base includes no metadata instance with `Applicable Context` property equal to any of the equivalent classes, Component Discovery queries the reasoner for direct parent classes of the specified context and recursively calls itself to gather versions applicable to parent context values. In the case of single inheritance, component version applicable to the direct parent context is assumed to be applicable to the child context as well. However, context ontology classes can have multiple inheritance and there may be multiple parents with applicable custom versions. Figure 9 displays all possible cases for a context class with multiple inheritance. The implications of multiple parents on the component discovery are as follows:

- Figure 9-(a): If there is no custom version applicable to any of the parent context values, standard UBL version is assumed to be applicable to the context.
- Figure 9-(b): If there is a custom version applicable to only one of the parent context classes, that version is assumed to be applicable to the context.
- Figure 9-(c): If there is a custom version applicable to multiple parent context classes, a new version, generated by merging those versions, is assumed to be applicable to the context.

The algorithm in Figure 10 displays the overall steps of how component discovery is executed.

UBLcmp: UBL component being searched
context: target context value

```
1.  reasoner.loadContextOntologies();
2.  equivalentCtxSet = reasoner.getEquivalentClasses(context);
3.  for all context class ctx ∈ equivalentCtxSet
4.      qry = (OriginalComponent=UBLcmp AND ApplicableContext=ctx);
5.      cmp = kBase.componentMetadata.query(qry);
6.      if (cmp ≠ null) then
7.        return cmp;
8.      end if
9.  end for
10.
11. //there is no custom version for exact context value; check parent classes
12. cmpList = new List();
13. directParentSet = reasoner.getDirectParentClasses(context);
14. for all context class parentCtx ∈ directParentSet
15.     cmp = discoverComponent(UBLCmp, parentCtx);
16.     cmpList.add(cmp);
17. end for
18.
19. if (cmpList.size) == 0) then // no parent with a custom version
20.     return null;
21. else if (cmpList.size) == 1) then // one parent with a custom version
22.     return cmpList[0];
23. else // multiple parents with a custom version
24.     cmp = mergeComponents(cmpList);
25.     return cmp;
26. end if
```

Figure 10. Component Discovery Algorithm

It should be noted that, since Component Discovery algorithm recursively calls itself for parent context classes, even component versions that are registered for higher level context classes may get merged to generate an applicable version for a particular context. Figure 11 displays an example ontology fragment where component versions linked with solid lines represent registered versions and those that are linked with dashed lines represent the influence of registered versions on lower lever context values.

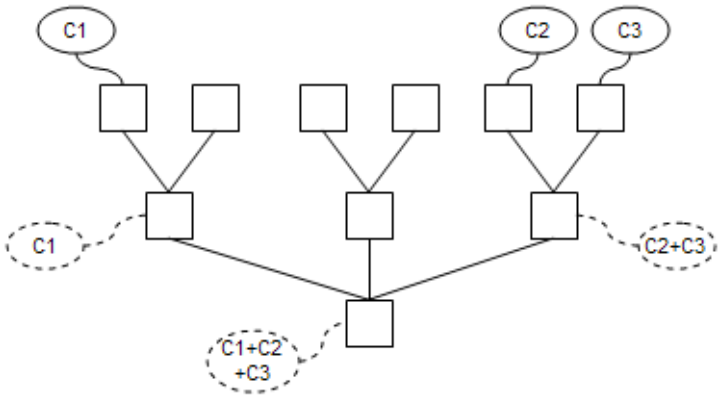


Figure 11. Influence of component versions on lower level context values

4.4. Document Schema Customization Service

Given a UBL document schema and a business context, the Document Schema Customization Service customizes the document schema by replacing the original UBL components with the customized components applicable for that particular context.

UBL document schemas are composed of several basic and aggregate components. Aggregate components themselves are collections of other basic and aggregate components in a recursive manner. Many aggregate components in this hierarchy are included by other components and UBL document schemas themselves. So customizing an intermediate component for a context implicitly customizes UBL document schemas including that component for the same context.

As an example, consider the greatly simplified *Order* and *Catalogue* document schemas in Figure 12. Both document schemas contain `Item` in their hierarchy: *Catalogue* includes it through `CatalogueLine` element and *Order* includes it through `OrderLine` element. Therefore, customizing `Item` for a context, for example by adding `expirationDate` for *Drugs and Pharmaceutical Products* context, has the effect of implicitly customizing the *Order* and *Catalogue* schemas for the same context. Whenever those document schemas are requested for the *Drugs and Pharmaceutical Products* context or any sub-context such as *Antibiotics*, the customized `Item` version replaces the original `Item` eventually modifying the schema of the relevant documents.

In order to customize a document schema for a particular business context, Document Schema Customization Service traverses the component hierarchy of the schema and gathers applicable custom versions through the Component Discovery Service. A business context value may consist of values from multiple context categories. For such values, Component Discovery is called once for each context

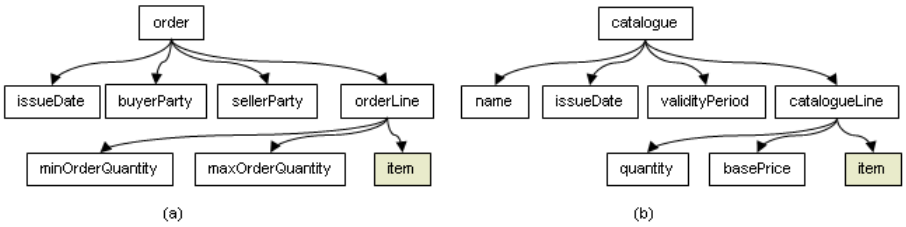


Figure 12. Simplified Order and Catalogue schemas

category and each of those calls may return a different version of the component. For such cases, Component Merge Service is called to merge multiple versions into one and that version is used to replace the standard UBL component in the customized document schema. The algorithm in Figure 13 displays the overall steps of how schema customization is executed.

schema: UBL document schema to customize
ctxSet: Set of context classes denoting the target context

```
1.  customSchema = schema.clone();
2.  for all component cmp ∈ schema
3.      componentList = new List();
4.      for all context class ctx ∈ ctxSet
5.        tmp = discoverComponent(cmp, ctx);
6.        componentList.add(tmp);
7.      end for
8.
9.      if (componentList.size) == 1) then
10.       replaceComponent(customSchema, cmp, componentList[0]);
11.     else if (componentList.size) > 1) then
12.       mergedCmp = mergeComponents(componentList);
13.       replaceComponent(customSchema, cmp, mergedCmp);
14.     end if
15. end for
```

Figure 13. Document Schema Customization Algorithm

4.5. Component Merge Service

Given the multiple versions of the same UBL component, Component Merge Service generates a unified version of the component that represents all different versions.

When a target context value includes multiple categories or has multiple parents or a mixture of both, component repository may contain multiple applicable versions of a particular component. For such cases, Component Merge Service extracts XSD derivations from individual versions, serializes them and successively applies to the corresponding UBL component.

As an example consider the simplified *item* in Figure 14-(a). Assume Figure 14-(b) is a customized version for the *Retail Trade* context that restricts the cardinality of *brandName* from `[0..unbounded]` to `[1..unbounded]` and removes *originCountry* by setting its cardinality to `[0..0]`. Similarly, assume Figure 14-(c) is another customized version for *Drugs and Pharmaceutical Products* context that restricts the cardinality of *brandName* from `[0..unbounded]` to `[0..5]` and extends *item* by adding a new element *ID*. When *item* is required for a context value such as *Drugs and Pharmaceutical Products* and *Retail Trade*, neither version can directly replace the standard *item*. Instead, a merged version, reflecting all derivations from customized versions needs to be generated. The algorithm in Figure 15 displays overall steps of how component merge is executed:

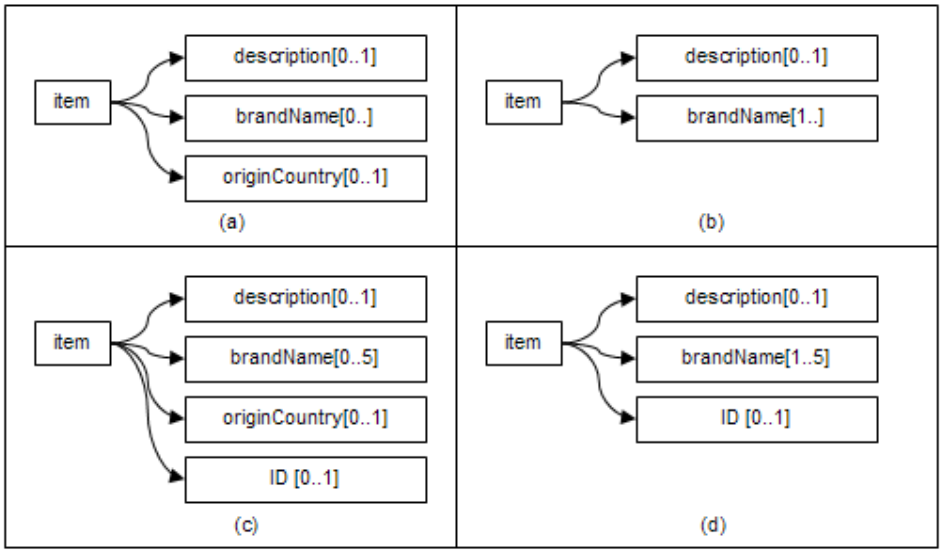


Figure 14. Merging two different versions of Item

When Component Merge Algorithm is called to merge *item* versions in Figure 14-(b) and (c), it executes through the following steps:

UBLcmp: standard UBL component custom versions derived from
versionList: list of custom versions to be merged

```
1.  extensionList = new List();
2.  restrictionList = new List();
3.  for all component version ∈ versionList
4.      extensionList.add(UBLcmp.gatherExtensions(version));
5.      restrictionList.add(UBLcmp.gatherRestrictions(version));
6.  end for
7.
8.  extensionList = eliminateRedundantExtensions(extensionList);
9.
10. if (hasConflictingCardinalities(restrictionList) == true) then
11.     terminate(); // require human assistance for resolving the conflict
12. end if
13.
14. mergedCmp = UBLcmp;
15. derivationList = extensionList ∪ restrictionList;
16. for all derivations d ∈ derivationList
17.     mergedCmp = mergedCmp.derive(d);
18. end for
19. return mergedCmp;
```

Figure 15. Component Merge Algorithm

- For *item* version in Figure 14-(b), it gathers two restriction operations: `brandName->[1..unbounded]` and `originCountry->[0..0]`
- For *item* version in Figure 14-(c), it gathers one restriction operation: `brandName->[0..5]` and one extension operation: `ID->[0..1]`
- These derivation operations are successively applied to the original *item* in Figure 14-(a), generating the version at (d).

Notice that the *item* version in Figure 14-(d) is a valid subset of the one in Figure 14-(a) in terms of XSD validation: that is, any instance document conforming to the schema in (d) also conforms to the original schema in (a) as mandated by the UBL.

Merge algorithm can handle all cases except when there are conflicting cardinalities. The most common case is the one in which one customization eliminates an element by specifying a `[0..0]` cardinality, and another mandates the same element with a `[1..1]` cardinality. When there is no valid intersection of cardinalities specified for an element, versions can not be merged automatically and require human assistance.

Removing Redundancy

Since components are customized by independent domain experts, it is reasonable to expect the creation of similar custom versions of components for different context values. When elements added by one extension are a subset of the elements added by another extension, merely merging such components would result in redundant elements in the resulting types. As an example, consider `Address` in Figure 16 and assume a type gets extended for one context by adding `Address` and another by adding `City` and `Country`. Both `City` and `Country` are included in `Address`, so there is no need to merge these versions; latter can be discarded.

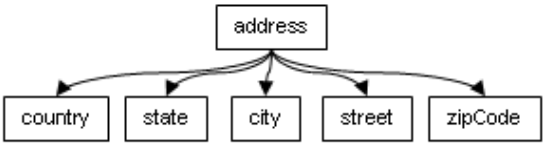


Figure 16. Address Component

In order to resolve redundancy, merge algorithm compares elements subject to extension operations with each other. The ones subsumed by others in respective type hierarchies are removed from `extensionList` before proceeding further.

5. Implementation Status

We provide a prototype implementation using the Java Programming Language and several third party libraries: XSD parsing and validation is accomplished using Apache Xerces [2], OWL-API [17] is used for accessing OWL-DL related functionality and Pellet [18] is used for reasoning.

Specialized converters are implemented in order to develop context ontologies representing classifications such as NAICS, NACE. Each of these converters parses a particular classification and generate corresponding OWL descriptions. Generated ontologies are available at [15].

Ontology development and testing is done using the Protege [20]. Prompt plugin for Protege [19] is used for accomplishing ontology alignment. Protege requires an external Description Logic reasoner for processing ontologies and RacerPro [21] is used for this purpose.

In order to allow users access to implemented functionality, various GUI tools are developed. Figure 17 displays a snapshot of the Component Customization Tool. The interface of this tool consists of a tabbed context browser to be used for specifying the context and a component browser to be used for specifying the UBL component to customize. Lower right text area is for entering XSD derivation operations. In Figure 17, the `ItemType` is being customized for the

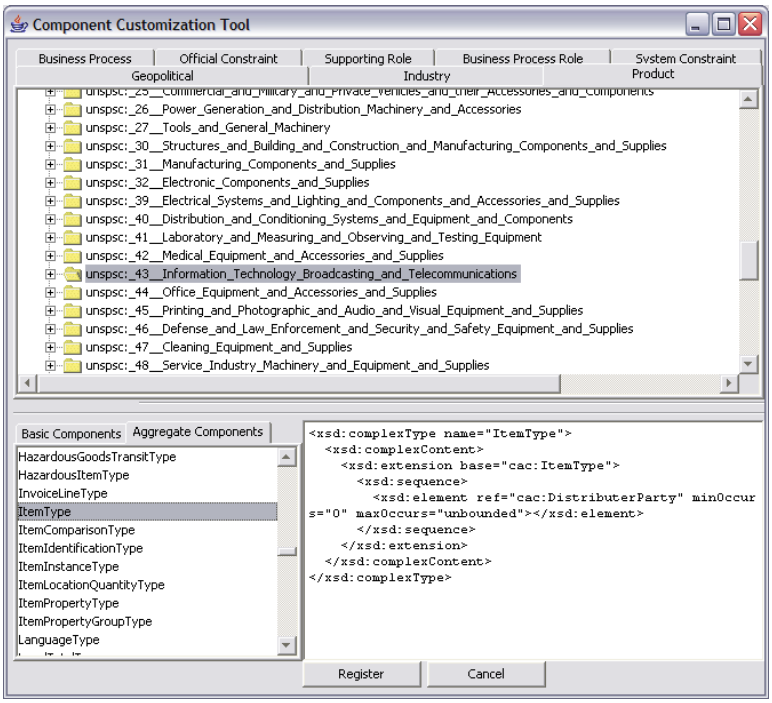


Figure 17. Component Customization Tool

unspsc:Information Technology, Broadcasting and Telecommunications context by adding the `DistributerParty` element.

Figure 18 displays a snapshot of the Document Schema Customization Tool. The interface of this tool is composed of a selectable list for specifying a UBL document schema and a context browser to be used for specifying values for individual context categories to compose the multi-category business context of the document schema. In Figure 18, the `Order` document schema is being customized for the *unspsc:Computer Equipment and Accessories*, *isic:Manufacturing* and *geo:Europe* context.

6. Related Work

There is a considerable amount of research about developing suitable context models to facilitate context sharing and interoperability of applications. [27] provides a survey on the subject and classifies existing context modeling approaches based on the schema of data structures they use to exchange contextual information: key-value models, markup scheme models, graphical models, object oriented models, logic based models and ontology based models. The work also identifies require-

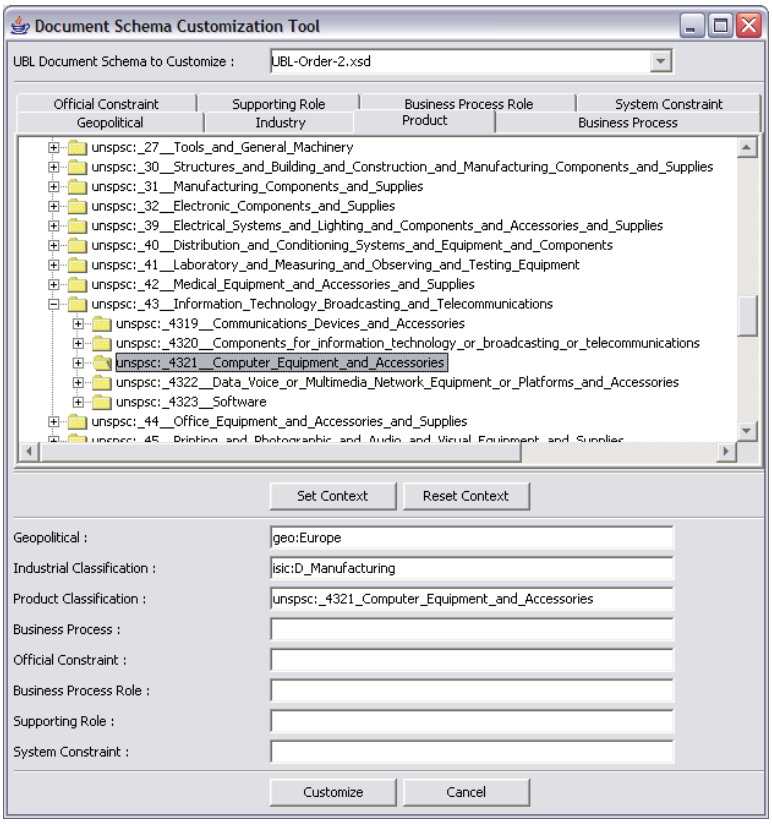


Figure 18. UBL Document Schema CustomizationTool

ments that a context modeling approach needs to meet in order to support ubiquitous computing: distributed composition, partial validation, richness and quality of information, incompleteness and ambiguity, level of formality and applicability to existing environments. Finally a comparative evaluation of the context modeling approaches based on identified requirements is provided, in which it is concluded that the most promising approaches are ontology based models.

One of the ontology based context modeling approaches is *Context Ontology Language (CoOL)* [28], which provides a uniform way for specifying the model's core concepts as well as an arbitrary amount of subconcepts and facts using ontologies. The contextual knowledge is evaluated using ontology reasoners.

Another ontology based context modeling approach is *Context Ontology (CONON)* [36], in which authors define an upper ontology to describe entities and concepts like location, user, activity, computational entity which they find most fundemental to represent context information as a part of their work to develop a context aware

service infrastructure. Then, they provide a quantitative analysis of the feasibility of using ontologies in context aware applications.

Both of the work provide general context models to address context aware service infrastructures and hence are different from UBL defined context domains for electronic business documents. Furthermore CONON develops an upper ontology whose lower levels have to be provided depending on the applications. In our case, we create ontologies from existing industrial taxonomies which are already used for core components. Then, we relate ontologies to one another through ontology alignment since this is necessary to facilitate the automated discovery of components.

7. Conclusions and Future Work

UBL specification provides a set of document schemas in an effort to standardize electronic business documents and is rapidly being adopted by both public and private sector organizations around the world. UBL also provides a customization mechanism to allow the tailoring of UBL schemas and components in response to user needs operating in different industry, geopolitical and regulatory contexts.

Even though this customization mechanism provides syntactic interoperability, it does not address semantic interoperability, that is, how users and applications can discover and reuse component and schema customizations provided by other users.

In this paper, we describe how a semantic formalization based on OWL ontologies can be developed to represent context domains and how these ontologies can be utilized to provide semantic interoperability for UBL. The innovative aspects of the paper are as follows:

- Taxonomies have been widely used by the industry. By developing converters to turn them into ontologies, taxonomies become machine processable and it becomes possible to formally express the correspondences between similar concepts in different ontologies through ontology alignment. Furthermore, reasoners can process ontologies to infer implicit relationships between concepts, that is, given a minimum sufficient set of relationships between ontologies, reasoning provides the complete set. This allows ontologies and in turn the taxonomies they represent to be semantically interoperable, that is, it becomes possible to use their content interchangeably.

- Using this capability for annotating custom UBL components allows us to express the context that components can be used in a semantically interoperable way. Together with a repository to store custom components and a knowledge base about repository content, this allows the development of automated processes that can intelligently search through custom components and are capable of schema customization.

- Automated customization capability leads to the ability to merge multiple versions of a component to generate additional versions, that is, given component

versions customized for individual context categories, new versions applicable to combinations of those categories can be automatically generated as needed.

As future work, we plan to develop the necessary mechanisms to express the semantics of components included in UBL document schemas. The ability to formally express the semantics of UBL components will lead to improvements in the quality of component merging activity, as it will allow identifying the components representing semantically similar concepts, hence the ability to avoid semantic redundancy in the merged components.

More important than that, the component semantics together with the context formalization and the ability to interpret context values presented in this paper lays the foundation for the development of a translation capability between documents conforming to different versions of a UBL schema customized for different business context values.

References

1. ANSI Accredited Standards Committee (ASC) X12, http://www.x12.org/
2. Apache Xerces, http://xerces.apache.org/
3. Bernstein, P.A., Levy, A.Y., Pottinger, R.A., "A Vision for Management of Complex Models", Microsoft Research Technical Report MSR-TR-2000-53.
4. Brun, M.H., Langg, C., "From 18 to 100 million UBL messages with a Service Oriented Infrastructure", In Proceedings of the XML Conference 2006, Boston USA. http://2006.xmlconference.org/proceedings/36/presentation.pdf
5. Central Product Classification Version 1.1, http://unstats.un.org/unsd/cr/family2.asp?Cl=16
6. Electronic Business using eXtensible Markup Language (ebXML), http://www.ebxml.org/
7. Fridman Noy, N. and Musen, M. A. 1999. An Algorithm for Merging and Aligning Ontologies: Automation and Tool Support. In Proceedings of the Workshop on Ontology Management at the Sixteenth National Conference on Artificial Intelligence (AAAI-99). Orlando, FL: AAAI Press.
8. Guidelines For The Customization of UBL v1.0 Schemas, http://docs.oasis-open.org/ubl/cd-UBL-1.0/doc/cm/wd-ubl-cmsc-cmguidelines-1.0.html
9. International Organization for Standardization, ISO 3166 Maintenance Agency http://www.iso.org/iso/en/prods-services/iso3166ma/index.html
10. International Standard Industrial Classification of All Economic Activities, Revision 3.1, http://unstats.un.org/unsd/cr/family2.asp?Cl=17
11. McGuinness, D., Harmelen, F., "OWL Web Ontology Language Overview", W3C Recommendation, February 2004, http://www.w3.org/TR/owl-features/
12. Medjahed1 B., Benatallah B., Bouguettaya1 A., Ngu A.H.H., Elmagarmid A.K., "Business-to-business interactions: Issues and Enabling Technologies", The International Journal on Very Large Data Bases, Vol. 12, No. 1, May 2003, pp 5985.
13. North American Industry Classification System, http://www.census.gov/epcd/www/naics.html
14. North American Product Classification System (NAPCS), http://www.census.gov/eos/www/napcs/napcs.htm
15. Ontologies generated from standard classifications, http://www.srdc.metu.edu.tr/ubl/contextOntology/
16. OWL Web Ontology Language 1.0 Reference http://www.w3.org/TR/2002/WD-owl-ref-20020729/ref-daml
17. OWL API, http://owl.man.ac.uk/api.shtml

18. Pellet OWL Reasoner, http://www.mindswap.org/2003/pellet/
19. Prompt plugin for Protege, http://protege.cim3.net/cgi-bin/wiki.pl?Prompt
20. Protege Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/
21. RacerPro: Renamed ABox and Concept Expression Reasoner, http://www.racer-systems.com/products/racerpro/index.phtml
22. RosettaNet, http://www.rosettanet.org/
23. Staab, S., Studer, R., Handbook on Ontologies, Springer, 2004.
24. Statistical Classification of Economic Activities in the European Community, Rev. 1.1, http://ec.europa.eu/comm/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_1_1&StrLanguageCode=EN
25. Statistical Classification of Economic Activities, Complete List and Correspondending ISIC-Classes, http://www.fifoost.org/database/nace/naceen_2002c.php
26. Statistical Classification of Products by Activity in the European Economic Community, 2002 version http://ec.europa.eu/comm/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=CPA&StrLanguageCode=EN
27. Strang T., Linnhoff-Popien C., "A Context Modeling Survey", Workshop on Advanced Context Modelling, Reasoning and Management associated with the Sixth International Conference on Ubiquitous Computing, 2004.
28. Strang T., Linnhoff-Popien C., Frank K., "CoOL: A Context Ontology Language to enable Contextual Interoperability", LNCS Distributed Applications and Interoperable Systems, Vol. 2893, 2003, pp 236-247.
29. Svefaktura (SwedInvoice), http://www.svefaktura.se/SFTI_Basic_Invoice20051130_EN/index.html
30. The Electronics Freight Management White Paper, http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/14246_files/14246.pdf
31. UN/CEFACT ebXML Core Components Technical Specification, http://www.unece.org/cefact/ebxml/CCTS_V2-01_Final.pdf
32. United Nations Directories for Electronic Data Interchange for Administration, Commerce and Transport (UN/EDIFACT), http://www.unece.org/trade/untdid/welcome.htm
33. Universal Business Language v2.0, http://docs.oasis-open.org/ubl/cs-UBL-2.0/UBL-2.0.html
34. Universal Standard Products and Services Classification (UNSPSC), http://www.unspsc.org
35. XML Common Business Library (xCBL), http://www.xcbl.org
36. Wang X.H., Zhang D.Q., Gu T., Pung H.K., "Ontology based context modeling and reasoning using OWL", Context Modeling and Reasoning Workshop at PerCom 2004, pp 18-22.

Energy-efficient Real-time Coordination and Routing in Wireless Sensor and Actor Networks

Ghalib A. Shah[†], Muslim Bozyiğit[‡]

[†]National University of Science and Technology, Rawalpindi
[‡]Middle East Technical University, Ankara
Email: [†]ghalib@ceme.edu.pk, [‡]bozyigit@ceng.metu.edu.tr

***Abstract*—In Wireless Sensor Actor Networks (WSAN), sensor nodes perform the sensing task and actor nodes take action based on the sensed phenomena. Generally, the network is configured to observe multiple events and to react upon each individual event accordingly. If sensor nodes are embedded with different sensing modules and energy levels, they are considered as heterogeneous. Time delay and energy efficiency are two important aspects of WSAN that require special attention. To ensure efficient and reliable operations of such heterogeneous WSAN, new communication protocols are imperative. In WSAN, a reliable operation is assumed as long as the events are reported to the appropriate actors within the event specific response time and the actors then react in the event region accordingly.**

In this paper, we propose an integrated energy-efficient real-time coordination and routing (*ERCR*) framework for WSAN. Sensor nodes coordinate with each other to form clusters which are adaptive to the energy and event reporting rates. Once the clusters are formed, cluster-heads coordinate with the actors to form *dynamic responsibility clusters*. ERCR ensures in-time event delivery in the *responsibility cluster* by presenting a delay-constrained energy aware routing (DEAR) protocol. It incorporates a novel real-time data aggregation (RDA) approach. In RDA, cluster-heads perform aggregation such that the packets deadlines are not affected. RDA is adaptive to the traffic conditions and provides fairness among the farther and nearer cluster-heads. In order to provide reliable action, actors coordinate with each other to respond to the event such that the whole field is targeted by the limited number of mobile actors. However, if an event region does not lie within the action coverage of any actor then ERCR ensures the action by moving an actor toward such region. Simulation experiments prove that ERCR achieves energy efficiency and honor the application-specific delay bound.

I. Introduction

Typically, the architecture of wireless sensor actor network (WSAN) consists of sensors which sense the phenomena, a sink that collects the data from the sensors to process and actors that act upon the commands sent by the sink. In the literature, such architecture is known as *semi-automated architecture*. An architecture in which sensor nodes send information to the actor nodes directly without the involvement of the sink node is called an *automated architecture* [1]. The communication path in *semi-automated architecture* introduces significant delay, which may not be acceptable for delay-sensitive applications. For example, consider a military application where sensors in the battlefield detect the movement of enemy forces and send the information to the sink employed part of a command and control center. The sink then sends action commands to actors for taking some measures in the reported area. In this case, unnecessary delay is introduced due to sensor-sink communication, which could be prevented if the actors can take localized actions without the involvement of the sink. The emergence of this new class of network raises several issues.

First, the most essential requirement at any level of communication protocol in WSANs is the energy conservation. With the collaboration of sensor nodes or sensor-sensor coordination, the event data can be efficiently reported to the interested clients. Second, WSAN imposes timing constraints in the form of end-to-end deadlines due to the active role of actors. Third, in order to reduce the communication delay, event reporting policy adopted in WSNs should be modified such that packets are routed directly to the actors rather than the sink node to support *automated architecture*. This requires sensor-actor coordination so that the event is reported to an appropriate actor capable of responding to it. Fourth, an actor receiving sensor readings may not always be the right choice to respond. It is either due to inaccessible region or overlapping action coverage of two or more actors. The former case is possible because actors have certain action range and actor-actor coordination is required to relocate one of the actors toward such regions. In the latter case, actor-actor coordination is required in order to balance the load or selecting the actor which can best perform in overlapping region. Intuitively, actor-actor coordination is imperative for the reliable and efficient response to the event. Fifth, although the automated architecture is advocated to be implemented in WSAN but it exposes the event data to the intruder. This would require some distributed security mechanism to authenticate both the sensors ans actors.

There have been considerable efforts to solve the nodes configuration and routing problem in wireless sensor networks [10], [13], [25], [8], [26], [17], [27]. However, these protocols do not consider the heterogeneity of WSAN. Moreover, none of these protocols provide sensor-actor/actor-actor coordination. Recently, coordination mechanisms [4], [5], [6], [7] are proposed for WSAN. The framework in [4] is an event-based reactive model of clustering. Cluster formation is triggered by an event so that clusters are created on-the-fly. The in-time packet delivery in terms of reliability is operated by the actor nodes. Therefore, sensor nodes react slowly to late

An earlier version of this paper appeared in [2].

traffic waiting for the feedback from the actor nodes in order to speed-up the delivery. Hence, the coordination framework is not suitable for time-critical events. Moreover, cluster to actor routing is done using greedy geographical approach. A packet forwarding node finds the next hop node according to the greedy approach failing to do so results into a packet loss as the packets enters into a void region. Since the work assumes that the network is dense therefore it does not propose any void region prevention or recovery mechanism.

In [5], the sensor field is divided into maps, where each map is represented by a sensor node which detects an event the earliest. For building a map, nodes which have detected an event but not reported yet, flood the event detection message. It applies aggregation hierarchically in the map and the representative node of the map collects data from all the nodes and reports it to the actuator. This is also an event-driven nodes configuration and aggregation that introduce high latency. This approach becomes more inefficient in terms of delay and energy when the event center moves frequently that requires rebuilding maps in the field. Moreover, it does not provide any mechanism to route data according to the specific response time i.e. no support of real-time routing. Delay energy-aware routing (DEAP) [6] is a geographical-based routing protocol proposed for WSANs that conserves the energy of nodes by an adaptive energy management scheme. It switches the node to sleep mode if the queue is empty and back to active mode when the packets are buffered. It extends the nodes active duration when there are buffered packets to avoid from latency. However, there is no explicit in-time delivery model to support real-time traffic and also does not provide any actor-actor coordination mechanism. Consequently, there exists no unified solution, which addresses the real-time routing, sensor-actor/actor-actor coordination for heterogeneous WSANs.

To address the above issues, we present energy-efficient real-time coordination and routing (ERCR) framework which provides an integrated solution in order to achieve an ultimate goal of timely reaction to the environment. The novelty of ERCR is forming clusters for nodes configuration such that the procedure is not only adaptive to energy of nodes (a common case) but also to their reporting rate and density in a heterogeneous environment. The energy is further conserved by presenting a novel real-time aggregation technique, which does not affect the deadlines of the packets. On the other hand, real-time routing and the coordination (sensor-actor/actor-actor) framework enhances the reliability of data delivery and in turn event response. A brief overview of the ERCR framework is provided in Section II-B.

The remainder of the paper is organized as follows. In Section II, we briefly present the solution to the above mentioned issues in WSAN. Energy efficient configuration of sensor nodes is described in Section III. We discuss the real-time routing and aggregation approach as well as sensor-actor/actor-actor coordination in Section IV. Performance evaluation and results are presented in Section V. Finally, the paper is concluded in Section VI.

II. ERCR Overview

In this section, we present the definitions and basic assumptions taken in the network model prior to proceeding for the discussion on framework and then contributions of ERCR framework are described.

A. Network Model

In this paper, we focus on a network model where a number of actors with heterogeneous capabilities are deployed in a sensor network field. For instance, while some actors are responsible for responding to sudden temperature changes, others might get responsibility of fire extinguishing. On the other hand, the number of sensor nodes is much higher in comparison to actor nodes and it is expected that an actor serves multiple sensors. Unlike sensors with limited resources, however, actors have abundant resources with no energy constraint. Actors can, therefore, use long range transmission to send control beacons (e.g. location announcement message) unlike the sensors that rely on hoping to report events. The IEEE 802.15.4 task group [34] has developed standard also known as ZigBee for low cost low power devices that best characterizes the wireless sensor and actor networks. According to ZigBee standard, sensor nodes can have transmission range from 10-100 meters with data rate capability of 250Kbps.

We assume that the sensor nodes have *a priori* knowledge of the actors action range. Action range represents the area in which an actor can reliably execute actions and has no relationship with the transmission range. There is no assumption about a particular application scenario. However a class of applications, in which real-time response is the primary objective, is the main focus of this framework. Similarly there is no bound on the density of nodes that may vary in different parts of the field. Links are symmetric, i.e., two nodes n1 and n2 can communicate using the same transmission power level.

In the network configuration, a *cluster* is a group of neighboring nodes, constructed in such a way that one of them gathers the readings observed by the other nodes and is responsible to relay them to the information sink or actor. The node responsible of collecting the data locally is known as *cluster-head* and the others are called *members* of the cluster. Whereas, under some circumstances, there may exist some nodes which have not joined any group or cluster are referred as *dangling nodes* (DN). One or more clusters and an actor form *responsibility cluster*, where the clusters report any event occured in their sensing area to the actor which is responsible of executing action in that area. A sensing region is said to be covered if one or more actors can respond in that region otherwise uncovered. We assume that the sensor nodes are at stationary positions in contrast to the actors, which are mobile and can move towards the sensors on occurrence of an event if the sensing region is uncovered.

B. ERCR Contributions

The contributions of ERCR are defused amongst different components of the framework; sensor-sensor, sensor-actor and actor-actor coordination as well as real-time aggregation and

routing. However, the security issue is not addressed in the framework and applications can implement any of the security algorithms proposed in the literature. In the following sections, we briefly present the novel solutions to the aformentioned issues in the form of incorporated components.

1) Efficient sensor-sensor coordination: ERCR provides a cluster-based configuration so called a multi-event adaptive clustering protocol (MEAC) in which sensor nodes coordinate with each other to form clusters as described in Section III. Among a group of nodes, a node is elected as a cluster-head which has higher energy level but lower traffic rate. Generally, the clustering protocols [17], [20] focus on the currently available energy of the nodes and periodically reorganize clusters to do energy balancing. However, this strategy is not practical when the nodes are sending traffic at different rates due to different events characteristics. If all the nodes have the same probability to become cluster heads then the nodes reporting events at higher rate will eventually loose their energy earlier than the others resulting into smaller network lifetime. Therefore, the load on a node, that is, its data rate is used as a key factor for cluster-head selection. Hence, MEAC distributes the energy usage of nodes by adapting to the residual energy and multiple events in the field in order to increase the stable period of the network.

2) Real-time routing: ERCR achieves the event specific delay bound τ of packets through the delay-constrained energy aware routing (DEAR) protocol presented in Section IV-A. The value of τ may differ for different events and DEAR is adaptive to consider the deadline of individual packets. The DEAR protocol exploits the hierarchical cluster-based configuration of nodes to relay the packets. Unlike greedy mode routing protocols [4], [5], [6], the selection of forwarding node in ERCR is based on the packet delay as well as the balanced energy consumption of sensor nodes. The greedy mode routing protocol becomes very cumbersome and incurs extra overheads for irregular or non-uniform deployment of sensor nodes requiring recovery mode operations in void regions. Cluster-heads in ERCR take localized decision about the in-time packet delivery unlike the coordination framework [4] in which sensor nodes wait for the actors to speed up the routing. Hence, DEAR responds to the late traffic locally and does not suffer in void region problem.

3) Data aggregation: It is argued that aggregation extends the queuing delay at the relaying nodes and becomes more challenging due to coexistence of mixed-type traffic resulting from concurrent events. ERCR incorporates a very simple yet practical real-time data aggregation (RDA) method. In RDA, only cluster-heads involve to aggregate data from their member nodes and keeps aggregating data as long as the end-to-end delay constraint is not violated. Moreover, it does not require building aggregation trees [28], [29], [30] since the nodes have already been configured in the form of clusters. It is worthwhile to note that RDA achieves the fairness implicitly since nearer cluster-heads have more time for aggregation due to shorter path delay and thereby reduce the traffic significantly near the destination.

4) Coordination Framework: In sensor-actor coordination, only the cluster-heads are responsible of coordinating with the actor nodes. A cluster-head selects an actor for coordination which can react in the clustered region and is attending least number of sensor nodes as explained in Section IV-B.

Actir-actor coordination considers two scenarios; overlapping region and inaccessible region. Generally, the existing studies [4], [5] consider only the first scenario and provide a solution to react in overlapping region. They have not discussed actor-actor coordination from the perspective of inaccessible region. However, ERCR moves an actor receiving request toward the inaccessible region if possible or trigger actor-actor coordination given in Section IV-D to relocate some other actor in that region. Hence, in ERCR, actors initiate A-A coordination at both occasions that enables the actors to execute certain actions no matter if the reporting region is in their range or not.

III. Clustering of Sensor Nodes in ERCR

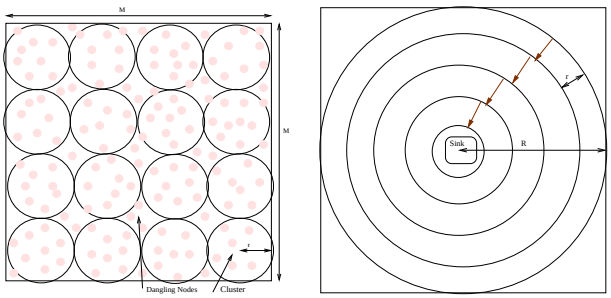
ERCR configures the sensor nodes hierarchically in the form of clusters to achieve energy effiency by using novel multi-event adaptive clustering (MEAC) protocol. MEAC conserves energy by considering three design factors; (1) limiting the number of clusters (*k*) in the network, (2) electing an appropriate node to function as a cluster-head, and (3) reducing the frequency of clusters reformation. To limit the number of clusters, it calculates the optimal number of clusters by using the given topology information i.e. dimension of the field, number of sensor nodes and their transmission radius. Nodes are then restricted to form approximately *k* number of clusters. Second, cluster formation is based on the weighting equation which includes the nodes energy, data rate and density of nodes. These parameters are given weight according to the application needs. Nodes exchange their aggregated weights with the neighbors and a node announces itself as a cluster-head subject to the highest aggregated weight. Cluster formation is not triggered periodically rather it defines threshold on weight and a cluster-head is withdrawn only if its weight goes down to threshold.

A. Optimal Number of Clusters (k_{opt})

MEAC computes optimal number of clusters (k_{opt}) such that it decreases the energy consumption, while providing high degree of connectivity. We devise a formula to find k_{opt} that is directly proportional to total number of nodes and the area of the network, while it is inversely proportional to the transmission range.

Let r be the transmission radius of each node regardless of its functioning. In the clustering process, there is some probability that a number of dangling nodes [1] (DN) may exist due to the deployment of nodes or coverage of the elected cluster-head. To find out the probability of such nodes, we map the sensor field ($M \times M$) to non-overlapping circles of radius r as shown in Fig 1(a) and assume that the nodes lying outside the boundary of the circle are DN nodes and the others are member nodes. These DN nodes affiliate to cluster-heads through the nodes inside the circle (member nodes) and

[1]The nodes which have not joined any cluster are referred as *dangling nodes*.



(a) (b)

Fig. 1. Network model to formulate the optimal clusters. Fig 1(a) represents the model to find the probability of DN nodes. Fig 1(b) illustrates the model of routing packets from cluster-heads to the sink node.

become the multi-hop members of cluster-heads. The square field M^2 can be packed by $M^2/(2r)^2$ non-overlapping circles of radius r. Thus, the probability P_{DN} of a multi-hop member is

$$P_{DN} = \frac{M^2}{(2r)^2} \times \frac{(2r)^2 - \pi r^2}{M^2} \approx 0.214$$

Let E_{elec} be the energy consumed by the electronic circuitry in coding, modulation, filtration and spreading of the signal. Whereas, $\epsilon_{amp} r^2$ is the energy consumed for signal amplification over a short distance r. Thus, the energy consumed by each member node is

$$E_{Member} = l(E_{elec} + \epsilon_{amp} r^2 (1 + P_{DN}))$$

where l is the size of data packet. The above equation can be simplified by taking the area as circle given in (16) of [18].

$$E_{Member} = l(E_{elec} + \epsilon_{amp} \frac{M^2(1 + P_{DN})}{2\pi k})$$

Let us assume that the sensory field is covered by a circle of radius R, where the sink node lies at the center of this circle as shown in Fig 1(b). This assumption is made for sending packets from cluster-heads to the sink. Cluster-heads do not extend their transmission range and, therefore, has the same radius r as member nodes. This adapts the multi-hop model proposed by [16] to route packets from cluster-head to the sink.

In the model, a circle is divided into concentric rings with the distance of r. The energy spent to relay the packet from outside ring toward inside ring is $l(2E_{elec} + \epsilon_{amp} r^2)$. The number of hops H_{CH-S} required to route packet from cluster-head to sink node can be calculated by $\frac{R}{r}(1 - P_{hops})$, where P_{hops} is the probability indicating the distance in terms of hops to the sink. This probability can be calculated by using the nodes distribution in the rings given in [16].

$$P_{hops} = \frac{r}{R} \sum_{i=1}^{R/r} \frac{R^2 - (ir)^2}{M^2}$$

Packets from cluster-heads that are far from the sink are relayed through intermediate nodes. Therefore, if N_s is the number of neighbors of the source node s then $N_s \times E_{elec}$ is the energy consumed by the electronic circuitry of the neighbors during the transmission of a data packet by s. The number of neighbors N_s of a node can be computed by $n\frac{\pi r^2}{M^2}$. Hence, the energy consumed in forwarding data from cluster-head to sink is measured as:

$$\begin{aligned} E_{CH-S} &= l(N_s E_{elec} + E_{elec} \\ &+ (2E_{elec} + \epsilon_{amp} r^2 + N_s E_{elec}) H_{CH-S}) \end{aligned}$$

The total energy dissipated by the network is

$$\begin{aligned} E_{total} &= l((n + nN_s)E_{elec} \\ &+ k(2E_{elec} + \epsilon_{amp} r^2 + N_s E_{elec}) H_{CH-S} \\ &+ n\epsilon_{amp} \frac{M^2(1 + P_{DN})}{2\pi k}) \end{aligned}$$

For $r < R$, the optimal value of k can be found by taking the derivative of above equation with respect to k and equating to zero

$$k_{opt} \approx \sqrt{\frac{n(1 + P_{DN})}{(2\pi(1 + \frac{2E_{elec}}{\epsilon_{amp} r^2} + \frac{N_s E_{elec}}{\epsilon_{amp} r^2})) H_{CH-S}}} \times \frac{M}{r} \quad (1)$$

The optimal value depends on the transmission range r. For long range transmission, the optimal number of clusters k_{opt} is small. For example, let $n = 100$, $M = 100$ and the sink is at the center of the field $(x = 50, y = 50)$. Then, the value of radius R is obtained by drawing a circle at $x = 50, y = 50$ to cover the field. The estimated value is $R = 60$ and let set the range r of individual nodes to 25. In this scenario, we obtain the value of $k_{opt} \approx 10$. By increasing the range of nodes to 40 meters, we obtain $k_{opt} \approx 7$, which is 10 in SEP [20].

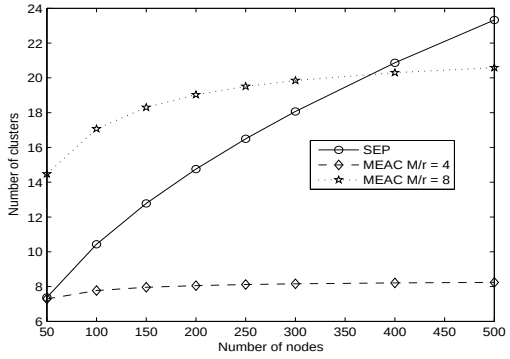


Fig. 2. Cluster formation in SEP and MEAC for different values of number of nodes.

Generally, a cluster is formed among a group of sensor nodes which hear the transmission of each other. The size of a group is large at higher density that results in a small number of clusters. Conversely, the size of cluster is small at lower density because a lesser number of nodes hear each other that results in a large number of clusters. The nodes density (μ) can be computed as

$$\mu(r) = \frac{n\pi r^2}{M^2}$$

For a given number of nodes n, μ varies by changing either the dimension of the field or the transmission radius.

If the nodes are deployed in a small area, then the density increases and decreases if the same number of nodes deployed in relatively large area. On the other hand, μ increases by increasing the transmission radius r and vice versa. However, for a given dimensions of the field and transmission radius, an increase in number of nodes increases only the group size. Eventually, the number of formed clusters remains the same. Fig. 2 shows the adaptiveness of MEAC to different density of nodes. At higher density for $100 \times 100\ m^2$ field and $r = 25\ m$, MEAC forms lesser number of clusters than SEP regardless of the number of nodes deployed. However, the number of clusters increases at lower density by setting the dimensions to $200 \times 200\ m^2$. It is shown in Fig. 2 that SEP is not adaptive to dimension of field and transmission radius but the number of clusters varies only with the number of deployed nodes.

It is proved in [11] that the number of clusters (k) is higher at lower cluster radius and lowers at higher cluster radius. Consequently the residual energy of cluster-heads is higher at higher cluster radius or lower number of cluster-heads. On the other hand, a very large value of radius resulting in very small value of k nearly approaching to 1, may converge to a network of single cluster that may become a bottleneck. It is therefore necessary to keep the optimal value k_opt.

B. Cluster Size

When the deployment is uniform, the optimal number of member nodes (N_{opt}) can be found by n/k_{opt} and restrict the clusters to not grow beyond this value. In practice, it is difficult to deploy the nodes uniformly and the number of potential member nodes vary greatly depending on the density in that zone. In order to adapt to the non-uniform deployment so that number of optimal clusters are ensured, we compute the minimum and maximum limits N_{Min} and N_{Max} respectively on the size of cluster. Each cluster-head limits its size according to these values resulting in approximately k_{opt} clusters. Let N_i be the number of neighboring nodes of any ith node. $Max(N_i)$ is the maximum number of neighboring nodes that any of the ith neighbor node have. We measure density of nodes in a particular zone by comparing the neighbor nodes N_i with N_{opt}. It can be concluded that the deployment is:

$$\begin{aligned} dense \quad if \quad N_i/N_{opt} \quad &> 1 \\ uniform \quad if \quad N_i/N_{opt} \quad &\approx 1 \\ sparse \quad if \quad N_i/N_{opt} \quad &< 1 \end{aligned}$$

Therefore, the number of nodes in a cluster can be constrained by setting the lower bound N_{Min} and upper bound N_{Max} as shown in Fig. 3 according to the deployment as:

$$N_{Max} \quad = \quad Max(N_{opt}, Max(N_i))$$

That is, the maximum of N_{opt} and maximum number of neighbors of any cluster-head at the time of cluster formation.

$$N_{Min} \quad = \quad N_{opt} \times Min(N_{opt}, Max(N_i))/N_{Max}$$

These limits allow the configuration to manage the dense as well as sparse deployment of nodes.

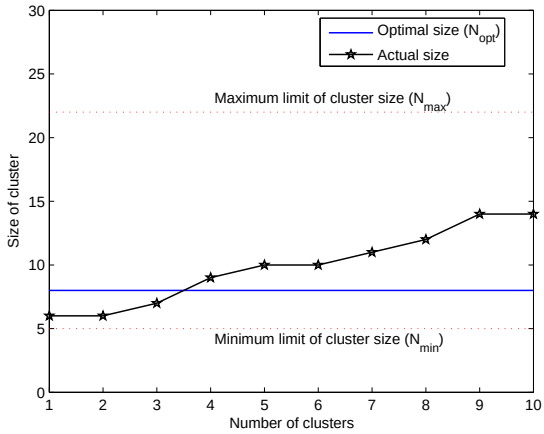


Fig. 3. Size of clusters in the network of 100 nodes deployed in $100 \times 100\ m^2$ field.

C. Cluster Formation

The cluster-head election procedure is based on calculating weight for each sensor node in the sensor field and a sensor node is elected as a cluster-head that has the maximum weight. We define weight threshold of the cluster-head to rotate the cluster-heads responsibility among all the potential nodes. A cluster is not strictly organized to 1-hop but it accepts the membership of a node that could not reach any cluster in the first phase of cluster formation. Therefore, a cluster can include d-hop members, for $d \geq 1$. Although the operations of the protocol starts after the first phase of cluster formation, there may still exist some DN nodes.

1) Cluster-head Election Procedure: Cluster-heads are elected by effectively combining the required system parameters with certain weighting factors. Every node calculates its weight based on its available power, data rate and the density of nodes. Let D_i be the average distance of node i to its neighbors, N_i be the total number of its neighbors, E_i be its available energy and ρ_i be its reporting rate. Node i computes its weight W_i as:

$$W_i = c_1 \frac{\rho_o}{\rho_i} \times \frac{E_i}{E_o} + c_2 \frac{D_i}{r} \times \frac{N_i}{N_{opt}} \tag{2}$$

where E_o and ρ_o represent the lowest energy level and data rate of the nodes respectively due to nodes heterogeneity. The coefficients c_1, c_2 are the weighting factors for the energy and data rate parameters. Node i announces itself a cluster-head if its weight is high among all its neighbors and sets its threshold $W_{Th} = c_w W_i$, where c_w is the threshold adjusting factor. It can be computed as

$$c_w = \frac{E[W]}{W_i} \tag{3}$$

where $E[W]$ is the mean weight of the neighbors of node i.

It is quite possible that a node receives announcement from multiple cluster-heads. However, it decides its membership according to the closeness and joins the cluster whose cluster-head is closer than the others. The pseudo-code of the op-

```
/*Pseudo-code executed by each node i in
each round.*/
ElectClusterhead(i)
//maximum weight among neighbors.
   W_max = 0;
//N_i is the list of neighbors of i.
   for(each j ∈ N_i)
     if (W_max < W_j)     W_max = W_j;
   end //for
   W_i = my_weight();
   if (status == NONE)
     if (W_i > W_max)
       announce_head();
       W_th = W_i × threshold_factor;
     end //if
   else if (status == HEAD)
     if (W_i < W_th)
       if (W_i < W_max)
         withdraw_head();
       else
         W_th = W_i × threshold_factor
       end //if
     end //if
   end //if
```

Fig. 4. Cluster-head election procedure.

erations executed by a sensor node in each round of cluster formation is reported in Fig. 4.

Values of the coefficients can be chosen according to the application needs. For example, power control is very important in CDMA-based networks, where the weight of the power factor can be increased. On the other hand, the distance weighting factor can be made larger if the density of nodes is high or the deployment is made in hostile environment. This ensures that a node is elected as a cluster-head that can receive the transmission from farther nodes and the number of clusters formed remains close to the optimal value.

It can be noted in (2) that a node having higher energy level than its neighbors is the potential candidate of becoming cluster-head. However, a node reporting events at higher rate is less likely to be elected as cluster-head. The weighting equation includes the ratio of data rate (ρ_o/ρ_i) to consider the multiple data rates due to different events. If a node i has $\rho_i > \rho_o$ then its weight is reduced. Therefore, it has lesser chances to become cluster-head than the other nodes. It is due to the fact that nodes sending packets at higher rates exhaust their energy soon and, thereby, the probability of becoming cluster-heads is reduced.

In order to achieve the goal of energy saving, it minimizes the frequency of cluster reformations. It is achieved by encouraging the current cluster-heads to remain cluster-heads as long as possible. MEAC reduces the frequency of clusters reformation by setting weight threshold at the time of cluster formation. In each round, each cluster-head recomputes its weight and compares with its threshold value. If W_i of cluster-head i is higher than its W_{Th} value then it keeps functioning as head. if $W_i < W_{Th}$ then it checks whether its W_i is also lower than any of its member node weight. If so, it withdraws itself from being cluster-head and cluster election procedure is initiated.

D. Inter-Clusters Connectivity

Packets may be routed through intermediate clusters since the sink or actors can be multi-hop away from the source clusters. As a result, clusters are linked with each other to provide multi-hop cluster routing. The link is provided either through some intermediate member nodes or they can directly hear each other. Member nodes, connecting the two cluster-heads, act as *gateways*. It is quite possible that the cluster-heads are connected through multiple *gateways*. However, cluster-heads keep track of all of these *gateways* in order to consider the alternate routes due to delay constraint or energy balancing.

We build a set of forwarding gateway nodes GS, for each cluster-head, for routing packets to neighboring clusters. Let SM_i be the set of members of cluster-head H_i and SM_j be the set of members of neighboring cluster-head H_j. H_i maintains a set of gateway nodes GS_i such that

$$GSi(Hi) = \{x \in SM_i / H_j \in N_x \; \forall\, y \in SM_j \;\wedge\; y \in N_x \;\forall\, i \neq j\}$$

Where N_x is the set of neighbors of node x. A member node x of cluster-head H_i belongs to the gateway set GS_i of head H_i if either H_j or some member y of H_j exists in the neighbors set of x. The attributes of the elements of GS_i are $\{AdjacentHead, Energy, Delay, Hops\}$. These attributes help the cluster-heads in selecting a particular item from the set GS. We will describe the selection criteria in detail in Section IV-A4.

Once the cluster formation is complete, each cluster gets the neighbor clusters list along with the gateways to reach them. The route computation is discussed in the next section.

IV. Coordination and Routing in ERCR

The main aim of *ERCR* framework is to provide real-time coordination and routing in WSAN with least energy consumption. This is achieved by incorporating a delay-constrained energy aware routing (DEAR) protocol. In ERCR, cluster-heads are responsible of determining their possible destination actors which is explained in Section IV-B. To conserve energy, ERCR constrains the traffic volume by implementing a novel real-time data aggregation (RDA) approach presented in Section IV-C in which cluster-heads perform aggregation such that the packets deadlines are not affected. In S-A coordination, it is not always possible that the cluster-heads have selected right actors to receive the event reports. Therefore, we provide A-A coordination in order to ensure reaction in the reported event. Actors coordinate with each other to respond to the event such that the limited number of mobile actors target the whole field. However, if an event region does not lie within the action coverage of any actor then ERCR ensures the action by implementing event targeting procedure explained in Section IV-D.

A. Delay-constrained Energy Aware Routing (DEAR) Protocol

DEAR works on top of the clustering protocol MEAC and delivers packets to the target nodes (Sink/Actors) respecting the end-to-end (E2E) deadline (τ) and balanced

energy consumption of the relaying nodes. Routes between the cluster-heads and actors are established through a backbone network, which is obtained by integrating the forward tracking and backtracking mechanism. Path from single/multiple hop members to cluster-heads is established during cluster formation. While the destination sink/actor initiates the route from cluster-heads to itself in backtracking manner as described in Section IV-A2. DEAR supports both the *semi-automated* as well as *automated architecture* of WSANs. In the presence of the sink, it adapts to the sink-based version of DEAR (S-DEAR) to coordinate with the actors through the sink if required. On the other hand, when there is no sink or ignoring its presence, it provides the distributed version of DEAR (D-DEAR) for coordination among sensors and actors.

1) Network Model: Before going into the details of the algorithm, we model the network as a connected directed graph $G = (V, E)$. The set of vertices V represents the sensor nodes, where $|V| = n$. E is the set of directed edges such that an edge $e(u \rightarrow v) \in E$ if $(u, v) \in V$. Two non-negative real value functions R(e), the available energy resource of node $v \in V$ on the outgoing link $e(u \rightarrow v) \in E$, and $\Delta(e)$, the delay experienced by the data packet on the corresponding link, are associated with the edges. These real values are used to compute the weight W(u,v) of the link $e(u \rightarrow v) \in E \vee (u, v) \in V$. The weight of an edge $e(u \rightarrow v) \in E$ can be defined as follows:

$$W(u, v) = R(e)/\Delta(e), \quad where \quad u, v \in V$$

Links are presumably asymmetrical because the R(e) and $\Delta(e)$ for the link $e(u \rightarrow v)$ may not be same while going in the opposite direction of this link $e(v \rightarrow v)$. The existence of alternative paths between a pair of vertices $u, v \in V$ provides the possibility of some paths being shorter than others in terms of their associated cost. We need to find out a minimum spanning acyclic subgraph of G having high total weight.

Let s be a source node and d be a destination node, a set of links $e_1 = (s, v_2), e_2 = (v_2, v_3), ..., e_j = (v_j, d)$ constitutes a directed path P(s,d) from $s \rightarrow d$. The weight of this path is given as follows:

$$W[P(s, d)] = \sum_{e \in P(s,d)} W(e)$$

Likewise, the E2E delay experienced by following the path P(s,d) is measured as:

$$\Delta[P(s, d)] = \sum_{e \in P(s,d)} \Delta(e)$$

After the formation of clusters, we can have a vertices subset H of the set V such that the elements in H are only the cluster-heads and have an associated integral function $hops[P(h \rightarrow target)], h \in H$. Similarly, we obtain the set $GS_h \quad \forall h \in H$ as the result of linking the clusters described in Section III-D. Each element h of set H maintains a set of outgoing links OUT_h subset of GS_h to the single destination node either sink or actor. In the next section, we describe the way of building the set $OUT_h \quad \forall h \in H$.

2) Establishing Routes: In order to compute the delay-constrained paths efficiently, we decompose G into a minimized acyclic subgraph $\bar{G} = (\bar{V}, \bar{E})$ constituting a large acyclic region within G. $\bar{V}$ is the set of nodes either in H or belong to the GS sets of cluster-heads i.e. $\bar{V} = H \cup GS_1 \cup GS_2 ... \cup GS_k$ for k number of clusters. $\bar{E}$ is the set of directed edges such that an edge $\bar{e}(u \rightarrow v) \in \bar{E}$ if $u, v \in \bar{V}$. The length of an edge $\bar{e}(u \rightarrow v) \in \bar{E}$ may be greater than one because the members in GS may be multi-hop far from heads. For instance, an edge $\bar{e}(u \rightarrow v) \in \bar{E}$ might exist due to some member node w such that $u \rightharpoonup w \rightharpoonup v), w \notin \bar{V}, (u, v) \in \bar{V}$. Here, $R(\bar{e})$ is the least available energy of any node visited while traversing the link $\bar{e}(u \rightarrow v)$ and $\Delta(\bar{e})$ is the cumulative delay experienced by the data packet on the corresponding link.

The decomposed minimized graph $\bar{G}$ is the backbone to establish the route from the source nodes to either the sink (*semi-automated architecture*) or the actor (*automated architecture*). In the next section, we look into the formation of the graph $\bar{G}$.

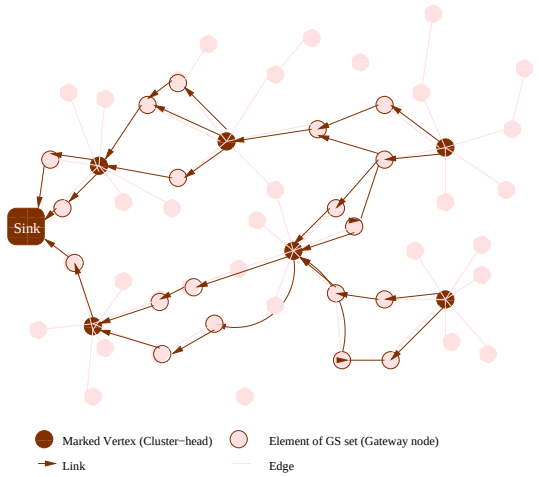


Fig. 5. Decomposition of graph G into the minimized acyclic subgraph $\bar{G}$ within the region G.

3) Sink-based DEAR (S-DEAR): In this section, we deal with the *semi-automated architecture* and its modification for *automated architecture* is provided in the following section. We assume that the sink node is stationary like sensor nodes and the path from cluster-heads to sink is built in proactive way. Sink is the destination for all the source nodes in *semi-automated architecture*. Source to sink path is divided into two phases; source to cluster head and cluster-head to sink. The first phase builds the path from source nodes to cluster-head that is done during the cluster formation in a forward tracking manner. The next phase deals with finding the path from cluster-heads to the sink using backtracking. It is activated initially by the sink during the network configuration phase and is updated periodically. To achieve this, the algorithm visits the graph G and marks all the vertices $h \in H$. A mark is associated with the life of the node, which is deleted as that vertice(node) expires. A vertex can be marked if $h \in H$ has not been already marked or the current path delay $\Delta[P(sink, h)]$ is less than the previously observed path delay.

Once all the elements h of set H are marked, we build a path $P(sink, h) \quad \forall h \in H$ in proactive fashion and each element $h \in H$ set its $hops[P(sink, h)] = |P(sink, h)|$.

When h is marked, h adds the incoming link $in(x \rightarrow h), x \in V$ to the set OUT_h in reverse-topological order $out(x \rightarrow sink)$. The incoming link in may be associated with the last marked element $g \in H$ in the marking process or $null$ if h is the first marked item and represents link to the root (sink node). This helps h to extend the set OUT_h by using the pre-determined set GS_h. The attributes of the elements of GS set contain the $AdjacentHead$ ID that corresponds to g. For each element $o(m \rightarrow g) \in GS_h$, it searches for the match of g with the attribute $AdjacentHead$ of o. If there exists such element $o(m \rightarrow g)$ then h adds the link as $o(m \rightarrow g)$ to OUT_h and associate an integral value $H(o)$ apart from the other two real value functions $R(o)$ and $\Delta(o)$. Hence, the edges set $\bar{E}$ of $\bar{G}$ can be obtained as $OUT_1 \cup OUT_2, ..., \cup OUT_k$ for k number of clusters.

Fig 5 illustrates the decomposed subgraph $\bar{G}$ with all the possible links to the sink node. We use the term link for set $\bar{E}$ rather than edge because vertices of set $\bar{V}$ may be connected by some intermediate vertices in V. The set OUT_h provides all the possible routes to the sink node and we exploit the multiple entries in OUT_h to provide delay-constrained energy aware routes and implicit congestion control. We describe the criteria of selecting the outgoing link in Section IV-A4. The cost of marking process is O(n) and, in fact, it is the actual cost of building route from source nodes to the sink node.

Implementation: The marking process is implemented by broadcasting *SACInfo* beacon in the network. That is, sink initiates the connection with the sensor nodes by broadcasting its *SACInfo* beacon periodically, where the length of period (life of mark) is larger than the $hello$ beacon. This periodic beacon helps to refresh the path because the topology of the sensor nodes is dynamic. A receiving node accepts this beacon if it meets one of the following conditions:

1) It has not already received this beacon or beacon has expired.
2) Delay of this beacon is smaller than the last received beacon.
3) The number of hops traversed by this beacon is small.

When a node receives a packet it calculates the delay and forwards the request in the direction of cluster-head. The cluster-head forwards it to its neighboring cluster-head. Hence, each cluster-head learns the loop free path to the sink node and gets the delay value and number of hops so far.

4) Alternative Path Selection: In addition to energy, E2E deadline τ is another constraint for real-time applications in wireless sensor and actor networks. Real-time event delivery is the first and foremost goal of DEAR. We have described the process of building the set of outgoing links OUT in the last section. The selection of a particular link $o(m \rightarrow g) \in OUT_h, g \in H$ originating from the member m by the cluster-head $h \in H$ is based on the criteria to balance the load in terms of delay and energy of its member nodes. Therefore, the cluster-head h selects an outgoing link $o(m \rightarrow g) \in OUT_h, g \in H$ for which $W[P(h, o)]$ is maximum.

The operations of the alternative gateway selection are outlined in the Algorithm 1.

Algorithm 1 Select Outgoing Link

Ensure: Delay-constrained energy aware outgoing link $out \in OUT_h$

```
Pseudo-code executed by source cluster-head h to select an outgoing link from the set OUT_h.
P = ∞
for all o(m → g) ∈ OUT_h, g ∈ H do
    if time_left/hops[P(sink, s)] < Δ(o) then
        if R(o) < P then
            P = R(n)
            out = o
        end if
    end if
end for
```

Cluster-head adds the $time_{left}$ field to its data packet that is set to τ by the source cluster-head. Each intermediate cluster-head looks for this $time_{left}$ field and selects the outgoing link accordingly by executing the above procedure. If the delay constraint can be meet through multiple links then it selects the one according to the criteria as described below:

"*The link, along which the minimum power available (PA) of any node is larger than the minimum PA of a node in any other links, is preferred*".

Every receiving node then updates the $time_{left}$ field as $time_{left} = time_{left} - delay_s$. It can be seen that the link selection criteria implicitly eliminates the congestion by alternating the links toward destination. Whenever a link is congested, the packet delay is increased and this delay is reported to the cluster-head in successive $hello$ beacon. The weight of this link is reduced and, eventually, the cluster-head reacts to it by selecting the other available link. Hence, the congestion is avoided in addition to the energy efficiency.

B. Sensor-Actor Coordination

The main communication paradigm in WSANs is based on the effective sensor-actor coordination. Right actions against the detected events cannot be performed unless event information is transmitted from sensors to actors. Therefore, the ultimate goal of any routing protocol in WSANs is to relay the event readings to the actors within a certain delay limit. In the classical *semi-automated architecture*, there is a central node that is responsible to collect the readings and issue action commands to the actors responsible for the action. Unlike this approach, *automated architecture* has also been realized due to the need of immediate action on the phenomena observed in the sensory field. In the former approach, sink is the destination of events reported by all the sources and is responsible to coordinate with actors. In the latter case, the mobile actors in *automated architecture* are the targets of the event readings observed by the sensor nodes and, hence, the coordination is local.

The distributed data routing approach is imperative due to the non-existence of central controller. Events detected by the

sensor nodes are directly routed to the actor nodes without the intervention of the sink node. To provide the distributed routing in the *automated architecture*, ERCR incorporates the distributed version of DEAR. In A-DEAR, we decompose the graph G into the m number of $\bar{G}$ subgraphs for each of m mobile actors. The idea is similar to S-DEAR described in detail in Section IV-A3 except that we have m possible destinations. The marking process is triggered independently by all the m actors to construct m number of $\bar{G}$ representing the paths $P(h, actor_1), (h, actor_2), ..., (h, actor_m) \quad \forall h \in H$. The cost of A-DEAR is O(mn).

In order to optimize the sensor-actor coordination in the distributed environment, the marking process also propagates the current $load$ factor of the actor. The $load$ represents the number of sources the actor is serving at the moment. The marking criterion in A-DEAR is modified such that h accepts the mark of an actor on the basis of its Euclidean distance. The nearest one is the best candidate for marking the element h of the set H. There might be the possibility that h lies within the action coverage of two or more actors. In such case, $load$ factor breaks the ties among such candidates and less-loaded actor is selected for coordination.

Actors are location aware mobile nodes. Whenever an actor moves, it triggers the construction of graph $\bar{G}$ in addition to the periodic reconstruction of graphs. The periodic update of graphs is required due to the highly dynamic topology of the wireless sensor and actor networks since sensor nodes may be deployed at any time or their energy deplete. Hence, the algorithm updates the path proactively to reduce the chances of path failure like the path establishment.

C. Real-time Data Aggregation (RDA)

Wireless sensor and actor networks are mostly designed to monitor and respond in various hostile environments. In order to increase the reliability of applications in such environments, sensors are deployed densely in the field. However, the dense deployment results in huge volume of traffic and creates hot spots in the network because the event data is synchronous by nature triggering all the nodes at once. The raw sensed data is typically forwarded to a sink or actors for processing which contains redundant event reports and unnecessarily consumes the scarce energy resource of the sensor nodes.

An important energy saving mechanism for sensor nodes is to exploit in-network data aggregation [28],[29],[30]. The main idea of in-network data aggregation is to eliminate unnecessary packet transmission by filtering out redundant sensor data and/or by performing an incremental assessment of the semantic of the data, e.g. picking the maximum temperature reading. It is argued that aggregation extends the queuing delay at the relay nodes and can thus complicate the handling of latency-constrained data. When we consider mixed-type traffic due to concurrent events where real-time and non-real-time traffic coexist, data aggregation becomes more challenging. In that case, we need to consider delay requirements for real-time data along with energy consumption of both real-time and non-real-time traffic.

ERCR incorporates a very simple yet practical aggregation method in which only cluster-heads involve to aggregate data from their member nodes and make sure that the end-to-end delay constraint is not violated. Hence, it does not require to build aggregation trees [28],[29],[30] because the nodes have already been configured in the form of clusters. To meet delay-constraint, cluster-head considers the event reports from its member node for aggregation as long as the deadline of any of the packet is not violated. For example at time $t = t_1$, a cluster-head h either detects an event itself or receives an event report from some of its member for the first time and has to send it to destination actor a. As it receives an event report, it checks for the condition

$$\Delta[P(h,a)] + \Delta(o) + T < \tau \tag{4}$$

where $o \in OUT_h$ is a potential *gateway* for a, T is the received packet delay and $\Delta[P(h,a)]$ is the expected route delay measured during path establishment in DEAR. Let us assume that at $t = t_2$, it receives a packet for which the above condition is not satisfied. It immediately transmits the aggregated packet. It is important to note that the head does not queue all the packets received during time t_1 to t_2 rather it aggregates the data as the packet is received. Consequently, it keeps the single aggregated packet for each different event and does not occupy the memory for aggregation. We apply three aggregation functions; maximum, minimum and average in order to better understand the event nature and thus the aggregated packets has the format $< \#reports, least_{time}, max_{value}, min_{value}, mean_{value} >$. where the number of reports represents the number of packet aggregated and $least_{time}$ is the least timestamps of all the aggregated packets.

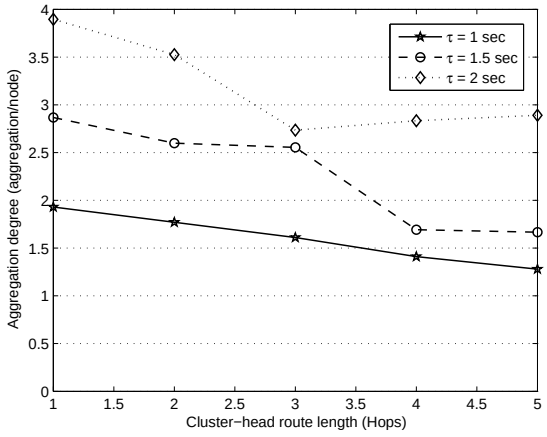


Fig. 6. Degree of aggregation applied by cluster-heads at different route lengths from the sink node at $\rho_o = 3\ samples/sec$.

It is worthwhile to note that in this approach the closer cluster-heads aggregate more packets than the farther cluster-heads as shown in Fig. 6. This achieves the fairness among them because the closing heads have more time for aggregation due to short path delay and thereby reduce the traffic significantly near the destination.

D. Actor-Actor (A-A) Coordination

Actor-actor coordination is usually required in two scenarios; *overlapping region* and *inaccessible region*. In case of *overlapping region*, actors may have overlapping action regions due to random deployment/movement of actors. In this case, cluster-heads select appropriate actors as described in Section IV-B by considering the distance and load on these actors. Since the load on actor changes dynamically, therefore, it is possible that an actor receiving the request might be busy in responding to other requests or have shortage of resources and can not respond to the source node. This requires the source head to update its coordinating actor. In order to update the coordinating actor, the receiving actor initiates actor-actor coordination by broadcasting the request to the nearby actors. An actor replies back to such requests if it is able to respond to the requesting source while attending its current sources as well. The source cluster-head then applies S-A coordination to all the replying actors and update its coordinating actor.

In second case of *inaccessible region*, actors are insufficient in the field to cover the whole region and therefore unattended regions are observed. As a result, an actor receiving request either needs to move itself toward such region if possible or trigger actor-actor coordination to relocate some actor in that region. This is a controlled targeting i.e. actor moves toward the requesting source such that the cluster-head is within the boundary of its action. However, this movement is possible only if the moving actor can attend its current request as well as the request from the inaccessible region. Hence, as an actor receives a request from a sensor node that lies outside of its coverage area, it starts the event targeting procedure.

1) Event Targeting: We assume that all the actors have the same circular action range of radius R, which is limited and different from the transmission range. When an actor a_i, residing at the point $I(x_i, y_i)$, receives a request from the source s at position $J(x_j, y_j)$, it calculates the distance $D(I,J) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ from the source. If $D(I,J) \leq R$, a_i is able to respond to the request. However, if $D(I,J) > R$, there are two possible scenarios. First, actor a_i needs to move $D(I,J) - R$ units of distance toward the target location at some new point $\bar{I}(\bar{x_i}, \bar{y_i})$ so that the coverage is provided to s as shown in Fig. 7. The new actor location $\bar{I}$ can be computed as follows

Taking $I(x_i, y_i)$ as the center of axis on the XY plane, we can find the coordinates $\bar{x_i}$ and $\bar{y_i}$ as

$$\bar{x_i} = x_i + (D(I,J) - R) \times cos(\alpha)$$
$$\bar{y_i} = y_i + (D(I,J) - R) \times sin(\alpha)$$

where α is the angle between the vector $X(R, 0)$ along x-axis and the vector $A(x_j - x_i, y_j - y_i)$.

To find out the value of α, first compute the length $|X|$ and $|R|$ of two vectors X and R.

$$|X| = R, \ |A| = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Normalize each vector (unit vector)

$$X_u = (R, 0)/|X|, \ A_u = (x_j - x_i, y_j - y_i)/|A|$$

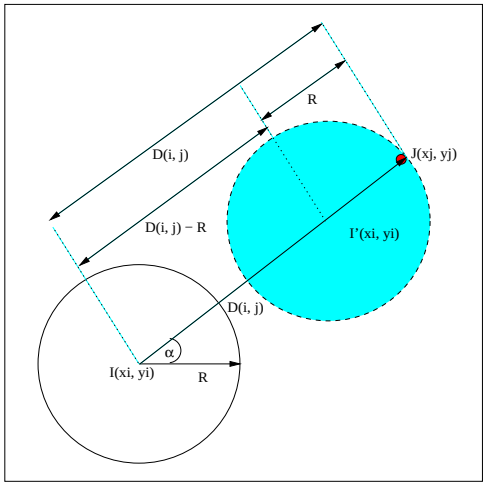


Fig. 7. Positioning actor toward uncovered source

```
   /*Pseudo-code executed by the actor a to
determine the coverage of source i.*/
   TargetEvent(i)
      d(i,a) = sqrt((x_a - x_i)^2 + (y_a - y_i)^2);
      C_a = coverSource((d(i,a) - R),(x_i,y_i));
      /* New center position C_a of a if it
can move d(i,a) - R toward (x_i,y_i).*/
      MoveOK = true;
      for(each s ∈ sources_a)
        MoveOK = canCover(C_a, R, (x_s,y_s));
        if (MoveOK = false)
         StartAACoordination(i);
         exit;
        end //if
      end //for
      moveTo(C_a);
```

Fig. 8. Algorithm of action coverage of a source.

Hence, the angle between X and A is:

$$\alpha = cos^{-1}(X_u.A_u) \quad (5)$$

However, the displacement is only possible if the movement of a_i does not leave the presently attended sources uncovered. In the second case, if the displacement of a_i is not valid then it initiates the actor-actor coordination algorithm.

The coordination among actors is triggered to move an actor in the uncovered region by broadcasting the *relocate* message that contains the location of the source s. Each actor receiving this message runs the procedure outlined in Fig. 8 to check for the possibility of attending the source. An actor a_j sends back the *relocation-ok* message to a_i if it is able to move toward the source. This message contains the residual energy of actor a_j as well as the total number of sources being attended. If the field is being operated by sufficient number of actors such that all the sources can get response from the actors then the actor a_i receives *relocation-ok* message from some of the neighboring actors. Consequently, a_i forwards the request to the actor covering lesser number of sources and has the highest residual energy.

When the actors have either limited action range or they are

small in number to cover the whole field, it is quite possible that some action requests remain unsatisfied. This situation arises when the intensity of events is too high to affect the entire field. We can compute the number of unattended sources χ in ideal case for $M \times M$ field as

$$\chi = \frac{M^2 - m \times \pi R^2}{M^2} \times n \tag{6}$$

where m is the number of mobile actors, n is the total number of deployed sensors. However, the value of χ depends on the density of nodes in the field. In case of non-uniform deployment when the action range is limited, actors might be stuck in sparse zone of the field that can not attend the nodes in dense part as described above, which greatly effects the optimal value of μ.

V. Performance Evaluation

The performance of the framework is evaluated by using the network simulator ns-2 [35]. The performance metrics comprises of average energy consumption, packet delay and delivery ratio for the evaluation of ERCR components; clustering, sensor-actor coordination, actor-actor coordination and aggregation. The initial energy E_o of energy constrained (EC) nodes is set to 0.5 $joules$. The transmission and reception power is set to 50 nJ/bit and sources produce traffic at 4 $kbps$.

A. Clustering Performance with Multiple Events

In order to study the impact of including data rate into cluster-head election procedure, the experiments are run by keeping the weighting factors of energy and reporting rate (RE-Weighted) high against high energy factors (E-Weighted) as is the case in the existing studies. The heterogeneity of sensor nodes in terms of energy (λ_{energy}) and data rate (λ_{rate}) is formulated in [3]. We consider 20% of the nodes have higher energy, i.e. $E_i > E_o$ and similar fraction has higher data rate likewise. Fig. 9 shows that ERCR has lesser energy consumption when the data rate of nodes is considered in calculating weight of the nodes as given in (2). Hence the consumption is lesser in RE-Weighted scenario than E-Weighted. It is obvious that the energy consumption is less in RE-Weighted even when there are no higher energy nodes ($\lambda_{energy} = 0$). Similarly, the network stability, in terms of minimum residual energy of any node, is higher in RE-Weighted case as shown in Fig. 10. Hence, by including data rate in the selection of cluster-heads along with the residual energy of nodes, ERCR conserves the energy of nodes, which extends the lifetime of network.

B. S-A Coordination

In these experiments, 200 sensor nodes and 4 actors are randomly deployed in 150×150 field with three different network configurations; *semi-automated*, *automated* with simply S-A coordination and *automated* with mobile A-A coordination. In *semi-automated* configuration, all the packets are transmitted to the actors through the centralized sink node. Contrarily, *automated* configuration delivers the packets from source nodes to the actors directly without the intervention of sink node. The *automated* configuration has further two setups; one lacking actor-actor coordination (immobile A-A coordination) and second with mobile A-A coordination.

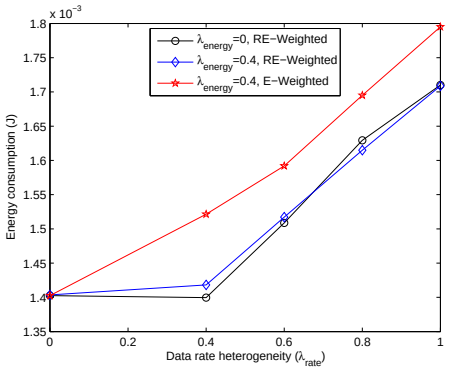


Fig. 9. Average energy consumption of sensor nodes.

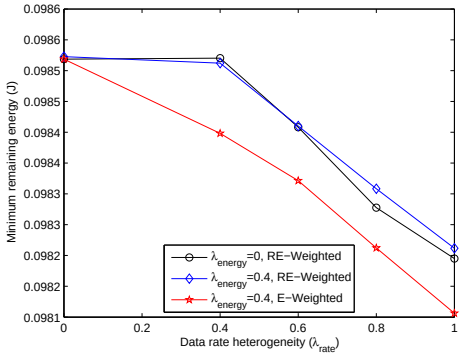
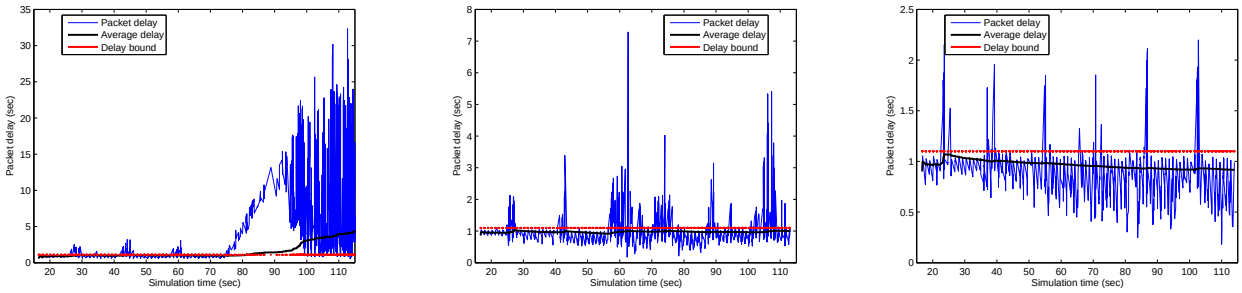


Fig. 10. Minimum residual energy of a node or network stability.

Fig. 11 illustrates the average delay at $\tau = 1$ sec for the three configurations. We assume 5% tolerance in delay and, therefore, plot the packet delay and average delay for a delay bound of 1.05 sec. The packet delay in *semi-automated* architecture hardly achieves the deadline and the network becomes congested after some time that results in extremely large packet delay as shown in Fig. 11(a). The average delay in *semi-automated* is 5 times greater than both the scenarios of *automated* architecture. However the two configurations of *automated* architecture differ in miss-deadline ratio. This ratio is about 14% in S-A coordination while just 5% with mobile A-A coordination. Hence, the adaptive mobility of actors in ERCR improves the performance further in addition to S-A coordination.

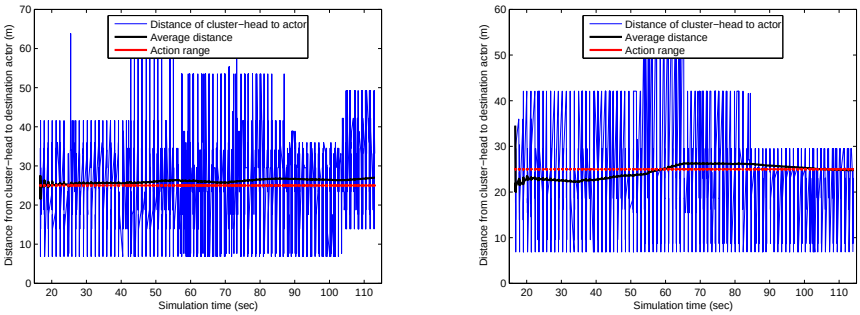
Similarly, the packet delivery ratio in *automated architecture* is higher than *semi-automated architecture*, which is approximately 30% higher as shown in Fig. 12.

We assume that the actors have limited action range and they can not respond to the sensor nodes if they lie out of their action coverage. Such requests may not be fulfilled and therefore assumed as unattended requests. The mobile A-A coordination is used to minimize such requests by moving the actors sufficiently toward these nodes. Intuitively, increasing the responsiveness of the actors. Fig. 13 shows the distance of the cluster-heads to the coordinated actors when the action



(a) Delay in semi-automated architecture (b) Delay in immobile A-A coordination scenario of automated architecture (c) Delay in mobile A-A coordination scenario of automated architecture

Fig. 11. Second scenario with action range of actors set to 25 meters.



(a) Distance of cluster-heads from coordinated actors in simple S-A coordination. (b) Distance of cluster-heads from coordinated actors with S-A/A-A coordination.

Fig. 13. 200 sensor nodes and 4 actors randomly deployed in 150×150 field.

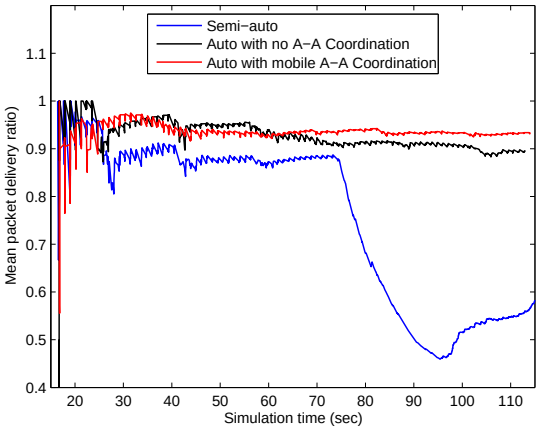


Fig. 12. Packet delivery ratio in *semi-automated architecture* and *automated architecture* for 2 samples/sec triggered by 3 phenomenon nodes of event radius 50 meters.

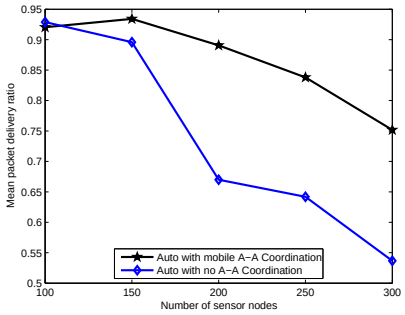
range is 25 meters. It is obvious that the average distance in immobile coordination given in Fig. 13(a) is beyond of action range. However, the average distance in mobile actors scenario is below the action range for most of the time. The unattended request ratio is 50% in the former case and 25% in the latter case, which is computed 11% by using (6). It happens when the number of actors operating in the field is lesser than the required number of actors to cover the whole field. However, with mobile A-A coordination, ERCR can significantly achieve better performance with smaller number of actors.

C. A-A Coordination

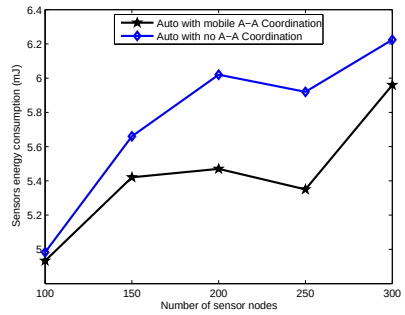
In this setup, we first keep the number of actors 4 and increase the number of sensors gradually from 100 to 300. and then varying the actor nodes between 2 and 7 for 200 sensor nodes in 150×150 m^2 field. These experiments are conducted to show the performance achieved when employing mobile A-A coordination, which is not possible with simple S-A coordination.

Fig. 14 shows the delivery ratio in first scenario. It is observed that the scalability is achieved by employing A-A coordination. The delivery ratio drops down to 17% by increasing the sensor nodes 3 times and dimensions of the field to 2 times. However, this value is approximately 50% when there is no A-A coordination and actors stay at their initial locations. Hence the gain in delivery ratio due to mobile A-A coordination is 3 times which is a worthwhile improvement. Similarly the energy consumption of not only the sensor nodes but also the actor nodes is reduced as shown in Fig. 14(b) and Fig. 14(c) respectively.

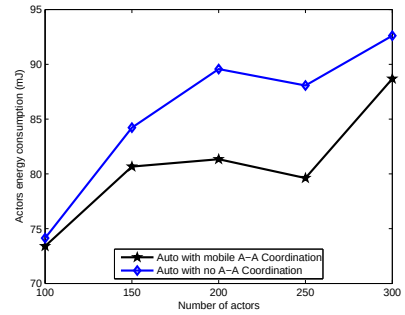
In the second scenario, we show that higher delivery ratio is obtained at smaller number of actors when employed A-A coordination. We achieve maximum delivery ratio of 90% at $m = 4$. The delivery ratio for the same number of actors in simple S-A coordination is approximately 68%, which is



(a) Packet delivery ratio with and without mobile A-A coordination.

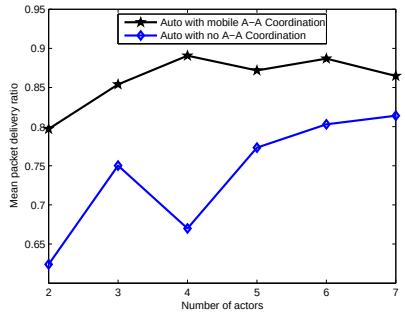


(b) Mean energy consumption of sensor nodes.

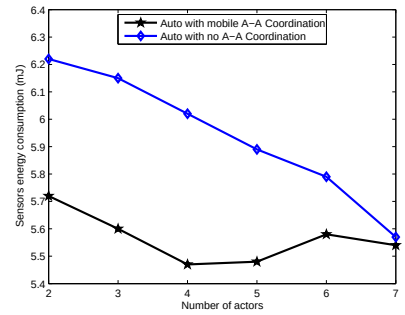


(c) Mean energy consumption of actor nodes.

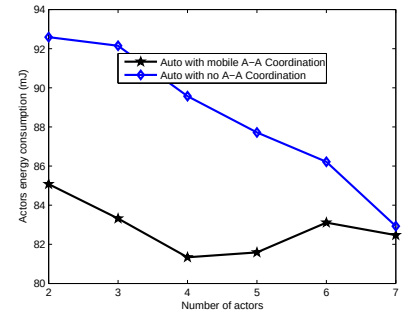
Fig. 14. Performance for different number of sensor nodes with 4 actor nodes randomly deployed.



(a) Packet delivery ratio with and without mobile A-A coordination.



(b) Mean energy consumption of sensor nodes.



(c) Mean energy consumption of actor nodes.

Fig. 15. Performance for different number of actor nodes deployed with 200 sensor nodes in 150×150.

significantly lower. Even the delivery ratio for $m = 7$ is 82% which is still lower than the delivery achieved in A-A coordination for $m = 4$. Intuitively, we achieve the delivery in A-A coordination at lower number of actors, which reduces the cost of application because the cost of actors is the major contribution in WSANs applications cost. Fig. 15(b) and 15(c) show the gain in energy of both sensor nodes and actors respectively. It is clear that we save the energy of sensors and actors by approximately 12%.

D. Aggregation

Data aggregation in WSANs is applied to reduce the volume of traffic in network and eventually conserve the energy of sensor nodes. Although, it achieves the energy efficiency but increases the packet delay which is usually not suitable for real-time applications. However the data aggregation approach presented in Section IV-C is adaptive to the application deadlines. Fig. 16 reflects the adaptiveness of RDA, where the average delay remains lower than τ. This is not applicable at shorter deadline and higher traffic rate ($\tau \leq 0.5\ sec,\ \rho_o = 4\ samples/sec$) where the average packet delay is much higher than τ. We consider the traffic in terms of event sampled by a sensor node in one second, where the size of a sample is kept 100 bytes. It is due to the fact that lesser or no packets are aggregated and larger number of packets would be transmitted. Therefore, sensor nodes can not deliver the packets at higher rate within the given short deadline because of congestion. This situation is improved at $\tau > 0.5\ sec$. Fig. 18 shows the effect of delay bound on the aggregation achieved and packet delay.

The adaptivity of RDA can be justified in Fig. 17 too. It shows that the mean aggregation is lower at smaller τ but higher at larger τ. Since the higher volume of traffic introduces larger packet delay, it must be considered in aggregating and

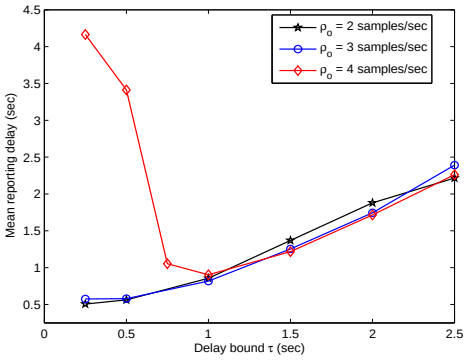


Fig. 16. Event reporting delay for different values of τ.

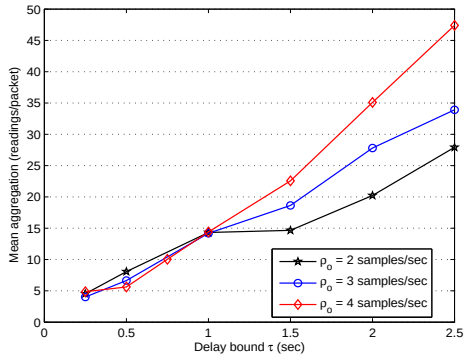


Fig. 17. Mean aggregation achieved at different values of τ.

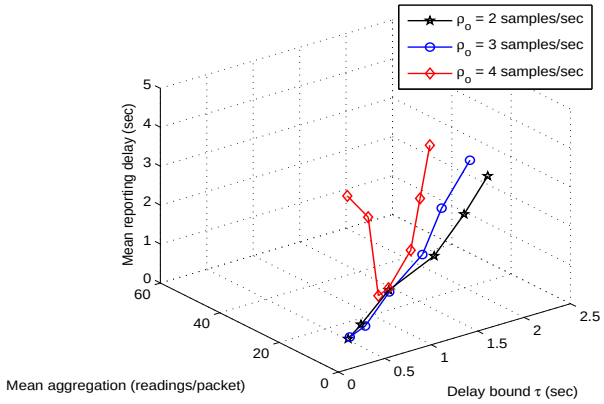


Fig. 18. Event reporting delay with different values of aggregation achieved applied for different values of τ.

routing real-time data. This effect is also supported in RDA and it is witnessed in Fig. 17 that by increasing traffic 2 times i.e. from $\rho_o = 2\ samples/sec$ to $\rho_o = 4\ samples/sec$, the aggregation is increased 60% since RDA gets less time for aggregation due to large communication delay. Hence, the integrated RDA approach of ERCR is not only adaptive to different deadlines of the packets but to the different data rates in the network as well.

At shorter deadlines, lesser number of packets are aggregated. Eventually the number of packets transmissions are increased that result in lower delivery ratio due to high

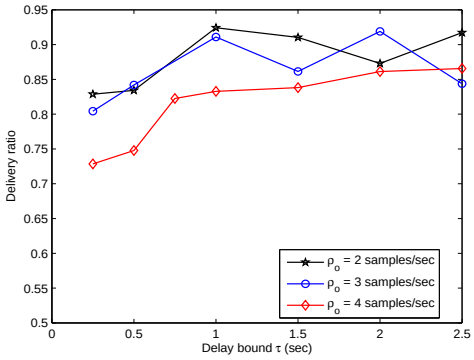


Fig. 19. Event delivery ratio for different values of τ.

volume of traffic. This fact is shown in Fig. 19. At $\rho_o = 2\ samples/sec$, the delivery ratio is initially $\approx 10\%$ higher than at $\rho_o = 4\ samples/sec$. However, as τ reaches to 0.75 sec, the delivery ratio is improved for $\rho_o = 4\ samples/sec$ and the difference of delivery ratio is reduced to 5%. By the suitable selection of τ, we can achieve the required packet delivery. Similarly the energy consumption is reduced by applying aggregation as shown in Fig. 20.

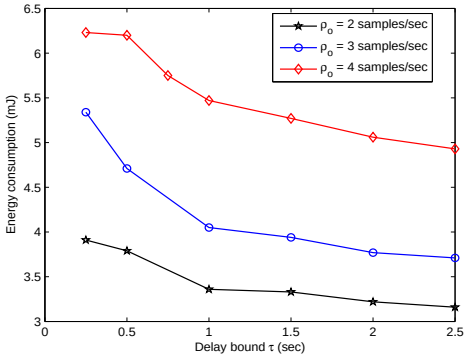


Fig. 20. Mean energy consumption for different values of τ.

VI. Conclusion

In this paper, we propose a new cluster-based real-time coordination and routing framework ERCR for WSAN. ERCR assumes energy efficiency and real-time data routing support at different levels as two very important performance metrics. To achieve energy efficiency, sensor nodes coordinate with each other to form clusters. A node is elected as a cluster-head which has higher energy level but lower traffic rate. The cluster-heads select the next forwarding node such that the delay-bound of packet can be met, as part of the real-time criteria. The real-time aspect is supported by data aggregation model. It helps in routing data within the given delay bound (τ) and conserving the energy by reducing the traffic volume.

Thirdly, it provides sensor-actor coordination so that the event reports are routed directly to the actors without involving the centralized sink node. This coordination does not only save the energy of sensor nodes but also reduces the packet delay.

At higher level when the actors receive data, they coordinate with each other to execute action in the event area. An actor moves toward the event area if it does not come in the action coverage of any other actor. As a result, future event reports are sent to the actor which can cover the event area. Hence, the responsiveness of the actors is improved, which in turn improves the performance of the entire application.

ERCR provides the real-time guarantees up to 95 % of the packets and it achieves the packet delivery ratio 60 % higher than that of *semi-automated architecture*. The mobile A-A coordination in ERCR enhances the performance with lesser number of actors compared to static A-A coordination. With all this increased performance, it reduces the cost of application as well. Moreover, not only the energy consumption of sensor nodes but that of the actor nodes is also reduced. Finally, simulation experiments prove that ERCR achieves its goal of real-time event delivery with realistic application-specific delay bound and energy efficiency.

REFERENCES

[1] I. F. Akyildiz and I. H. Kasimoglu, "Wireless Sensor and Actor Networks: Research Challenges," *Ad Hoc Networks, Vol. 2, Issue 4, pp. 351-367, October 2004*.

[2] G. A. Shah, M. Bozyigit, Ö. B. Akan and B. Baykal, "Real-time Coordination and Routing in Wireless Sensor and Actor Networks," *in Proc. 6th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN), May 2006*.

[3] G. A. Shah, Ö. B. Akan and M. Bozyigit, "Multi-Event Adaptive Clustering (MEAC) Protocol for Heterogeneous Wireless Sensor Networks," *in Proc. 5th Annual Mediterranean Ad Hoc Networking Workshop (MedHoc-Net 2006), June 2006*.

[4] T. Melodia, D. Pompili, V. C. Gungor and I. F. Akyildiz, "A Distributed Coordination Framework for Wireless Sensor and Actor Networks," *ACM MobiHoc'05, May 2005*.

[5] E.C.H. Ngai, M.R. Lyu and J. Liu, "A Real-Time Communication Framework for Wireless Sensor-Actuator Networks," *in Proc. of IEEE Aerospace Conference, pp. 1-9, March 2006*.

[6] Arjan Durresi, Vamsi Paruchuri and Leonard Barolli, "Delay-Energy Aware Routing Protocol for Sensor and Actor Networks," *in Proc. of the 2005 11th International Conference on Parallel and Distributed Systems (ICPADS'05)*.

[7] H. Peng, W. Huafeng, M. Dilin and G, Chuanshan, "ELRS: An Energy-Efficient Layered Routing Scheme for Wireless Sensor and Actor Networks," *In Proc. of the 20th Intl Conference on Advanced Information Networking and Applications, pp. 452-460, 2006*.

[8] T. He, J.A. Stankovic, C. Lu and T. Abdelzaher, "SPEED: A Real-Time Routing Protocol for Sensor Networks,"*Proc. IEEE International Conference on Distributed Computing Systems, May 2003*.

[9] C. Lu, B. M. Blum, T. F. Abdelzaher, J. A. Stankovic and T. He, "RAP: A Real-Time Communication Architecture for Large-scale Wireless Sensor Networks,"*in Proc. IEEE RTAS, San Jose, CA, USA, Sep 2002*.

[10] K. Akkaya, M. Younis, "An Energy-Aware QoS Routing Protocol for Wireless Sensor Networks," *23rd ICDCSW, pp. 710-715, 2003*.

[11] S. Younis and O. Fahmy, "HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks," *IEEE Transactions on Mobile Computing, vol. 3, no. 4, 2004*.

[12] S. Basagni, "Distributed Clustering for Ad Hoc Networks," *Proc. of IEEE Intl. Conference of Vehicular Technology, 1999*.

[13] D. Tian and N.D. Georganas, "Energy Efficient Routing with Guaranteed Delivery in Wireless Sensor Networks," *Mobile Computing and Communications Review, 2001*.

[14] I. Chatzigiannakis, A. Kinalis and S. Nikoletseas, "An Adaptive Power Conservation Scheme for Heterogeneous Wireless Sensor Networks with Node Redeployment," *Proc. of the 17th annual ACM SPAA'05, Jul 2005*.

[15] E. J. Duarte-Melo and M. Liu, "Analysis of Energy Consumption and Lifetime of Heterogeneous Wireless Sensor Networks," *IEEE GLOBECOM'02. vol. 1 pp. 21-25, Nov 2002*.

[16] V. Mhatre and C. Rosenberg, "Homogeneous vs heterogeneous clustered sensor networks: A comparative study," *IEEE ICC'04 , vol. 27, no. 1, pp. 3646-3651, Jun 2004*.

[17] M. J. Handy, M. Haase and D. Timmermann, "Low Energy Adaptive Clustering Hierarchy with Deterministic Cluster-Head Selection," *IEEE International Conference on Mobile and Wireless, 2002*.

[18] W. R. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "Energy efficient communication protocol for wireless microsensor networks," *Proc. of the 33rd Hawaii Intl. Conference on System Sciences (HICSS-33), Jan 2000*.

[19] M. Chatterjee, S.K. Das and D. Turgut, "A Weight Based Distributed Clustering Algorithm for Mobile Ad hoc Networks," *HiPC 2000: pp. 511-521*.

[20] G. Smaragdakis, I. Matta and A. Bestavros "SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks," *2nd Intl. Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004)*.

[21] W. R. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "Minimizing communication costs in hierarchically-clustered networks of wireless sensors," *Computer Networks, vol. 44, no. 1, pp. 1-16, Jan 2004*.

[22] G. Gupta and M. Younis , "Performance Evaluation of Load-Balanced Clustering of Wireless Sensor Networks," *10th Intl. Conference on Telecommunications (ICT'03), vol. 2, Mar 2003*.

[23] A.D. Amis, and R. Prakash, "Load-Balancing Clusters in Wireless Ad Hoc Networks", *Proc. of ASSET, Mar 2000*.

[24] A.D. Amis, R. Prakash, T.H.P. Vuong and D.T. Huynh, "Max-Min-D-Cluster Formation in Wireless Ad hoc Networks," *Proc. IEEE Infocom 2000, Tel Aviv, Israel, March 2000*.

[25] C. Intanagonwiwat, R. Govindan and D. Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," *Proc. ACM MobiCom, March 2000*.

[26] V. Rodoplu and T.H. meng, "Minimum Energy Mobile Wireless Networks," *IEEE JSAC, August 1999*.

[27] W. R. Heinzelman, J. Kulik and H. Balakrishnan, "Adaptive Protocol for Information Dissemination in Wireless Sensor Networks," *Proc. ACM MobiCom, 1999*.

[28] Y. Hu, N. Yu and X. Jia, "Energy efficient real-time data aggregation in wireless sensor networks," *in Proc. of the International Wireless Communications and Mobile Computing Conference (IWCMC 2006), pp. 803-808, July 2006*.

[29] K. Akkaya, M. Younis, and M. Youssef, "Efficient aggregation for delay-constrained data in wireless sensor networks,"' *in Proc. of Internet Compatible QoS in Ad Hoc Wireless Networks, 2005*.

[30] T. He, B. M. Blum, J. A. Stankovic and T. Abdelzaher, "AIDA: Adaptive Application-Independent Data Aggregation in Wireless Sensor Networks," *ACM Transactions on Embedded Computing Systems, vol. 3, no. 2, pp. 426-457, May 2004*.

[31] Q. Sun and H. Langendörfer, "A new distributed routing algorithm for supporting delay-sensitive applications," *Journal of Computer Communications vol. 9 no. 6 May 1998*.

[32] R. Gawlick, A Kamath, S. Poltkin and K. G. ramkrishnan "Routing and admission control in general topology networks," *Tech. Report STAN-CS-TR-95-1548, Stanford University, CA, 1995*.

[33] S. Bandyopadhyay and E.J. Coyle, "An Energy Efficient Hierarchical Clustering Algorithm for Wireless Sensor Networks," *IEEE INFOCOM 2003*.

[34] "IEEE 802.15 WPAN Task Group 4 (TG4)" *http://www.ieee802.org/15/pub/TG4.html, 2006*.

[35] UC Berkeley, LBL, USC/ISI and Xerox PARC, "The Network Simulator ns-2," *The VINT Project, http://www.isi.edu/nsnam/ns/*.

Adaptive Pull-Push based Event Tracking in Wireless Sensor Actor Networks

Ghalib A. Shah[1], Muslim Bozyiğit[2] and Demet Aksoy[3]

[1,2] Department of Computer Engineering
[3] Department of Computer Science
[1] National University of Science and Technology, Rawalpindi Pakistan
[2] Middle East Technical University, Ankara Turkey 06531
[3] University of California, Davis USA
Email: [1] ghalib@ceme.edu.pk, [2] bozyigit@ceng.metu.edu.tr, [3] aksoy@cs.ucdavis.edu

***Abstract*—Wireless sensor networks have attracted significant interest for various scientific, military, and e-health applications. Recently a new class of sensor networks "sensor/actor networks" has been introducing new research challenges due to the unique coordination requirements. In sensor/actor networks, actors are nodes that have the capability to move in the field equipped with powerful devices and can thus respond to emergencies. With this capability autonomous operation of the network is possible without a centralized control mechanism. This, however, requires the network to apply cooperative organization in response to emergencies to decide when and how monitoring is done and how it will be responded. In this regard, little work is done in terms of co-existing Push and Pull data flows in the network.**

In this paper, we propose an Adaptive Pull-Push (APP) based Event Tracking approach that allows sensor-to-actor communication and actor coordination in response to emergencies. APP proposes two models of organization: region-based organization (RAPP) and neighbor-based organization (NAPP) to alert nodes in the vicinity of reported emergency. APP exploits the mobility of actor nodes to form dynamic responsibility clusters, thus ensures an event specific response time to emergencies. Simulation results suggest significant performance improvements in terms of response time and energy conservation.

***Index Terms*—Adaptive Pull-Push, Routing with Adaptive Targeting, Sensor-Actor Coordination, Heterogeneous WSAN**

I. Introduction

RECENT advances in wireless technology and the development of small, low-cost, low-energy devices has enabled deployment of large scale highly distributed and intelligent sensor networks. Increasing onboard processing capability of individual sensors allow reaction to changes in application and network requirements. Such capability allows scientists and engineers to deploy autonomous and untethered networks. Various scientific and engineering applications may utilize such intelligent sensor networks [1] in reconnaissance, surveillance, environmental and habitat monitoring, and wildlife tracking. In practice, it is not only important to monitor the environment, but also to react to it. Wireless sensor and actor networks (WSANs) have been proposed for this purpose [1]. WSANs consist of sensor nodes that observe the environment and actor nodes that will react to changes in the environment, such as autonomous mobile robots, light switches or climate regulators.

In WSANs, unlike sensor nodes that have severe resource constraints, actors are mobile and resourceful, i.e., they have higher energy, more processing and storage capabilities and they can communicate using long range transmissions. In this regard, actors can take an active role in the network towards autonomous operation without coordinating with the sink node. The main goal of traditional sensor networks has been to communicate collected data to a stationary sink node [1]. In WSANs, communication of collected data can be destined to a sink or directly to an actor node that has the responsibility and also the capability to respond to emergencies.

There are two ways to communicate the event data observed by sensor nodes, namely Push-based and Pull-based [15] approaches. In the Push policy, event information is sent out to the sink without explicit requests for it. In WSANs, this requires continuous monitoring and thus, it consumes more energy. In contrast, with Pull based approach events are reported in response to explicit requests received from the sink/actor. In practice, it is unnecessary to report a particular event scenario which is not of interest to the sink or the actors. Therefore Pull can help regulating the rate at which observations are made. Although this approach saves energy but it may introduce higher event observation latency.

There have been a number of studies developing Push-based dissemination in WSN [16], [17], [19], [26] as well as Pull based [27]. Some hybrid Push-

An earlier version of this paper appeared in [3].

Pull techniques [28], [29] are also suggested to overcome the drawbacks of Push and Pull based approaches in WSNs. A number of coordination protocols for WSANs have been proposed [2], [8], [9], [10], [11], [12]. To the best of our knowledge, there exists no integrated Pull-Push solution for WSANs to organize the sensor and actor networks to respond the emergencies in the field.

In this paper, we propose a multi-level solution to Pull-Push coordination such that the network can respond to emergencies with/without coordinating with a centralized control. In particular, the proposed Adaptive Pull-Push (APP) approach is aimed to allow energy efficient sensor-actor coordination in response to emergencies. Routing in APP is based on Routing by Adaptive Targeting (RAT) protocol [3]. RAT relays the packets to dynamically relocated actors (targets) such that delay constraints can be met while energy consumption of forwarding nodes is balanced.

The coordination in APP is organized by actors in two ways; Neighbor-based APP (NAPP), and Region-based APP (RAPP). Actor nodes organize communication by exploiting neighborhood information of reporting node in NAPP and geographical responsibility region in RAPP such that the spread-outs can be observed in a timely manner. APP exploits the mobility of actor nodes to form dynamic *responsibility clusters*. A responsibility cluster in the field consists of a group of sensors observing events and at least one actor responding to the events. The whole field is divided into such responsibility clusters to collectively achieve the goal of the application. To maintain such responsibility clusters in the field, the actor position may need to be arranged to keep them close to emergency areas and reduce the response time. Hence, APP enables relocating the actor nodes in response to received event reports. Simulation results suggest significant performance improvements in terms of response time and energy conservation.

The remainder of the paper is organized as follows. In Section II, we describe the timing constraints in WSANs and introduce an analysis to study the lower bounds of the expected latency. In Section, III we describe the proposed Adaptive Pull-Push approach in WSAN and actors distribution and organization in response to emergencies. Performance evaluation and results are considered in Section IV. Finally, the paper is concluded in Section V.

II. Delay-Constrained Routing in APP

Sensor networks are proposed to play a major role in detecting and characterizing Chemical, Biological, Radiological, Nuclear, and Explosive (CBRNE) material and other emergency cases to prevent disastrous conditions. Obviously, an immediate action is imperative to encounter the threats in such applications and available resources should be exploited whenever possible. For applications that require time critical actions, such as monitoring a strategic area, it is, therefore, more efficient to report the abnormal event readings directly to nearby actors that have the responsibility to take action. Otherwise, when emergencies are reported to a stationary sink node as in traditional WSNs, after the delivery is complete the sink node still needs to communicate with the actor node in the area to take action. This would intuitively increase the response time. Hence, we assume that all communication is targeted at the actors rather than to a stationary sink node.

In WSAN, multiple actors, placed or moved at different geographical positions in the field, can be the potential destinations of the sensor readings. In this regard, a major component in our study is to adaptively select the target among one or more actor nodes patrolling the field. Routing by Adaptive Targeting (RAT) [3] relies on delay-constrained geographical routing to relay an incoming data packet to the neighbor who is the closest to the selected destination among its one-hop neighbors. In this study, we assume that both the sensors and the actor nodes are location-aware and the approach is not dependent on a particular localization protocol since there are a wide-range of localization protocol [20], [22] [21], [23], [24].

RAT maintains delay constrained single hop neighborhood table based only on the location information of the neighboring nodes, and the destination actor node. This allows efficient processing that is highly desirable for resource constrained nodes. We define the delay tolerance of events as τ and try to minimize $\tau - \epsilon$, where ϵ represents the time that is required by the actor to respond to sensor input. The value of ϵ depends on two factors; 1) the time required by an actor to get ready and take certain

action, and 2) how many action requests are in the waiting queue. In the rest of the section, we describe how to obtain the delay constrained forwarding table and provide analysis for meeting packet deadlines.

A. Delay-constrained Forwarding Set (DFS) and RAT

A sensor node initially keeps all its neighbors in the forwarding set (FS) along with an approximation of packet delay given in [3] and records in FS. However, FS is modified according to the response time demanded by the application and energy efficiency. The process of forwarding the data packet toward an actor consists of two phases

- Construct a delay-constrained forwarding subset (DFS) from the set FS such that the given delay constraint can be met. The source node sets the time to live (TTL) field in the data packet to $\tau - \epsilon$ and each forwarding node updates the TTL field by deducting the traversed hop delay. The packet delivery is considered unsuccessful whenever TTL reduces to 0.
- Balance the load on the nodes in the DFS in terms of energy consumption by selecting the forwarding node from the set DFS that has the highest energy level.

Hence, a forwarding node is selected such that the response time can be met as well as energy consumption is balanced. For each possible destination actor *a*, node *i* decomposes its set FS into subsets DFS_i^a as:

$$DFS_i^a = \{j |\ j \in FS_i \wedge D(i,a) > D(j,a) \wedge \frac{D(i,a)}{D(i,j)} < \frac{TTL}{T(j)}\}$$

where FS_i is the forwarding set at node *i*, *D(n1,n2)* is the distance between the node *n1* and *n2*, *a* is the actor node that is the ultimate destination of this communication and *T(j)* is the time required to relay the packet to node *j*. For this purpose, node *i* evaluates the expected number of hops to actor *a* based on the distance to node *j* as the unit. It then considers the time it takes to communicate within a single hop *T(j)* to see if the evaluated number of hops would result in a latency that is less than the deadline.

In essence, routing is dynamically modified to adapt to different circumstances in the network. To illustrate the performance benefits, we plot the mean packet delivery delay of our delay-constrained geo-routing policy in comparison to greedy routing in Fig. 1. In this scenario 100 nodes are deployed in a 100×100 field. In the experiment 70 nodes report observations at 2 samples/second rate while 30 sensors report an event of interest at 8 samples/second rate. The deadline to respond to emergencies is set to 1 second. We have 4 actors patrolling the field and the experiment is run for 200 seconds. We report the average packet delay of RAT in comparison to greedy routing. In greedy routing, nodes always select the next forwarding node as the closest node to the actor among its single-hop neighbors. On the other hand, RAT considers packet deadline and the energy consumption of the forwarding nodes resulting in the orders of magnitude improvement observed in Fig. 1. The energy consumption of RAT is observed at 43.4405 mJ as compared to 47.9105 mJ during this process. At top of all, the packet deadline miss ratio for greedy is observed to be at 7.4% while RAT missed no deadline (0% *packet deadline miss rate*) for this setting. In the following section we provide an analytical analysis on delay bounds in general on a sample scenario.

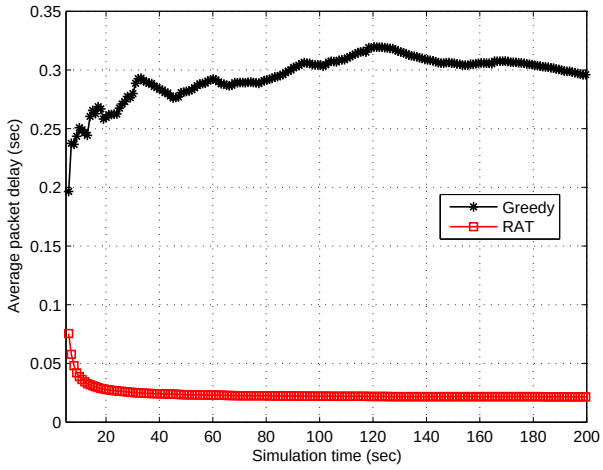


Fig. 1. Delay in RAT vs pure greedy routing.

B. Delay Analysis

In this section we provide an analysis on the performance bounds of APP. For simplification of the analysis, we consider a particular WSAN snapshot. We assume the sensor field is circular with radius *M* and the actor is lying at the center of the circle to collect data. Let us consider a random uniform deployment and a total of N_s number of sensor

nodes in the region are deployed. Furthermore, let us assume that the sensor nodes have transmission radius *r* and two nodes can communicate only if their distance falls within *r*. In Fig. 2, we plot network model where nodes within an outer ring forward their packets to the nodes within the inner ring until they reach at the destination actor present at the center of the field.

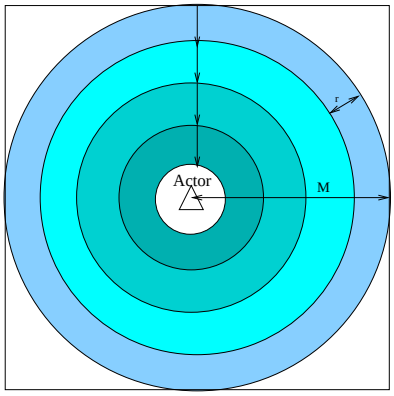


Fig. 2. Model of wireless sensor actor network.

The traffic model for the network may be described as follows. Each sensor node in the network can be a source or relay of packets. We assume that each node generates packets with rate λ_e packets/sec as the reporting rate. The size of each packet is constant and equals to *l* bits. We further assume that when a node forwards a packet, each one of its neighbors is equally likely to receive the packet. Nodes lying within the πr^2 transmission area of node n are considered its immediate (one hop) neighbors. Based on uniform deployment assumption, the number of neighboring nodes $N(n)$ of n are computed by $N_s \times (\pi r^2/\pi M^2)$. Hence, the probability that node j is in the neighbors of node n is (r^2/M^2).

In order to determine the worst case delay in multihop communications, we assume that the model circle is divided into concentric rings of thickness r as shown in Fig. 2. Now if an event is sampled by a node in the i^{th} ring, the data packet has to be relayed through the nodes in inner rings to reach the destination actor. We determine the delay incured by the packet originating in the i^{th} ring, where i may vary from 1 to M/r. Since we assume a uniform distribution of sensor nodes, the average number of sensor nodes lying outside the i^{th} ring are $N_s \times (\pi M^2 - \pi(ir)^2)/\pi M^2$. We expect that the i^{th} ring is populated by $N_s \times (\pi(ir)^2 - \pi((i-1)r)^2)/\pi M^2$ number of nodes. Intuitively, $N_s \times (\pi M^2 - \pi(ir)^2)/\pi M^2$ number of packets are forwarded by $N_s \times (\pi(ir)^2 - \pi((i-1)r)^2)/\pi M^2$ number of nodes. Let us represent the average number of packets by L_i, which a node in the i^{th} ring forwards as

$$L_i = \frac{N_s(\pi M^2 - \pi(ir)^2)/\pi M^2}{N_s(\pi(ir)^2 - \pi((i-1)r)^2)/\pi M^2} \tag{1}$$

which can be simplified as

$$L_i = \frac{M^2 - (ir)^2}{r^2(2i-1)} \tag{2}$$

In addition to forwarding L_i number of packets, the forwarding node has also its own packet to transmit. Therefore, a node transmits $L_i + 1$ packets in one cycle. Consequently, the effective arrival rate at the node is obtained

$$\lambda_i = \lambda_e(L_i + 1) \tag{3}$$

Let $1/\gamma_i$ be the mean service time at the node i. The utilization factor, denoted by ρ_i, is given by

$$\rho_i = \lambda_i/\gamma_i \tag{4}$$

In practice, when a packet is relayed from i^{th} ring to the actor node, it would require at least *i* number of hops. To satisfy this minimality condition, the nodes density in the field should be sufficiently large. It has been shown that the sensors covering a region are connected as long as their communication range is not less than twice the sensing range [14]. In order for two sensors to communicate they have to be within r distance of each other. In other words, one sensor within a distance of r is a must requirement to remain connected. For N rings (M/r) where each ring is of depth r, there has to be N sensors along a radius of $N \times r$. So, approximately the lower limit on the sensor density per unit area of r^2 will be one. For an area of πr^2 with $\mu(r) = N_s \times \pi r^2/\pi M^2$ nodes in it, we approximate the number of nodes along the diameter as $\sqrt{\mu(r)}$. Hence, the nodes along a distance of r is $\sqrt{\mu(r)}/2$. We expect $\sqrt{\mu(r)}/2$ to be greater than 1 for connectivity. Therefore, the chances of nodes being disconnected (σ) can be approximated as

$$\sigma = \begin{cases} 0 & \text{if } \sqrt{\mu(r)} \geq 2 \\ 1/\sqrt{\mu(r)} & \text{otherwise} \end{cases}$$

The number of hops ($H(i)$) required to relay the packet from i^{th} ring to the actor can be computed as

$$H(i) = i(1 + \sigma) \tag{5}$$

Therefore, the average number of hops $E[H]$ is obtained as

$$E[H] = \frac{r}{M} \sum_{x=1}^{M/r} H(x) \tag{6}$$

Bisnik et. al. [5] has modeled the multihop wireless ad hoc networks by using diffusion approximation to solve an open G/G/1 queuing network. All the nodes in the network are considered single server with first-come first serve (FCFS) service policy. Let $D(i)$ represents the packet delay when traveling from i^{th} ring toward the inner $(i-1)^{th}$ ring. This includes the medium access delay as well as the queuing delay. We obtain the average one hop packet delay $\bar{D}(i)$ as given in [5]

$$\bar{D}(i) = \frac{\bar{\rho_i}}{\lambda_i(1 - \hat{\rho})} \tag{7}$$

The average end-to-end delay can be hence formulated as $E[H] \times E[D]$. where $E[D]$ is the mean one hop delay of all the rings. The packet deadline τ for given topology parameters N_s, r, M and λ_e must satisfy

$$\tau \geq E[D] \times H(M/r)$$

However, if the above condition can not be met then there are two possible solutions; either the number of hops must be reduced or the average delay should be reduced. The former case can be implemented by deploying more than one actor in the field, which results in distributing the field among multiple destinations. This provides sensor nodes the flexibility to transmit packets to their closer destination actor. While, the second case can be implemented by limiting the event sampling rate or increasing the channel utilization. However, this is not a practical approach since it puts constraints on applications in terms of reporting rate. Therefore, we advocate the first case and formulate a relationship between the τ and the respective number of actors (N_a) which make sure that the event response time is achievable. Therefore,

$$\tau \geq \frac{E[D] \times H(M/r)}{N_a}$$

In particular, we allow the actor nodes to relocate in the field in response to emergencies as descried in section III-C.

III. Adaptive Pull-Push (APP) Coordination

In this section, we present the adaptive coordination approach APP in order to track the critical events efficiently. In this model, the Push corresponds to the case where sensors continuously monitor the environment to make sure that they do not miss any activity in the field. In the Pull mode, on the other hand, actors instruct the sensors when to sample the environment. This helps in defining *duty cycle* of the sensor nodes. *The duty cycle is defined as the ratio of wakeup period to sleep period.* The Pull mode allows sensors to sleep at times when no monitoring is required and, therefore, allows energy conservation. However, consider a scenario where the network is deployed to observe different emergency situations that can not be predicted a-priori. In such a case, it is possible that an event occurs prior to the wakeup time and is missed by the sensor node. This motivates to integrate the Push and Pull to achieve the benefits of both.

APP first starts in Pull mode to conserve energy. It then adapts to Push mode if an emergency situation is reported and hence, called as Adaptive Pull-Push mode. That is, the nodes are instructed to skip their sleep cycle when an emergency is likely to happen otherwise they follow their duty cycle. This aims at conserving energy like Pull but behaving like Push when necessary, i.e., when an emergency is detected.

In APP, as actors move within a geographical delta they report their location, *event of interest*, *report interval* (γ) and *action response capabilities* including ϵ using a long-range in-network subscription message. Sensors receiving such subscription messages are expected to publish their observations of interest to the actor nodes, resulting in a publish/subscribe relationship. APP is implemented in two versions: 1) Neighborhood-based APP (NAPP), where the neighborhood, e.g. close vicinity of an observed emergency, are alerted, and 2) Region-based APP (RAPP), where the sensor nodes in a larger area, referred to as region, are alerted when

necessary. In the following subsection, we describe these two approaches in more detail.

A. Neighbor-based APP (NAPP)

In NAPP, an actor node determines the neighborhood of a node i reporting an emergency and alerts the nodes in the vicinity for possible follow up events. The neighborhood is based on the assumed event radius ($\Re$) centered at the location $C(x_i, y_i)$ of the reporting sensor node i. Actor estimates $\Re$ by considering the intensity of event based either on the past statistics or the locations of reporting nodes. This zone is declared as alert zone by the actor. The actor node broadcasts the alert message with a location-based description ($\Re$ and $C(x_i, y_i)$) of the neighborhood for a sufficiently long enough time (T_a) to make sure that each sensor in the neighborhood will receive the message during their duty cycle.

For instance, if the asleep period is T_s seconds and awake period is T_w seconds then T_a should prolong for $T_s < T_a < T_s + T_w$ seconds in order to make sure that all the nodes in the alert zone hear the message. However, the frequency of this message during T_a is set to $1/T_s$ in order to avoid unnecessary messages. For instance, if the nodes are asleep for 30 seconds and then awake for 10 seconds, the duration of the broadcast should be more than 30 seconds to guarantee that nodes will be awake to receive the alert. This is required since each node applies the duty cycle on an independent clock. In other words, there is no predetermined or synchronized wake up and sleep period of the nodes

In Fig. 3, we demonstrate a sensor network with an actor node, represented by a triangle, patrolling the field. In Fig. 3(a), an emergency occurs in the field and is observed by a single sensor in the impact area having two nodes. The reporting node is marked with a darker shade. The other sensor is asleep during this time and is unable to observe the event. When the awake node reports its observation to the actor node, the actor node estimates an area of impact for possible spread-outs using a circle with radius $\Re$ as shown in Fig. 3(b). All nodes within this region are alerted to stay awake for an emergency that might follow.

Note that NAPP has some expected latency depending on the duty cycle of the nodes. At the extreme case, if all nodes in the neighborhood are asleep for a long duration of time, they would not be able to receive the alert from the actor and to monitor the follow-up events during that time. If the emergency spreads at a faster rate, it might be more efficient to alert nodes within a larger region to increase the number of nodes that will receive the alert to observe subsequent emergencies.

B. Region-based APP (RAPP)

NAPP is employed when the event speed is slower such that a number of subsequent events occur in the emergency zone of radius $\Re$. At faster speed of event, it would be possible that the event is missed and actors will have to declare new emergency zone in order to track the event or if it happens frequently then the emergency zone can be expanded by increasing the value of $\Re$. At larger value of $\Re$, more number of nodes will remain in wakeup mode thereby consuming more energy. Intuitively, the NAPP approach will behave similar to the pull approach as the value of $\Re$ approaches to the field radius. In order to overcome the inefficiency of NAPP at higher event speed, we propose a Region-based APP (RAPP). The actor will switch between the two approaches by observing the event speed. Eventually NAPP will be employed at lower event speed and RAPP at higher event speed.

In RAPP, rather than confining the emergency zone to a particular region of radius $\Re$, the actor disseminates the alert message to a much larger region i.e., its responsibility cluster or the whole field. A single parameter, an alert threshold (κ) between $[0, 1)$, limits the number of nodes that will stay awake within this region to save from the overall energy consumption. The value of κ is computed according to the observed reliability and can be adjusted to ensure certain reliability. It is worthwhile to note that higher value of $kappa$ would alert larger number of nodes resulting in more energy consumption. Actor nodes compute it dynamically according to required reliability of the application and observed reliability tracked by the actors. Here, the reliability is considered as the number of reports per event presumably received within the event specific deadline.

Each node generates a random number between $[0, 1)$, referred to as p_{alert}, when the alert message is received from the actor carrying κ. If the value of p_{alert} is within the range $[0, \kappa)$, the node switches to

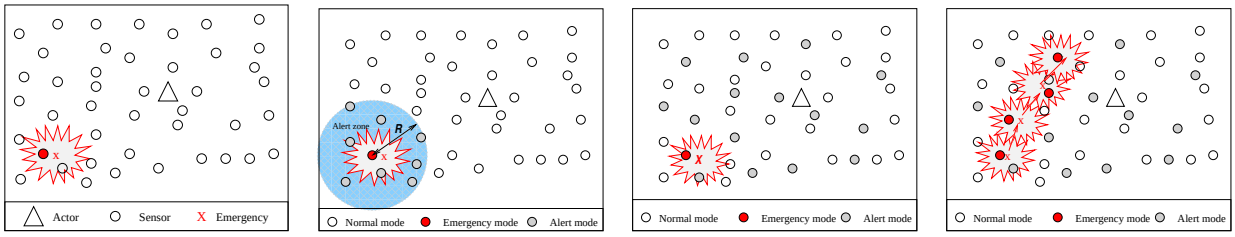


Fig. 3. An example scenario where an event originates at point *X* in the field in Fig. 3(a), NAPP approach to alert nodes about emergency in Fig. 3(b), RAPP approach to alert nodes about the emergency in Fig. 3(c) and emergency detection in RAPP when the event moves in Fig. 3(d).

alert mode otherwise, the message is ignored. The value of κ is computed as follows; We assume that the radius $\Re$ of emergency event can be determined or estimated.

With uniform distribution of nodes, we can compute the number of nodes in the emergency area by $N_s \times \pi \times \Re^2/\pi M^2$. Assuming that the emergency reported by any single node in the emergency zone is sufficient for learning about an abnormal activity then we can determine the value of κ as

$$\kappa = \eta \frac{M^2}{N_s \times \Re^2}, \qquad \eta >= 1 \tag{8}$$

where η is the factor that controls the number of nodes to stay awake within the emergency zone. Fig. 3(c) and 3(d) demonstrates the case where RAPP alerts the nodes in the region with a parameter κ of 0.5; the shaded nodes correspond to nodes that stay awake as a result of this communication. As can be seen RAPP aims at alarming a minimal number of nodes within the region by tuning the κ parameter.

C. Field Coverage and Actor Coordination

Based on the capabilities of an actor node, it can be able to respond to a particular emergency in the field, e.g., spraying fire extinguisher in an area, etc. In this regard, each actor has an action range (*R*) which is different from the transmission range (*r*). An actor can respond to emergencies within distance R to its current location without relocating itself. A sensor node is said to be covered by an actor if its distance to that actor lies within the limit *R* otherwise it is assumed to be uncovered. If an actor does not cover an event area, that is it can not respond to the sensor node reporting an event from that area without relocating, it can coordinate with the other actors so that the uncovered field is targeted by some actor.

Actors are assumed to be able to attend to an emergency anywhere in the field. For this purpose, an actor is moved towards an emergency zone if the source node is not within its action range. We refer to this as a controlled targeting, i.e. actor moves toward the requesting source such that the source node is within its action range.

Sensor nodes send their readings to the nearest actor. However, an actor receiving an emergency report may not be able to act on the event itself due to its small action range or low response capability. This requires exploiting the coordination between the actor nodes to track the target area. For this purpose, each actor node maintains a list of actors in order to trigger coordination. As an actor receives a request from a sensor node that lies outside its action area, it needs to determine if it should respond to this emergency or if it should delegate another actor node for response.

1) Event Targeting: In our analysis, we assume that all the actors have the same action range, represented by a circle of radius *R*. When an actor a_i, residing at the point $I(x_i, y_i)$, receives a request from the source s at position $J(x_j, y_j)$, it calculates the distance $D(I,J) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ from the source. If $D(I,J) \leq R$, a_i is able to respond to the request. However, if $D(I,J) > R$, there are two possible scenarios. In the first scenario, the actor a_i does not have any commitments, i.e., no other event suggesting an emergency is reported from a different location. In the second scenario, the actor is already alarmed to control a possible emergency when a new report from a different location arrives. An actor a_i needs to move $D(I,J) - R$ units of distance toward the target location at some new point $\bar{I}(\bar{x_i}, \bar{y_i})$ so that the coverage is provided to s as shown in Fig. 4. The new actor location $\bar{I}$ can be computed as follows

Taking $I(x_i, y_i)$ as the center of XY plane, we can

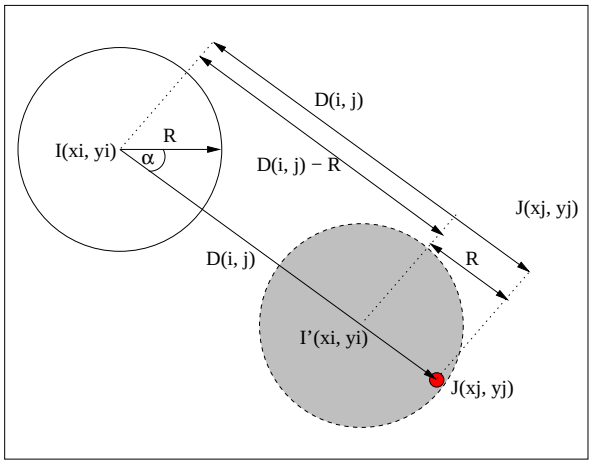


Fig. 4. Positioning actor toward uncovered source

```
   /*Pseudo-code executed by the actor a to
determine the coverage of source i.*/
   TargetEvent(i)
      d(i,a) = √((x_a − x_i)^2 + (y_a − y_i)^2);
      C_a = coverSource((d(i,a) − R), (x_i, y_i));
      /* New center position C_a of a if it
can move d(i,a) − R toward (x_i, y_i).*/
      MoveOK = true;
      for(each s ∈ sources_a)
        MoveOK = canCover(C_a, R, (x_s, y_s));
        if (MoveOK = false)
         StartAACoordination(i);
         exit;
       end //if
     end //for
     moveTo(C_a);
```

Fig. 5. Algorithm of action coverage of a source.

find the coordinates $\bar{x_i}$ and $\bar{y_i}$ as

$$\bar{x_i} = x_i + (D(I,J) - R) \times cos(\alpha)$$
$$\bar{y_i} = y_i + (D(I,J) - R) \times sin(\alpha)$$

where α is the angle between the vector $X(R,0)$ along x-axis and the vector $A(x_j - x_i, y_j - y_i)$.

To find out the value of α, first compute the length $|X|$ and $|R|$ of two vectors X and R.

$$|X| = R,\ |A| = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

Normalize each vector (unit vector)

$$X_u = (R,0)/|X|,\ A_u = (x_j - x_i, y_j - y_i)/|A|$$

Hence, the angle between X and A is:

$$\alpha = cos^{-1}(X_u.A_u) \tag{9}$$

However, the displacement is only possible if the movement of a_i does not leave the presently attended sources uncovered. In the second scenario, if the displacement of a_i is not not feasible due to a previous alarm that will be left off uncovered if the actor moves to this emergency, then it initiates the actor-actor coordination algorithm.

2) Actor-Actor Coordination: The coordination among actors is triggered by broadcasting the *relocate* message that contains the location of the source *s*. Each actor receiving this message runs the procedure outlined in Fig. 5 to check for the possibility of attending to the reporting source. An actor a_j sends back the *relocation-ok* message to a_i if it is able to move toward the source. This message contains the residual response power of actor a_j, i.e., remaining power extinguisher source, as well as the total number of sources being attended. If the field is being operated by sufficient number of actors such that all the sources can get response from the actors then the actor a_i receives *relocation-ok* message from some of the neighboring actors. Consequently, a_i forwards the request to the actor covering lesser number of sources and has the highest residual energy or resources.

When the actors have either limited action range or they are small in number to cover the whole field, it is quite possible that some action requests remain unsatisfied. This situation arises when the intensity of events is too high to affect wider area. We can approximate the number of unattended sources α in uniform deployment for $M \times M$ field as

$$\alpha = \frac{M^2 - N_a \times \pi R^2}{M^2} \times N_s \tag{10}$$

where N_a is the number of mobile actors, N_s is the total number of deployed sensors. However, the value of α depends on the density of nodes in the field. In case of non-uniform deployment when the action range is limited, actors might be stuck in sparse zone of the field that can not attend the nodes in dense part as described above, which greatly effects the value of α.

Fig. 6 illustrates the target tracking in APP. Source *Y* publishes its emergency report to the actor *A* since it is the nearest one and expect response from *A*. However, *Y* is not in the action coverage of *A* and would require A-A coordination in order to be served. If *A* moves to cover the event area of *Y* then it leaves the current sources uncovered. Therefore it coordinates with *B* and finds that *B* is able to respond in area of *Y*. Although $D_{ay} < D_{by}$ but *B* is

the right actor because the position of its existing sources permit it to attend *Y*. This is depicted in shaded area covering *Y*.

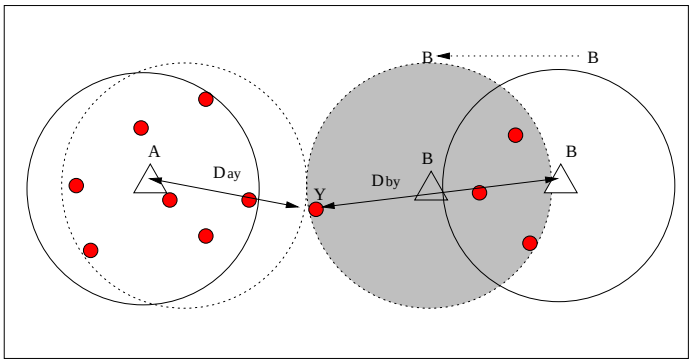


Fig. 6. A-A coordination to target the event area.

IV. Performance Evaluation

The performance of APP is evaluated using network simulator *ns-2* [33]. The experiments are run to study the response time of actors to emergency events and successful emergencies observed in APP. Energy consumption is also considered as an important parameters in addition to response time. *The emergency event in the scenario is considered a critical pop-up event which appears in the field at some point and may change its impact area with the passage of time*. For instance, an intruder may enter in the restricted premises and subsequently move to different parts of the field. Therefore, the sensor nodes sampling the casual events like light, temperature etc periodically, are also assumed to observe the emergency events. We compare the performance of NAPP and RAPP with the traditional data flow models namely, Push and Pull as described in Section III.

TABLE I
Parameters setup for the first scenario

Parameter	Value
τ (*deadline*)	250 ms
r (*transmission radius*)	20 m
sampling interval	0.5 sec
emergency interval	1 sec
R (*action range*)	30 m
$\Re$ (*event radius*)	20 m
κ (*alert factor*)	0.5

A. Response Time

Response time is evaluated using the parameters selected as given in Table I. In this setup, 150 sensor nodes are randomly deployed in $100 \times 100m^2$ area with different number of actor nodes varying from 1 to 4. Sensor nodes report the emergency observed during their wakeup period to the subscribed actor allowing the actor nodes to take full responsibility such that they relocate or organize other actors to respond to reported emergencies.

Fig. 7 shows the average response time in four approaches; Push, Pull, NAPP and RAPP. It is observed that the response time in Pull is much larger as compared to other approaches, which is approximately 6 times higher than the Push and 12 times higher than NAPP and RAPP. It is due to the fact that emergency happens in a region where the nodes are in sleep mode and miss that event resulting in higher response time.

In contrast, nodes continuously monitor the field in Push mode and never go to sleep mode. This helps in catching all the possible events. However, it produces extra traffic which might congest the network and makes the event delivery late or sometimes packets are dropped. This is implied in Fig. 8, where the deadlines miss-ratio is higher as compared to other approaches.

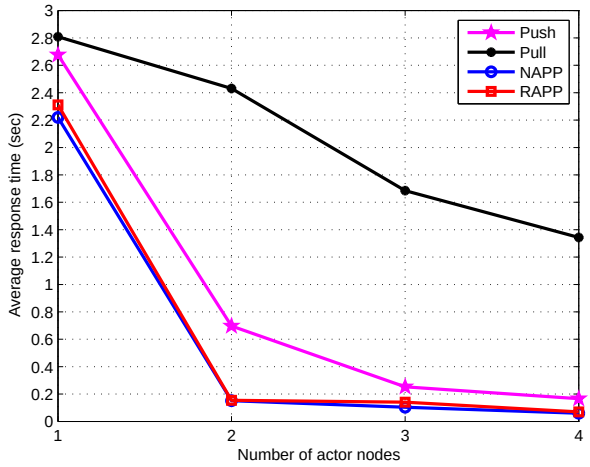


Fig. 7. Average response time at different number of actor nodes.

Initially, with a single actor in the entire field, the whole traffic rushes to that actor and results in significantly higher deadlines-miss ratio. The value in NAPP is 50 % lower than Push, although 25 % higher than Pull but reponse time is much lesser than Pull as described earlier. However, the response-time and miss-ratio is reduced as we increase the number of actors in the field. As analysed in Section II-B that if the deadline is shorter and is not possible

to meet with single destination then increasing the number of actors would improve the performance. The miss-ratio in RAPP is slightly higher than NAPP and Pull but we will see that it is an adaptive approach which tradeoff between energy and reliability. Hence, Pull in terms of response time and Push in terms of miss-ratio are outperformed by NAPP and RAPP within the whole range. As expected the difference between the approaches diminishes as we increase the number of actors in the field. This is due to the fact the communication to the actor node becomes a major bottleneck in the system. Note that RAPP and NAPP achieves their deadline with only two actors that is not possible with other two, reducing the application cost too.

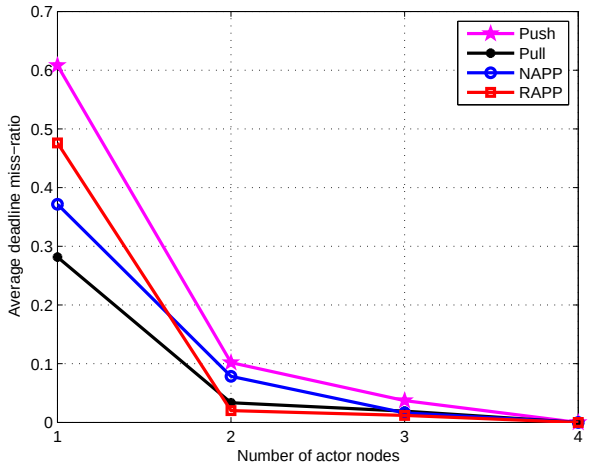


Fig. 8. Average miss-ratio at different number of actor odes for deadline of 250ms.

Similarly, if emergency event is to be observed by all the nodes in the event impact area for the reliability purpose then Pull has the worst performance in comparison to the others as seen in Fig. 9, which plots the average observation reports for the same experiment. However, if a single report from the impact area is sufficient then Pull is the potential approach in terms of energy conservation.

To study the energy consumption in the process we plot the average energy consumption per node for the same experiment in Fig. 10. As suggested Pull is the model with the least energy consumption since nodes sleep during the times they are not supposed to be reporting any activity to the actor nodes. At the other extreme, Push has the highest energy consumption since nodes monitor the environment continuously. While the energy consumption in NAPP and RAPP is in between Pull

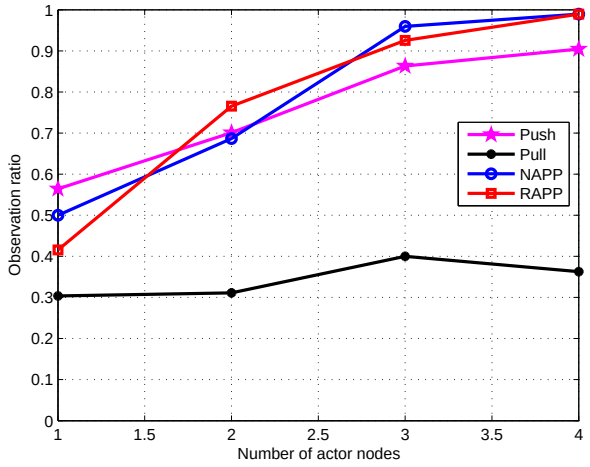


Fig. 9. Average emergency reports at different number of actor nodes.

and Push, that is 25 % higher than Pull but with same value lower than Push.

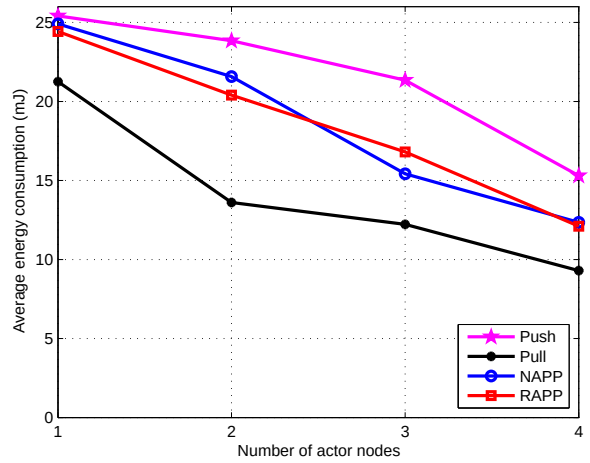


Fig. 10. Average energy consumption per node at different number of actor odes.

B. *Emergency Observations*

In order to model the emergency events occurring in the field we exploit a synthetic *event* model as follows. We use a random walk model with a varying velocity that is between 0-10 meters/sec. It allow the event to be monitored only at specified intervals. For instance, if the event interval is set to 2 seconds, the event is allowed to be observed by the nodes every 2 seconds. In order to observe the emergency observation delay and successfully emergencies observed, we select the sampling interval 3 seconds and emergency interval varies from 1 to 5 seconds. In this scenario, 100 sensor nodes,

deployed randomly in $100 \times 100 m^2$ area and 4 actors are also placed in the field at random locations. We change event radius $\Re$ and action range R in this scenario to observe the successfully observed events.

Fig. 11 illustrates the mean delay observed to report the emergencies at various intervals. At lower emergency intervals, the emergency is likely to spread within the vicinity of previous observations. As the emergency interval is increased, the emergency moves to further away points than previously observed locations. We observe Pull has the highest latency as expected, which is approximately 10 times higher than the NAPP and Push. In particular, Pull enforces a fixed sampling period in the network regardless of observations being reported to the actors. As a result it is possible that events are missed during the sleep cycles of the nodes. We see that the delay of Pull increases exponentially. In contrast, we observe a flattening out in the delay observed by the RAPP approach suggesting that for higher intervals of emergency, the difference between RAPP and Pull will become even more significant. NAPP provides a performance that is almost as good as Push within the whole range.

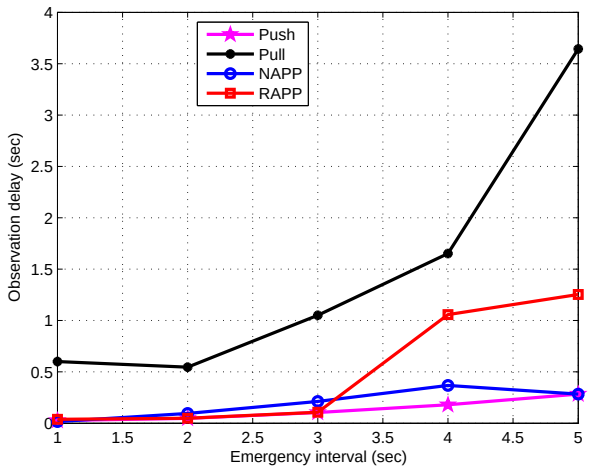


Fig. 11. Observation Delay as the emergency interval is enlarged.

When we analyze the energy consumption for the same experiment, as plotted in Fig. 12, we see that Push which was providing the lowest latency has the highest energy consumption of all. As a result we conclude that it is not the best approach considering factors other than latency. In the figure we see that Pull has the lowest energy consumption. Yet it had the worst latency performance as demonstrated in the previous figure. Adaptive Pull-Push techniques, NAPP and RAPP are shown to have reasonably close energy consumption as of Pull. And we have observed both of these approaches have better performance in terms of latency in comparison to Pull. In this regard, they provide a good trade-off between latency and energy consumption.

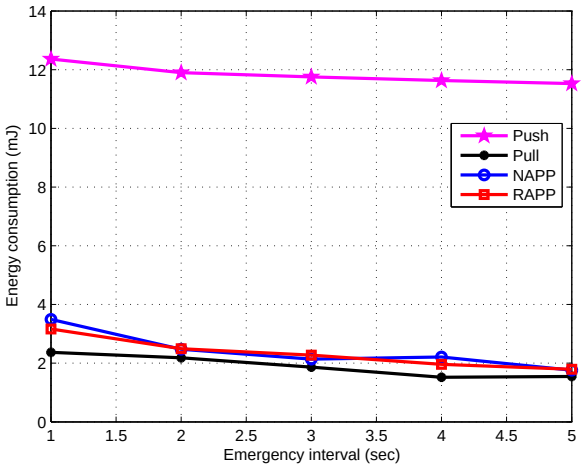


Fig. 12. Observation Delay as the emergency interval is enlarged.

In our next set of experiments, we study how APP scales to larger networks. For this purpose, we gradually increase the number of nodes from 100 to 500 while spreading out the network, and measure the observation ratio while reporting the emergencies to the same number of actor nodes. As seen in Fig. 13, APP's performance drops as the number of nodes in the network increase. However, its performance is still better than Pull within the whole range.

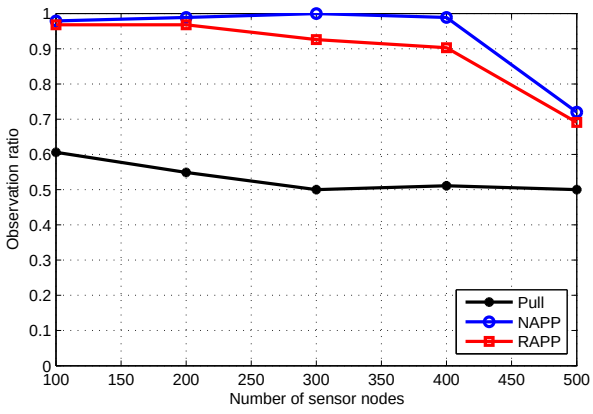


Fig. 13. Observation ratio at emergency and sampling interval 1 sec with different number of sensor nodes.

We plot the energy consumption measured for

the same experiment in Fig. 14. As can be seen the energy consumption of both APP approaches are close to Pull and significantly lower than Push within the whole range.

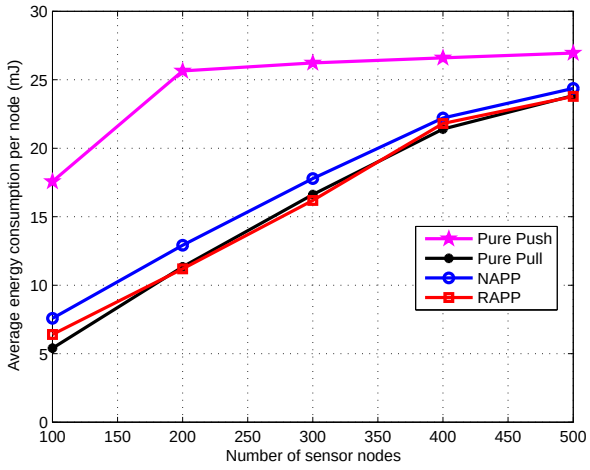


Fig. 14. Average node energy consumption at emergency and sampling interval 1 sec with different number of sensor nodes.

Obviously, the energy consumption of RAPP can be adjusted by the alert threshold (κ) parameter, which is the key in the design of this approach. To study the impacts of this parameter we varied it between 0 and 100 % of the total number of nodes as shown in Fig. 15. At 0 no nodes are alerted to watch for emergencies, and the observations are based on the duty cycles of the nodes, i.e., when a node is awake. That is, an emergency is reported to the corresponding actor only if a node is in wakeup state in the emergency impact area. We have studied three different event radii in this experiment from 10m to 30m. For impact area with 10 m radius, the number of nodes that in this area can observe the emergency is the lowest in comparison to the other two cases. As a result, we observe the lowest observation ratio for 10 m; followed by the other two cases. It is due to the fact that there will be low probability that the nodes will detect the emergency causing the emergency to escape from being observed. It is unlike with the higher event radius because larger number of nodes will lie in the impact area increasing the probability of detecting emergency.

On the other hand, emergency observation is also controlled with the value of κ in RAPP. With lower value of κ, lesser nodes will be alerted resulting in low emergency observation. However, with smaller value of $\Re$, increasing the κ improves the observation ratio. As a result, actors can set κ to larger value for smaller impact area and vice versa. There is a tradeoff between κ and energy consumption as plotted for the experiment in Fig. 16.

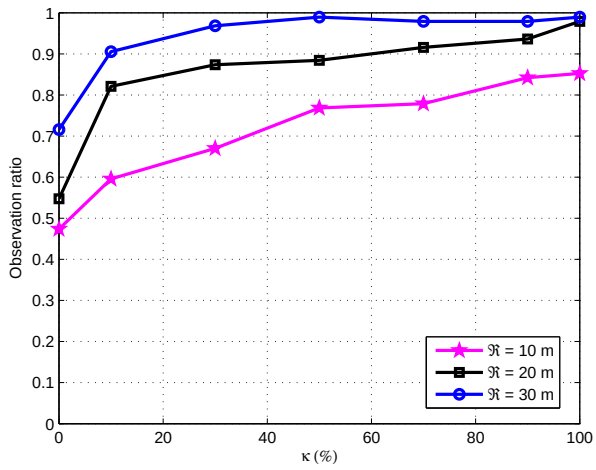


Fig. 15. Observation ratio at different values of κ in RAPP at emergency and sampling interval 1 sec.

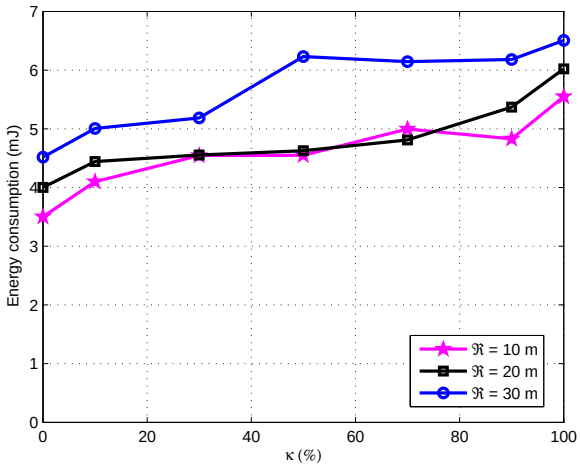


Fig. 16. Average node energy consumption at different values of κ in RAPP at emergency and sampling interval 1 sec.

V. Conclusion

We have presented an Adaptive Push-Pull (APP) based event tracking mechanism in wireless sensor and actor networks. Pull and Push are the two traditional methods of event dissemination in WSNs. However, they are not efficient when the nodes are assigned the tasks to observe both the casual events at some predefined intervals as well as emergency events which may occur at any time, anywhere and spread-out in the entire field. The Pull approach

has the drawback of higher latencies in detecting the events since the nodes follow duty cycle and may be asleep when an emergency situation occurs. On the other hand, the pure Push approach results in high energy consumptions due to the continuous monitoring and is not attractive in WSANs.

APP is an adaptive approach which integrates the salient features of both by providing better response time with lower energy consumption. It first starts in Pull mode to conserve energy and then adapts to Push mode if an emergency situation is reported and hence, called as Adaptive Pull-Push mode. That is, the nodes are instructed to skip their sleep cycle when an emergency is likely to happen otherwise they follow their duty cycle. Furthermore, APP exploits the mobility of actor nodes to relocate them in the vicinity of emergency zones such that the emergency area is under the action coverage of actors. This helps in providing immediate response and thus increases the reliability of applications.

APP is implemented in two versions: 1) Neighborhood-based APP (NAPP), where the neighborhood, e.g. close vicinity of an observed emergency, are alerted, and 2) Region-based APP (RAPP), where the sensor nodes in the entire responsibility cluster area, are alerted when necessary. RAPP allows to control the number of alerted nodes through its alert threshold parameter (κ) making the approach more attractive due to its controlled energy consumption. Simulation results prove that the response time in APP is much closer to Push with lower energy consumption like Pull. Moreover, the performance in APP can be achieved with lesser number of actors and thus reducing the application cost too.

REFERENCES

[1] I. F. Akyildiz and I. H. Kasimoglu, "Wireless Sensor and Actor Networks: Research Challenges," *Ad Hoc Networks, Vol. 2, Issue 4, pp. 351-367, Oct 2004*.

[2] G. A. Shah, M. Bozyiğit, Ö. B. Akan and B. Baykal, "Real-time Coordination and Routing in Wireless Sensor and Actor Networks," *in Proc. of 6th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN), May 2006*.

[3] G. A. Shah, M. Bozyigit and Demet Aksoy, "RAT: Routing by Adaptive Targeting in Wireless Sensor/Actor Networks" *in Proc. of 2nd IEEE/ACM International Conference on COMmunication System softWAre and middlewaRE (COMSWARE), January 2007*.

[4] S. Kapadia1, B. Krishnamachari, "Comparative Analysis of Push-Pull Query Strategies for Wireless Sensor Networks," *in Proc. of International Conference on Distributed Computing in Sensor Systems, Jun 2006*.

[5] N. Bisnik and A. Abouzeid, "Queuing network models for delay analysis of multihop wireless ad hoc networks," *In Proc. of International Wireless Communications and Mobile Computing Conference (IWCMC), July 2006*.

[6] V. Mhatre and C. Rosenberg, "Design guidelines for wireless sensor networks: communication, clustering and aggregation," *Ad Hoc Networks, vol. 2, no. 1, pp. 45-63, January 2004*.

[7] L. Zhang, Y. Shu, O. Yang and G. Wang, "Study of Medium Access Delay in IEEE 802.11 Wireless Networks," *IEICE Transactions on Communications vol. E89-B no. 4, pp. 1284-1293, 2006*.

[8] T. Melodia, D. Pompili, V. C. Gungor and I. F. Akyildiz, "A Distributed Coordination Framework for Wireless Sensor and Actor Networks," *in Proc. of ACM MobiHoc'05, May 2005*.

[9] E.C.H. Ngai, M.R. Lyu and J. Liu, "A Real-Time Communication Framework for Wireless Sensor-Actuator Networks," *in Proc. of IEEE Aerospace Conference, pp. 1-9, Mar 2006*.

[10] H. Peng, W. Huafeng, M. Dilin and G, Chuanshan, "ELRS: An Energy-Efficient Layered Routing Scheme for Wireless Sensor and Actor Networks," *In Proc. of 20th International Conference on Advanced Information Networking and Applications, pp. 452-460, 2006*.

[11] X. Z. Yang, Z. GuangSheng, D. WenHua and F. Qi, "A Hop-Bounded Single-actor Selection Algorithm for Wireless Sensor and Actor Networks," *In Proc. of International conference on Communications and mobile computing, pp. 1067-1072, Jul 2006*.

[12] Arjan Durresi, Vamsi Paruchuri and Leonard Barolli, "Delay-Energy Aware Routing Protocol for Sensor and Actor Networks," *In Proc. of 11th International Conference on Parallel and Distributed Systems (ICPADS'05)*.

[13] K. Liu and N. Abu-Ghazaleh, "HGR: Hybrid Geographic and Virtual Coordinate Routing for WSN," *http://arxiv.org/abs/cs.NI/0511059 Nov 2005*.

[14] Xiaorui Wang, Guoliang Xing, Yuanfang Zhang, Chenyang Lu, Robert Pless and Christopher Gill, "Integrated coverage and connectivity configuration for energy conservation in sensor networks," *ACM Transactions on Sensor Networks, vol. 1, issue 1, pp. 36-72, August 2005*.

[15] D. Aksoy, M. Leung, "Pull vs. Push: A Quantitative Comparison," *in Proc. of IEEE GLOBECOM, December 2004*.

[16] J. Kulik, W. Heinzelman, and H. Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks," *Wireless Networks, Vol. 8, pp. 169-185, 2002*.

[17] K. Akkaya, M. Younis, "An Energy-Aware QoS Routing Protocol for Wireless Sensor Networks," *in Proc. of 23rd ICDCSW, pp. 710-715, 2003*.

[18] D. Tian and N.D. Georganas, "Energy Efficient Routing with Guaranteed Delivery in Wireless Sensor Networks," *Mobile Computing and Communications Review, 2001*.

[19] Y. Yu, R. Govindan and D. Estrin, "Geographical and Energy Aware Routing (GEAR): a recursive dissemination protocol for wireless sensor networks," *UCLA/CSD-TR-01-0023, Tech. Rep., 2001*.

[20] M. Wong, and D. Aksoy, "Relative Accuracy based Location Estimation in Wireless Ad-Hoc Sensor Networks," *in Proc. IEEE ICC (International Conference on Communications), Glasgow, Scotland, UK, June 2007*.

[21] B. Karp and H. T. Kung, "GPSR: Greedy Perimeter Stateless Routing for Wireless Networks," *in Proc. of MobiCom, 2000*.

[22] K. Langendoen and N. Reijers, "Distributed Localization in WSN:A Quantitative Comparison", *Elsevier, Computer Networks, 2003*.

[23] X. Sheng and Y. H. Hu, "Collaborative Source Localization in Wireless Sensor Network System". *in Proc. of IEEE Globecom, 2003*.

[24] Ling Xuan Hu and David Evans, "Localization for Mobile Sensor Networks", *in Proc. of ACM Mobicom, pp. 45-57, 2004*.

[25] H. Peng, G. Chuanshan, Y. Min, M. Dilin and Y. Bo, "ECLS: An Efficient Cooperative Localization Scheme For Wireless Sensor and Actor Networks," *In Proc. of 5th International Conference on Computer and Information Technology (CIT'05) pp. 396-400, 2005*.

[26] T. He, J.A. Stankovic, C. Lu and T. Abdelzaher, "A Real-Time Routing Protocol for Sensor Networks," *In Proc. of IEEE International Conference on Distributed Computing Systems, May 2003*.

[27] C. Intanagonwiwat, R. Govindan and D. Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks," *In Proc. of ACM MobiCom, March 2000*.

[28] X. Liu, Q. Huang and Y. Zhang, "Combs, Needles, Haystacks: Balancing Push and Pull for Discovery in LargeScale Sensor Networks," *in Proc. of ACM Sensys, Nov 2004*

[29] N. Trigoni, Y. Yao, A. Demers and J. Gehrke, "Hybrid Push-Pull Querying Processing for Sensor Networks,"*In Proc. of the Workshop on Sensor Networks (WSN) 2004*

[30] Y. Xinqing, W. Wei, L. Wenfeng and C. Dingfang, "A Publish/Subscribe Architecture for Sensor Networks and Applications", *in Proc. of IEEE International Conference on Mechatronics and Automation, 2006*.

[31] P. Costa and G. P. Picco , "Publish-Subscribe on Sensor Networks: A Semi-probabilistic Approach", *In Proc. of MASS 2005*,

[32] A. Günther and C. Hoene, “Measuring Round Trip Times to Determine the Distance between WLAN Nodes,” *In Proc. of Networking 2005, Waterloo, Canada, May 2005*.
[33] UC Berkeley, LBL, USC/ISI and Xerox PARC, “The Network Simulator ns-2,” *The VINT Project, http://www.isi.edu/nsnam/ns/*.