Proceedings of the First International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)

Held at the 9th Extended Semantic Web Conference (ESWC 2012)

May 28th, Heraklion, Greece

Preface

The 1st International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012) will be held in Crete in conjunction with 9th Extended Semantic Web Conference (ESWC 2012).

FEOSW is unique in targeting both Semantic Technologies and Financial and Economic Research, providing an opportunity for researchers in both fields to exchange opinions and experiences or delving into further needs in both domains on how to foster the development of financial information through new Semantic Technologies breed such as Linked Data. The aim of the workshop is to harness the benefits for different fields bringing along both research directions. In the following, we will discuss where the main synergies can be found.

“There are three ways to make a living in this business: be first, be smarter, or cheat”. These are words quoted by fictional character John Tuld, in “Margin Call”, a 2011 American independent drama film written and directed by J.C. Chandor about a loosely-modeled on Lehman-Brothers investment bank collapse during the financial crisis of 2007-2008. The movie ponders on how financial information, particularly historical volatility levels and excessive leverage in a firm assets (in this case, Mortgage Based Securities (MBS)) can lead to the destruction of the firm. Interestingly enough, the alarm is fired by a trader who accidentally checks a number of financial models and finds out what has been happening recently in the firm and how critical is the risk they are taking. Even though, the movie is fiction, it clearly shows and discusses a very key issue in the last economic downturn: financial information is not widespread, it is difficult to interpret and it should be more accessible and understandable not only for the financial markets experts but also for the layman investor.

In FEOSW, we would like to discuss how Semantic Technologies can provide a formal and powerful basis to make financial information more understandable and easy to interpret for Information Systems. Recent evolutions of these technologies, such as Linked Data provide a solid underlying framework to manage, store and query financial information more efficiently. Shifting from 10-K reports with confusing Balance Sheets and Income Statements (these being enforced through XBRL by the Securities Exchange Commission (SEC) in a near future) to RDF-based reports will enable intelligent management of vital information and markets, together with a broad lattice of financial extraction and reasoning services that could foster a new era for smart financial systems.

Many hearty thanks to all our contributors and participants at FEOSW 2012 and also the Programme Committee whose very valuable and insightful feedback has resulted in a challenging and fruitful collection of papers, posters and demos, providing added value to current leading edge research.

We are positive that the FEOSW series of workshops will continue over time in future Semantic Technologies related events and they will be able to bridge the gap between current financial information caveats and new emerging technologies which could impact dramatically in the state of the art, marshaling a new era for Financial and Economic Information Management.

Finally, we want to thank you for their support to the Ministry of Economy and Competitiveness with FLORA project (TIN2011-27405) and SoftLab group of University Carlos III of Madrid. Also, a special gratitude to Dr. Sean O'Riain from DERI for accepting our invitation and participate as keynote speaker of FEOSW 2012.

May 2012

Ángel García-Crespo
Juan Miguel Gómez-Berbís
Alejandro Rodríguez-González
Brahmananda Sapkota

Organization

Organizing Committee

Ángel García-Crespo (Universidad Carlos III de Madrid, Spain)
Juan Miguel Gómez-Berbís (Universidad Carlos III de Madrid, Spain)
Alejandro Rodríguez-González (Universidad Carlos III de Madrid, Spain)
Brahmananda Sapkota (University of Twente, The Netherlands)

Program Committee

Ricardo Colomo Palacios, PhD, University Carlos III of Madrid, Spain
Ioan Toma, PhD, Semantic Technology Institute (STI), Austria
Fernando Guldris Iglesias, Banco Santander, Spain
Knud Moller, PhD, Digital Enterprise Research Institute (DERI), Ireland
Rafael Valencia Garcia, PhD, University of Murcia, Spain
Giner Alor Hernandez, PhD, Orizaba Technology Institute, Mexico
PHUC V. Nguyen, PhD (ABD), Arkansas State University, Jonesboro, Arkansas, USA
Mateusz Radzimski, MSc, ATOS Research & Innovation, Spain
Patricia Ordoñez de Pablos, PhD, University of Oviedo, Spain
Sean O'Riain, PhD, Digital Enterprise Research Institute (DERI), Ireland
Jose Maria Alvarez Rodríguez, PhD (ABD), Univerity of Oviedo, Spain
Enrique Jimenez Domingo, MsC, University Carlos III of Madrid, Spain
Jose Luis Sanchez Cervantes, MsC, University Carlos III of Madrid, Spain

Sponsorship

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ministry of Economy and Competitiveness under the project FLORA (TIN2011-27405)

Universidad Carlos III de Madrid

University Carlos III of Madrid. SoftLab group.

Keynote Speaker Talk

Title:

Towards the construction of Financial Data Spaces: Challenges, Approaches and Trends

Abstract:

The advent of Open Data and regulatory adopted standards such as XBRL, have hastened the availability of semantically rich Business and Financial data for general consumption. Market pressures, competition and the do-more-with-less philosophy demands increasing levels of flexibility when dealing with data and its interoperability. Familiar with integrated views of corporate data, enterprises are now expected to flexibly cater for different information views that move beyond their traditional boundaries. Lacking full semantic integration this emergent Data Space perspective better represents a data co-existence environment, where traditional data management strategies have to be re-visited. This talk will cover major challenges and report on practical experiences and insights gained from the implementation of financial data space components such as search and query, multilingual knowledge access and regulatory auditability. Lastly we will touch on developments relating to business standards activity, EU legislation, emergent areas and remaining challenges.

Keynote Bio:

Sean O'Riain is Lead of the DERI eBusiness Unit. Before joining DERI Sean was a senior research with Hewlett Packard's Semantic Infrastructure Research Group, Galway, Ireland focusing on semantic based content analytics. His research interests include the use of Web Science (Linked Data, provenance, entity consolidation), natural language query mechanism for Linked Data, business taxonomic standards (XBRL) for enhanced semantic information access and Linked Data technology adoption patterns within enterprises. Sean's Masters and Doctoral topics relate to distributed information retrieval and business filings content analysis.

Table of Contents

Using SPIN to Formalise Accounting Regulations on the Semantic Web

Dennis Spohr[†], Philipp Cimiano[†], John M^{c}Crae[†] and Seán O'Riain[*]

[†]Cognitive Interaction Technology Centre of Excellence
Semantic Computing Group, University of Bielefeld, Germany
{dspohr,cimiano,jmccrae}@cit-ec.uni-bielefeld.de
[*]Digital Enterprise Research Institute
NUI Galway, Ireland
sean.oriain@deri.org

Abstract. The eXtensible Business Reporting Language (XBRL) has standardised financial reporting and provide a machine-interpretable format that makes financial and business reports easier to access and consume. Leveraging XBRL with Open Linked Data for purposes such as multi-dimensional regulatory querying and investigation requires XBRL formalisation as RDF. This paper investigates the use of off-the-shelf Semantic Web technologies to formalise accounting regulations specified in XBRL jurisdictional taxonomies. Specifically the use of the SPARQL Inferencing Notation (SPIN) with RDF to represent these accounting regulations as rule constraints, not catered for in the RDF abstract model is investigated. We move beyond previous RDF to XBRL transformations and investigate how SPIN enhanced formalisation enables inferencing of financial statement facts associated with financial reporting concepts and sophisticated consistency checks, which evaluate the correctness of reported financial data with respect to the calculation requirements imposed by accounting regulation. The approach illustrated through two use cases demonstrates the use of SPIN to meet central requirements for financial data and regulatory modelling.

1 Introduction

Despite the proliferation of financial information available from sources such as company websites, institutions and regulating bodies there remains a lack of transparency with regard to financial information. Two areas that can contribute to enhancing transparency are the adoption of a standard data representation formalism and greater levels of interoperability with and between different financial sources.

Multiple heterogeneous formats (e.g. HTML, PDF, CSV), ensure that data usage and interpretation typically has a dependency on manual intervention with a knock on effect for accurate and timely analysis of for example, financial reports [9]. Overcoming transparency and re-use issues associated with heterogeneous formats requires that financial information providers move towards data provision in machine-interpretable and interoperable formats. XBRL [1], has been adopted by regulatory agencies for consolidated financial filings and is gaining acceptance for general business reporting. Within the U.S. the Securities and Exchange Commission (SEC) has mandated XBRL use by all financial filers by 2014[2].

XBRL, an XML-based format defines financial concepts and their relations based on jurisdictional Generally Accepted Accounting Practices (GAAP'S). Relations include calculation rules for monetary concepts – for example, the value for the financial element *Assets* is calculated from the sum of *Current assets* and *Non-current assets* – in addition to other more complex business rules expressed through XBRL formula.

XBRL offers automated processing of financial reports and increased interoperability between reporting instances. Even though XBRL provides a common interoperable format its document-centric nature has been identified as an inhibitor to integration of financial information from multiple sources [6] and [1]). XBRL's abstract model also remains unclear as to how instances and concepts can be linked with other data sources. Semantic Web formalisms such as RDF with a well defined and understood abstract model has been gaining popularity for multiple data source integration using the data mash-up approach. Attempts to make XBRL interoperable with other Web based information through its transformation to RDF or OWL have gained momentum in recent years (e.g. [6] and [1]). While the approaches adequately represent financial statements contained within XBRL formatted financial reports, they do not formalise the semantics inherent in the calculation rules. As the calculation rules represent country specific regulatory instruction for financial instrument calculation, inability to adequately express them will result in a lack of conformance and regulatory checking capability against the accounting standards from which they have been reported. Formalising regulatory information especially when transferring to an alternate representation is therefore important.

In this paper, we present an approach to implementing XBRL semantics using off-the-shelf Semantic Web technologies and specifically SPIN[3],) to semantically model regulatory requirements mentioned. The resulting representation can be used to infer values for monetary concepts, and consistency check reported values without the need for customised XBRL software. The SPIN vocabulary, developed for business rule representation, can capture the intended semantics of accounting regulations in a transparent and intuitive way. SPIN adoption is supported by tools such as TopBraid Composer[4]), in progress standardisation efforts and an open-source Java API [5].

[1] XBRL V2.1 Taxonomy Specification Recommendations `http://www.xbrl.org/SpecRecommendations/`.

[2] See `http://xbrl.sec.gov/`.

[3] Specification at `http://www.spinrdf.org/`

[4] See`http://topquadrant.com/products/TB_Suite.html`

[5] See `http://www.spinrdf.org/faq.html` for more details.

After a brief introduction to XBRL, financial reporting using XBRL taxonomies and SPIN, we position our work with respect to recent efforts from García and Gil [6] and Bao et al. [1]. Section 3 presents the general translation of XBRL to RDF/OWL with particular focus on transforming calculation rules to SPIN (Section 3.2). Section 4 then illustrates how resulting representations are capable of inferring values for reporting concepts, and checking that the values of reported concepts conform to the rules as defined in the respective accounting standard. In Section 5, we relate this SPARQL-based approach to generally available rule-based approaches. Finally Section 6 outlines how emergent XBRL-related developments integrate with the approach presented here.

2 Background and related work

This section introduce the main aspects of XBRL and financial reporting as relevant to this work (Section 2.1), as well as the basic features of SPIN (Section 2.3). Section 2.2 discusses related work, firstly with respect to transforming XBRL data to its RDF equivalent, and secondly regarding the use of SPARQL for business-related modelling issues.

2.1 Financial reporting in XBRL

XBRL is an XML-based formalism which aims to replace dependency on proprietary formats usage in financial reports preparation and compilation [7]. XBRL targets increased interoperability across different companies, thereby reducing the manual effort required to create and consume financial information. At its core is the notion of *taxonomies* and *instance documents* (see Fig. 1). The instance document represents the financial report [6], stating financial instrument data facts such as its monetary value and units. Each fact is linked to a reporting context, which additionally specifies an entity – commonly the company which issued the report – as well as the period to which the fact applies. In XBRL these are termed dimensions. The example (1), taken from the 2009 SAP annual report, specifies that SAP had *Cash and cash equivalents* of €1.88 billion on December 31, 2009.

(1)
```
<context id="FYp0Qp0e">
 <entity>
  <identifier scheme="http://www.sec.gov/CIK">0000943042</identifier>
 </entity>
 <period>
  <instant>2009-12-31</instant>
 </period>
</context>
 ...
<ifrs:CashAndCashEquivalents contextRef="FYp0Qp0e" decimals="-6"
 unitRef="EUR">1884000000</ifrs:CashAndCashEquivalents>
```

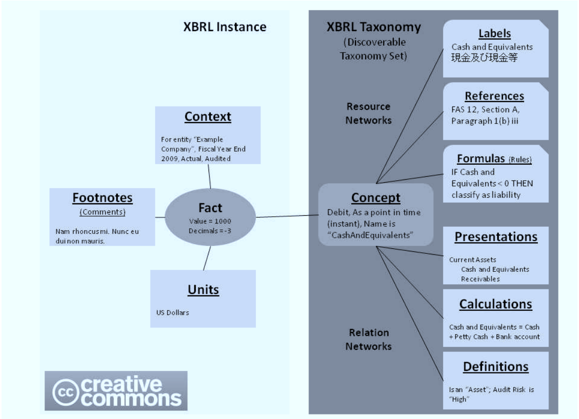

XBRL Instance
Context
For entity "Example Company", Fiscal Year End 2009, Actual, Audited
Footnotes
(Comments)
Nam rhoncus mi. Nunc eu dui non mauris.
Fact
Value = 1000
Decimals = -3
Units
US Dollars
creative commons
XBRL Taxonomy
(Discoverable Taxonomy Set)
Resource Networks
Relation Networks
Concept
Debit, As a point in time (instant), Name is "CashAndEquivalents"
Labels
Cash and Equivalents
現金及び現金等
References
FAS 12, Section A, Paragraph 1(b) iii
Formulas (Rules)
IF Cash and Equivalents < 0 THEN classify as liability
Presentations
Current Assets
Cash and Equivalents
Receivables
Calculations
Cash and Equivalents = Cash + Petty Cash + Bank account
Definitions
Is an "Asset"; Audit Risk is "High"

Fig. 1: High-level model of XBRL by Charles Hoffman[7]

As shown in (1), values in an instance document refer to concepts defined in XBRL taxonomies, which specify, for example, that *Cash and cash equivalents* is a monetary concept with balance debit, which is measured at a point in time – as opposed to over a duration of time. In addition to adopting internationally standardised taxonomies like the International Financial Reporting Standards (IFRS), companies can also provide their own taxonomy extensions, in instances where they need to report a value for a concept which is not covered by the standard. For example, the concept *Software revenue* is not found in the 2009 IFRS taxonomy, but is provided as a custom taxonomy extension by SAP, with the facts in the instance document. Each concept in a taxonomy is linked to a set of XLink *linkbases*[8] (called *resource* and *relation networks* in the figure). These specify labels for the concepts, as well as other information such as, how the values of the concepts should be displayed in different statement types . For example, the International Financal Accounting Standard [9] *"Statement of financial position, current/non-current"* specifies that the concept *"Assets"* should be displayed above *"Non-current assets"* and *"Current assets"* in a consolidated filing, whereas the financial instrument *"Statement of financial position, order of liquidity"* places *"Assets"* above *"Property, plant and equipment"*, *"Investment property"*. To express such relationships, XBRL uses XML Linking Language (XLink:[10]) arcs and extended links, which can be used to group any number of

[6] Note: XBRL has also been used by the Global Reporting Initiative `https://www.globalreporting.org/` to report on sustainability issues.

[7] `http://xbrl.squarespace.com`

[8] e.g. the U.S. 2009 GAAP taxonomy contains 450 linkbases.

[9] `http://www.ifrs.org/XBRL/IFRS+Taxonomy/IFRS+Taxonomy+2011/IFRS+Taxonomy+2011.htm`.

[10] `http://www.w3.org/TR/xlink/`.

arcs and link them to other resources For the case just described, the XBRL presentation linkbase specifies how the concepts are linked using *parent-child* arcs, and that a set of presentation arcs are associated with a particular accounting standard using an extended link role.

For the purpose of our investigations we focus on the XBRL calculation linkbases, which defines how concept values are calculated as defined by specific accounting standards and general business rules (see sections 3.2 and 6. Example (2) shows the XBRL representation of a calculation arc taken from the calculation linkbase of the IFRS taxonomy. The XBRL formula linkbase is outside the research scope and left for future work considerations.

```
(2) <loc xlink:type="locator" xlink:label="ifrs_CashAndCashEquivalents"
      xlink:href="http://xbrl.iasb.org/taxonomy/2009-04-01/
                  ifrs-cor_2009-04-01.xsd#ifrs_CashAndCashEquivalents" />
    <calculationArc xlink:type="arc"
      xlink:arcrole="http://www.xbrl.org/2003/arcrole/summation-item"
      xlink:from="ifrs_CurrentAssets"
      xlink:to="ifrs_CashAndCashEquivalents" order="1" weight="1" />
```

The arc states that the concept *"CurrentAssets"* is linked to *"CashAndCashEquivalents"* through a *"summation-item"* relation. The arc weight of 1 indicates that the concept values are added and a weight of -1 that the values are subtracted. Currently XBRL calculation links only allow for the summation of items. Clearly Semantic Web formalisms offer a far more compact and intuitive representation for expressing such statements. To that end we next discuss related approaches that convert XBRL data to its RDF equivalent.

2.2 Related work

Transforming XBRL to Semantic Web formalisms. Bao et al. [1] have presented the most recent approach to transforming XBRL data to a Semantic Web standard. They present an OWL-based model that "faithfully preserves the implicit semantics in XBRL" (see [1], p. 144). In fact, however, due to their focus on making the semantics of linkbase arcs explicit, their approach omits a considerable amount of the semantics described in XBRL documents. One of these concerns the use of extended link roles, which Bao et al. [1] refer to as "non-semantic". These link roles (Section 2.1)limit the scope of assertions in an XBRL linkbase e.g. to a particular type of statement. Such information seems to be lost in the representation of Bao et al., making their arcs become globally applicable.

In addition to this, their strategy for representing linkbase arcs is based on the assumption that the intended interpretation of arcs holds between instances of the respective concepts linked by a particular arc. The XBRL Specification does not note this as being the intended interpretation, but instead states that arcs relate "one XBRL concept to one other XBRL concept"[11]. While the assumption seem reasonable for concrete-numeric-concepts, abstract concepts by

[11] See http://www.xbrl.org/Specification/XBRL-RECOMMENDATION-2003-12-31+Corrected-Errata-2008-07-02.htm#_3.5.3.9.

definition do not have instances. Irrespective they are still related by means of `parent-child` arcs, and it is not apparent how [1] caters for those cases.

Bao et al. are not "mechanical" in their preservation of the structural properties of XBRL while other approaches by Declerck and Krieger [3] and García and Gil [6] are. Adhering to the latter approaches we preserve relevant aspects of the structural information in XBRL, while adding further interpretation that address the shortcomings of Semantic Web vocabularies to semantically model mathematical relations contained in the XBRL calculation [8].

Using SPARQL in the context of business information. Fürber and Hepp [5] have presented an approach to using SPARQL in order to detect data quality problems. The use of SPIN to model consistency constraints in different scenarios is discussed and how detected inconsistencies are flagged, illustrated in TopBraid Composer. We note the overlap with our approach through the use of some of their SPIN constraints that can also be applied to XBRL data. We however further the use of SPIN in a more advanced scenario that requires performing constraint checking on data which have been inferred through iterative rule application.

2.3 SPARQL Inferencing Notation (SPIN)

According to its developers, the SPARQL Inferencing Notation (SPIN) "is the de-facto industry standard to represent SPARQL rules and constraints on Semantic Web models"[12]. SPIN has been developed out of the necessity to perform calculations on property values, a task which is largely unsupported in current Semantic Web formalisms, as well as the need for constraints checking with closed-world semantics (see Section 5).

SPIN provides the syntax to attach SPARQL queries to resources in an RDF-compliant way using RDF properties `spin:rule`, `spin:constraint`, and super-property `spin:query`. The `spin:rule` property accepts SPARQL CONSTRUCT queries as value and can be used to infer new triples on the basis of the statements in the query's WHERE clause. A basic example is provided in (3), and the corresponding SPIN representation in Turtle syntax in (4).

```
(3) CONSTRUCT { ?this a ?c2 . }
    WHERE { ?c1 rdfs:subClassOf ?c2 .
            ?this a ?c1 . }
(4) [ a sp:Construct ; sp:templates ([ sp:subject spin:_this ;
                        sp:predicate rdf:type ;
                        sp:object _:b1 ])
         sp:where ([ sp:subject _:b3 ;
                        sp:predicate rdfs:subClassOf ;
                        sp:object _:b1
                     ]  [ sp:subject spin:_this ;
                       sp:predicate rdf:type ;
           sp:object _:b3 ]) ]
```

[12] See `http://www.spinrdf.org`.

The example, based on that available from TopBraid's SPIN website[13], formalises the semantics of `rdfs:subClassOf` and illustrated variable use. While variables in a SPARQL query are generally mapped to blank nodes, in SPIN RDF notation, the variable `?this` refers to the resource `spin:_this`. This variable, like the corresponding keyword in object-oriented programming languages, refers to an instance of the class to which the respective rule has been attached. As a result, if the rule in Example (3) was attached to `owl:Thing`, it would be applied to every instance of `owl:Thing` satisfying the statements in the WHERE clause.

The `spin:constraint` property can be used to model consistency constraints, using the SPARQL ASK queries. Where an ASK query evaluates to true, the respective instance is indicated as violating the constraint. Finally, the general property `spin:query` can be used to generally attach SPARQL queries to RDF resources, i.e. also SELECT queries. As will be shown in Section 3, we make use primarily of CONSTRUCT and ASK queries in order to capture the intended semantics of accounting regulations.

In addition to standard SPARQL operators like UNION, OPTIONAL and FILTER, SPIN supports SPARQL extensions such as the ARQ keyword LET[14], which allows for value assignment to variables, as well as the possibility to define custom functions.

With SPIN having recently started standardisation activities as a W3C member submission and the fact that SPARQL is already the query language of choice in numerous Semantic Web applications – there is a solid basis for its wide-spread adoption by the Semantic Web community.

3 Transforming XBRL to RDF

This section discusses the conversion of XBRL to RDF, with focus on the SPIN-based representation of accounting regulations. After a brief introduction to the general underlying ideas in Section 3.1, we discuss the representation of calculation rules (3.2) and consistency constraints used (3.3) to transform the accounting regulations from XBRL to RDF.

3.1 General strategy

Our approach adheres to what Bao et al. [1] refers to as a representation of the "logical model" of XBRL, which preserves structural information from the original data. The motivation for doing so is to have a representation of XBRL that is interoperable with other data on the Semantic Web and also enables inferences and consistency checking, while at the same time allowing users to query the structure itself (e.g. "what is the *"Assets"* concept hierarchy in the *"Statement of financial position"*)?

[13] `http://topbraid.org/spin/owlrl-all.html`

[14] See e.g. `http://jena.sourceforge.net/ARQ/assignment.html`.

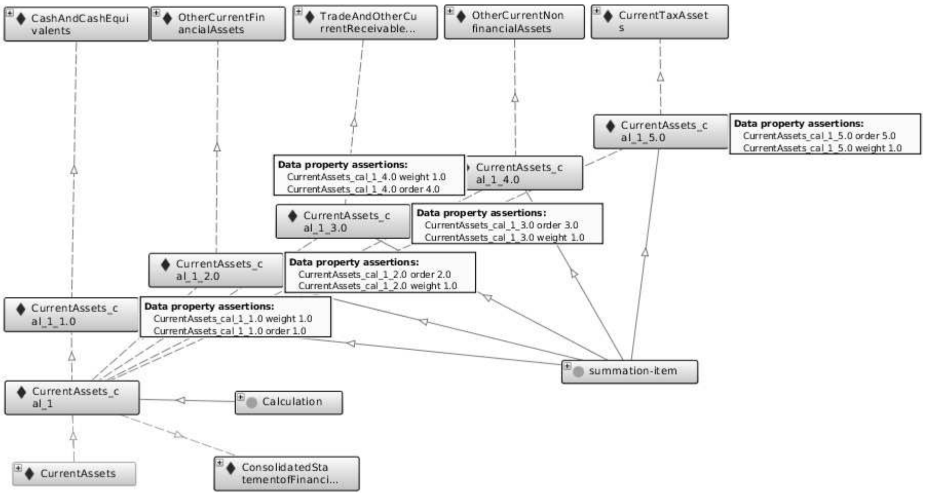

CashAndCashEquivalents
OtherCurrentFinancialAssets
TradeAndOtherCurrentReceivable...
OtherCurrentNonfinancialAssets
CurrentTaxAssets
CurrentAssets_cal_1_5.0
Data property assertions:
CurrentAssets_cal_1_5.0 order 5.0
CurrentAssets_cal_1_5.0 weight 1.0
Data property assertions:
CurrentAssets_cal_1_4.0 weight 1.0
CurrentAssets_cal_1_4.0 order 4.0
CurrentAssets_cal_1_4.0
CurrentAssets_cal_1_3.0
Data property assertions:
CurrentAssets_cal_1_3.0 order 3.0
CurrentAssets_cal_1_3.0 weight 1.0
CurrentAssets_cal_1_2.0
Data property assertions:
CurrentAssets_cal_1_2.0 order 2.0
CurrentAssets_cal_1_2.0 weight 1.0
CurrentAssets_cal_1_1.0
Data property assertions:
CurrentAssets_cal_1_1.0 weight 1.0
CurrentAssets_cal_1_1.0 order 1.0
summation-item
CurrentAssets_cal_1
Calculation
CurrentAssets
ConsolidatedStatementofFinanci...

Fig. 2: Structural representation of the calculation of `ifrs:CurrentAssets` (generated with the OntoGraf Protégé 4 plug-in)

Figure 2 shows the structural representation of the calculation of `ifrs:CurrentAssets`. The bottom left-hand corner of the figure shows an instance representing the XBRL concept which is linked to an instance of the class `Calculation`, as well as to several instances linking the individual concepts, of which all are part of the calculation. Each of these instances represents a reification of a calculation arc from the original XBRL format, and carries the weight and order attributes as values of datatype property statements.

In order to add to this structural representation and capture the intended semantics of XBRL, we make use of the meta-modelling facilities of OWL2 DL, namely *punning*. Punning allows us to refer to an entity, both a class and an individual, allowing them be treated as distinct on the level of OWL and remain within DL. For SPARQL querying however, resources with the same URIs are interpreted as referring to the same entity. For XBRL this allows concepts defined in an XBRL taxonomy be referred to as OWL individuals in order to express the relations that hold among them. They can also be modelled as classes, enabling values reported in XBRL instance documents be used to *instantiate* the concepts. OWL restrictions can then be added to the concepts (e.g. that monetary concepts have integer or double values only), which are then inherited by their instances.

3.2 SPIN rules for calculations

In order to model the regulatory calculation of monetary concepts in a Semantic Web compliant way, SPIN rules based on the data contained in XBRL calculation linkbases were generated. specifically calculation arcs such as those shown in example (2), are converted into their SPIN representation (below), which represents the calculation of `ifrs:CurrentAssets`.

(5)
```
CONSTRUCT { ?this xbrlrdf:calculatedValue ?cvalue . }
WHERE { ?this xbrli:contextRef ?context; xbrli:unitRef ?unit .
 ?x0 a ifrs:CurrentTaxAssets ;
       xbrli:contextRef ?context; xbrli:unitRef ?unit;
       xbrlrdf:calculatedValue ?cv0 .
 ?x1 a ifrs:OtherCurrentNonfinancialAssets ;
       xbrli:contextRef ?context; xbrli:unitRef ?unit;
       xbrlrdf:calculatedValue ?cv1 .
 ?x2 a ifrs:TradeAndOtherCurrentReceivables ;
       xbrli:contextRef ?context; xbrli:unitRef ?unit;
       xbrlrdf:calculatedValue ?cv2 .
 ?x3 a ifrs:OtherCurrentFinancialAssets ;
       xbrli:contextRef ?context; xbrli:unitRef ?unit;
       xbrlrdf:calculatedValue ?cv3 .
 ?x4 a ifrs:CashAndCashEquivalents ;
       xbrli:contextRef ?context; xbrli:unitRef ?unit;
       xbrlrdf:calculatedValue ?cv4 .
LET (?cvalue := 1.0 * ?cv0 + 1.0 * ?cv1 + 1.0 * ?cv2 +
                1.0 * ?cv3 + 1.0 * ?cv4 ) . }
```

Each of the graph patterns in the query represents one calculation arc. References to URIs for the context and units, ensure that the values of relevant instances are only taken into account. This excludes cases where a particular value refers to different entities or different segments of the same entity, as well as cases in which values are reported for different time periods. This is a normal occurrence as financial statements generally contain figures for both the current and preceding reporting periods. Finally, the `LET` clause specifies how the values of the individual concept instances should be combined. For accounting rules, this is limited to summation and subtraction. XBRL provides a single arc role `summation-item` for this purpose, and uses the value of the *weight* attribute – either 1 or -1 (example (2) above) – to indicate whether the value of a particular concept should be added or subtracted. For further more complex calculations possible using SPIN we refer the reader to the SPIN vocabulary specification [15].

The example illustrates how rules can make use of previously calculated values to calculate further values. This allows value calculation for composite monetary concepts (i.e. those whose values are calculated on the basis of the values of other concepts) by specifying values for atomic concepts and then applying the rules iteratively (see Section 4.1 for more details).

Moreover, it should be noted that in our RDF representation, rules are not modelled as blank nodes, but instead carry a URI. This has the benefit of enabling other instances to dereference the rules, allowing a particular calculation rule be reused across different financial reports of the same accounting standard. It also further allows attachment of additional properties to a rule and a reference to the type of financial statement to which the rule applies. Therefore, in

[15] `http://www.spinrdf.org`.

addition to the actual rule, the instance representing the rule in (5) is the subject of a triple relating it to the URI representing SAP's *Consolidated Statement of Financial Position* by means of `xbrlrdf:roleRef`.

As mentioned, rules can be executed iteratively, making use of previously inferred values. In order to make sure that a particular calculation rule with atomic concepts (i.e. those concepts which lack regulatory calculation rules attached) can also be applied, we add the default calculation shown in (6) to atomic monetary concepts. This rule then assigns the reported value of the respective instance to `xbrlrdf:calculatedValue`.

```
(6) CONSTRUCT { ?this xbrlrdf:calculatedValue ?value . }
    WHERE { ?this xbrlrdf:value ?value } .
```

3.3 SPIN constraints for consistency checking

On the basis of the SPIN rules presented above, each monetary concept which participates in some calculation and has reported values in the respective context is assigned a calculated value. The next step in modelling the accounting regulation is to specify that calculated values need to match reported values. As SPIN rules and constraints are applied to all instances of the class to which they have been attached, as well as to the instances of its subclasses, this can be achieved by attaching a single SPIN constraint to the top monetary concept:

```
(7) ASK WHERE { ?this xbrlrdf:value ?value ;
                      xbrlrdf:calculatedValue ?cvalue .
                FILTER (?value != ?cvalue) }
```

Additionally, SPIN constraints can be used to formalise more general constraints imposed by the XBRL specification. Below, we illustrate this using a constraint (restriction) which states that if two concepts have the same balance type (i.e. credit or debit), they can only be added to one another, not subtracted. In other words, the value of the XBRL weight attribute, which is preserved in our structural representation of XBRL, has to be positive:

```
(8) ASK WHERE { ?this xlink:from ?from ; xlink:to ?to ;
                      xbrlrdf:weight ?weight .
                ?from a ?balance . ?to a ?balance .
                ?balance rdfs:subClassOf xbrlrdf:BalanceType .
                FILTER (?weight < 0) }
```

4 Specific use cases

The SPIN rules and constraints given in Section 3 above can be used to infer values for instances of the respective reporting concepts, as well as check their consistency. We next illustrate how these representations integrate with TopBraid Composer, taking as example the SAP 2009 annual report reported against its custom extension of the IFRS 2009 taxonomy.

4.1 Inferring values of monetary reporting concepts

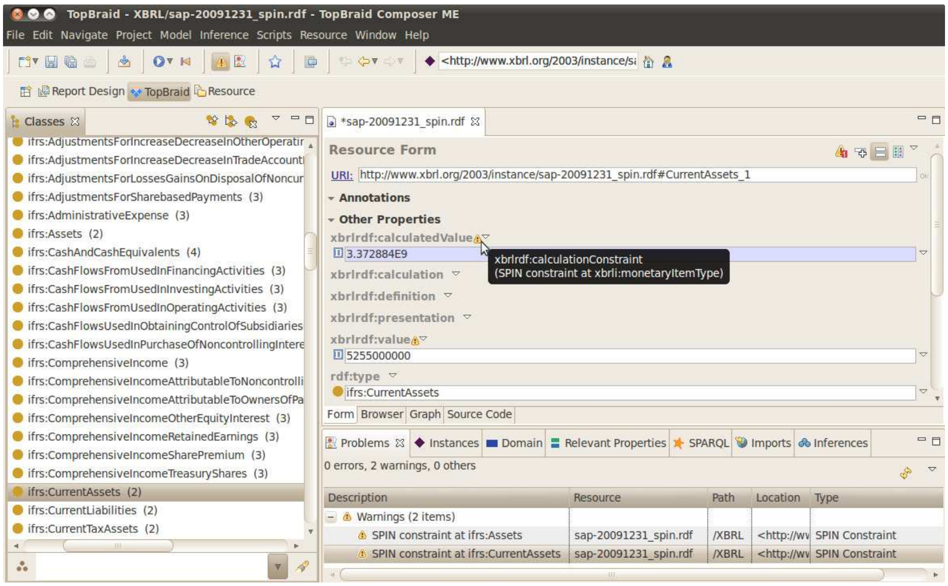
In order to test whether the rules explained in Section 3 actually behave as desired, we have first generated a modified version of the report such that it contained reported values for *atomic* monetary concepts only. All composite monetary concepts were assigned their values through iterative application of the SPIN calculation rules. This allowed evaluation as to whether the available information triggered the application of all rules necessary to calculate the missing information and whether the calculated figures corresponded to those reported in the original filing. Table 1 summarises the results, with the figures from the original report recorded in parentheses.

	absolute	relative
Concepts in IFRS 2009 taxonomy and SAP extension	3,021	100.00%
Regulatory calculation rules	492	100.00%
Reported monetary values	351 (482)	72.82%
Monetary concepts with reported values	129 (171)	75.44%
Inferred monetary values	458	95.02%
Monetary concepts with inferred values	167	97.66%
Monetary values inferred by default rule	343	74.89%
Monetary values inferred by regulatory rules	115	25.11%
Regulatory rules applied	42	8.54%
Total number of monetary values	458 (482)	95.02%
Total number of correct monetary values	458	100.00%

Table 1: Reported and inferred values in the modified annual report 2009 of SAP

Tuples 3 and 4 of the table detail that the modified report contains 351 reported values for 129 monetary concepts, compared to 482 and 171, respectively, from the original report. After applying the calculation rules, values are inferred for 97.66% of these 171 concepts, indicating that the modified report contains 458 values, as opposed to 482 original report values. 25.11% of the 458 inferred values are due to regulatory rules, outlining that the remaining 343 values have been inferred for atomic concepts by means of the default rule. Over all 8.54% of all the regulatory rules available in the IFRS 2009 taxonomy and the SAP 2009 extension have been applied.

The figures illustrate that for 4 of the monetary concepts no value could be inferred. Analysis revealed that this is due to values for `ifrs:BasicEarningsLossPerShare` and `ifrs:DilutedEarningsLossPerShare`, being missing from the XBRL report instance, despite being part of a composite concept. As a result, the corresponding rules could not be applied (see Section 5 for a discussion regarding the optionality of calculation arcs).



ifrs:AdjustmentsForIncreaseDecreaseInOtherOperatir
ifrs:AdjustmentsForIncreaseDecreaseInTradeAccounti
ifrs:AdjustmentsForLossesGainsOnDisposalOfNoncur
ifrs:AdjustmentsForSharebasedPayments (3)
ifrs:AdministrativeExpense (3)
ifrs:Assets (2)
ifrs:CashAndCashEquivalents (4)
ifrs:CashFlowsFromUsedInFinancingActivities (3)
ifrs:CashFlowsFromUsedInInvestingActivities (3)
ifrs:CashFlowsFromUsedInOperatingActivities (3)
ifrs:CashFlowsUsedInObtainingControlOfSubsidiaries
ifrs:CashFlowsUsedInPurchaseOfNoncontrollingIntere
ifrs:ComprehensiveIncome (3)
ifrs:ComprehensiveIncomeAttributableToNoncontrolli
ifrs:ComprehensiveIncomeAttributableToOwnersOfPa
ifrs:ComprehensiveIncomeOtherEquityInterest (3)
ifrs:ComprehensiveIncomeRetainedEarnings (3)
ifrs:ComprehensiveIncomeSharePremium (3)
ifrs:ComprehensiveIncomeTreasuryShares (3)
ifrs:CurrentAssets (2)
ifrs:CurrentLiabilities (2)
ifrs:CurrentTaxAssets (2)

URI: http://www.xbrl.org/2003/instance/sap-20091231_spin.rdf#CurrentAssets_1

xbrlrdf:calculatedValue
3.372884E9

xbrlrdf:calculationConstraint
(SPIN constraint at xbrli:monetaryItemType)

xbrlrdf:calculation
xbrlrdf:definition
xbrlrdf:presentation
xbrlrdf:value
5255000000
rdf:type
ifrs:CurrentAssets

0 errors, 2 warnings, 0 others

Description	Resource	Path	Location	Type
Warnings (2 items)				
SPIN constraint at ifrs:Assets	sap-20091231_spin.rdf	/XBRL	<http://wv	SPIN Constraint
SPIN constraint at ifrs:CurrentAssets	sap-20091231_spin.rdf	/XBRL	<http://wv	SPIN Constraint

Fig. 3: View of an instance in TopBraid Composer after introducing an incorrect value

4.2 Consistency control of reported values

Table 1 results detail that all inferred values were correct, which in turn reflects XBRL's contribution to overall financial reporting consistency. The results also indicate that the functionality of the SPIN constraint were not evaluated by this method. To address this situation the reported value of the atomic concept `ifrs:CashAndCashEquivalents` in the original report was changed, and the change tracked to see whether the change would be propagated along the calculation "hierarchy", and yield inconsistent composite concepts. Result displaying an instance of `ifrs:CurrentAssets` after rule application and constraint checking in TopBraid Composer are reported in Figure3.

The figure demonstrates that TopBraid Composer correctly flags the instance where the calculated value differs from the reported one and where inconsistency have arisen at concept that has `ifrs:CurrentAssets` as calculation component (i.e. `ifrs:Assets`).

5 Discussion

The previous sections detail how the SPARQL CONSTRUCT rules and ASK queries capture the semantics of the XBRL data in an intuitive and transparent way. However the XBRL does allow for calculation arcs and if so to have an assumed value of zero. This usage of default values, cannot be naturally handled with monotonic logics such as OWL and would normally require the use of a formalism such as Reiter's default logic [12]. A naive approach would be to develop a SPIN rules by the use of the OPTIONAL keyword available in SPARQL, for example:

```
(9) CONSTRUCT { ?this xbrlrdf:calculatedValue  ?cvalue . }
    WHERE { ?x0 a ifrs:CurrentTaxAssets ;
                xbrlrdf:calculatedValue ?cv0 .
            OPTIONAL {
             ?x1 a ifrs:CashAndCashEquivalents ;
                 xbrlrdf:calculatedValue ?cv1 .  }
            LET(?cvalue := 1.0 * ?cv0 + 1.0 * ?cv1) }
```

If it is assumed that that rules can be applied in any order, this rule would be applied before the rule that calculates `?cv1`, and a different value produced. We therefore take the position that every value should be defined or explicitly stated as not undefined. While OWL has no specific vocabulary to state that a value has not been defined we can achieve the same result using an OWL class axiom.

For example, in order to state that SAP did not report a value for `ifrs:BasicEarningsLossPerShareFromDiscontinuedOperations` in units `iso4217:EUR` year ending December 2009[16], the following would need to be asserted.

$$\begin{aligned}(10)\quad &\mathsf{ifrs{:}BasicEarningsLossPerShareFromDiscontinuedOperations} \sqsubseteq \\ &\quad \neg((\exists \mathsf{xbrli{:}contextRef}.\{\mathsf{FYp0YTD}\}) \sqcap (\exists \mathsf{xbrli{:}unitRef}.\{\mathsf{iso4217{:}EUR}\}))\end{aligned}$$

With the non-existence of such an instance explicitly asserted it can be combined with the SPIN rules by providing a default rule that specifies the value of the property if it is known not to exist as follows:

```
(11) OPTIONAL { ?x4 a ifrs:CashAndCashEquivalents;
                    xbrli:contextRef ?context; xbrli:unitRef ?unit;
                    xbrlrdf:calculatedValue ?cv4 } .
     OPTIONAL { NOT EXISTS { ?x4 a ifrs:CashAndCashEquivalents;
                                 xbrli:contextRef ?context;
                                 xbrli:unitRef ?unit; } .
                LET (?cv4 := 0) . }
```

Here the assumption is made that `NOT EXISTS` predicate evaluates to true where it is provable that the dataset does not contain the predicate. Interesting future work could look to determine whether a default logic rule language such as RIF-SILK[17] could be used.

[16] In SAP's XBRL instance document, the period context represented by FYp0YTD.

[17] See `http://silk.semwebcentral.org/RIF-SILK.html`.

6 Conclusion and Future Work

The paper outlines SPARQL's capability to meet complex query requirements, central to modelling financial data, and specifically their accounting regulations on the Semantic Web. The approach transforms financial reports represented in XBRL to RDF, and uses the RDF-compatible SPARQL Inferencing Notation to capture the regulatory rules expressed by the XBRL calculation linkbase. The resulting representation was evaluated against XBRL financial data, both with respect to inferring values for instances of monetary concepts and checking their consistency. Additionally the use of the representation to formalise additional constraints to address the well-formedness and high quality of the data was discussed.

The approach taken can be extended to cater for the more complex mathematical operations of the XBRL formula specification[18]. For example the formula specification defines that value calculations apply to instances that refer to *identical* contexts, and more generally to concept instances which are *p(arent)-equal*, *c(ontext)-equal* and *u(nit)*-equal. p-equality and u-equality have been previously shown through rule attachment to the composite class and including the reference to the unit in the rule. Alternatively c-equality could be inferred beforehand, by specifying that two contexts which share the same entity and period are linked by `owl:sameAs`. When applying the rules iteratively to a repository that is OWL-aware, the rules shown above can be applied as is.

Arguments for financial information integration include the ability to conduct financial metrics comparison [2] and querying of heterogeneous data sets to gain wider holistic insight [4]. For Linked Data driven information systems, data abstraction presents challenges for financial integration [11] and financial values comparison. Semantic Web offers a level of interoperability between data sources that would assist comparability based on the representational transformation of financial data, such as XBRL to RDF. The approach is not new, having been previously applied to areas such as investment funds analysis [8] or more recently promoted for wider financial ecosystems evolution [10]. Financial standards interoperability also faces additional challenges from different jurisdictional and regulatory rules. Use of an ontology architecture to accommodate multiple XBRL formats from the business community have only been proposed [13]. Within this context the ability of Semantic Web vocabularies to model regulatory relations contained within XBRL reporting formats will become increasingly important.

We have demonstrated SPIN's viability but others rule languages such as the Semantic Web Rule Language [19] suggested by [13], could also be investigated as part of a best practises recommendation for Semantic Web rule format representation.

[18] `http://www.xbrl.org/Specification/formula/REC-2009-06-22/formula-REC-2009-06-22.html`.

[19] `http://www.w3.org/Submission/SWRL/`

7 Acknowledgements

The work presented in this paper has been funded by the Multilingual Ontologies for Networked Knowledge (MONNET) project under EU Grant No. 248458, and Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Bao, J., Rong, G., Li, X., Ding, L.: Representing Financial Reports on the Semantic Web: A Faithful Translation from XBRL to OWL. In: Semantic Web Rules, Lecture Notes in Computer Science, vol. 6403, pp. 144–152. Springer (2010)
2. Debreceny, R.: Feeding the information value chain : Deriving analytical ratios from xbrl filings to the sec (2010)
3. Declerck, T., Krieger, H.U.: Translating XBRL into description logic. An approach using Protégé, Sesame & OWL. In: Business Information Systems (BIS). pp. 455–467. Klagenfurt, Germany (2006)
4. Freitas, A., Curry, E., Oliveira, J.a.G., O'Riain, S.: Querying heterogeneous datasets on the linked data web: Challenges, approaches, and trends. IEEE Internet Computing 16(1), 24–33 (2012)
5. Fürber, C., Hepp, M.: Using SPARQL and SPIN for Data Quality Management on the Semantic Web. In: Business Information Systems, Lecture Notes in Business Information Processing, vol. 47, pp. 35–46. Springer (2010)
6. García, R., Gil, R.: Publishing XBRL as Linked Open Data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web (LDOW) (2009)
7. Hoffman, C., Strand, C.: XBRL Essentials. American Institute of Certified Public Accountants, New York (2001)
8. Lara, R., Cantador, I., Castells, P.: Xbrl taxonomies and owl ontologies for investment funds. In: 1st International Workshop on Ontologizing Industrial Standards at the 25th International Conference on Conceptual Modelling. pp. 6–9 (2006)
9. Mueller, D., Raggett, D.: Report for the W3C Workshop on Improving Access to Financial Data on the Web (2009), http://www.w3.org/2009/03/xbrl/report.html
10. O'Riain, S., Curry, E., Harth, A.: Xbrl and open data for global financial ecosystems: A linked data approach. Int J Account Inf Syst (2012), doi:10.1016/j.accinf.2012.02.002
11. ORiain, S., Harth, A., Curry, E.: Linked data driven information systems as an enabler for integrating financial data. Information Systems for Global Financial Markets, Emerging Developments and Effects pp. 239–270 (2011)
12. Reiter, R.: A logic for default reasoning. Artificial Intelligence 13(1), 81–132 (1980)
13. Wenger, M., Thomas, M., Jr, J.B.: An ontological approach to XBRL financial statement reporting. In: 17th Americas Conference on Information Systems. Michigan, USA (2011)

Using Semantic Web Technologies to Facilitate XBRL-based Financial Data Comparability

Héctor Carretié[1], Beatriz Torvisco[1], Roberto García[2]

Universidad Rey Juan Carlos
Paseo Artilleros. 28032 Madrid, Spain
{hector.carretie, beatriz.torvisco}@urjc.es

Universitat de Lleida
Jaume II, 69. 25001 Lleida, Spain
roberto.garcia@udl.cat

Abstract. The XML Business Reporting Language (XBRL) is a standard for business and financial reporting. Many institutions are making available or requiring data in this format, e.g. the US SEC or the Spanish CNMV. However, XBRL data is loosely interconnected and it is difficult to mix and compare, especially when it is based on different accounting principles. Our contribution is based on converting XBRL reports into semantic data and then using Semantic Web technologies to formalise equivalences among terms from different accounting standards. This approach has been evaluated in a particular scenario and it is available online.

Keywords. Business, accounting, finance, interoperability, comparability, Semantic Web, ontology.

1 Introduction

There are many attempts to move existing data to the Semantic Web domain, especially relevant due to the amount of data being mapped are those around the Linked Data initiative [1]. The main motivation to do so is that usually this data is not offering its full potential because it is isolated, i.e. not connected to other external pieces of data that enrich them. It might even be the case that the data is loosely interconnected internally, because it lacks formal semantics. Most of the time this is due to the fact that the technological solutions used to publish that data do not make it easy to interconnect it internally and to other external data sources.

Business reporting is a domain where the need for a common data format for reports has already been identified. XBRL (eXtensible Business Reporting Language) is an XML language intended for modelling, exchanging and automatically processing business and financial information. XBRL is gaining a lot of momentum, especially thanks to the support of some regulators and government agencies worldwide. It is especially significant the importance of the XBRL program promoted by the U.S. Securities and Exchange Commission (SEC). Currently, all companies filing to the SEC are doing so using XBRL following the Government Information Transparency Act, which requires federal agencies to collect their data in a uniform, searchable format using XBRL thereby simplifying mandatory financial reporting for companies that receive federal funds.

However, despite the great success in the adoption of XBRL, we have observed some limitations in its support for cross analysis of financial information in XBRL tools and applications, as it is detailed in Section 2, that might threaten its usefulness. These limitations are not just among data based on different accounting principles, which are represented in XBRL using taxonomies. It even happens when comparing filings for different companies based on the same taxonomies or filings for the same company based on different versions of the taxonomies.

We argue that this limitation is inherited from the technologies underlying XBRL, especially XML. XML takes a document-oriented approach, where each document presents a tree structure. This makes it difficult for XML-based tools to provide functionalities that blur this separation into documents and that overcome the limitations of a tree structure when mashing-up data from different sources. Moreover, XBRL does not provide formal semantics that might help to integrate different taxonomies using logic reasoners.

In any case, the integration of XBRL data into comparable information is a strong requirement for the analysis of business and financial information at a global scale. This might increase the efficiency and effectiveness of the decision-making processes relying on this kind of information. For instance, bankruptcy prediction and other tasks related to the assessment of the solvency of a firm, a business sector or set of interrelated companies. Many have already pointed to this issue and propose Semantic Web technologies as a natural choice for XBRL data integration, cf. Section 2.

Despite these potential benefits, currently, financial and business data is being produced using XBRL and it seems that more and more XBRL data is going to be available in the future. XBRL is been promoted by regulators and government agencies like the US SEC, as it has been shown before, but also other bodies like the European Union or the Spanish Securities Commission (CNMV) [2].

Consequently, our opinion is that the best short-term approach to enjoy the benefits of Semantic Web technologies when working with financial data is not to propose and alternative language based on these technologies, but to apply methods to map existing XBRL to semantic metadata.

The rest of this paper is organised as follows. The next subsections introduce the structure of XBRL and Section 2 presents the related work. Then, in Section 3, the approach for generating semantic data from XBRL is presented. It is based on a transformation from XML data to RDF using the XBRL to RDF mapping, which is described in Section 3.1. Then, the second step is to map the XML Schemas that structure XBRL data to OWL ontologies using the XBRL Schema to OWL mapping detailed in Section 3.2.

The results of the previous mappings, as detailed in Section 4, are a set of OWL ontologies for the main XBRL taxonomies used by the US SEC and based on the US GAAP[1]. Based on these ontologies, it has been possible to map the XBRL instance documents sent to the US SEC since 2009 resulting in more than 100M triples availa-

[1] Generally Accepted Accounting Principles, http://en.wikipedia.org/wiki/Generally_Accepted_Accounting_Principles_(United_States)

ble from the LOD Cloud as the Semantic XBRL dataset[2]. Some preliminary experiments have also been done with XBRL data based on the International Financial Reporting Standards (IFRS) and the Spanish PGC (Plan General Contable) accounting regulations.

Section 5 presents the main evaluations done so far. First of all, there are the results of a basic logical evaluation of the resulting ontologies. Then, we present a deeper evaluation of the overall approach through an scenario where comparability between two XBRL reports for the same company but based on different accounting principles is attained using Semantic Web technologies once they have been mapped to semantic data. Finally, Section 0 presents the conclusions and the future work.

1.1 XBRL

XBRL is based on two kinds of documents, instance documents and taxonomies. Instance documents report business facts and point to a set of taxonomies, which define the meaning of these facts, e.g. under what accounting principles they hold, what other facts they related to or what kind of things do they refer to.

1.1.1 Instances

More concretely, a XBRL instance document contains business Facts. An example of a Fact could be "sales in the last quarter". If the Fact is simple valued, like "the long term debt is 350,000" whose value is just a number, it is called Item. If the Fact has a more complex value, like "for the *preferred stock*, the *preferred stock par value per share* is 0 and the *preferred stock shares authorized* is 2000", it is called Tuple.

Items are represented in XBRL as a single XML element with the value as its content while Tuples are represented by XML elements containing nested Items or Tuples, i.e. subelements.

However, facts are not isolated entities and it is not enough to provide their values, it is also necessary to contextualize them. Consequently, four more entities are introduced in the XBRL model:

- **Context**: it defines the *entity* (e.g. company or individual) to which the fact applies, the *period* of time the fact is relevant and an optional *scenario*. The period of time can have zero length for instance and its value is based on ISO 8601 for date and time values. Scenarios provide further contextual information about the facts, such as whether the business values reported are actual, projected or budgeted. Contexts are referenced from Facts using the "contextRef" attribute, which specifies that the given Fact is valid for an *entity*, *period* and *scenario*.
- **Unit**: it defines a unit of measure, such as "USD" or "shares". They are referenced from Facts using the "unitRef" attribute, which specifies that the numeric or fractional value of the Fact is based on that unit of measure. Complex units can also be defined, like "USD per share". Currency units are based on ISO 4217.
- **Reference**: The kinds of facts under consideration are defined by taxonomies, which specify their meaning in the context of some accounting principles or purpose, e.g. Facts relevant for banking and savings institutions. These kinds of facts

[2] http://thedatahub.org/dataset/semantic-xbrl

are then used in instance documents in order to specify actual values for them. However, they are linked to their definition in the taxonomies, typically through schema references, in order to be able to retrieve their meaning.

- **Footnote**: it contains some additional support content and it is associated to Fact using XLink.

Table 1 shows part of an instance document from the EDGAR program that contains a Context element, which defines a company, a time period and the scenario “unaudited”. Then, there is a Fact that holds in that context. The Fact references the Context and the value unit, while their content is the fact numeric value.

Table 1. Context and facts examples from an EDGAR filing

```
…
<context id="From20080301-To20080530_EnterpriseSolutions_Unaudited">
    <entity>
      <identifier scheme="http://www.sec.gov/CIK">796343</identifier>
      <segment><adbe:EnterpriseSolutions /></segment>
    </entity>
    <period>
      <startDate>2008-03-01</startDate>
      <endDate>2008-05-30</endDate>
    </period>
    <scenario><adbe:Unaudited /></scenario>
</context>
…
<adbe:EnterpriseSolutionsRevenue decimals="-6"
contextRef="From20080301-To20080530_EnterpriseSolutions_Unaudited"
unitRef="USD">54400000</adbe:EnterpriseSolutionsRevenue>
…
```

1.1.2 Taxonomies

Taxonomies are the other kind of XBRL document. A taxonomy defines a hierarchy of concepts, basically kinds of Facts, and captures part of their intended meaning. In XBRL there is a set of base taxonomies that define the core concepts and other ones that extend them in order to particularize these concepts for concrete accounting principles, application domains, etc. Additionally, it is possible to extend existing taxonomies and accommodate them to particular needs.

Taxonomies are based on XML Schemas, which provide the taxonomy building primitives and the extension mechanisms. Moreover, there are also “linkbases”, which allow establishing links beyond the taxonomy tree structure using XLink.

- **Schemas** define concepts that are instantiated as Items or Tuples, depending on their complexity, in the instance documents. They are based on XML Schema elements (xsd:element). A concept definition provides the fact name, whether it is a tuple or an item and its value data type (such as monetary, numeric or textual).
- **Linkbases** define links from concepts in a taxonomy to labels, pieces of content or other concepts. The XBRL specification defines five different kinds of linkbases.
    - **Label Linkbase**: set of links that provides human readable strings for concepts, potentially in multiple languages.
    - **Reference Linkbase**: these links associate concepts with citations of some body of authoritative literature.

- **Calculation Linkbase**: these are links that associate a set of values of concepts in taxonomies with a mathematical calculation that must be checked for consistency, for instance that a set of concepts with percentage values sum up 100%.
- **Definition Linkbase**: it provides semantic relations between concepts like is-a, whole-part, etc.
- **Presentation Linkbase**: This linkbase associates concepts with other concepts so that the resulting relations can guide the creation of a user interface, rendering, or visualisation.

2 Related Work

The U.S Securities and Exchange Commission (SEC) offers some online tools that allow interacting with the data available in XBRL form. There is a tool called Interactive Financial Reports that allows viewing and charting companies financial information. It also provides some functionality that allows comparing different filings and different companies, though it is hard to use and prone to even the slightest differences between the compared filing facts, even when there is just a name change for facts from filings of the same company.

There is also the Financial Explorer, which presents company financial data through very informative diagrams but just from one company at a time, and the Executive Compensation tool. The later allows comparing just two facts, Public Market Capitalization and Revenue, across all filed companies.

Apart from the SEC tools, there are some other XBRL tools, most of them proprietary and with quite high licensing cost. Among them, the Fujitsu XBRL Tools[3] should be highlighted because they are one of the most popular tool sets and it is available for XBRL Consortium members and academic users. The tools comprise taxonomy and instance editors, viewers and validators.

The most powerful tool in this set, though still in beta and with many usability problems, is the Instance Dashboard. This application can consume multiple instance documents and, by specifying a base taxonomy, users can perform some comparison analysis, though limited to facts in a taxonomy that appears in all the filings.

As it can be noted from the previous analysis, the main limitation of XBRL tools is their limited support for cross analysis of financial information, not just among data based on different taxonomies, even when comparing filings for different companies based on the same taxonomies.

This limitation is inherited from the technologies underlying XBRL, especially from XML. XML takes a document-oriented approach, where each document presents a tree structure. This makes it difficult for XML-based tools to provide functionalities that blur this separation into documents and that overcome the limitations of a tree structure when mashing-up data from different sources.

Consequently, Semantic Web tools are being considered by people like Charles Hoffman, the father of XBRL: "*This field [W3C semantic standards] is rich with pos-*

[3] Fujitsu XBRL Tools, http://www.fujitsu.com/global/services/software/interstage/xbrltools/

sibilities and stands as the next logical step in the natural progression of information technology to seek a higher value proposition" [3].

This interest is materializing, and the combination of XBRL and the Semantic Web has been receiving some attention in different blogs [4,5], mailing lists and web groups[4]. The first attempts to combine both technologies focused on specific for some parts of XBRL. For instance, there is an ontology about financial information based on XBRL that is specific for investment funds [6] and, though it is generated using a generic XBRL taxonomy to OWL ontology algorithm, there is not and equivalent tool that maps generic XBRL instance data.

Another quite specific tool maps quarterly and semester accounting information submitted to the Spanish securities commission (CNMV) to RDF [2]. Both approaches are based on procedural code specially developed in order to extract specific patterns from the XBRL data. Consequently, they are difficult to scale to the whole XBRL specification and sensible to minimal changes in it.

More recent attempts have widened and generalised their scope. For instance, eTEN was an European Community programme providing funds to help make e-services available throughout the European Union. This programme ended in 2006. Within this programme there was the WINS project: Web-based Intelligence for common-interest fiscal Networked Services.

WINS provides a Web-based Business Intelligence (BI) Service to public and private Financial Institutions by integrating BI products and knowledge discovery tools to produce new financial knowledge on companies from information gathered through interoperable information services. Within the WINS context, Declerk and Krieger [7] pointed out some limitations encountered in the XBRL schema documents mainly due to the lack of reasoning support over XML-based data. They proposed the "ontologization" or process to translate XBRL taxonomies into OWL to overcome these limitations.

The "ontologization" starts from the WINS information extraction (IE) task, which gathers financial facts from PDF files and converts them into XBRL documents. From these document, the process continues based on a hand-made translation of XBRL facts into OWL ontologies that then helps classifying the facts into higher-level concepts like Balance Sheet or Statement of Income. However, the ontologies are not exploited beyond this point in order to facilitate the comparability of the financial facts across different accounting standards.

Another example of mapping from XBRL to Semantic Web technologies is OpenLink XBRL Sponge, which maps generic XBRL instance data to RDF [8]. However, in this case, there is not and associated mapping from the taxonomies instance data is based on to ontology languages. Therefore, it is not easy to facilitate the comparability of the financial facts by working at the conceptual level provided by the ontologies.

Bao et al. [9] do consider the comparability issue and they point out the tremendous human cognitive effort that must be done when comparing financial data written in XBRL. Their proposal is to overcome this problem by defining the *logic model* of XBRL reports using the Web ontologies language OWL to design ontologies that cap-

[4] XBRL Ontology Specification Group, http://groups.google.com/group/xbrl-ontology-specification-group

ture the meaning of the reports beyond just their structure. They transform concepts into classes and arcroles into properties. However, the possibilities of the logic models generated are not put into practice in comparability scenarios that involve different accounting regulations.

Finally, latest approaches start to focus on comparability and attempt to profit from Semantic Technologies and Linked Data principles to attain it [10]. For instance, the XBRL European Business Registry (xEBR) is an XBRL Europe project to create a list of concepts, which are common across the various European business registries. The concepts encompass basic financial data as well as company profiles. However, this Project is still limited by the fact that there is no common regulation for Business Registries in Europe. Therefore, many Registries in Europe have built their own set of taxonomies.

Our proposal, as detailed in the next sections, focuses on facilitating comparability at the semantic level, where it is easier to establish the equivalences among financial facts independently of the particular taxonomies and associated accounting standards they come from. In order to do that, we propose an approach that, instead of directly processing XBRL data, takes profit from the fact that it is expressed using XML and specified using XML Schemas. The instance XML documents are translated into RDF that models the financial facts and refers to the concepts modelled in ontologies generated from the schemas. From this point, it is now possible to establish equivalences that facilitate comparability at the ontology level use inference to benefit from this knowledge at the instance level.

3 Approach

The proposed approach is based on the transfer of existing XBRL taxonomies and instance data to Semantic Web technologies. This transfer is based on the XML Semantics Reuse methodology [11,12] and the XML Schema to OWL and XML to RDF tools implemented in the ReDeFer project[5].

This methodology combines an XML Schema to web ontology transformation, XSD2OWL, with a transparent translation from XML to RDF, XML2RDF. The ontologies generated by XSD2OWL are used during the XML to RDF step to generate semantic metadata that takes into account the XML Schema intended meaning.

This approach differs from other attempts to move metadata from the XML domain to the Semantic Web. Some of them just model the XML tree using the RDF primitives [13]. Others concentrate on modelling the knowledge implicit in XML languages definitions, i.e. DTDs or the XML Schemas, using web ontology languages [14,15]. Finally, there are attempts to encode XML semantics integrating RDF into XML documents [16,17].

However, none of them facilitate an extensive transfer of XML metadata to the Semantic Web in a general and transparent way. Their main problem is that the XML Schema implicit semantics are not made explicit when XML metadata instantiating this schemas is translated. This is so because the RDF data produced from XML in-

[5] ReDeFer project, http://rhizomik.net/redefer

stance data looses its links to the XML Schemas that structure them and model the relations among different XML entities.

These relations among different XML entities are what carry the XML Schema implicit semantics. They capture part of the meaning intended by the schema developer that, though XML Schema does not provide a way to encode semantics, is recorded in the way XML Schema constructs are used. For instance, by modelling that element "father" is a *subtitutionGroup* for element "parent", it is possible to interpret that "parent" is more general than "father" and that "father" can appear where "parent" appears. More details about the implicit semantics of XML Schema constructs as compared to OWL ones are provided in Section 3.2.

Therefore, the previous transformations from XML to RDF do not take profit from the meaning encoded in XML Schemas and produce RDF metadata almost as semantics-blind as the original XML. Or, on the other hand, they capture this semantics but they use additional ad-hoc semantic constructs that produce less transparent metadata.

3.1 XML2RDF

The XML to RDF transformation follows a structure-mapping approach [13] and tries to represent the XML metadata structure, i.e. a tree, using RDF. The RDF model is based on the graph so it is easy to model a tree using it. Moreover, we do not need to worry about the loss of semantics produced by structure-mapping. We formalised the underlying semantics into the corresponding ontologies and we will attach them to RDF metadata using the instantiation relation *rdf:type*.

The structure-mapping is based on translating XML metadata instances to RDF that instantiates the corresponding constructs in OWL. The more basic translation is from *xsd:elements* and *xsd:attributes* to *rdf:Properties* (*owl:ObjectProperties* for node to node and *owl:DatatypeProperties* for node to value relations).

Values are kept during the translation as simple types and RDF blank nodes are introduced in the RDF model in order to serve as the source and destination for properties. They will remain blank until they are enriched with semantic information.

The resulting RDF graph model contains all that we can obtain from the XML tree. It is already semantically enriched thanks to the *rdf:type* relation that connects each RDF property to the *owl:ObjectProperty* or *owl:DatatypeProperty* it instantiates. It can be enriched further if the blank nodes are related to the *owl:Class* that defines the package of properties and associated restrictions they contain, i.e. the corresponding *xsd:complexType*. This semantic decoration of the graph is formalised using *rdf:type* relations from blank nodes to the corresponding OWL classes.

At this point we have obtained a semantically enabled representation of the input metadata, a representation that makes the meaning intended by the XML and XML Schema modelers explicit from a computer point of view. The instantiation relations can now be used to apply OWL semantics to metadata. Therefore, the semantics derived from further enrichments of the ontologies, e.g. integration links between different ontologies or semantic rules, are automatically propagated to instance metadata thanks to inference.

Focusing on XBRL data, what we get by applying this triplification process of the corresponding XML data is summarised in Fig. 1. This figure shows the XBRL core concepts as they are modeled in the resulting RDF data. The report is modelled as an instance of the class "ReportType" and facts are modelled as instances of "FactType".

In fact, if a direct modelling of the underlying XML tree was performed, facts should be modelled as RDF Properties because they correspond to XML elements. However, in order to make the resulting RDF data more usable as it is more intuitive to view a fact as class instance than as a relation one, we have introduce a modification in the basic XML2RDF algorithm as it is detailed in the next subsection.

Then, continuing from the "FactType" instance, there are relations to the actual value of the financial fact modelled using rdf:value and two properties stating the decimals and unit used for that value. There is also a property linking the fact to its context, which details the involved entity, the time period and the scenario.

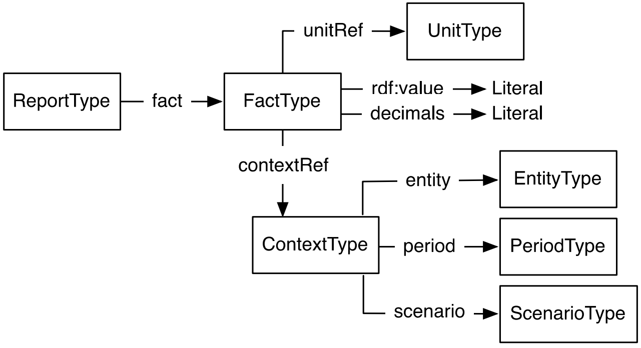


Fig. 1. RDF model for the core XBRL concepts generated using XML2RDF and XSD2OWL (boxes correspond to classes and arrows to properties having them as domain/ranges)

3.2 XBRL Schema to OWL Mapping

The XML Schema to OWL transformation is responsible for capturing the schema implicit semantics, which is determined by the combination of XML Schema constructs. The transformation is based on translating these constructs to the OWL ones that best capture their intended meaning. These translations are detailed in Table 2.

The XML Schema to OWL transformation is quite transparent and captures a great part XML Schema semantics. The same names used for XML constructs are used for OWL ones, although in the new namespace defined for the ontology. XSD and OWL constructs names are identical; this usually produces uppercase-named OWL properties because the corresponding element name is uppercase, although this is not the usual convention in OWL. Therefore, XBRL Schema to OWL produces OWL ontologies that make explicit the semantics of the corresponding XBRL taxonomies.

The only caveats are the implicit order conveyed by *xsd:sequence* and the exclusivity of *xsd:choice*. For the first problem, *owl:intersectionOf* does not retain its operands order, there is no clear solution that retains the great level of transparency that has been achieved. The use of RDF Lists might impose order but introduces ad-hoc constructs not present in the original metadata.

Table 2. XBRL Schema to OWL translations for the XML Schema constructs

XML Schema	OWL	Mapping motivation
element[@substitutionGroup= "xbrli:item"]	owl:Class	Facts, though elements, are mapped to classes
element | attribute	rdf:Property owl:DatatypeProperty owl:ObjectProperty	Named relation between nodes or nodes and values
element@substitutionGroup="xbrli:item"	rdfs:subClassOf	The corresponding element is mapped to a owl:Class rdfs:subClassOf xbrli:item
element@substitutionGroup	rdfs:subPropertyOf	Relation can appear in place of a more general one
element@type	rdfs:range	The relation range kind
complexType|group |attributeGroup	owl:Class	Relations and contextual restrictions package
complexType//element	owl:Restriction	Contextualised restriction of a relation
extension@base | restriction@base	rdfs:subClassOf	Package concretises the base package
@maxOccurs @minOccurs	owl:maxCardinality owl:minCardinality	Restrict the number of occurrences of a relation
sequence choice	owl:intersectionOf owl:unionOf	Combination of relations in a context

Moreover, as it has been demonstrated in the Semantic Web community, the element ordering does not contribute much from a semantic and knowledge representation point of view [18] in most cases and when it is a requirement it is more convenient to explicitly represent it using some sort of order attribute or property. For the second problem, *owl:unionOf* is an inclusive union, the solution is to use the disjointness OWL construct, *owl:disjointWith*, between all union operands in order to make it exclusive.

4 Results

First of all, we have generated an ontological infrastructure for the XBRL core, currently XBRL 2.1. It is composed by the ontologies resulting from mapping the XBRL core XML Schemas using the XBRL Schema to OWL mapping: XBRL Instance, XBRL Linkbase, XBRL XL and XBRL XLink. These ontologies have been adapted to accommodate the changes introduced by XBRL to RDF that make the output semantic data more usable, basically by making facts classes and no longer properties.

Apart from the previous schemas, the following schemas have been also mapped in order to be able to map the XBRL data submitted to the US SEC.

From US GAAP (Generally Accepted Accounting Principles) the schemas, and corresponding ontologies, are: Primary Terms Elements (USFR-PTE), Primary Terms Relationships (USFR-PTR), Financial Services Terms Elements (USFR-FSTE), Financial Services Terms Relationships (USFR-FSTR) and Investment Management Terms Relationships (USFR-IME). For specific industries: Banking and Savings Institutions (US-GAAP-BASI), Commercial and Industrial (US-GAAP-CI), Insurance (US-GAAP-INS) and Investment Management (US-GAAP-IM).

There are also some non-GAAP schemas that have been also mapped to OWL ontologies: Accountants Report (USFR-AR), Management Discussion and Analysis (USFR-MDA), Management Report (USFR-MR) and SEC Certifications (USFR-SECCERT).

The same approach has been followed to map the IFRS taxonomies and the ones used by the Spanish securities commission (CNMV). Most of the previous ontologies are available from the BizOntos Business Ontologies web page[6] and the semantic data for all the processed filings can be queried and browsed from the Semantic XBRL site[7]. Currently, more than 25 thousand filings have been processed from the US SEC, plus some from the CNMV. The combination of all these filings once mapped to RDF amounts slightly more than 100 million triples. At this step, it is possible to take profit from semantic web technologies in order to improve the interconnectedness of the dataset by means of semantics-enabled data integration.

5 Evaluation

The proposed approach has been evaluated using two input XBRL reports for the same company but based on different accounting principles, and consequently different taxonomies. The input data is from Telefonica S.A., one of the reports was submitted to the Spanish CNMV and the other to the US SEC[8], more specifically the consolidated Balance Sheet for the years 2009 and 2008.

The motivation is that Telefonica is one of the few Spanish corporations that files their financial statements to the Spanish securities commission (CNMV) in XBRL format and also to the American Securities Exchange Commission (US SEC). The 2009 period was the last period available in the CNMV and SEC websites, at the time of the elaboration of the present evaluation.

The elaboration of the financial statements for the CNMV has been done under the Spanish GAAP regulations[9], i.e. Plan General de Contabilidad, issued in 2007 and based on IFRS. Meanwhile, financial information filled to the US SEC was elaborated under the IFRS, following SEC's provisions for foreign corporations.

Therefore, it could be expected that both XBRL financial reports would be the same or at least quite similar. However, as the Table 4 shows, there are some differences mainly due to different levels of disaggregation. The totals for assets or liabilities coincide but not the figures contained under these main sections.

Fig. 2 highlights the accounts where quantity differences are found. For instance, in the 2009 balance sheet for the SEC (on the left), "Non-current financial assets"

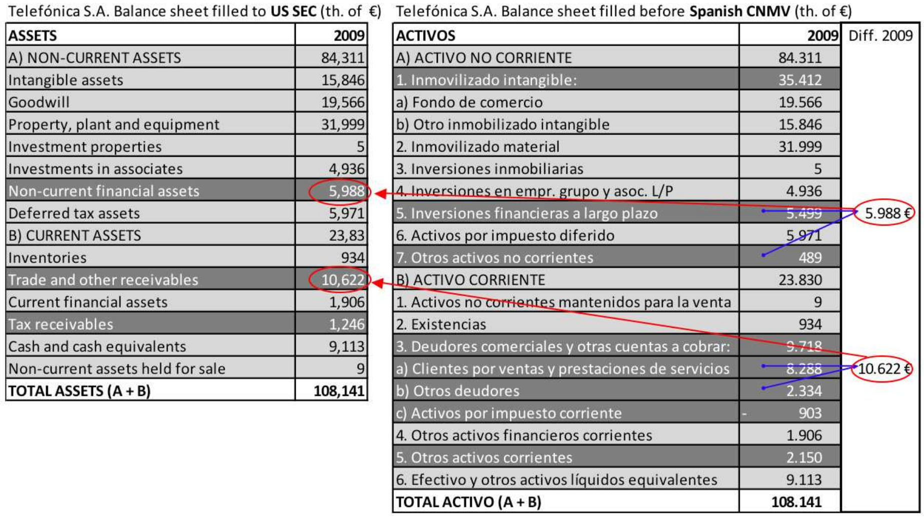
[6] BizOntos, http://rhizomik.net/ontologies/bizontos

[7] SemanticXBRL, http://rhizomik.net/semanticxbrl

[8] Telefonica's report to the CNMV is available from http://www.cnmv.es/ipps/default.aspx and the one sent to SEC is available from http://www.sec.gov/Archives/edgar/data/814052/000095010310000881/dp16939_20f.htm

[9] Models recently modified by Ministerial Oder JUS/1698/2011 of June 13, approving the model for presentation at the Mercantile Registry of the consolidated financial statements

amounts 5,988 millions of euros, meanwhile in balance sheet for the CNMV (on the right) “Inversiones financieras a largo plazo” (long-term financial investments) amounts 5,499 millions and “Otros activos no Corrientes” (Other non-current assets) amounts 489 millions. Both accounts sum up 5,988 millions, so the sum of “Inversiones financieras a largo plazo” and “Otros activos no Corrientes”, two terms specific to CNMV taxonomies, is equivalent to the IFRS term “Non-current financial assets”.



Telefónica S.A. Balance sheet filled to **US SEC** (th. of €)

ASSETS	2009
A) NON-CURRENT ASSETS	84,311
Intangible assets	15,846
Goodwill	19,566
Property, plant and equipment	31,999
Investment properties	5
Investments in associates	4,936
Non-current financial assets	5,988
Deferred tax assets	5,971
B) CURRENT ASSETS	23,83
Inventories	934
Trade and other receivables	10,622
Current financial assets	1,906
Tax receivables	1,246
Cash and cash equivalents	9,113
Non-current assets held for sale	9
TOTAL ASSETS (A + B)	**108,141**

Telefónica S.A. Balance sheet filled before **Spanish CNMV** (th. of €)

ACTIVOS	2009	Diff. 2009
A) ACTIVO NO CORRIENTE	84.311	
1. Inmovilizado intangible:	35.412	
a) Fondo de comercio	19.566	
b) Otro inmobilizado intangible	15.846	
2. Inmovilizado material	31.999	
3. Inversiones inmobiliarias	5	
4. Inversiones en empr. grupo y asoc. L/P	4.936	
5. Inversiones financieras a largo plazo	5.499	5.988 €
6. Activos por impuesto diferido	5.971	
7. Otros activos no corrientes	489	
B) ACTIVO CORRIENTE	23.830	
1. Activos no corrientes mantenidos para la venta	9	
2. Existencias	934	
3. Deudores comerciales y otras cuentas a cobrar:	9.718	
a) Clientes por ventas y prestaciones de servicios	8.288	10.622 €
b) Otros deudores	2.334	
c) Activos por impuesto corriente	- 903	
4. Otros activos financieros corrientes	1.906	
5. Otros activos corrientes	2.150	
6. Efectivo y otros activos líquidos equivalentes	9.113	
TOTAL ACTIVO (A + B)	**108.141**	

Fig. 2**.** Assets section Telefonica’s Balance Sheet filled to the US SEC

Other equivalences requiring the addition of different account are also highlighted in Fig. 2 and marked using dark grey. The accounts marked with light grey have direct equivalences between the taxonomies used by the US SEC and the CNMV. For instance, “Intangible assets” in the US SEC document is equivalent to “Otro inmobilizado intangible”.

The instance document from Telefonica filed to the Spanish CNMV includes terms specific to the Spanish terminology, defined in the “ipp-gen” namespace in the XBRL instance documents, but with an equivalent term in IFRS. Other terms reuse the international standard, and thus are in the “ifrs-gen” namespace, but in some cases they do not coincide with the terms specified in the IFRS taxonomy. Finally, some elements are specific to the CNMV, e.g. “ipp-gen:TotalActivoNiif”.

Both, numerical and terminological differences, dramatically decrease the comparability of the two consolidated balance sheets. However, it is possible to establish equivalences at the conceptual level, and arithmetic operations among them when there is not a direct equivalence. This is easily achievable thanks to Semantic Web technologies once the involved taxonomies have been mapped to OWL and the corresponding instance documents to RDF. The next section presents some examples.

5.1.1 Mappings between Spanish PGC and IFRS

Table 3 shows some of the semantic mappings generated for the Telefonica scenario between the ontologies corresponding to the IFRS taxonomies and thus for the CNMV taxonomies.

Table 3. Mappings between Spanish PGC and IFRS

Spanish CNMV (PGC taxonomies)	**US SEC (IFRS taxonomies)**	**Semantic Mappings**
ipp-gen: ActivoNoCorrienteNiif 84.311 €	ifrs: NoncurrentAssets 84.311 €	ipp-gen:ActivoNoCorrienteNiif owl:**equivalentClass** ifrs:NoncurrentAssets

ifrs-gp:
TradeAndOtherReceivablesNet
Current =
ipp-gen:
ClientesVentasPrestaciones
Servicios +
ipp-gen:
OtrosDeudores
8.288€ + 2.334€
	ifrs:
	TradeAndOtherCurrentReceivables
	10.622€
		ifrs-gp:TradeAndOtherReceivablesNetCurrent
		owl:equivalentClass
		ifrs:TradeAndOtherCurrentReceivables
		
		CONSTRUCT {
		[] a ifrs-gp:
		TradeAndOtherReceivablesNetCurrent;
		xbrli:contextRef ?context;
		xbrli:unitRef ?unit;
		xbrli:decimals ?decimals;
		rdf:value ?value. }
		WHERE {
		?cvps a ipp-gen:
		ClientesVentasPrestacionesServicios;
		xbrli:contextRef ?context;
		xbrli:unitRef ?unit;
		xbrli:decimals ?decimals;
		rdf:value ?cvps-value.
		
		?od a ipp-gen:OtrosDeudores;
		xbrli:contextRef ?context;
		xbrli:unitRef ?unit;
		xbrli:decimals ?decimals;
		rdf:value ?od-value.
		BIND(?cvps-value+?od-value AS ?value) }

The approach is to model the accounts determined to be equivalent, because the are in the same part of the balance sheet and correspond to the same quantity, as equivalent at the ontology level using the *equivalentClass*[10] OWL construct. When the relation is more complex than a simple equivalence, for instance when the value for a term in one vocabulary is the sum of more than one value in other vocabularies, then the approach is to use a *Construct*[11] SPARQL query that computes the combined value, for instance the sum, and creates the computed fact.

The complete set of mappings is available from an online demo[12], where they are also put into practice using a Semantic Web repository that includes an inference an inference and a SPARQL engine that "execute" these mappings. For the demo, just the CNMV XBRL document is loaded into the repository and the mappings are used to generate most of the IFRS version of the assets part of the balance sheet using the semantic mappings.

[10] OWL Equivalent Class, http://www.w3.org/TR/owl-ref/#equivalentClass-def

[11] SPARQL Construct, http://www.w3.org/TR/sparql11-query/#construct

[12] SemanticXBRL Demo, http://rhizomik.net/semanticxbrl-demo/

6 Conclusions and Future Work

As it has been shown, it is possible to map the XML data for XBRL filings in order to generate semantic data that keeps all the original information and structure. This mapping also includes the involved XML Schemas that structure the XML data. These schemas are mapped to Web ontologies, which make all the semantics implicit in the original XML Schemas explicit and available when querying semantic data.

Moreover, it is also possible to take profit from Web ontology primitives in order to semantically integrate different filings following different XML Schemas, i.e. XBRL taxonomies. Once mapped to ontology concepts and relations, the XBRL contexts, facts and other resources defined for different filings can be related as more specific, more general or equivalent.

This approach has been put into practice in the context of the US SEC's XBRL program. It has been possible to apply the previous XML to RDF and XML Schema to Web ontology mappings to filings sent to the US SEC and some from the Spanish CNMV. More than 100 million triples have been obtained, which are structured by the ontologies generated from the corresponding taxonomies.

Moreover, the benefits of the approach have been validated in a real scenario where it is possible to generate an XBRL report following the IFRS taxonomies starting from one based on the Spanish CNMV taxonomies using semantic mappings established at the ontology level.

Future work focuses on, once we establish more semantic mappings at the conceptual level that can be reused to map instance documents for different companies, obtaining financial statement analysis ratios, taking profit from the semantic data already available.

For instance, to compute the debt ratio (equivalent to total liabilities / total assets), and current ratio (equivalent to total liabilities / total assets) by analysing the balance sheets, or the Return on Sales (ROS, equivalent to net income / sales revenue). From these ratios and the semantic mapping, we will be able to create a ranking showing the best-positioned international companies for each ratio mixing the data they submit to different regulators.

Acknowledgements

The work described in this paper has been partially supported by Spanish Ministry of Science and Innovation through the Open Platform for Multichannel Content Distribution Management (OMediaDis) research project (TIN2008-06228).

References

[1] Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems. 5, 1–22 (2009).

[2] Núñez, S., de Andrés, J., Gayo, J.E., Ordoñez, P.: A Semantic Based Collaborative System for the Interoperability of XBRL Accounting Information. Emerging Technologies and Information Systems for the Knowledge Society. pp. 593–599. Springer, Berlin/Heidelberg, DE (2008).

[3] Hoffman, C.: Financial Reporting Using XBRL: IFRS and US GAAP Edition. Lulu.com (2006).

[4] Raggett, D. XBRL and RDF. In: Dave Raggett's Blog, (2008). Available from http://people.w3.org/~dsr/blog/?p=8

[5] DuCharme, B. Changing my mind about XBRL again. In: Bob DuCharme's weblog, bobdc.blog, (2008). Available from http://www.snee.com/bobdc.blog/2008/08/changing_my_mind_about_xbrl_ag.html

[6] Lara, R., Cantador, I., Castells, P.: Semantic Web Technologies For The Financial Domain. In: Cardoso, J. and Lytras, M. (eds.) The Semantic Web: Real-World Applications from Industry. pp. 41–74. Springer, New York, NY, USA (2008).

[7] Declerck, T., Krieger, H.: Translating XBRL into Description Logic: an approach using Protege, Sesame and OWL. In: Abramowicz, W. and Mayr, H.C. (eds.) Proceedings of the 9th International Conference on Business Information Systems, BIS'06. pp. 455–467. GI, Bonn, DE (2006).

[8] Erling, O., Mikhailov, I. RDF Support in the Virtuoso DBMS. In: Pellegrini, T., Auer, S., Tochtermann, K. and Schaffert, S. (eds.) Networked Knowledge - Networked Media. pp. 7-24. Springer, Berlin/Heidelberg, DE (2008).

[9] Bao, J., Rong, G., Li, X., Ding, L.: Representing Financial Reports on the Semantic Web: A Faithful Translation from XBRL to OWL. In: Dean, M., Hall, J., Rotolo, A., and Tabet, S. (eds.) Semantic Web Rules. pp. 144–152. Springer, Berlin/Heidelberg, DE (2010).

[10] O'Riain, S., Curry, E., Harth, A.: XBRL and open data for global financial ecosystems: A linked data approach. International Journal of Accounting Information Systems. In Press (2012).

[11] García, R.: Chapter 7: XML Semantics Reuse. In: García, R. A Semantic Web Approach to Digital Rights Management. VDM Verlag, Saarbrücken, Germany (2010).

[12] García, R., Gil, R.: Linking XBRL Financial Data. In: Wood, D. (ed.) Linking Enterprise Data. pp. 103–125. Springer, New York, NY, USA (2010).

[13] Klein, M.C.A.: Interpreting XML Documents via an RDF Schema Ontology. Proceedings of the 13th International Workshop on Database and Expert Systems Applications, DEXA'02. pp. 889–894. IEEE Computer Society, Washington, DC, USA (2002).

[14] Amann, B., Beeri, C., Fundulaki, I., Scholl, M.: Ontology-Based Integration of XML Web Resources. Proceedings of the 1st International Semantic Web Conference, ISWC 2002. pp. 117–131. Berlin/Heidelberg: Springer (2002).

[15] Cruz, I., Xiao, H., Hsu, F.: An Ontology-based Framework for XML Semantic Integration. Eighth International Database Engineering and Applications Symposium, IDEAS'04. pp. 217–226. IEEE Computer Society, Washington, DC, USA (2004).

[16] Lakshmanan, L., Sadri, F.: Interoperability on XML Data. Proceedings of the 2nd International Semantic Web Conference, ICSW'03. pp. 146–163. Springer, Berlin/Heidelberg, DE (2003).

[17] Patel-Schneider, P.F., Simeon, J.: The Yin/Yang web: XML syntax and RDF semantics. Proceedings of the 11th International World Wide Web Conference, WWW'02. pp. 443–453. ACM Press (2002).

[18] Berners-Lee, T. Why RDF model is different from the XML model. W3C Dessign Issues, (1998). Available from http://www.w3.org/DesignIssues/RDF-XML.html

FLORA – Publishing Unstructured Financial Information in the Linked Open Data Cloud

Mateusz Radzimski, José Luis Sánchez-Cervantes, Alejandro Rodríguez-González, Juan Miguel Gómez-Berbís, Ángel García-Crespo

[1] Departamento de Informática
Universidad Carlos III de Madrid, Spain
{mradzims, joseluis.sanchez, alejandro.rodriguez, juanmiguel.gomez, angel.garcia}@uc3m.es

Abstract. In the world, where computers assist humans in information processing in almost every aspects of our lives, there are still huge gaps of unsurveyed areas, where data exists in an unstructured or unprocessable form limiting its usefulness and requiring extra human effort. Many times such data is extremely useful for many parties, as is the case of financial data. This paper describes an ongoing work of the FLORA system that aims at transforming unstructured financial data into Linked Data form and interlinking it with other relevant datasets of LOD initiative in order to provide a financial knowledgebase for financial data analysis framework.

Keywords: linked open data, financial data, data integration, data publishing

1 Introduction

With increasing number of financial data sources being published on the web, still doesn't come the easiness of analysis and retrieving relevant information. Documents containing financial statements appear to be structured, however many times only its content is structured, but not the data itself. Therefore any analysis or further processing of such datasets are limited by high cost of transformation in order to be fit into existing analytic models and tools [1]. This situation is also keeping the bar of multisource data integration very high. On the other hand we experience the blossoming development of Linked Open Data [2] cloud that offers best practices of sharing the data across the Web with great integration capabilities. Apart from bringing transparent access, Linked Open Data allows for easy combination of many information sources thus allowing for better data analysis. Financial information in such form could be of use by many entities, such as regulatory bodies detecting market anomalies, banks analyzing the risk of held assets or investors making better informed decisions.

Such semantic information integration starts to play crucial role in many domains such as bioinformatics, medical domain and other life sciences with growing amount of data gathered in repositories such as Bio2RDF [3] or Linked Life Data [5]. Integrating those dataset using Linked Data approach opens new possibilities for better data discovering, querying and visualization.

This paper presents a high level overview of an ongoing work in the FLORA project. FLORA aims at bringing the advantages of Linked Data (and Linked Open Data) to lower the integration obstacles [4] and to transform financial data into the form which can be used in automated environments in the interoperable way. Employing NLP techniques for information extraction will foster transforming unstructured data from public companies' statements dealing with accounting and financial perspective. Linked Data cloud of computation results is not only a suitable format for browsing and querying, but it also forms an input for further analysis services for decision support, multi-faceted data presentation [6] and visualization [7].

2 Related Work

There are several initiatives related with extracting unstructured information in the financial domain. Some of these works have obtained outstanding results through applying semantic technologies. In this section some of this works are described briefly.

The MONNET project [8] proposes a solution to the cross-language information access problem by using a novel combination of Machine Translation and Semantic Web Technology [9] for the public and financial sector. MONNET achieves this through semantically aware term translation based on a novel approach that integrates ontology-based domain semantics with linguistic information from the domain lexicon. MONNET project provides several benefits of scientific innovation and scientific impact as: ontology-lexicon model, multilingual ontology localization, cross-lingual ontology-based information extraction, cross-lingual knowledge access, presentation framework, formal model for multilingual, lexicalized knowledge representations, ontology localization services, methodology for developing cross-lingual information access applications, integrated approach to ontology localization, cross-lingual ontology-based information extraction, ontology-based language-independent information access, to mention a few.

The FIRST project [10] addresses the challenges of dealing with financial data in a near real-time with vast and constantly growing amounts of heterogeneous sources from financial markets. It aims at providing a large-scale information extraction and integration infrastructure for supporting financial decision-making process. The main result, Integrated Financial Market Information System, is based on a pluggable open architecture framework for non-ICT skilled end-users for on-demand information access and highly scalable execution of financial market analyses.

The XLite [11] project's main goal is to develop technology to monitor and aggregate knowledge that is currently spread across mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence. By combining modern computational linguistics, machine learning, text mining and semantic technologies with the purpose to deal with the following two key open research problems: the first is extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases, and the second to adapt linguistic techniques and crowd-sourcing to deal with irregularities in informal language used primarily in social media.

LOD2 project [12] aims at developing technologies for scalable management of Linked Data collections in the many billions of triples and progress the state of the art of Semantic Web in data management, both commercial and open-source. It assumes RDF data representation as a viable choice for organizations worldwide and a premier data management format [13]. The LOD2 project contributes high-quality interlinked versions of public Semantic Web data sets, promoting their use in new cross-domain applications by developers across the globe. LOD2 also develops a suite of tools for data cleaning, linking and fusing that will help bootstrapping creation of datasets for new domains[1].

In [14] the lack of a well-documented software library for access and publication of data throughout the lifecycle of Linked Open Data and the problems related with the amount and quality sparse of the links among Linked Open Data Sources because of its growth were mentioned. The LATC project provides an alternative of solution to the problems previously mentioned. To support interested parties in Linked Data publication and consumption, LATC publishes and maintains a publication & consumption tools library along with screen-casts and tutorials. To provide an in-depth test-bed for data intensive applications, LATC publishes data produced by the European Commission, the European Parliament, and other European institutions as Linked Data.

In [15], the authors discuss how semantics can improve XBRL (Extensible Business Reporting Language) characteristics of expressiveness and interoperability beyond plain XML data representation. In a practical sense XBRL provides a potential platform for wide acceptance and adoption of Semantic Web. Finally the knowledge representation and Semantic Rules on the Web were mentioned.

These initiatives offer alternatives of solution to different problematic situations as: provide a solution for cross-language information access problem in public and financial sector, offer a large-scale information extraction and integration infrastructure for supporting financial decision-making process, develop technology to monitor and add knowledge that are spread across mainstream and social media, the develop a technology that allows the management of increased of Linked Data collections and the management of quality of the links among Linked Open Data Sources, to mention a few.

In comparative with the previously mentioned proposals, the main idea of our approach is based in obtaining large sets of unstructured financial information with the aim of use it in the creation of financial knowledge after the appliance of natural language processing techniques to filter out such information and get more accurate information which will be offered to stakeholders through after its publication in Linked Open Data cloud.

3 Conceptual Model

Financial data, both public and undisclosed, is traditionally represented in a vast number of different formats, ranging from textual descriptions to XML documents,

[1] LOD2 project objectives, as of February 2011: http://lod2.eu/WikiArticle/Project.html

such as XBRL format for public financial disclosures [16]. Although the terms and meaning is well known to those financial analysts that are dealing with it on a daily basis, it is still an enormous obstacle for machines that results in big integration efforts. Activities such as moving data between systems, getting data from quarterly reports for fundamental analysis or cross-domain data breakout need substantial manual effort of analyst. Even though there are numerous systems and formats dealing with financial data, they still need proper transformations for enabling interoperability. The idea of having data that is only usable within specific system is becoming out of fashion. On the other hand FLORA proposes data-driven approach for financial data integration, by following Linked Data principles, where the data instead of systems is the center of the overall process.

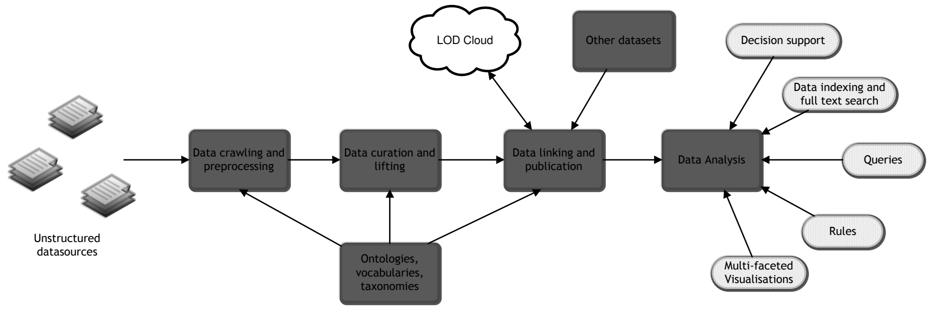

LOD Cloud
Other datasets
Decision support
Data indexing and full text search
Data crawling and preprocessing
Data curation and lifting
Data linking and publication
Data Analysis
Queries
Unstructured datasources
Ontologies, vocabularies, taxonomies
Rules
Multi-faceted Visualisations

Figure 1: Overview of the FLORA information extraction process

Transforming unstructured data into the structured form (as presented on Figure 1) comprise multiple steps that usually devise a similar high-level pattern that can be described as follows:

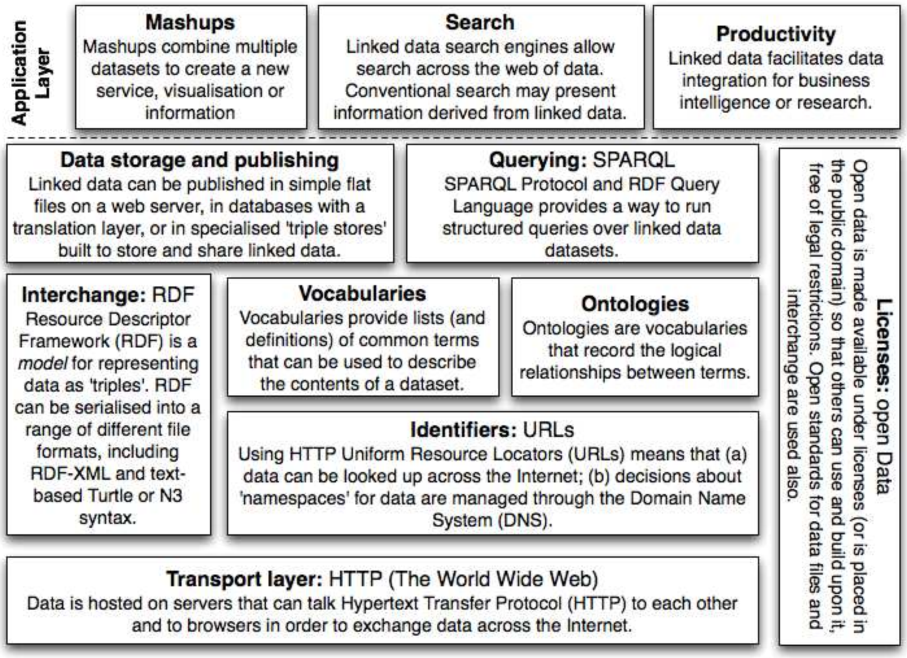
- Raw data is acquired and preprocessed in order to capture and extract the structure and isolate relevant data. Documents are cleaned and filtered according to input constraints. In case of documents having no defined structure, further NLP processes and ontology based information retrieval techniques, e.g. using GATE [17] for information retrieval.
- The data is lifted to the semantic form, using established ontologies and vocabularies that describe the dataset's domain. At this point, user-assisted data curation might be needed in order to improve the overall quality of data.
- Data can be published and interlinked with other datasets on the basis of using the same ontologies or describing same concepts. Linking and locating corresponding concepts across different datasets can be done automatically or semi-automatically, following approaches of LIMES [18] or SILK [19] projects.
- For improved data discovery, an indexing service is constantly traversing whole dataset facilitating full text search and results ranking.

Once published, the access to LOD dataset is realized by exposed SPARQL endpoint for data querying and HTTP access for data navigation. As it might be useful for lightweight data browsing and analysis, more sophisticated use cases might need the access to the whole RDF dataset. Such analysis services running on top of the dataset may provide further features, such as multi-faceted financial data visualizations (e.g. financial cockpit), decision support and data mining algorithms execution.

Adding reasoning capabilities on top of the RDF dataset will further improve of the data, as any data inconsistency can be easily detected and contradicting data removed or corrected.

4 Linked Data infrastructure for data-driven integration

Publishing results from previous steps in a form of Linked Data (and Linked Open Data in case of data with open license) forms an important part of the overall architecture. Most important building blocks of the Linked Data stack are show in the Figure 2. FLORA project covers all infrastructure blocks for making data available in the LOD cloud, following best practices in data publishing [20].



Application Layer

Mashups Mashups combine multiple datasets to create a new service, visualisation or information

Search Linked data search engines allow search across the web of data. Conventional search may present information derived from linked data.

Productivity Linked data facilitates data integration for business intelligence or research.

Data storage and publishing Linked data can be published in simple flat files on a web server, in databases with a translation layer, or in specialised 'triple stores' built to store and share linked data.

Querying: SPARQL SPARQL Protocol and RDF Query Language provides a way to run structured queries over linked data datasets.

Interchange: RDF Resource Descriptor Framework (RDF) is a *model* for representing data as 'triples'. RDF can be serialised into a range of different file formats, including RDF-XML and text-based Turtle or N3 syntax.

Vocabularies Vocabularies provide lists (and definitions) of common terms that can be used to describe the contents of a dataset.

Ontologies Ontologies are vocabularies that record the logical relationships between terms.

Identifiers: URLs Using HTTP Uniform Resource Locators (URLs) means that (a) data can be looked up across the Internet; (b) decisions about 'namespaces' for data are managed through the Domain Name System (DNS).

Transport layer: HTTP (The World Wide Web) Data is hosted on servers that can talk Hypertext Transfer Protocol (HTTP) to each other and to browsers in order to exchange data across the Internet.

Licenses: open Data Open data is made available under licenses (or is placed in the public domain) so that others can use and build upon it, free of legal restrictions. Open standards for data files and interchange are used also.

Elements of a Linked Open Data Stack - 2nd May 2011. CC BY-SA-NC
Draft sketch by Tim Davies (@timdavies / tim@practicalparticipation.co.uk) for IKM Working Paper on Linked Open Data for Development. Comments welcome. Search 'linked open data stack' on http://www.opendataimpacts.net for latest version.

Figure 2: Building blocks of the Linked Data stack[2]

[2] Elements of the Linked Open Data stack, rev 3, May 2011 http://www.practicalparticipation.co.uk/odi/2011/05/whats-in-the-linked-open-data-stack/

Combining the data from different datasets is possible by reusing the same vocabularies and ontologies for defining financial statements. Establishing links between corresponding concepts allows for augmenting financial reports with contextual information. In certain cases, such “data fusion” may result in the inference of the new knowledge, not explicitly stated before. Therefore the value of FLORA is higher that only the sum of its data.

For the data annotation, established financial ontologies are analyzed in order to be reused. Due to the broad scope of financial aspects covered by FLORA an upper ontology is considered to bridge different domains in the financial area. Further step is to provide necessary mappings to the concepts from to the core LOD dataset, DBpedia [21]. On top of the LOD stack, FLORA will provide services for data exploration and analysis, tailored to the financial domain.

7 Conclusions and Future Work

This article presents a complex, unified process of transforming unstructured financial data into an interlinked, navigable knowledge base for financial information management and information discovery. We described the system that is using ontology-based information extraction and data annotation for extracting relevant data that is further interlinked with appropriate LOD datasets. The whole underlying data infrastructure serves as a basis for providing services for data exploration, querying, discovery and visualizations.

In the broader sense this work will facilitate the financial data reuse and integration by following established data publication techniques. As presented in this paper, Linked Data-based approach adopted by FLORA is slowly becoming de-facto standard for structured data publication.

In the future work we aim at implementing envisaged system architecture and evaluate results based on such metrics as data quality and accuracy of the information extraction. We are also working on developing market surveillance and financial accounting use cases based on the FLORA results.

8 Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the project FLORA (TIN2011-27405).

References

1. Ciccotelloa, S.C. & Wood, R.E. An investigation of the consistency of financial advice offered by web-based sources, Information Systems, 10(1-4), pp. 5-18, 2001.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22

3. Belleau, François; Nolin, Marc-Alexandre; Tourigny, Nicole; Rigault, Philippe & Morissette, Jean: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics, Vol. 41 , Nr. 5 (2008) , S. 706-716.
4. O'Riain, S., Harth, A., & Curry, E.: Linked Data Driven Information Systems as an Enabler for Integrating Financial Data. (A. Yap, Ed.) Information Systems for Global, 239-270. 2011, IGI Global.
5. Momtchev, V., Peychev, D., Primov, T., & Georgiev, G. (2009). Expanding the Pathway and Interaction Knowledge in Linked Life Data. International Semantic Web Challenge, 2009.
6. David F. Huynh, David R. Karger, and Robert C. Miller. 2007. Exhibit: lightweight structured data publishing. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 737-746.
7. Oren, E. Delbru, R. Decker, S. Extending Faceted Navigation for RDF Data. Proceedings of the International Semantic Web Conference (ISWC06). Athens, Georgia. 2006.
8. Declerck, T., Krieger H. U., Thomas S. M., Buitelaar P., O'Riain S., Wunner T., Maguet G., McCrae J., Spohr D., & Montiel-Ponsoda E. Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe, Internal Financial Control Assessment Applying Multilingual Ontology Framework, Chaper 4, 67-76, 2010.
9. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. Challenges for the multilingual Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web, 11(2), 63-71, 2012.
10. Grčar, M., Häusser, T., & Ressel, D. FIRST-Large scale information extraction and integration infrastructure for supporting financial decision making. September 201.
11. Grobelnik, M. Fact Sheet: XLike - Cross-lingual Knowledge Extraction. January 2012
12. Michael Hausenblas LOD2 Creating Knowledge out of Interlinked Data http://lod2.eu/WikiArticle/Project.html Retrieved February 20-2012
13. Klyne, G., & Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, Available at: http://www.w3.org/TR/rdf-concepts
14. Hausenblas, M.: Project Fact Sheet - The LOD Around-The-Clock (LATC) (2011).
15. Grosof, B. Opportunities for Semantic Web knowledge representation to help XBRL, Workshop on Improving Access to Financial Data on the Web. XBRL International and World Wide Web Consortium (W3C), 2009.
16. Roger Debreceny, Glen L. Gray, The production and use of semantically rich accounting reports on the Internet: XML and XBRL, International Journal of Accounting Information Systems, Volume 2, Issue 1, January 2001, Pages 47-74.
17. H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.
18. Axel-Cyrille Ngonga Ngomo and Sören Auer: LIMES – A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. Proceedings of IJCAI 2011.
19. Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov: Silk – A Link Discovery Framework for the Web of Data . 2nd Workshop about Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009.
20. Heath T., Bizer C., Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool.
21. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Zachary Ives: DBpedia: A Nucleus for a Web of Open Data, 6th Int'l Semantic Web Conference, Busan, Korea, 2007

Semantic-Based Sentiment analysis in financial news

Juana María Ruiz-Martínez[1], Rafael Valencia-García[1], Francisco García-Sánchez[1]

[1]Facultad de Informática. Universidad de Murcia.
Campus de Espinardo. 30100 Espinardo (Murcia). España
Tel: +34 86888 8522, Fax: +34 86888 4151
{jmruymar, valencia, frgarcia}@um.es

Abstract. Sentiment analysis deals with the computational treatment of opinions expressed in written texts. The addition of the already mature semantic technologies to this field has proven to increase the results accuracy. In this work, a semantically-enhanced methodology for the annotation of sentiment polarity in financial news is presented. The proposed methodology is based on an algorithm that combines several gazetteer lists and leverages an existing financial ontology. The financial-related news are obtained from RSS feeds and then automatically annotated with positive or negative markers. The outcome of the process is a set of news organized by their degree of positivity and negativity.

Keywords: opinion mining, sentiment analysis, financial news, ontologies, semantic web.

1 Introduction

The success of Web 2.0 technologies along with the growth of social content available online have stimulated and generated many opportunities for understanding the opinions and trends, not only of the general public and consumers, but also of companies, banks, and politics. Many business-related research questions can be answered by analyzing the news and, for this reason, sentiment analysis and opinion mining is a burning issue, specifically in the financial domain.

Opinion mining, a subdiscipline within data mining and computational linguistics, refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content. Sentiment analysis is often used in opinion mining to identify sentiment, affect, subjectivity, and other emotional states in online texts [1].

Originally, the task of sentiment analysis was performed on product reviews by processing the products' attributes [2-4]. However, nowadays sentiment polarity analysis is used in a wide range of domains such as for example the financial domain [5-7]. Millions of financial news are circulating daily on the Web and financial markets are continuously changing and growing. In this scenario, as Ahmad et al. [5]

point out, the creation of a framework with which sentiments can be extracted without relying on the intuition of the analysts as to what is good or bad news is both a necessity and a challenge.

In this paper, we present a semantic-based algorithm for opinion extraction applied to the financial domain. The proposed methodology is supported by natural language processing methods to annotate financial news in accordance with a financial ontology. Then, the annotated financial news are analyzed by passing them through a number of gazetteer lists, which results in two separate sets, one with positive financial news and the other with negative financial news.

The rest of paper is organized as follows. Some relevant related works are shown in Section2. Section 3 presents the technological background necessary for the development of the methodology. In Section 4, the platform and the way it works is described in detail. In Section 5, the experimental results of the evaluation are shown. Finally, some conclusions and future work are put forward in Section 6.

2 Related works

In the literature, a number of methods for the automatic sentiment analysis from financial news streams have been described. The proposal of [6] uses theories of lexical cohesion in order to create a computable metric to identify the sentiment polarity of financial news texts. This metric is readapted in [5] to Chinese and Arabic financial news. The analysis of financial news is a particularly relevant topic in the prediction of the behaviour of stock markets. For example, in [7] the authors use some simple computational linguistic techniques, such as bag of words or named entities, together with support vector machine and machine learning techniques to assist in making stock market predictions. In fact, in real life, stock market analysts' predictions are usually based on the opinions expressed in the news.

Semantic technologies have been around for a while, offering a wide range of benefits in the knowledge management field. They have revolutionized the way that systems integrate and share data, enabling computational agents to reason about information and infer new knowledge [8]. The accuracy results of opinion mining and sentiment polarity analysis can be improved with the addition of semantic techniques, as shown in [9]. In that work, some semantic lexicons are created in order to identify sentiment words in blog and news corpora. Then, a polarity value is attached to each word in the lexicon and such polarity is revised when a modifier appears in the text.

The FIRST project[1] provides an information extraction, information integration and decision making infrastructure for information management in the financial domain. The decision making infrastructure includes a module responsible for the sentiment annotation from financial news and blog posts. Its main aim is to classify the polarity of sentiment with respect to a sentiment object of interest [10]. These sentiment objects are classified by means of an ontology-guided and rule-based information extraction approach. Even though the ontology contains the financial-domain related relevant objects, the classification process is carried out entirely using

[1] http://project-first.eu/

JAPE rules. Therefore, it can be concluded that this approach does not leverage the reasoning capabilities of the ontology.

3 Technological background

The methodology proposed here is based on two main elements, namely, ontologies and natural language processing tools. In this section, the key features of these technologies are pointed out.

3.1 Ontologies and the Semantic Web

Ontologies constitute the standard knowledge representation mechanism for the Semantic Web [8]. The formal semantics underlying ontology languages enables the automatic processing of the information and allows the use of semantic reasoners to infer new knowledge. In this work, an ontology is seen as “a formal and explicit specification of a shared conceptualization” [8]. Ontologies provide a formal, structured knowledge representation, and have the advantage of being reusable and shareable. They also provide a common vocabulary for a domain and define, with different levels of formality, the meaning of the terms and the relations between them. Knowledge in ontologies is mainly formalized using five kinds of components: classes, relations, functions, axioms and instances [11].

Ontologies are thus the key for the success of the Semantic Web vision. The use of ontologies can overcome the limitations of traditional natural language processing methods and they are also relevant in the scope of the mechanisms related, for instance, with Information Retrieval [12], Semantic Search [13], Service Discovery [14] or Question Answering [15].

Next, the financial ontology that has been developed for the purposes of this work is described.

3.1.1 Financial Ontology

The financial domain is becoming a knowledge intensive domain, where a huge number of businesses and companies hinge on, with a tremendous economic impact in our society. Consequently, there is a need for more accurate and powerful strategies for storing data and knowledge in the financial domain. In the last few years, several finances-related ontologies have been developed. The BORO (Business Object Reference Ontology) ontology is intended to be suitable as a basis for facilitating, among other things, the semantic interoperability of enterprises' operational systems [16]. On the other hand, the TOVE ontology (Toronto Virtual Enterprise) [17], developed by the Enterprise Integration Laboratory from the Toronto University, describes a standard organization company as their processes. A further example is the financial ontology developed by the DIP (Data Information and Process Integration) consortium, which is mainly focused on describing semantic web services in the stock market domain [18]. Finally, the XBRL Ontology Specification Group, developed a set of ontologies for describing financial and economical data in RDF for sharing and interchanging data. This ontology is becoming an open standard means of electronically communicating information among businesses, banks, and regulators [19].

As part of this work, a financial ontology has been developed on the basis of the above referred ontologies, with the focus set on the stock exchange domain. The ontology, created from scratch, has been defined in OWL 2. This ontology covers three main financial concepts (see figure 1):

- A financial market is a mechanism that allows people to easily buy and sell financial assets such us stocks, commodities and currencies, among others. The main stock markets such as New York Stock Exchange, NASDAQ or London Stock Exchange have been modelled in the ontology as subclasses of the Stock_market class.
- The Financial Intermediary class represents the entities that typically invest on the financial markets. Examples of such entities are banks, insurance companies, brokers and financial advisers.
- The Asset class represents everything of value on which an Intermediary can invest, such as stock market indexes, commodities, companies, currencies, to mention a few. So, for instance, enterprises such as Apple Inc., General Electric or Microsoft belong to the Company concept and currencies such as US dollar or Euro are included as individuals of the Currency concept.

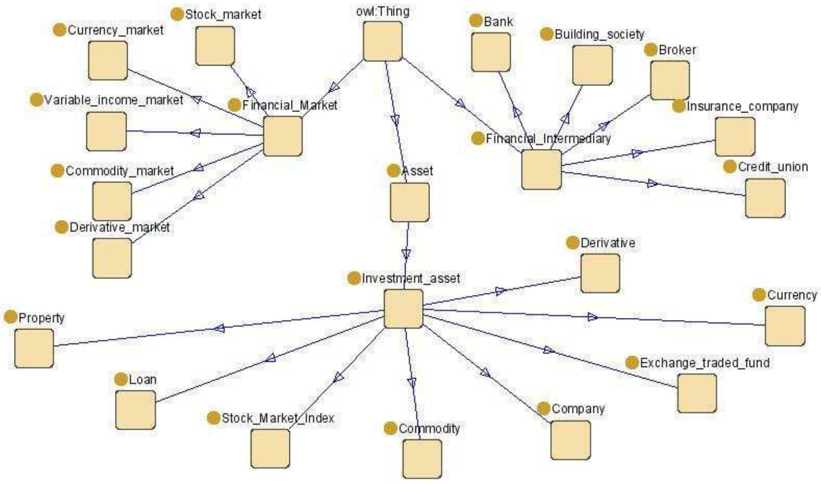


Figure 1. An excerpt of the financial ontology

3.2 Natural Language Processing and Sentiment Analysis

Sentiment annotation can be seen as the task of assign positive, negative or neutral sentiment values to texts, sentences, and other linguistic units [20]. In this work, the values positive, negative and neutral have been assigned to general terms, which express some kind of sentiment (e.g. '*benefit*', '*positive*', '*danger*') and to financial terms (e.g. '*risk capital*', '*rising stock*', '*bankruptcy*'). Moreover, terms pertaining to the financial domain have been semantically annotated as '*risk premium*', '*capital market*' or '*Ibex35*' for example.

The open source software GATE[2] carries out sentiment and semantic annotation by means of gazetteers lists. GATE is an infrastructure for developing and deploying software components that process human language. One of the GATE's key components is gazetteer lists. A gazetteer list is a plain text file with one entry (a term, a number a name, etc.), which permits to identify these entries in the text. In this work, the lists have been developed using BWP Gazetteer[3]. This plugin provides an approximate gazetteer for GATE, based on Levenshtein's Edit Distance for strings. Its goal is to handle texts with noise and errors, in which GATE's default gazetteers may have difficulties. The implemented lists are based on the linguistic particularities of the financial domain.

Grishan and Kittredge [21] define a sublanguage as the specialized form of a natural language that is used within a particular domain or subject matter. A sublanguage is characterized by a specialized vocabulary, semantic relationships, and in many cases specialized syntax [22]. The boundaries of financial news domain are non very sharply defined [22]. For example, "*Euribor rates rise after ECB interest warnings*" or "*Portugal needs the luck of Irish*" are both headline of financial news, although the second one does not contain any financial term or a particular syntactic structure. Nevertheless, it is possible to define a wide set of financial specialized vocabulary (e.g. '*Euribor*', '*Ibex35*', '*investors*') which coexists with frequently used non-specialized terms (e.g. '*to rise*', '*unemployed*', '*construction*').

In this work, the semantic and sentiment gazetteers developed are employed to mark up all sentiment words and associated entities in our ontology. Six different kinds of gazetteers have been developed on the basis of the common characteristics and vocabulary of financial domain. The lists are used by the system in order to create three different types of annotations, that is, semantic annotations, sentiment annotations and modifier annotations. Semantic annotation refers to financial terms that are present in the financial ontology. Sentiment annotation indicates the polarity of selected terms. Modifiers annotation refers to elements that can invert or increase the polarity of the previously annotated terms. For each kind of annotation a gazetteer category has been created. Thus, semantic, sentiment and modifiers gazetteers have been developed. Each gazetteer category consists of one or more gazetteer lists, as explained below.

i. Semantic gazetteer

[2] http://gate.ac.uk/

[3] http://gate.ac.uk/gate/doc/plugins.html#bwp

a. Financial domain vocabulary gazetteer. This gazetteer contains the most relevant domain terms and entities. It has been directly mapped onto the ontology classes and individuals and their corresponding labels including synonyms. Examples in this category are '*Annual Percentage Rate*' (APR), '*Compound Interest*', '*Dividend*', '*Income Tax*', '*Apple*' and '*BBVA*'. This list is used for the semantic annotation and it does not contain any information related with opinions.

ii. Sentiment gazetteer

a. Positive sentiment gazetteer. It contains general terms that imply a positive opinion such as, for example, '*growth*', '*trust*', '*positive*' or '*rising*'.
b. Negative sentiment gazetteer. It contains general terms that imply a negative opinion such as, for example, '*danger*', '*doubts*' or '*to cut*'.
c. Financial positive sentiment gazetteer. It contains terms related to the financial domain that imply a positive opinion. For example, '*earning*', '*profitability*' or '*appreciating asset*'.
d. Financial negative sentiment gazetteer. It contains terms related to financial domain that imply a negative opinion. For example, '*depreciation*', '*Insufficient Funds*' or '*creditor*'.

iii. Modifier gazetteer

a. Intensifier gazetteer. It contains terms that are used to change the degree to which a term is positive or negative such as, for example, '*very*', '*most*' or '*extremely*'.
b. Negation gazetteer. It contains negation expressions such as, for example, '*no*', '*never*' or '*deny*'.
c. Temporal sentiment gazetteers. They contain temporal expressions that imply a modification in the whole news. These expressions appear in conjunction with positive or negative linguistic expressions modifying their meaning. They usually increase or decrease negative or positive sentiment. There are two temporal gazetteers, one with long-term expressions and the other with short-term expressions. "*Last year*", "*trimester*" or "*several weeks*" are examples of the first type, while "*this morning*", "*today*" "*this week*" are examples of the second type. The following sentences show an example of the modification capacity of temporal terms in the financial domain:
(1) Apple shares have risen around 17% in the last month.
(2) Apple shares have fallen 4.5% this morning.
Here, "*last month*" and "*this morning*" can relativize the weight of the global meaning. In general, long-term positive or negative opinions are more reliable than short-term opinions. That is, if the user searches for the general status of Apple shares and the system retrieves these two entries, then the general opinion should be positive.

4 Platform Architecture

The architecture of the platform is shown in figure 2. The architecture is composed of four main components: the financial news extraction module, the semantic annotation module, the opinion-mining module and the search engine. Next, these components are described in detail.

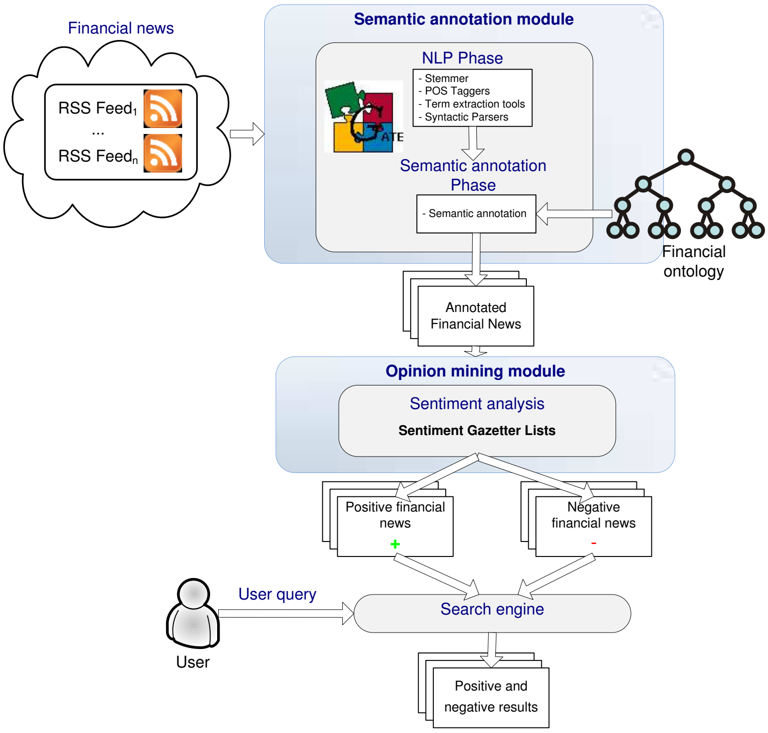


Figure 2. Architecture of the system.

4.1 Financial news extraction module

This module manages the list of RSS feeds. RSS is a family of Web feed formats used for syndicating content from blogs or Web pages and is commonly used by newspapers. RSS is an XML file that summarizes information items and links to the information sources [23]. Once the resources have been selected, this module generates a set of abstracts, which will be used as input for the system. An example list of financial news-related RSS feeds is shown in table 1.

Table 1. Example of RSS feeds

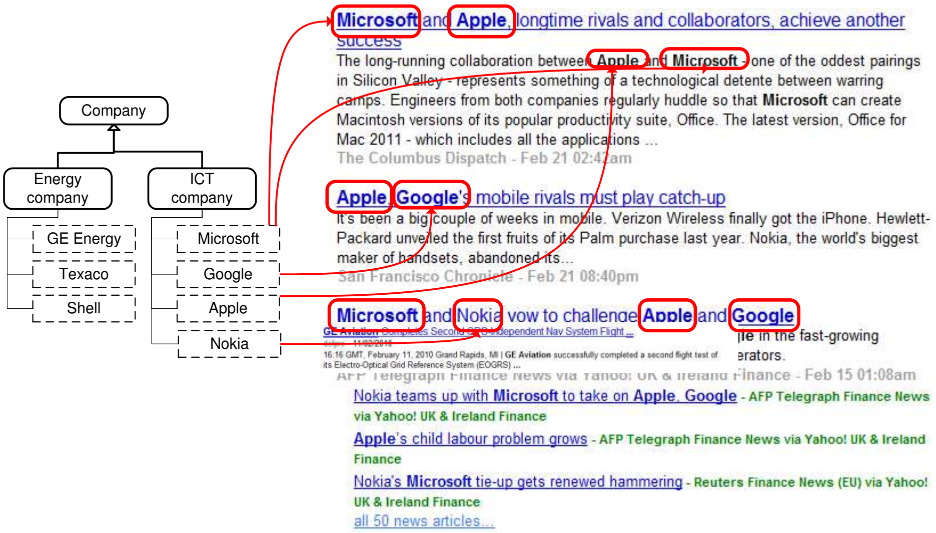
http://www.economist.com/feeds/print-sections/75/europe.xml
http://feeds.reuters.com/reuters/USpersonalfinanceNews
http://feeds.nytimes.com/nyt/rss/Business
http://feeds.bbci.co.uk/news/business/rss.xml

For each RSS source the last news are obtained and stored in a database. The information that is retrieved from each news is the date of publication, the information source, the url and the abstract. Abstracts constitute the corpus from which the system extracts the information. We only consider the abstract and the headline because they usually condense the polarity of news. Indeed, the analysis of the whole text can induce to error, since the sentiment polarity of an entire document is not necessarily the sum of its parts.

4.2 Semantic annotation module

This module identifies the most important linguistic expressions in the financial domain using the previously described semantic gazetteer. For each linguistic expression, the system tries to determine whether the expression under question is an individual of any of the classes of the domain ontology. Next, the system retrieves all the annotated knowledge that is situated next to the current linguistic expression in the text, and tries to create fully-filled annotations with this knowledge.

Each class in the ontology is defined by means of a set of relations and datatype properties. Then, when an annotated term is mapped onto an ontological individual, its datatype and relationships constitute the potential information which is possible to obtain for that individual. For example, a company has associate relationships such as '*Moody'sRate*', '*tradeMarket*' or '*isLegalRepresentativeFor*'. In figure 3, an example of the annotation process of financial news using GATE is depicted.



Microsoft and Apple, longtime rivals and collaborators, achieve another success
The long-running collaboration between Apple and Microsoft - one of the oddest pairings in Silicon Valley - represents something of a technological detente between warring camps. Engineers from both companies regularly huddle so that Microsoft can create Macintosh versions of its popular productivity suite, Office. The latest version, Office for Mac 2011 - which includes all the applications ...
The Columbus Dispatch - Feb 21 02:42am

Apple, Google's mobile rivals must play catch-up
It's been a big couple of weeks in mobile. Verizon Wireless finally got the iPhone. Hewlett-Packard unveiled the first fruits of its Palm purchase last year. Nokia, the world's biggest maker of handsets, abandoned its...
San Francisco Chronicle - Feb 21 08:40pm

Microsoft and Nokia vow to challenge Apple and Google
GE Aviation Completes Second GPS-Independent Nav System Flight ...
16:16 GMT, February 11, 2010 Grand Rapids, MI | GE Aviation successfully completed a second flight test of its Electro-Optical Grid Reference System (EOGRS) ...
le in the fast-growing
erators.
AFP Telegraph Finance News via Yahoo! UK & Ireland Finance - Feb 15 01:08am

Nokia teams up with Microsoft to take on Apple, Google - AFP Telegraph Finance News via Yahoo! UK & Ireland Finance

Apple's child labour problem grows - AFP Telegraph Finance News via Yahoo! UK & Ireland Finance

Nokia's Microsoft tie-up gets renewed hammering - Reuters Finance News (EU) via Yahoo! UK & Ireland Finance

all 50 news articles...

Figure 3. Example of knowledge entities identified in financial news.

4.3 Opinion mining module

The main objective of this module is to classify the set of news obtained in the previous module according to their polarity: positive, negative or neutral. For any retrieved news which has been annotated, the sentiment orientation or sentiment polarity value is computed. For this, the module makes use of the previously described gazetteer lists.

The sentiment polarity (SP) value for each news item is calculated by summing the polarity values of all annotated terms in the news. In this process, the system must consider both the terms polarity included in the positive and negative gazetteers and the contextual valence shifters included in the negation and intensifier gazetteers.

For any annotated term (***at***) in a sentence s∈S, its SP value (**SP(at)**) is computed as follows:

1. If *at*∈ GeneralPositive^k, SP(at) = Positive1
2. If *at*∈ DomainPositive^k, SP(at) = Positive2
3. If *at*∈ GeneralNegative^k, SP(at) = Negative1
4. If *at*∈ DomainNegative^k, SP(at) = Negative2
5. If within the relevant cotext of *at*, there is a term *at'*∈Negation, SP(at)= -SP(at)
6. If within the relevant cotext of *at*, there is a term at'∈Intensifier, SP(at) = 2xSP(at)
7. When within the relevant cotext of *at*, there is a term at'∈Temporal, if…
    7.1. at'∈LongTerm, SP(at) = 2xSP(at)

7.2. at'∈ShortTerm + Negative(SP), SP(at) = 2xSP(at)
7.3. at'∈ShortTerm + Positive(SP), SP(at) = 1xSP(at)

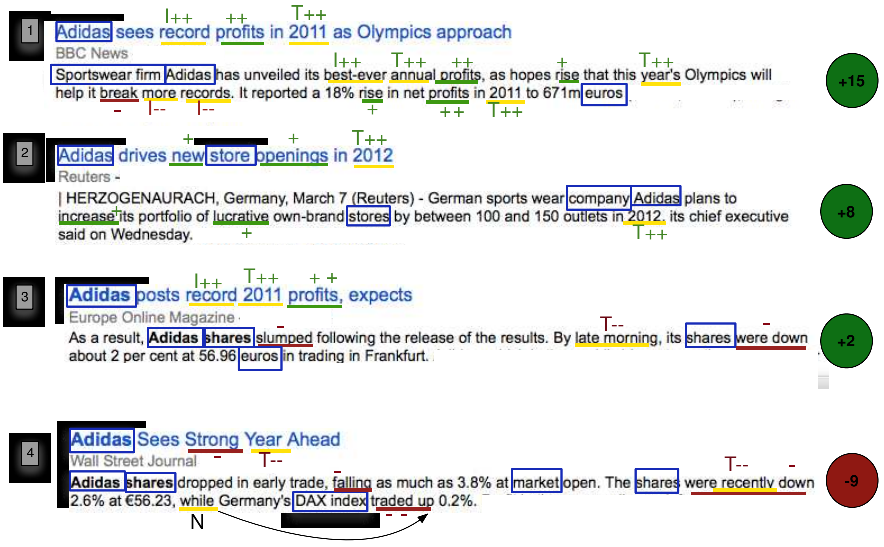
Then the polarity of each news item is represented as the sum of all SP(at) present in such news item (n):

$$f^{k}SP(n)^{k} = \sum_{at \in n} SP(at)$$

In the above algorithm, the term '*cotext*' refers to the linguistic set that surrounds an annotated term within the limit of a sentence, i.e. the rest of annotated terms present before and after it and pertaining to the same sentence. '*Positive1*' and '*Positive2*' refer to the degree of positivity of an annotated term, while '*Negative1*' and '*Negative2*' refer to the degree of negativity of an annotated term.

When a long-term temporal expression is found, its value is calculated taking into account the *at* pertaining to its cotext. If a positive *at* is found, then its value is 2. On the contrary, if a negative *at* is found its value is -2. Sort- term temporal expressions are calculated in the same way for negative value, i.e adding -2. However, for positive value the system only adds 1positive. This is because we consider that financial short-term positive values change too frequently to consider them at the same level as long-term values.

Next, if the semantic polarity value of a news is less than 0, the news is labelled as negative. In contrast, if the value is higher than 0, the news is labelled as positive. Finally, if the sum of all values is 0 the news is labelled as neutral. An example of how the algorithm works is shown in figure 4.



1 Adidas sees record profits in 2011 as Olympics approach
BBC News
Sportswear firm Adidas has unveiled its best-ever annual profits, as hopes rise that this year's Olympics will help it break more records. It reported a 18% rise in net profits in 2011 to 671m euros.

2 Adidas drives new store openings in 2012
Reuters -
| HERZOGENAURACH, Germany, March 7 (Reuters) - German sports wear company Adidas plans to increase its portfolio of lucrative own-brand stores by between 100 and 150 outlets in 2012, its chief executive said on Wednesday.

3 Adidas posts record 2011 profits, expects
Europe Online Magazine
As a result, Adidas shares slumped following the release of the results. By late morning, its shares were down about 2 per cent at 56.96 euros in trading in Frankfurt.

4 Adidas Sees Strong Year Ahead
Wall Street Journal
Adidas shares dropped in early trade, falling as much as 3.8% at market open. The shares were recently down 2.6% at €56.23, while Germany's DAX index traded up 0.2%.

Figure 4. Semantic Polarity annotation example

Let us suppose that a user searches for the company '*Adidas*'. In the example depicted in figure 4, four different news items are retrieved. In the figure, semantic annotations are the elements surrounded by a rectangle, which have been mapped onto ontology instances. GeneralPositive are indicated with one '+' sign and DomainPositive with two, '++'. On the other hand, GeneralNegative are indicated with one '–' sign and DomainNegative with two, '--'. The modifiers Negative, Temporal and Intensifier are indicated with 'N', 'T', 'I' respectively, together with the corresponding positive or negative symbol.

The outcome of the process is three positive and one negative news items. In this particular example, the presence of long-term temporal expressions, such as '*2012*' or '*year*', in conjunction with positive annotated terms, gives to the news a high positive value. The user can organize the final results in accordance with their degree of positivity and negativity.

4.4 Semantic search engine

In OWL-based ontologies, '*rdfs:label*' is an instance of '*rdf:property*' that may be used to provide a human readable version of a resource name. In this work, all the resources in the ontology have been annotated with the '*rdfs:label*' descriptor. By considering that, the main objective of this module is to identify the financial news items that are related to the query issued by a user. Besides, this module is responsible for classifying and sorting the results in accordance with the sentiment classification that was described in the previous section.

The system is constantly crawling news information from RSS feeds and creating semantic annotations for the news pages. If no annotations are created for a news item, then such news item is not stored in the database. On the other hand, the news items that have been successfully annotated are processed to obtain their sentiment classification, which is also stored in the database. For example, let us suppose that the ontology contains the taxonomy presented in figure 3. There are two kinds of companies, namely, "Energy company" and "ICT company". Each of these classes contains a set of individuals such as "Microsoft" and "GE energy", respectively. If the user is searching for news about "Microsoft", the system will certainly return all the news annotated with the individual Microsoft. Moreover, news related to other ICT companies could be relevant to the user, so the system also shows other news about companies such as Google, Apple and Nokia. If the user queries the system for "Energy companies", then the result will include all the news that contains the concept "Energy company" and therefore the news related to the "GE Energy", "Texaco" and "Shell" companies will be retrieved. Furthermore, if the query is such a general word as "Company", the user is given the possibility of filtering the results according to the subclasses of "Company", namely, "Energy company" and "ICT company".

5 Evaluation

In this section, the experimental results obtained by the proposed method in the financial news domain are presented. The corpus of the experiment contains 57.210

words and comprises 900 abstracts of financial news (512 negative and 388 positive). This corpus has been extracted from the RSS feeds shown in table 1 and each news item has been manually labelled, either as a positive news or a negative one, by two different annotators. This constitutes the baseline for the evaluation, which works as follows: if the result displayed by the system fits in with the manually annotated news, the result is considered correct, otherwise, incorrect. In the sentiment analysis field, it is agreed that human-based annotations are around 70-80% precise (i.e. 2 different humans can disagree in 20-30% of cases). However, for the purposes of this experiment, the news items that have been source of disagreement between annotators have been removed.

In the experiment, a total of five queries are issued to the system to find information in the financial domain. The results of the experiment are shown in table 2. It is possible to observe that the sentimental analysis accuracy results are very promising, with an aggregate accuracy mean of 87%. These results take into account the system's final decision (positive or negative) and not the process that the system carries out to produce such decision.

Table 2. Hits results in information retrieval.

Query		Baseline	Our approach	Accuracy
1	Pos	33	28	84.85%
	Neg	11	9	81.82%
	Total	44	37	84.09%
2	Pos	13	13	100%
	Neg	36	34	94.44%
	Total	49	47	95.92%
3	Pos	15	14	93.33%
	Neg	29	24	82.76%
	Total	44	38	86.36%
4	Pos	25	21	84%
	Neg	97	86	88.66%
	Total	122	107	87,70%
5	Pos	66	55	83.33%
	Neg	14	12	85.71%
	Total	80	67	83.75%
Total		678	592	87.32%

6. Conclusions

This paper proposes an algorithm for opinion extraction in financial news. Different gazetteer lists have been created as specialized lexicons in financial sentiment. The sentiment algorithm assigns different degrees of positivity or negativity to relevant annotated terms and calculates what the polarity of the news is.

This approach contributes to the research on financial sentiment annotation, and the development of decision support systems (1) by proposing a novel approach for financial sentiment determination in news which combines ontological resources with natural language processing resources, (2) by describing an algorithm for assigning differential degrees of positivity or negativity to classifier results on different categories identified by the classifier, and (3) by proposing a set of resources, i.e. gazetteer lists and an ontology, for sentiment annotation.

Acknowledgements

This work has been supported by the Spanish Government through project SeCloud (TIN2010-18650).

References

1 Chen, H., Zimbra, D.: AI and opinion mining. Intelligent Systems, IEEE. 25(3), pp. 74-80 (2010)

2 Popescu, A.M., Etzioni, O.: In: Extracting product features and opinions from reviews. Proceedings of the conference on human language technology and empirical methods in natural language processing; Association for Computational Linguistics, pp. 339-46 (2005)

3 Ding, X., Liu, B.: The utility of linguistic rules in opinion mining. In Proceedings of 30th Annual International ACM Special Interest Group on Information Retrieval Conference (SIGIR'07), Amsterdam, The Netherlands (2007)

4 Balahur, A., Montoyo, A.: Determining the semantic orientation of opinions of products- a comparative analysis. Procesamiento del lenguaje natural, 41, pp. 201-8 (2008)

5 Ahmad, K., Cheng, D., Almas, Y.: Multi-lingual Sentiment Analysis of Financial News Streams. In: Proceedings of the Second Workshop on Computational Approaches to Arabic Script-based Languages, Linguistic Society of America, Linguistic Institute, Stanford University, pp. 1-12 (2007)

6 Devitt, A., Ahmad, K.: Sentiment analysis in financial news: A cohesionbased approach. In Proceedings of the Association for Computational Linguistics (ACL), pp. 984–991 (2007).

7 Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems 27, pp.1–19 (2009)

8 Studer R, Benjamins V.R., Fensel D.: Knowledge engineering: Principles and methods. Data Knowledge Engineering. 25(1-2), pp.161-97 (1998)

9 Godbole, N., Srinivasaiah, M., Skiena, S.: Largescale sentiment analysis for news and blogs: In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM) (2007)

10 Klein A., Häusser T., Altuntas O., Grcar M., Large scale information extraction and integration infrastructure for supporting financial decision making. Deliverable: D4.1 First semantic information extraction prototype, http://project-first.eu/content/d41-first-semantic-information-extraction-prototype (2012)

11 Gruber TR.: A translation approach to portable ontology specifications. Knowledge Acquisition. 5(2), pp.199-220 (1993)

12 Valencia-García, R. Fernández-Breis, J.T., Ruiz-Martínez, J.M., García-Sánchez, F. and Martínez-Béjar, R.: A knowledge acquisition methodology to ontology construction for information retrieval from medical documents. Expert Systems: The Knowledge Engineering Journal 25(3), pp.314-334 (2008)

13 Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J.T., Camón-Herrero, J.B.: Financial news semantic search engine. Expert systems with applications 38(12) pp. 15565-15572 (2011)

14 García-Sánchez, F., Valencia-García, R., Martínez-Béjar, R., Fernández-Breis, J.T.: An ontology, intelligent agent-based framework for the provision of semantic web services. Expert Systems with Applications 36(2) Part 2, pp.3167–3187 (2009)

15 Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J.T.: OWLPath: an OWL ontology-guided query editor: IEEE Transactions on Systems, Man, Cybernetics: Part A, vol 41(1), pp. 121 – 136 (2011)

16 Partridge C.: The role of ontology in integrating semantically heterogeneous databases. Report No.: LADSEB-CNR Technical Report 05/2002 (2002)

17 Fox, M.S., Gruninger, M.: Enterprise modeling. AI magazine. 19(3):109 (1998)

18 Corcho, O., Losada, S., Martínez Montes, M., Bas, J.L., Bellido, S.: Financial Ontology. DIP deliverable D10.3 (2004)

19 Bonsón, E., Cortijo, V., Escobar, T.: Towards the global adoption of XBRL using international financial reporting standards (IFRS). International Journal of Accounting Information Systems, 10(1), pp. 46-60 (2009)

20 Andreevskaia, A., Bergler, S.: When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. Proceedings of ACL-08: HLT, pp- 290-298 (2008)

21 Grishman R, Kittredge R.: Analyzing language in restricted domains: Sublanguage description and processing. Lawrence Erlbaum, (1986).

22 Grishman R: Adaptive information extraction and sublanguage analysis. In Kushmeric N (ed.) Proceedings of Workshop on Adaptive Text Extraction and Mining at Seventeenth International Joint Conference on Artificial Intelligence. WA: Seattle. http://nlp.cs.nyu.edu/pubs/papers/grishman-ijcai01.pdf, (2001)

23 Murugesan, S.: Understanding web 2.0.: IT professional. 9(4), pp. 34-410, (2007)

Zoè Lacroix
Edna Ruckhaus
Maria-Esther Vidal (Eds.)

RED'12

Fifth International Workshop on REsource Discovery

Workshop co-located with the 9th Extended Semantic Web Conference (ESWC 2012)

Heraklion, Greece, May 27th, 2012

Proceedings

Editors' addresses:
Arizona State University
{zoe.lacroix}@asu.edu
Universidad Simón Bolívar
Department of Computer Science
Valle de Sartenejas
Caracas 1086, Venezuela

{ruckhausl | mvidal}@ldc.usb.ve

Preface

This volume contains abstracts from the technical program of the Fifth International Workshop on REsource Discovery, held on May 27th, 2012. After four successful events, first in Linz, Austria, joined to IIWAS (2008), then in Lyon, France, collocated with VLDB (2009), next in Pontoise, France, joined again to IIWAS (2010), and the fourth edition in conjunction with ESWC11. Finally, the fifth International Workshop on REsource Discovery (RED 2012) was run again together with ESWC in Heraklion, Greece.

A resource may be a data repository, a database management system, a SPARQL endpoint, a link between resources, an entity in a social network, a semantic wiki, or a linked service. Resources are characterized by core information including a name, a description of its functionality, its URLs, and various additional Quality of Service parameters that express its non-functional characteristics. Resource discovery is the process of identifying, locating and selecting existing resources that satisfy specific functional and non-functional requirements; also, resource discovery includes the problem of predicting links between resources. Current research includes crawling, indexing, ranking, clustering, and rewriting techniques, for collecting and consuming the resources for a specific request; additionally, processing techniques are required to ensure an efficient and effective access of the resources.

The Fifth International Workshop on Resource Discovery aimed at bringing together researchers from the database, artificial intelligence and semantic web areas, to discuss research issues and experiences in developing and deploying concepts, techniques and applications that address various issues related to resource discovery. This fifth edition focused on techniques to efficiently collect and consume resources that are semantically described. Approaches of special interest contribute to solve the resource discovery problem such as query rewriting in Databases, service selection and composition in Service Oriented Architectures, social network navigational techniques, link prediction techniques, and strategies to process queries against Linked Data or SPARQL endpoints.

We received seven submissions, out of which we selected five for inclusion in the digital and printed proceedings. We set up an exciting program which included three invited talks. The first on Semantic Source Modeling given by our invited speaker, José Luis Ambite; the second, on the advantages of using semantic annotations in medical image visualization given by Alexandra La Cruz; finally, Edna Ruckhaus presented Probabilistic Models and Reasoning Techniques to Detect Inconsistencies in Linked Data. We organized two sessions, one on Techniques for Resource Discovery and another section on Applications of Resource Discovery.

We thank the 25 members of our Program Committee, the invited speakers and the authors for their valuable contribution to the workshop. We are also grateful to ESWC organizers for their support in making this meeting successful. We kindly acknowledge the National Science Foundation (grant IIS 0944126), and the DID-USB.

May 2012 Zoé Lacroix, Edna Ruckhaus, Maria-Esther Vidal

Workshop Chairs and Organizing Committee

Zoè Lacroix, Arizona State University, USA
Edna Ruckhaus, Universidad Simón Bolívar, Venezuela
Maria-Esther Vidal, Universidad Simón Bolívar

Program Committee

Maribel Acosta, AIFB, Karlsruhe Institute of Technology, Germany.
José Luis Ambite, University Southern California, USA.
Yudith Cardinale, Universidad Simón Bolívar, Venezuela.
Oscar Corcho, Universidad Politecnica de Madrid, Spain.
Jose Cordeiro, Polytechnic Institute of Setubal, Portugal.
Valeria De Antonelis, Universita degli Studi di Brescia, Italy.
Alberto Fernandez, Universidad Juan Carlos I, Spain.
Norbert Fuhr, University of Duisburg, Germany.
Manolis Gergatsoulis, Ionian University, Greece.
Marlene Goncalves, Universidad Simón Bolívar, Venezuela.
Andreas Harth, AIFB, Karlsruhe Institute of Technology, Germany.
H.V. Jagadish, University of Michigan, USA.
Nikos Kiourtis, National Technical University of Athens, Greece.
Birgitta Koning-Ries University oj Jena, Germany.
Gunter Ladwig, AIFB, Karlsruhe Institute of Technology, Germany.
Maria Maleshkova, KMI, The Open University, United Kingdom.
Anja Metzner, University of Applied Scinces, Augsburg, Germany.
Pascal Moli, Nantes University, LINA, France.
Fatiha Sais, LRI (Paris-Sud 11 University & CNRS), France.
Sherif Sakr, National ICT Australia (NICTA) and University of New South Wales (UNSW), Australia.
Miguel-Angel Sicilia, University of Alcala, Spain.
Hala Skaf-Moli, Nantes University, LINA, France.
Dimitrios Skotas, University of Hannover, Germany.
Andreas Thor, Universitat Leipzig, Germany.
Maciej Zaremba, DERI and National University of Ireland, Ireland.

Table of Contents

SDDS based Hierarchical DHT Systems for an Efficient Resource Discovery in Data Grid Systems

Riad Mokadem, Abdelkader Hameurlain, Franck Morvan

Institut de Recherche en Informatique de Toulouse (IRIT)
118, route de Narbonne, Toulouse, France
{mokadem, hameur, morvan}@irit.fr

Abstract. Despite hierarchical Distributed Hash Table (DHT) systems have addressed flat overlay system problems, most of the existing solutions add a significant overhead to large scale systems. In this paper, we propose a hierarchical DHT solution based on scalable distributed data structures (SDDS) for an efficient data sources discovery in data Grids. Our solution deals with a reduced number of gateway peers running a DHT protocol. Each of them serves also as a proxy for second level peers in a single Virtual Organization (VO), structured as an SDDS. The proposed solution offers good performances especially for intra-VO resource discovery queries since they are completely transparent to the top level DHT lookups. The analysis results proved significant system maintenance save especially when nodes join/ leave the system.

Keywords: Resource discovery, Data Grid, Peer to peer system, Distributed hash table, Scalable distributed data structure, Super peer models.

1 Introduction

A resource discovery consists to discover resources (e.g., computers, data) that are needed to perform distributed applications in large scale environments [21]. It constitutes an important step in a query evaluation in such environments. Throughout this paper, we focus on the discovery of metadata describing data sources in data Grid systems.

Several research works have adopted the Peer-to-Peer solutions to deal with resource discovery in Grid systems [19] and [26]. P2P routing algorithms have been classified as structured or unstructured [27]. Although the good fault tolerance properties in P2P unstructured systems (e.g., KaZaa [13]), the flooding –used in each search- is not scalable since it generates large volume of unnecessary traffic in the network. Structured Peer-to-Peer systems as DHT are self-organizing distributed systems designed to support efficient and scalable lookups in spite of the dynamic properties in such systems. Classical flat DHT systems organize peers, having the same responsibility, into one overlay network with a lookup performance of O(log(N)), for a system with N peers. However, the using of a flat DHT do not consider neither the autonomy of virtual organizations and their conflicting interests nor the locality principle, a crucial consideration in Grids [10]. Moreover, typical structured P2P systems as Chord [25] and Pastry [24] suffer not only from temporary unavailability of some of its components but also from churn. It occurs in the case of the continuous leaving and entering of nodes into the system. Recent research works as [21] proved that hierarchical overlays have the advantages of faster lookup times, less messages exchanged between nodes, and scalability. They are valuable for small and medium sized Grids, while the super peer model is more effective in very large Grids [30]. In this context, several research works [5], [6], [12], [17], [18], [20] and [31] proved that hierarchical DHT systems based on the super peer concept can be advantageous for complex systems. A hierarchical DHT employ a multi level overlay network where peers are grouped according to a common property such as resource type or locality for a lookup service used in discovery [5]. In this context, a Grid can be viewed as a network composed of several, proprietary Grids, virtual organizations (VO) [18] where every VO is dedicated to an application domain (e.g., biology, pathology). Within a group, one or more peers are selected as super peers to act as gateways to peers in the other groups. Furthermore, most existing hierarchical DHT solutions neglect the churn effect and deal only with the improving performance of the overlay network routing. They mainly generate significant additional overhead to large scale systems. Several proposals for reducing maintenance costs, have also appeared in the literature [7], [9], [14], [16], [23] and [32]. Despite a good strategy to manage a churn in [14] through a lazy update of the network access points, inter-organizations lookups were expensive because of the complex addressing system. [16] proposed the SG-1 algorithm, based on the information exchange between super peers through a gossip protocol [1], to find the optimal number of super peers in order to reduce maintenance costs. However, most of these solutions add significant load at some peers which generates an additional overhead to large scale systems.

In this paper, we propose a scalable distributed data structure (SDDS) based Hierarchical DHT solution (SDDS- HDHT) for an efficient resource discovery in data Grids. It combines SDDS routing scheme [15] with DHT systems and aims to improve both lookup and maintenance costs while minimizing the overhead added to the system. Our solution consists of a two level hierarchical overlay network dealing with super peers (called also gateways) and second level peers. Gateway peers establish a structured DHT based overlay. Only one peer per VO is considered as a gateway. Then, each of them serves as a proxy for second level peers in a single VO, structured as an SDDS. SDDS were among the first research works dealing with structured P2P systems. [29] noted numerous similarities between Chord and the best known SDDS scheme: LH* (Linear Hashing) [15]. Both implement key search and have no centralized components. Resource discovery queries, in our system, are classified into intra-VO and inter-VO queries. The intra-VO discovery consists to apply the principle of locality by favoring the metadata discovery in a local VO through the efficient LH* routing system. Key based queries in LH*, in its $LH*_{RS}^{P2P}$ versus, need at most two hops to find the target when the key search in a DHT needs O(log N) hops, N is the number of peers in the system [29]. In fact, super peers are not concerned by intra-VO queries unlike previous solutions as [31] which put super peers more under stress. Regarding Inter-VO queries, they are first routed to the reduced DHT overlay which permits to locate the gateway peer affected to the VO containing the resource to discover. Then, another $LH*_{RS}^{P2P}$ lookup is done in order to discover metadata of this resource. The proposed solution takes also into account the continuous leaving and joining of nodes into the system (dynamicity properties of Grid environments). Only the arrival of a new VO requires the DHT maintenance. The connection/ disconnection of gateways do not require excessive messages exchanged between peers in order to maintain the system. This is done through a lazy system update which avoids high maintenance costs [14].

A simulation analysis evaluates performances of the proposed solution through comparison with previous solution performances. It shows the reduction of lookups costs especially for intra-VO queries. It also provides a significantly maintenance costs reduction, especially when peers frequently join/leave the system. The rest of the paper is structured as follows. Section 2 recalls hierarchical DHT and SDDS principles. Section 3 presents our resource discovery solution through the proposed protocol. It also describes the maintenance process. The simulation analysis study section shows the benefit of our proposition. Section 5 details related work. The final section contains concluding remarks and future works.

2. Preliminaries

2.1 Scalable Distributed Data Structure

Scalable Distributed Data Structures (SDDS), designed for P2P applications, are a class of data structures for distributed systems that allow data access by key in constant time [29]. Many variant of SDDS were proposed. In this paper, we deal with LH*$_{RS}$^{P2P} scheme which improves later LH* variants (LH*$_{RS}$, LH*$_{g}$…). We assume that the reader is familiar with a linear hashing algorithm LH* as presented in [15]. Each node stores records in a bucket which splits when the file grows. Every LH* peer node is both client and, potentially, data or parity server which interacts with application using the key based record search, insert, update or delete query or a scan query performing non key operations.

Each record in LH * is identified by its key whivh determines the record location according to the linear hashing Algorithm described in [29]. The file starts with one data bucket and one parity bucket. It scales up through data bucket splits, as the data buckets get overloaded. It can be occurred when a peer splits its data bucket. In old SDDS scheme, one peer acted as a coordinator peer. It was viewed as the single node knowing the correct state of the file or relation. However, [29] ameliorates this scheme. Split coordinator does not constitute a centralized node for the SDDS scheme. It intervenes only to find a new data server when a split occurs and never in the query evaluation process. Any other peer uses its local view ‘image’, which may be not adjusted, to find the location of a record given in the key based query. The peer server applies another algorithm LH*$_{RS}$^{P2P} described in [29]. It first verifies whether its own address is the correct one. If needed, the server forwards this query. The query always reaches the correct bucket in this step. Then, it sends an Image Adjustment Message (IAM) informing the initial sender that the address was incorrect and the sender adjusts its image reusing the LH* image adjustment algorithm described in [29]. Hence, the most important property here is that the maximal number of forwarding messages for key-based addressing is one. Another advantage of using SDDS is the possibility to support range queries very well and the less vulnerability in the presence of high churn [29].

2.2 Principles of Hierarchical Distributed Hash Tables

Structured systems such as DHT offer deterministic query search results within logarithmic bounds as sending message complexity. In systems based on DHT as Chord [25], Pastry [24] and Tapestry [33], the DHT protocol provides an interface to retrieve a key-value pair. Each resource is identified by its key using cryptographic hash functions such SHA-1. Each peer is responsible to manage a small number of peers and maintains its location information. In this paper, we have focused on a Pastry DHT system [24]. But, our method can be applied to other DHT systems. Pastry DHT system offers deterministic query search results within logarithmic bounds. It requires Log_B (N) hops, where N is the total number of peers in the system and B typically equal to 4 (which results in hexadecimal digits). Pastry system also notifies applications of new peers arrivals, peer failures and recoveries. Unlike Chord peers, Pastry peer permits to easily locate both the right ad left neighbors in the DHT. These reasons motivate us to choose the Pastry routing system. Hierarchical DHT systems partition its peers into a multi level overlay network. Because a peer joins a smaller overlay network than in flat overlay, it maintains and corrects a smaller number of routing states than in flat structure. In such systems, one or more peers are often designated as super peers. They act as gateways to other peers organized in groups in second level overlay networks. Throughout this section, we interest to two previous hierarchical DHT solutions which we consider comparable to our solution.

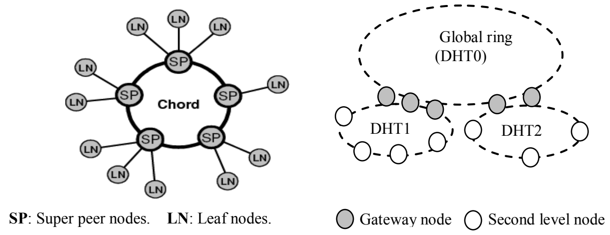


Fig. 1. SP-HDHT (left) and MG-HDHT (right) solutions.

In Fig. 1-left, super peers establish a structured DHT overlay network when second level peers (called leaf nodes) maintain only connection to their super peers. This corresponds to the Super Peer HDHT (SP-HDHT) solution [31]. However, [17] proved that this strategy can maintain super peers more under stress by maintaining pointers between super peers and their leaf nodes. Furthermore, a super peer stores information's of all leaf nodes which it is responsible and acts as a centralized resource for them. Then, performances depend on the ratio between super peer's number and the total number of peers in the system. Multi-Gateway Hierarchical DHT (MG-HDHT) solution [18] is another example of 2-levels hierarchy system having multiple gateways by VO (Fig. 1- right). The system forms a tree of rings (DHTs in this example). Typically, the tree consists of two layers, namely a global ring as the root and organizational rings at the lower level. A group identifier (*gid*) and a unique peer identifier (*pid*) are assigned to each peer. Groups are organized in the top level as DHT overlay network. Within each group, nodes are organized as a second level overlay. This solution provides administrative control and autonomy of the participating organizations. Unlike efficient intra-organization lookups, inter-organization lookups are expensive since the high maintenance cost of the several gateway peers. Hence, there is a trade-off between minimizing total network costs and minimizing the added overhead to the system.

3. Resource Discovery through SDDS based Hierarchical DHT Systems

A resource discovery is a real challenge in unstable and large scale environments. It constitutes an important step in the evaluation of a query in Grid environment [22]. The fact that users have no knowledge of the resources contributed by other participants in the grid poses a significant obstacle to their use. Hence, a centralized scheme forms naturally a bottleneck for the system [20]. The duplicated approach forces the update in every peer which will result in flooding the network. The distributed approach is more appropriate in such systems [19]. In this context, distributed Peer to Peer techniques are used to discover resources in data Grids. Furthermore, Grid environment is likely to scale to millions of resources shared by hundreds of thousand of participants. In consequence, the fact that peers frequently leave/join the system generates high maintenance costs especially on the presence of a churn effect. We have first study a flat DHT resource discovery solution. When one searches a peer responsible for some resource, the typical number of hops in DHT is $O\ (log_B(N_T))$ when N_T is the total number of nodes in the system. However, value of N_T can be a greater number and the maintenance of the DHT will be more complex. More, this solution does not take into account the autonomy of organizations. One solution to this problem is to deal with a super peer model. However, a super peer acts as a centralized resource for a number of peers which depend on the availability of the super peer. Also, a single point of failure of this peer constitutes a serious problem. We have study some previous hierarchical DHT solutions. Existing solution as [31] improves significantly the routing performance. But, complex algorithms are suitable to manage connection between nodes and performances depend on the ratio between super peers and total number of peers.

3.1 Architecture

Instead to adopt one of these solutions, we propose an SDDS based hierarchical DHT solution for resource discovery in data Grids. It aims to reduce both lookup and maintenance costs while minimizing overhead added to the system. Resource Discovery through our solution deals with two different classes of peers: gateways (called also super peers) and second-level peers. A Grid can be viewed as a network composed of several, proprietary Grids, virtual organizations (VO) [11] as shown in Fig. 2. Every VO is dedicated to an application domain (e.g., biology, pathology) [14]. It permits to take into account the locality principle of each VO [10]. Within a VO, one peer is selected as a super peer. It acts as a gateway (or a proxy) for other peers, called second level peers, in the other VOs. Gateways communicate with each other through a DHT overlay network. Each of them knows, through the $LH*_{RS}^{P2P}$ routing system, how to interact with all second level peers belonging to the same VO. In this context, [5] proved that a DHT lookup algorithm required only minor adaptations to deal with groups instead of individual peers. In order to make a resource in VO$_i$ visible through the top level DHT, hash join *H* is applied to this resource, when it joins the system, to generate a group identifier *gid*. Then, an other hash function *h* is applied to this resource in order to generate a peer identifier *pid*. This permits to associate each resource to its VO [17]. We may assume that gateway peers are relatively more stable than second level peers. In contrast, gateways establish a structured DHT based overlay when each VO -regrouping second level peers- is structured as an SDDS. We consider here the peers as homogenous. Recall also that we have not interesting on the assignment of a joining second level peer to an appropriate gateway, i.e., loads balancing. We defer these issues to future work.

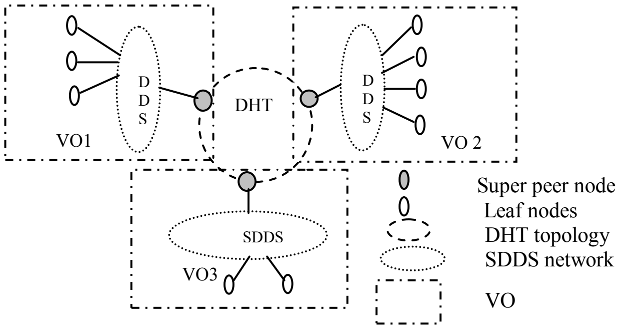


Fig. 2. SDDS based hierarchical DHT architecture.

3.2 Resource Discovery Protocol

In this section, we describe the resource discovery protocol used in the proposed SDDS-HDHT solution. Suppose that a second level peer p_i € VO$_i$ wants to discover a resource *Res* through a resource discovery query *Q*. Let the peer p_J the peer responsible for *Res*. Let Gp_i the gateway peer responsible for VO$_i$, *Gp$_i$_list* the list of its neighbors in the top level DHT (e.g., the left and right neighbor) and *Response* the metadata of *Res.* Thus, a lookup request for *Res* implies locating the peer responsible for *Res*. Hence, we distinguish two scenarios classifying resource discovery queries:

(i) Peers p_i and p_j belong to the same VO. Then, the query Q corresponds to an intra-VO resource discovery query.

(ii) Peers p_i and p_j are in different VOs. Then, the query Q corresponds to an inter-VO resource discovery query.

Intra-VO resource discovery queries are evaluated through a classical LH*$_{RS}^{P2P}$ routing system which is completely transparent to the top level DHT. Generally, users often access data in their application domain, i.e. in their VO. In consequence, it is important to search metadata source first in the local VO$_i$ before searching in other VOs. This solution favors principle of locality [10]. Recall that finding a peer responsible of metadata of the searched resource requires only two messages. Finally, the peer p$_J$ sends metadata describing Res (if founded) to p$_i$, the peer initiator of Q.

When the researched resource *Res* is not available in the local VO$_i$, resource discovery is required in other VOs. This corresponds to an inter-VO resource discovery process. Before introducing the resource discovery process, let's recall that we have defined a certain period of time (e.g. Round- Trip Time RTT) as in [21]. The manner in which the RTT values are chosen during lookups can greatly affects performances under churn. [23] has demonstrates that a RTT is a significant component of lookup latency under churn. In fact, requests in peer to peer systems under a churn are frequently sent to a peer that has left the system. At the same time, A DHT rooting has several alternate paths to complete a lookup. This is not the case when a failure concerns the gateway peer. In our solution, a RTT is mainly useful to maximize time to discover resources when a failure occurred in a gateway peer. In this case, p_i do not expect indefinitely. When RTT is exceeded, it considers that Gp_i is failed and consults the gateway neighbours list *Gp$_i$_list* received in the connection step. Then, p_i sends its query to one of the peers founded in *Gp$_i$_list*. Let's recall that in the connection step of any gateway peer Gp$_i$, this latter sent its list neighbors *Gp$_i$_list* to p$_0$ in its VO. Then, p$_0$ forwards *Gp$_i$_list* to all other second level peers. It i the nearest second level peer s done just on the connection step.

Let now examine an inter-VO lookup cost in SDDS-HDHT solution. When *Res* is not found in VO$_i$, the query is propagated to the gateway Gp$_i$. The localisation of the gateway responsible for the VO$_J$ containing *Res* requires $Lc_G=O(log_B(N_G))$ hops. After that, another lookup through the LH*$_{RS}^{P2P}$ routing system is required to search metadata of *Res* in VO$_J$. It requires two additional hops at most. Then, the total lookup cost for an inter-VO resource discovery query is $Lc=O(log_B(N_G))+4$ messages. In summary, the resource discovery process is defined in four steps:

(i) The peer p_i routed the query to the gateway Gp_i. If a Gp_i failure is detected (RTT is elapsed), it requests one neighbor of Gp_i, already received.

(ii) Once the query reaches a gateway peer Gp_i, a hash function *H* is applied to *Res* in order to discover the gateway responsible for the VO that containing *Res*. The query arrives at some Gp_J. This is valid whenever a resource, matching the criteria specified in the query, is found in some VO$_J$.

(iii) Using the LH*$_{RS}^{P2P}$ routing system in the founded VO$_J$, Gp_J routes the query to the peer p_J Є VO$_J$ that is responsible for Res.

(iv) Metadata of *Res* are sent to Gp_j which forward it to p_i via the reversing path.

3.3 System Maintenance

The continuous leaving and entering of nodes into the system is very common in Grid systems (dynamicity proprieties). In consequence, updating the system is required. Peer departures can be divided into friendly leaves and peer failures. Friendly leaves enable a peer to notify its overlay neighbors to restructure the topology accordingly. Peer failures possibility seriously damages the structure of the overlay with data loss consequences. Remedying this failure generates additional maintenance cost. In structured peer-to-peer systems, such as Pastry [24] used in our system, the connection / disconnection of one peer generates $2B*Log_B(N_T)$ messages [24]. Furthermore, the maintenance can concern the connection/ disconnection of one or more peers. Throughout this section, we explore the different factors that affect the behavior of hierarchical DHT under churn (super peer failure addressing, timeouts during lookups and proximity neighbor selection) [23]. Then, we discuss the connection/ disconnection of both gateways and second level peers.

Second Level Peer Connection/ Disconnection. The connection/ disconnection of a second level peer p_i do not affect lookups in other peers except the possible split of a bucket if this latter gets overloaded. Let's discuss the only one required maintenance. When p_i joins some VO_i, it asks its neighbor about Gp_i_list. In consequence, only two messages are required. This process avoid that several new arrival peers asked simultaneously the same gateway which can constitute a bottleneck as in SP-HDHT solution. In other terms, when a new second level peer arrives, it searches its gateway (only one) and neighbors of this one. This process permits also to reduce messages comparing to the complex process in the MG- HDHT solution in which the new second level peer should retrieve all gateways.

Gateway Peer Connection/ Disconnection. For this aim, we propose a protocol in order to reduce the overhead added to the system. When a gateway peer connection/ disconnection occur, we distinguish two types of maintenance: (i) maintenance of the DHT and (ii) maintenance of the neighbour's lists. We will not discuss the first maintenance since it corresponds to a classical DHT maintenance [25]. In the other hand, without any maintenance protocol, a disconnection or a failure of a gateway peer paralyzes access to all second level peers which is responsible for them. Addressing this failure generates additional maintenance cost. Before describing the maintenance process, let's analyze the connection of a gateway peer Gp_i to VO_i.

(i) Gateway peer Gp_i sent its list neighbors *Gp_i_list* (the left and right neighbor) to the nearest second level peer p_0 in VO_i.

(ii) Peer p_0 contacts peers in *Gp_i_list* to inform them about its existence (in order to have an entry to VO_i in the case of Gp_i failure).

(iii) Peer p_0 sent this list to all second level peers in VO_i via a multicast message. Recall that other second level peers do not report their existence to neighbors of Gp_i.

Recall also that this process is done just once at the initial connection of Gp_i and only p_0 periodically executes a *Ping/Pong* algorithm with iGp_i. It sends a *Ping* message to Gp_i and this one answers with a *Pong* message in order to detect any failure in Gp_i. Let us discuss the case of a gateway failure/ update. When Gp_i is replaced by another, the process of maintenance (after the DHT maintenance) is:

(i) The new gateway Gp_{New} contacts the nearest (only one) second level peer p_0 and gives him its neighbor's list *Gp_{New}_list.*

(ii) Peer p_0 inform peers in *Gp_{New}_list* about its existence. But, it does not inform other second level peers about *Gp_{New}_list* (lazy update).

Remark that the peer p_0 do not sent description of the new gateway peer Gp_{New} and its updated *Gp_{New}_list* to other second-level nodes at this moment. A lazy update is adopted. When Gp_i does not respond after a RTT period, a second level peer consults its old *Gp_i_list* to reach other VOs. Thus, it rejoins the overlay network in spite of a gateway failure. The update of this list is done during the reception of the resource discovery result as in [14]. Also, a failure of p_0 does not paralyze the system since the new gateway peer always contacts its nearest second level peer. The entry to the VO can also be done through peer p_0 since this one reported its existence in the connection step. This process allow a robust resource discovery process although the presence of dynamicity of peers. This is not the case in MG-HDHT solution when failures of all gateways in some VO paralyze the input/ output to/ from this VO. Recall also that one of the limitations that our solution suffers from: the failure of both a gateway peer and its neighbors in *Gp_i_list*. A solution consists on enrich the neighbors list of the any gateway node.

4. Performance Analysis

Experimental results based on a simulation of the suggested resource discovery solution are presented in this section. We based on a virtual network as 10000 nodes to prove the efficiency of our solution in large grid networks. We deal with a simulated environment since it is difficult to experiment thousands of nodes organized as virtual organization in a real existing platform as Grid'5000 [8]. We based our experiments on a platform having four features: (i) emulation of nodes, ii) emulation of network, (iii) using FreePastry [4], one implementation of the Pastry DHT and (iiii) $LH*_{RS}^{P2P}$ SDDS prototype implemented by Litwin's team in Dauphine University [2]. Variables used bellows are defined as follows: N_T is the number of nodes in the system, N_G the number of super peers, NSL the number of second level nodes and α the super peer ratio. It is the ratio between gateways and the total number ($N_G= \alpha. N_T$). Key of the discovered resource corresponds to a relation name in our experiments. For the detection of failed peers, we set a TTL to 1 sec. We simulate performances of (a) a flat DHT solution in order to measure the benefits hierarchical systems and previous hierarchical DHT solution b) SP-HDHT solution [31] in which gateways establish a DHT overlay network when each leaf peers maintains a connection to its gateway, (c) MG-HDHT solution in which several gateways are maintained between hierarchical levels. Then, we compare theirs performances.

Throughout this section, we deal with three classes of experiments: (i) Lookup performances experiments in which we interest to elapsed times which includes the query processing and communication costs. (ii) maintenance overhead experiments in which we simulate a join/leave peers scenario and interest to the required update messages and (iii) experiments to find the optimal ratio between gateway and second level peers in order to evaluate the impact of the gateway ratio in performances. For this aim, we have varied N_G but the total number of peers always stay constant.

4.1 Lookup Performances Analysis

First experiments simulate a flat DHT solution in which all peers run a DHT protocol. Thus, specify the equivalence between such systems and SDDS-DHT systems when

$N_T/N_G=1$. When we nalyze the hops number required to discover one resource in both solutions, our results are always better when it concerns an intra-VO resource discovery query. In fact, $LH^*_{RS}{}^{P2P}$ lookup requires a maximum of two (2) messages when this number is always $log_B(N_T)$ in flat DHT solutions. For inter-Vo queries, we have showed in last sections that the theoretically worse case corresponds to $O(log_B(N_G))+4$ hops with SDDS-HDHT scheme. By a simple calculation, we deduce that flat DHT performances are better when our DHT overlay is composed by more than 1000 gateways. In other terms, from 10 leaf peers/VO ($\alpha<1\%$), our results are better. This is due to the fact that adding new second level peers do not influences $LH^*_{RS}{}^{P2P}$ lookup performances. However, these results correspond to theoretical numbers of hops for only one resource discovery query. In the case of simultaneous resource discovery messages, the results should take into account that all messages are forward to the same gateway (in one VO). This generates some congestion in this peer. To confirm this, we have experiment systems with (i) 2000 gateways (5 leaf peers/ VO, $\alpha=20\%$) and (ii) 500 gateways (20 leaf peers/VO, $\alpha=5\%$). We also interest to the number of simultaneous resource discovery queries. It is useful since it shows if the SDDS-HDHT solution is also scalable in the presence of high number of messages. Fig. 3-left shows elapsed response times for resource discovery queries (intra and inter-VO queries). It confirms that our performances are always better when queries constitute intra-VO resource discovery queries. Elapsed response times are 50% better than flat DHT solution. This is due to the reason mentioned above. Let analyze performances of inter-VO queries. When we experiment with $\alpha=20\%$, performances are almost similar for a reduced simultaneous discovery queries. But, elapsed responses time increase from 20 queries/sec. It is due to the fact that all queries transit by the same gateway in each VO. However, a great leaf peers number ($\alpha=5\%$) improves significantly our performances which are better. The save is close 10% compared to the flat DHT solution in spite of the simultaneous messages. It provides from the gain in the DHT lookup. In fact, the probability to find the searched resource in a local VO is greater.

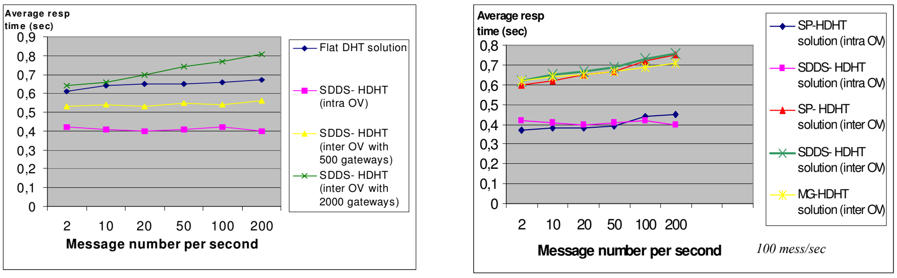


Fig. 3. SDDS-HDHT performances vs. Flat DHT (left) and SP-HDHT (right) performances

We have also compared our results to both SP-HDHT and MG-HDHT results. [31] proved that best performances are obtained with small number of gateways. We simulate a network with 100 VOs (with 100 level peers/ VO). Fig. 3-right shows that the SP-HDHT solution is slightly better for intra-VO queries when less simultaneous messages are used. From 70 messages/ second, our solution is 10% better than SP-HDHT solution. We explain this by the fact that intra-VO lookups are done without any gateway peer intervention when a bottleneck is generated in each gateway in the compared SP-HDHT solution. This is the reasons why the simultaneous messages influenced significantly the SP-HDHT results. We remark that the average response time is almost constant when we have several simultaneous messages in both SDDS-HDHT and MG-HDHT solution. We conclude that the save can be better if we experiment with great number of simultaneous discovery queries. Note that these experiments do not include the more costly connection step.

For inter-V0 queries, simultaneous resource discovery queries influences performances of both solutions. Bottleneck is generated since all queries transit by the same gateway peer which increases response times in SP-HDHT and SDDS-HDHT solutions. Then, SP-HDHT results are slightly better when we have less than 70 messages per second. From this value, results are almost close for the two solutions with slight advantage to SDDS-HDHT solution since intra-VO queries always precede inter-VO queries. We conclude that in inter-VO queries, we have dependence between performances and simultaneous queries for these two solutions. The same impact is observed with a reduced gateway ratio α. In the other hand, performances of MG-HDHT solution are better (rate of 5%) especially for high simultaneous messages since queries are propagated through the several gateways in the same VO.

4.2 Maintenance Analysis

We measure the impact of the join/ leave peers in the system. We interest to the total messages number required when a peer joins/leaves the network. We tabulate churn in an event-based simulator which processes transitions in state (*down*, *available*, and *in use*) for each peer as in [7]. We simulate a churn phase in which several peers join and leave the system but the total number of peers N_T stays appreciatively constant. The maintenance costs are measured by the number of messages generated to maintain the system when peers join/leave the system.

Lets a system with a peers distribution as {N_G=100 and 100 peers/ VO}. This configuration corresponds to average results in inter-VO discovery queries performances. In these experiments, when a number of new connections/ disconnections exceed 20 peers, 10% of them concern gateway peers. Fig. 4-left shows the impact of peers connection/ disconnection in the total messages number in the system. Flat DHT solution generates the greater number of messages in the connection /disconnection of peers. Compared to our solution, the messages number ratio is 1.1 (resp 4.5) for the connection of one leaf peer (resp 100 peers). It is clear that maintaining a flat DHT generates greatest costs especially when several peers join/leave the system. When a gateway join/leave the system in our solution, it generates $2BLog_B(N_G)$ messages. It corresponds to only two messagse for a connection of a second level peer and three messages for a connection of a new gateway without any update in the gateway's DHT. We compare these results to the SP-HDHT performances. The numbers of update messages are closes when we have only second level peers connections/disconnections. It corresponds to the case when less than 10 peers join the system. In fact, all new peers must contact their super peer in SP-HDHT solution. Increasing the number of connection/ disconection of second level peers can generates a bottleneck. Our solution offers a significant maintenance cost gain when the update occurs in gateways. As the number of gateways connection increase as the gain is important since the required update messages is less with our solution. The save is 59% for the connection of 90 leaf peers and 10 gateways. Certainly, update DHT messages concern both solutions. But, in the SP-HDHT solution experiments, the new gateway establishes connections with all its leaf nodes. It is also the case in the MG-HDHT solution. The fact that new second level peers in MG-HDHT must contact several gateways generates additional messages. It is not the case in our solution. A new second level peer contacts only its neighbour and the connection of a new gateway generates only two additional messages.

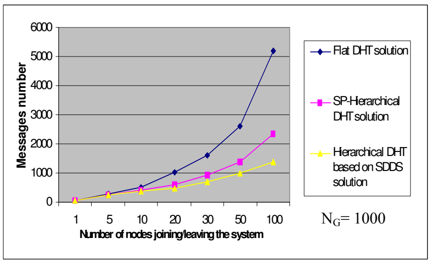


Fig. 4. Impact of the connection/ disconnection nodes in the messages number exchanged in the system.

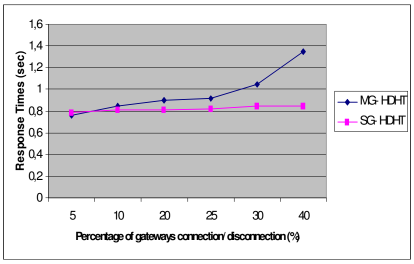


Fig. 5. Impact of the percentage of the gateways connection/ disconnection in the total response time.

We also experiment the impact of the percentage of the gateways arrival/ departure in the total response time as shown in Fig. 5. It corresponds to resource discovery process under a high churn. When only 5% of gateways are replaced by other gateways, MG-HDHT solution has slightly better results than SDDS-HDHT performances. However, when this percentage increases, SDDS-HDHT performances remain stable since second level peers used the gateway neighbor's list to reach other gateways in the DHT when they used, in MG-HDHT solution, the other not failed gateways in the same VO pending the update of the new gateways. From 25% gateways connection/ disconnection in the system, MG-HDHT curve increase significantly. Recall that we have deliberately ensured that not all gateways in the same VO are failed in MG-HDHT solution. Otherwise, a second level peer in some VO_i will be not able to contact any gateway of other VO_j ($i \neq j$) until. It is not the case in our solution in which second level peers can use the *Gp_i_list*. But, recognize that if all peers in the *Gp_i_list* failed, consequences are also the same as above.

4.3 Impact of the Gateway Ratio in Performances

Through these experiments, our goal is to determine optimal configurations on the three compared solutions. In first experiments, without any peer arrival/departure to the system, a centralized overlay network with only one super peer in SP-HDHT solution generates the lowest traffic costs. The reason is that only lookup and *Ping/ Pong* messages are exchanged between the super peer and its second level nodes. Also, same performances are obtained with the configuration (α=100%) in the three experimented solution since all peers participate in a flat DHT overlay. If the number of gateways increases ($N_G>1$), we notice increased lookup costs for the three compared solution. This cost is most important in SDDS-HDHT and SP- HDHT solution, mostly caused by the bottleneck in the only one gateway. Indeed, it is due to the fact that all queries transit by the same gateway when the several gateways are less in stress on the MG-HDHT solution. This cost decrease from α=20% in the SP-HDHT and SDDS-HDHT solutions. It is from α=10% in the MG-HDHT solution. We conclude that MG-HDHT solution constitutes the better solution when we have not or very little departures/ arrivals of peers in the system. Good performances obtained from α=10% with our solution. We also deal with experiments taking into account the arrival/ departure of peers to the system. We deal with the connection/ disconnection of 10% of the gateways in the system and 10% of second level nodes in each VO. From α=1%, the maintenance cost of the MG-HDHT solution is always the most important since each gateway inform all its second level nodes in each arrival/ departure. It is also the case with the SP- HDHT solution with better results. This is not the case in SDDS- HDHT which has the best results with □between 1 and 50%. It is due to the fact that second level nodes used a lazy update to update their neighbor's gateway list. For each value of α between 1 and 50%, the SDDS-HDHT solution generates the lowest total cost. It is valuable for the case when the major maintenance cost is generated by the departure/ arrival of second level nodes but also for the case when the departure/ arrival of gateways constitutes the major maintenance cost. We conclude that the best results are of SDDS-HDHT solution are obtained with $\alpha \in [1\%, 20\%]$ which is close to real grid systems with several VOs.

5. Related Work

Many research works [5], [6], [12], [17], [18] and [31] presented advantages of hierarchical DHT systems based on the super peer concept. However, most of them add a significant overhead to the system. [5] proposed a two-tier hierarchy using chord for the top level to reduce the lookup costs, but only with the goal of improving performance of the overlay network routing. [28] demonstrated the high maintenance state needed (memory, CPU and bandwidth) when all peers in the overlay are attached to different levels of the hierarchy. [18] explored the using of multiple Chord systems in order to reduce latency of lookups. Nevertheless, it neglects the churn effects. [31] gives a cost-based analysis of hierarchical P2P overlay network with super peers forming DHT and leaf nodes attached to them. However, super peers are put more under stress for both intra and inter-VO resource discovery queries especially if the leaf nodes number increase. Moreover, performances depend on the ratio between super peer's number and the total number of peers in the system. [12] presented a two-layer structure 'Chord2' to reduce maintenance costs in Chord. The lower layer is the regular Chord ring when the upper layer is a ring for maintenance constructed from super peers. On the other hand, several algorithms [7], [9], [16], [23] and [32] were proposed to resolve these problems. We cite the Bamboo protocol [23] designed to handle networks with high churn efficiently and the self organizing distributed algorithm [32] in which all decisions taken by the peers are based on their partial view in the sense that the algorithm became fully decentralized and probabilistic. Hence, there is trade-off between minimizing total network costs and minimizing the added overhead to the system. For these reasons, we have proposed to combine DHT and SDDS structures in order to minimize these costs without excessive overheads.

6. Conclusion and Future Works

We have proposed a hierarchical DHT solution for data sources discovery in data Grid systems. It deals with both the reduction of lookup costs and the managing of churn while minimizing additional overhead to the system. It also takes into account the content/path locality of organizations in Grids. Our solution combines DHT systems to scalable distributed data structures SDDS in its LH*$_{RS}^{P2P}$ variant. Only fewer nodes are mapped on a DHT. Each of them acts as a super peer for leaf-nodes and can serves a Virtual Organization (VO), structured as an SDDS, in a Grid. The first contribution is the improvement of lookup query complexity to discover metadata of any data source especially for intra-VO queries since these queries are transparent to the top level DHT lookup. Also, only the arrival of a new VO requires the DHT maintenance. Our solution addresses also other super peer problems as a single point of failure by using a minimum of messages. In fact, leaf nodes update theirs super peer neighbours during resource discovery queries. The performance analysis shows the benefit of our proposition through comparisons of our performances to those of previous solutions. It shows the improvement of lookup query performances especially when we have an important number of simultaneously resource discovery messages. It also shows a significantly maintenance saves especially in presence of dynamicity of nodes.

Our method can be useful in large scale grid environment since our solution generates less traffic network. Further work includes more performance studies in more realistic large grid environments with a high number of nodes. Also, we would like include more realistic models of churn as to scale traces of sessions times [3] collected from deployed networks to produce a range of churn rates with a more realistic distribution. Also, we would like to study the effects of alternate routing table neighbours as in [33].

7. References

1. M.S. Artigas, P. García and A. F. Skarmeta. "Deca: A Hierarchical Framework for Decentralized Aggregation in DHTs". LNCS, Volume 4269/2006, 246-257. 2006.

2. http://lamsade.dauphine.fr/~litwin/default.html

3. T. Fei, S. Tao, L. Gao, and R. Guerin. How to select a good alternate path in large peer-to-peer systems? In Proc. of the int. conf. IEEE INFOCOM 2006.

4. http://Freepastry.org/FreePastry/.

5. L. Garces-Erice, E. W. Biersack, K. W. Ross, P. A. Felber, and G. Urvoy-Keller. Hierarchical Peer to Peer Systems. In Proc. of ACM/IFIP Intern. Conf. Euro-Par'03.

6. P. Ganesan, K. Gummadi, and H. Garcia-Molina. Canon in g major: designing DHTs with hierarchical structure. Intern. Conf. on Distributed Computing Systems'04, pp 263–272.

7. P. B. Godfrey, S. Shenker, and I. Stoica. Minimizing Churn in Distributed Systems. Int. Conf. SIGCOMM. pp 147–158, Italy 2006.

8. GRID'5000. www.grid5000.org
9. I. Gupta, Ken Birman, P. Linga, A. Demers & R.V Renesse. Kelips: Building an Efficient and Stable P2P DHT through Increased Memory and Background Overhead. Lecture notes in computer science, 2003. Springer.
10. N Harvey, M Jones, S Saoiu, M. Theimer & A. Wolman. Skipnet: A Scalable Overlay Network with Practical Locality Properties. In Proc of USITIS 2003, Seattle, USA.
11. A. Iamnitchi, I. Foster, "A peer-to-peer approach to resource location in grid environments", Proc. of HPDC'02, Edinburgh, UK, August 02.
12. Y Joung, J-C Wang. "Chord^2: A two-layer Chord for reducing Maintenance Overhead via Heterogeneity". Computer Networks, vol. 51, no. 3, pp. 712–731, 2007.
13. Kazaa. http://www.kazaa.com/.
14. I. Ketata, R. Mokadem, F. Morvan. Resource Discovery Considering Semantic Properties in Data Grid Environments. Proc. of Inter. Conf. Globe 2011, Toulouse, Springer, LNCS 6864.
15. W. Litwin. "Linear hashing: A new tool for file and table addressing". VLDB 1980. Reprinted in Readings in Database Systems, Stonebreaker ed, 2nd Ed, Morgan Kaufmann'95.
16. A. Montresor, "A Robust Protocol for Building Superpeer Overlay Topologies," in IEEE International Conference on Peer-to-Peer Computing (P2P 2004).
17. I. Martinez, R. Cuevas, C. Guerrero, A. Mauthe. Routing Performance in a Hierarchical DHT-based Overlay Network. Euromicro Intern. Conf. PDP'08, 508-515, Toulouse.
18. A. Mislove and P. Druschel. "Providing Administrative Control and Autonomy in Structured Overlays". In Proceedings of IPTPS'04, pp 162- 172. San Diego, CA, Feb 2004.
19. E. Meshkova & al. A survey on Resource Discovery Mechanisms, Peer to Peer and Service Discovery Frameworks Computer Networks. Science Direct. Elsevier'08 (2097- 2128).
20. R. Mokadem , A. Hameurlain, A. Min Tjoa. Resource Discovery Service while Minimizing Maintenance Overhead in Hierarchical DHT Systems. In Intern. Conf. on Information Integration and Web-based Applications & Services (iiWAS'10), Paris, France.
21. Mastroianni C., Talia D. and Verta O. "Evaluating Resource Discovery Protocols for Hierarchical and Super-Peer Grid Information Systems". 19th Euromicro Intern. Conf. PDP'07.
22. E. Pacitti, P Valduriez & M Mattosso. "Grid data management: Open Problems and News Issues"; In Intl. Journal Grid Computing. Springer, 2007, Vol. 5, pp. 273-281.
23. S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz, "Handling churn in a DHT". in Proceedings of the General Track : 2004 Usenix Annual Technical Conference, Boston, USA.
24. A. Rowston & P. Druschel. "Pastry: Scalable Distributed object location and routing for large-scale peer-to-peer systems". Proceeding of the 18th IFIP/ACM international conference on Distributed Systems Platforms. Vol 2218, 2001, pp 329-350.
25. I. Stoica, Morris, Karger, Kaashoek, Balakrishma. CHORD : A scalable Peer to Peer Lookup Service for Internet Application. SIGCOMM'O, August'01, San Diego, USA
26. P. Trunfio, D Talia, H Papadakid, P Fragoupoulou, M mordachini, M Penanen, P Popov, V Valssov and S Haridi. Peer-to-Peer resource discovery in Grids: Models and systems. Future Generation Computer Systems (2007).
27. P. Valduriez P & E. Pacitti. "Data Management in Large-Scale P2P Systems". VECPAR 2004. M Daydé & al. (eds). LNCS 3402. pp 104-118. Springer-Verlag. 2005.
28. Z. Xu, R. Min, and Y. Hu. "HIERAS: a DHT Based Hierarchical P2P Routing Algorithm". Proceedings of Intern. Conf on Parallel Processing (ICPP'03), pp 187– 194, 2003.
29. H. Yakouben, W. Litwin, T. Schwarz. "LH*$_{RS}^{P2P}$: a Scalable Distributed Data Structure For the P2P Environment". Int conf. on new technologies of Distributed Systems. France, 2008.
30. B. Yang and H. Garcia-Molina. Designing a Super-Peer Network. Proc. of intern. conf. on Data Engineering ICDE'03, Bangalore, India.
31. S. Zöls, Z Despotovic, W Kellerer. "Cost-Based Analysis of Hierarchical DHT Design". Intern. Conf. P2P'06. Cambridge, IEEE Computer Society 2006 pp 233-239.
32. S. Zöls, Q. Hofstatter, Z. Despotovic, W. Kellerer. "Achieving and maintaining Cost-Optimal Operation of a Hierarchical DHT System". Proc. of Inter. Conf. ICC 2009, Germany.
33. B. Zhao, Kobiatowicz & A. Joseph. Tapestry: A resilient global scale overlay for service deployment. IEEE journ. on selected Areas in communications, 22 vol 1,2004.

Towards Service-Oriented Resource Discovery by means of Semantic Web Reasoning

Alexey Cheptsov

High Performance Computing Center Stuttgart (HLRS), University of Stuttgart, Nobelstrasse 19, 70569 Stuttgart, Germany

cheptsov@hlrs.de

Abstract. Reasoning is one of the essential tools of the modern Semantic Web. A number of applications for resource discovery on the Web such as random indexing enjoy a prominent place in face of the novel Semantic Web Reasoning trends. However, the reasoning algorithms are dealing with significant challenges when scaled up to the problem sizes addressed by the modern Semantic Web application. As such, they are not well-optimized to be applied to the emerging Internet-scale knowledge bases. We introduce a solution to building highly efficient and scalable reasoning applications based on the Large Knowledge Collider – a service-oriented incomplete reasoning platform breaking the scalability barriers of the existing solutions. We discuss the application of incomplete reasoning for the resource discovery tasks and demonstrate a service-oriented realization for the query expansion and subsetting algorithms based on the random indexing knowledge extraction technique.

Keywords: Random Indexing, Semantic Web Reasoning, Large Knowledge Collider.

1 Introduction

The large- and internet-scale data applications is a primary challenge for the Semantic Web, and in particular for reasoning algorithms, used for processing exploding volumes of data, exposed currently on the Web. Reasoning is the process of making implicit logical inferences from the explicit set of facts or statements, which constitute the core of any knowledge base. The key problem for most of the modern reasoning engines such as Jena [1] or Pellet [2] is that they can not efficiently be applied for the real-life data sets that consist of tens, sometimes of hundreds of billions of triples (a unit of the semantically annotated information), which can correspond to several petabytes of digital information. Whereas modern advances in the Supercomputing domain allow this limitation to be overcome, the reasoning algorithms and logic need to be adapted to the demands of rapidly growing data universe, in order to be able to take advantages of the large-scale and on-demand infrastructures such as high performance computing or cloud technology. On the other hand, the algorithmic principals of the reasoning engines need to be reconsidered as well in order to allow for very large volumes of data. Service-oriented architectures (SOA) can greatly contribute to this goal, acting as the main enabler of the newly proposed reasoning techniques such as incomplete reasoning [3]. This paper focuses on a service-oriented solution for constructing Semantic Web applications of a new generation, ensuring the drastic increase of the scalability for the existing reasoning applications, as elaborated by the Large Knowledge Collider (LarKC)[1] EU project.

The paper is organized as follows. In Section 2, we collect our consideration towards enabling the large-scale reasoning and its application for the resource discovery tasks. In Section 3, we discuss LarKC – a service-oriented platform for development of fundamentally new reasoning application, with much higher scalability barriers as by the existing solutions. In Section 4, we introduce some successful resource discovery applications implemented with LarKC, such as Random Indexing. In Section 5, we discuss our conclusions and highlight the directions for future work in highly scalable semantic reasoning.

2 Semantic Reasoning on the Web Scale

Despite the majority of data on the Web is available as an unstructured text, e.g. generated from the content kept in RDBM, the application areas of the modern Semantic Web spawn a wide range of domains, from social networks to large-scale Smart Cities projects in the context of the future internet [4][5]. However, data processing in such applications goes far beyond a simple maintenance of the collection of facts; based on the explicit information, collected in datasets, and simple rule sets, describing the possible relations, the implicit statements and facts can be acquired from those datasets.

Many data collections as well as application built on top of them allow for rule-based inferencing to obtain new, more important facts. The process of inferring logical consequences from a set of asserted facts, specified by using some kinds of logic description languages (e.g., RDF/RDFS and OWL[2]), is in focus of semantic reasoning. The goal is to provide a technical way to determine when inference processes is valid, i.e., when it preserves truth. This is achieved by the procedure which starts from a set of assertions that are regarded as true in a semantic model and derives whether a new model contains provably true assertions.

The latest research on the Internet-scale Knowledge Base Technologies, combined with the proliferation of SOA infrastructures and cloud computing, has created a new wave of data-intensive computing applications, and posed several challenges to the Semantic Web community. As a reaction on these challenges, a variety of reasoning methods have been suggested for the efficient processing and exploitation of the semantically annotated data. However, most of those methods have only been approved for small, closed, trustworthy, consistent, coherent and static domains, such as synthetic LUBM [6] sets. Still, there is a deep mismatch between the requirements on the

[1] http://www.larkc.eu/

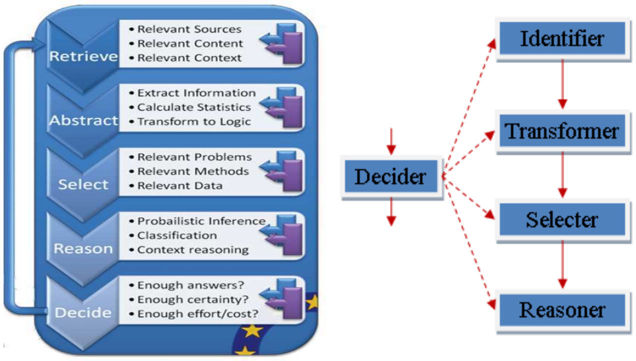
[2] http://www.w3.org/TR/owl-ref/

real-time reasoning on the Web scale and the existing efficient reasoning algorithms over the restricted subsets.

Whereas unlocking the full value of the scientific data has been seen as a strategic objective in the majority of ICT- related scientific activities in EU, USA, and Asia [7], the “Big Data” problem has been recognized as the primary challenger in semantic reasoning [8][9]. Indeed, the recent years have seen a tremendous increase of the structured data on the Web with scientific, public, and even government sectors involved. According to one of the recent IDC reports [10], the size of the digital data universe has grown from about 800.000 Terabytes in 2009 to 1.2 Zettabytes in 2010, i.e. an increase of 62%. Even more tremendous growth should be expected in the future (up to several tens of Zettabytes already in 2012, according to the same IDC report [10]).

The “big data” problem makes the conventional data processing techniques, also including the traditional semantic reasoning, substantially inefficient when applied for the large-scale data sets. On the other hand, the heterogeneous and streaming nature of data, e.g. implying structure complexity [11], or dimensionality and size [12], makes big data intractable on the conventional computing resource [13]. The problem becomes even worse when data are inconsistent (there is no any semantic model to interpret) or incoherent (contains some unclassifiable concepts) [14].

The broad availability of data coupled with increasing capabilities and decreasing costs of both computing and storage facilities has led the semantic reasoning community to rethink the approaches for large-scale inferencing [15]. Data-intensive reasoning requires a fundamentally different set of principles than the traditional mainstream Semantic Web offers. Some of the approaches allow for going far beyond the traditional notion of absolute correctness and completeness in reasoning as assumed by the standard techniques. An outstanding approach here is interleaving the reasoning and selection [16]. The main idea of the interleaving approach (see Fig. 1a) is to introduce a selection phase so that the reasoning processing can focus on a limited (but meaningful) part of the data, i.e. perform incomplete reasoning.



a) b)

Fig. 1. Incomplete reasoning, the overall schema (a) and the service-oriented vision (b)

As discussed before, the standard reasoning methods are not valid in the existing configurations of the Semantic Web. Some approaches, such as incomplete reasoning, offer a promising vision how a reasoning application can overcome the “big data” limitation, e.g. by interleaving the selection with the reasoning in a single “workflow”, as shown in Fig. 1a. However the need of combining several techniques within a single application introduces new challenges, for example related to ensuring the proper collaboration of team of experts working on a concrete part of the workflow, either it is identification, selection, or reasoning. Another challenge might be the adoption of the already available solutions and reusing them in the newly developed applications, as for example applying selection to the JENA reasoner [1], whose original software design doesn’t allow for such functionality. The SOA approach can help eliminate many of the drawbacks on the way towards creating new, service-based reasoning applications. Supposed that each of the construction blocks shown in Fig. 1a is a service, with standard API that ensures easy interoperability with the other similar services, quite a complex application can be developed by a simple combination of those services in a common workflow (see Fig. 1b).

Resource discovery is an essential feature of the Semantic Web, which involves tasks of decentralized and autonomous control, distributed service discovery etc. Reasoning can greatly contribute to solving these issues by for example improving the fine-grained service matchmaking, resource ranking, etc. in typical resource discovery workflows [29].

Although utilizing reasoning in the resource discovery workflows is not a new concept for the Semantic Web [17][18], there was quite a big gap in realizing the single steps of the reasoning algorithms (Fig. 1b) as a service. This was due to many reasons, among them complexity of the data dependency management, ensuring interoperability of the services, heterogeneity of the service’s functionality. Realizing a system where a massive number of parties can expose and consume services via advanced Web technology was also a research highlight for Semantic Web. An example of very successful research on offering a part of the semantic reasoning logic as a service is the SOA4ALL[3] project, whose main goal was to study the service abilities of development platforms capable of offering semantic services. Several useful services wrapping such successful reasoning engines as IRIS [19] and several others had been developed in the frame of this project. Nevertheless, the availability of such services is only an intermediate step towards offering reasoning as a service, as a lot of efforts were required to provide interoperability of those services in the context of a common application. Among others, a common platform is needed that would allow the user to seamlessly integrate the service by annotating their dependencies, manage the data dependencies intelligently, being able to specify parts of the execution that should be executed remotely, etc.

An outstanding effort to develop such a platform was performed in the LarKC (Large Knowledge Collider) [20] project. In the following sections, we discuss the main ideas, solutions, and outcomes of this project.

[3] http://www.soa4all.eu/

3 Large Knowledge Collider Approach

In order to create a technology for creation of trend-new applications for large-scale reasoning, several leading Semantic Web research organizations and technological companies have joined their efforts around the project of the Large Knowledge Collider (LarKC), supported by the European Commission. The mission of the project was to set up a distributed reasoning infrastructure for the Semantic Web community, which should enable application of reasoning far beyond the currently recognized scalability limitations [22], by implementing the interleaving reasoning approach. The current and future Web applications that deal with "big data" are in focus of LarKC.

The LarKC's design has been guided by the primarily goal to build a scalable platform for distributed high performance reasoning. Fig. 2 shows a conceptual view of the LarKC platform's architecture and the proposed development life-cycle. The architecture was designed to holistically cover the needs of the three main categories of users – semantic service (plug-in) developers, application (workflow) designers, and end-users internet-wide. The platform's design ensures a trade-off between the flexibility and the performance of applications in order to achieve a good balance between the generality and the usability of the platform by each of the categories of users.

Below we introduce some of the key concepts of the LarKC architecture and discuss the most important platform's services and tools for them.

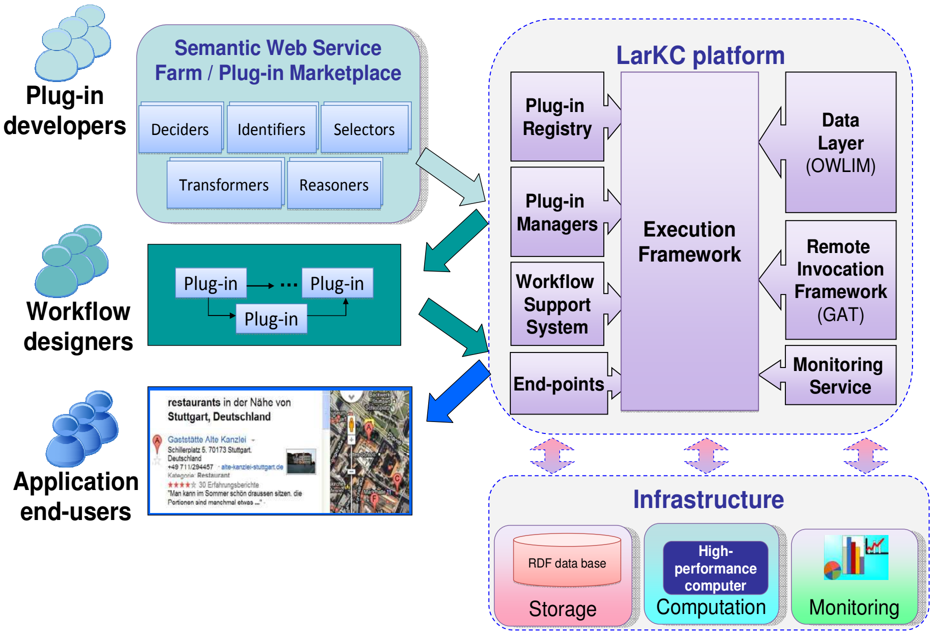


Fig. 2. Architecture of LarKC.

1. Plug-ins

Plug-ins are standalone services implementing some specific parts of the reasoning logic as discussed previously, whether it is selection, identification, transformation, or reasoning algorithm, see more at [21]. In fact, plug-ins can implement much broader functionality as foreseen by the incomplete reasoning schema (Fig. 1), hence enabling the LarKC platform to target much wider Semantic Web user community as originally targeted, e.g. for machine learning or knowledge extraction. The services are referred as plug-ins because of their flexibility and ability to be easily integrated, i.e. plugged into a common workflow and hence constitute a reasoning application. To ensure the interoperability of the plug-ins in the workflows, each plug-in should implement a special plug-in API, based on the annotation language [23]. Most essentially, the API defines the RDF schema (set of statements in the RDF format) taken as input and produced as output by each of the plug-ins. The plug-in development is facilitated by a number of special wizards, such as Eclipse IDE wizard or Maven archetype for rapid plug-in prototyping. The ready-to-use plug-ins are uploaded and published on the marketplace – a special web-enabled service offering a centralized, web-enabled repository store for the plug-ins[4].

2. Workflows

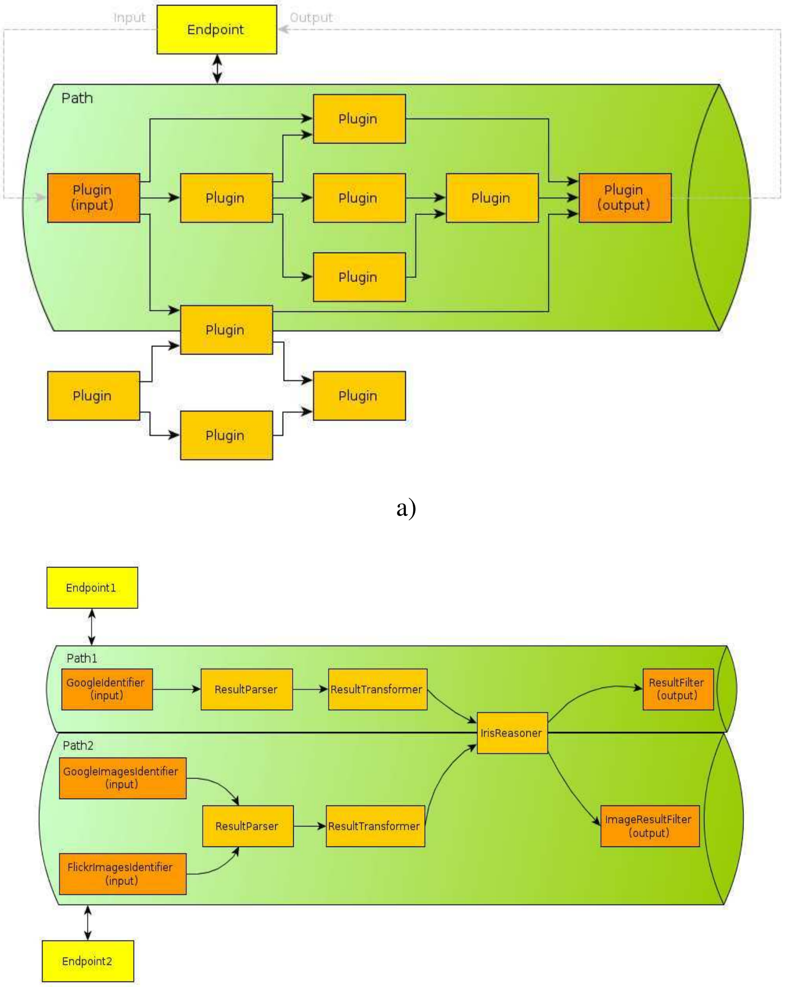
The workflow designers get access to the Marketplace in order to construct a workflow from the available plug-ins, combined to solve a certain task. In terms of LarKC, workflow is a reasoning application that is constructed of the (previously developed and uploaded on the Marketplace) plug-ins. The workflow's topology is characterised by the plug-ins included in the workflow as well as the data- and control flow connections between these plug-ins.

The complexity of the workflow's topology is determined by the number of included plug-ins, data connections between the plug-ins (also including multiple splits and joins such as in Fig. 3a or several end-points such as in Fig. 3b), and control flow events (such as instantiating, starting, stopping, and terminating single plug-ins or even workflow branches comprising several plug-ins). Same as for plug-ins, the input and output of the workflow is presented in RDF, which however can cause compatibility issues with the user's GUI, which are not obviously based on an RDF-compliant representation. In order to confirm the internal (RDF) dataflow representation with the external (user-defined) one, the LarKC architecture foresees special end-points, which are the adapters facilitating the workflow usage in the tools outside of the LarKC platform. Some typical examples of end-points, already provided by LarKC, are e.g. SPARQL end-point (SPARQL query as input and set of RDF statements as output) and HTML end-point (HTTP request without any parameters as input and HTML page as output).

For the specification of the workflow configuration, a special RDF schema was elaborated for LarKC, aiming at simplification of the annotation efforts for the work-

[4] Visit the LarKC Plug-in Marketplace at http://www.larkc.eu/plug-in-marketplace/

flow designers. Fig. 4a shows a simple example of the LarKC workflow annotation. Creation of the workflow specification can greatly be simplified by using upper-level graphical tools, e.g. Workflow Designer that offers a GUI for visual workflow construction (Fig. 4b) [28]. The elaborated schema makes specification of the additional features such as remote plug-in execution extremely simple and transparent for the users and can be used for tuning the front-end graphical interfaces of the applications to adapt them to the user needs.



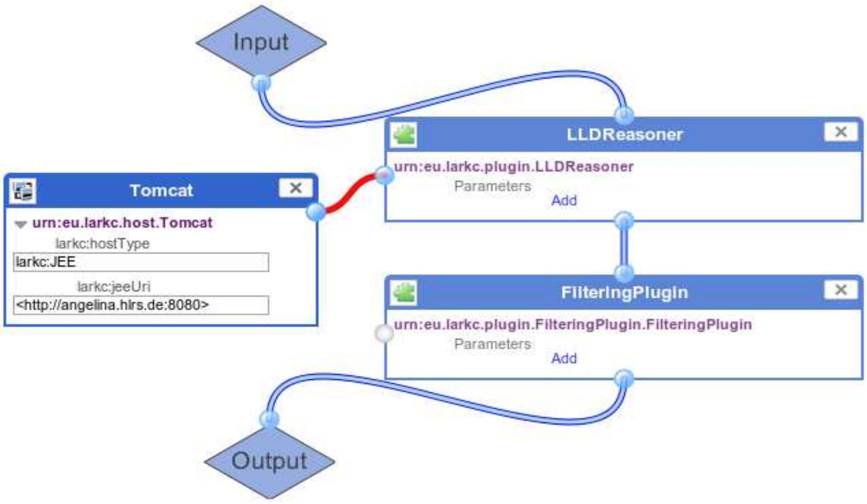
a)

b)

Fig. 3. Examples of LarKC workflows: a) workflow with non-trivial branched dataflow (containing multiple splits/joins), b) workflow with multiple end-points

```

# Define plug-ins
_:plugin1 a <urn:eu.larkc.plugin.LLDReasoner> .
_:plugin1 a <urn:eu.larkc.FilteringPlugin.FilteringPlugin>
_:plugin1 larkc:runsOn _:host1 .

  # Define hosts
  _:host1 a <urn:eu.larkc.host.Tomcat> .
  _:host1 larkc:hostType larkc:JEE .
  _:host1 larkc:jeeUri <http://angelina.hlrs.de:8080> .

# Define a path to set the input and output of the workflow
_:path a larkc:Path .
_:path larkc:hasInput _:plugin1 .
_:path larkc:hasOutput _:plugin1 .

# Connect an endpoint to the path
_:ep a <urn:eu.larkc.endpoint.sparql.SparqlEndpoint> .
_:ep larkc:links _:path .
```

a)



urn:eu.larkc.plugin.LLDReasoner

urn:eu.larkc.host.Tomcat

larkc:hostType: larkc:JEE

larkc:jeeUri: <http://angelina.hlrs.de:8080>

urn:eu.larkc.plugin.FilteringPlugin.FilteringPlugin

b)

Fig. 4. Further example of LarKC workflows: a) RDF schema for workflow annotation, b) Workflow Designer GUI with the specification of the remote host

3. Applications

Workflows are already standalone applications that can be submitted to the platform and executed by means of such tools as Workflow Designer discussed above. Nevertheless, workflows can also be wrapped into much more powerful user interfaces, adapted to the needs of the targeted end-user communities, e.g. Urban Computing [24], and using LarKC as a back-end engine. The service-oriented approach makes possible hiding the complexity of the LarKC platform, by enabling its whole power to the end-users through such interfaces. We present an exemplarily LarKC application in Section 4.

4. Platform services

All above-described activities related to plug-in creation, workflow design, and application development are facilitated by an extensive set of the platform services, as shown in Fig. 2. A detailed description of the main LarKC services can be found in our previous publication [21].

4 Application Scenario – Random Indexing

Random indexing [25] is a distributional statistic technique used in resource discovery for extracting semantically similar words from the word co-occurrence statistics in the text data, based on high-dimensional vector spaces (Fig. 5).

Random indexing offers new opportunities for a number of large-scale Web applications performing the search and reasoning on the Web scale [26]. Prominent application using random indexing is subsetting (Fig. 6a) and query expansion (Fig. 6b).

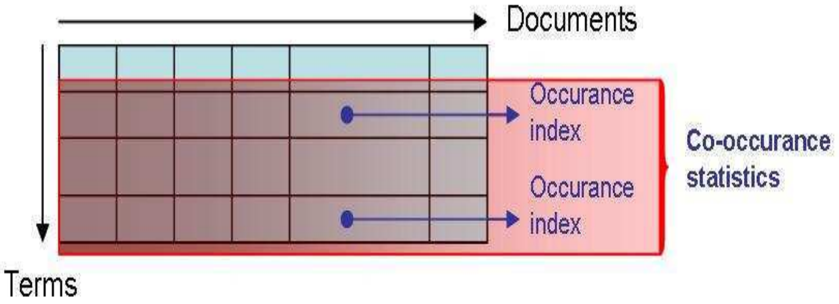
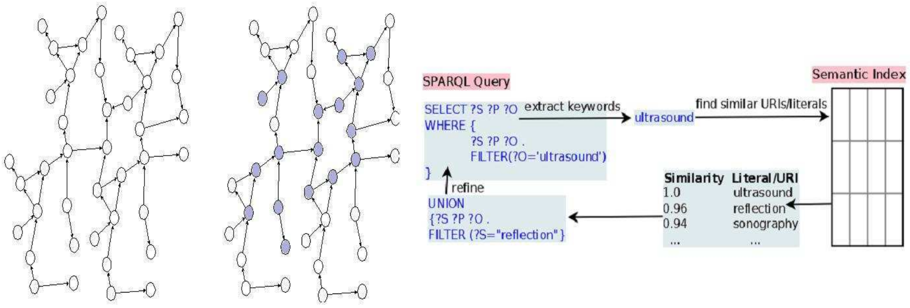


Fig. 5. Schema of the co-occurrence statistical analysis of text corpora.



```
SELECT ?S ?P ?O
WHERE {
        ?S ?P ?O .
        FILTER(?O='ultrasound')
}
```

```
UNION
{?S ?P ?O .
FILTER (?S="reflection"}
```

Similarity	Literal/URI
1.0	ultrasound
0.96	reflection
0.94	sonography
...	...

a) b)

Fig. 6. Application of Random Indexing: a) subsetting b) query expansion.

Query expansion [30] is used in information retrieval with the aim to expand the document collection returned as a result to a query, thus covering the larger portion of the documents. Subsetting (also known as selection) [31], on the contrary, deprecates the unnecessary items from a data set in order to achieve faster processing. Both presented problems are complementary, as change properties of the query to best adapt it to the search needs.

The main complexity of the random indexing algorithms lies in the following:

- High dimensionality of the underlying vector space.

A typical random indexing search algorithm performs traversal over all the entries of the vector space. This means, that the size of the vector space to the large extent defines the search performance. The modern data stores, such as Linked Life Data or Open Phacts consolidate many billion of statements and result in vector spaces of a very large dimensionality. Random indexing over such large data sets is computationally very costly, with regard to both execution time and memory consumption. The latter is of especial drawback for use of random indexing packages on the mass computers. So far, only relatively small parts of the Semantic Web data have been indexed and analyzed.

- High call frequency.

Both indexing and search over the vector space is typically a one-time operation, which means that the entire process should be repeated from scratch every time new data is encountered.

The implementation as a LarKC plug-in allows random indexing to take advantages of the LarKC data and execution model, being seamlessly integrated with the other plug-ins and building up a common workflow. This allows random indexing to be coupled with reasoners to improve the resource discovery algorithm. On the other hand, the reasoning process can also benefit from the integration, for example by using random indexing to expand the initial query and improve the quality of the obtained results, such as shown in Fig. 7.

LarKC is the technology that not only enables the large-scale reasoning approach for the already existing applications, but also facilitates their rapid prototyping with low initial investments, leveraging the SOA approach through the unique platform solutions. Furthermore, LarKC delivers a complete eco-system where the researches from very different domains can team up in order to develop new challenging mashup-applications, e.g. for the resource discovery, hence having a dramatic impact on a lot of problem domains.

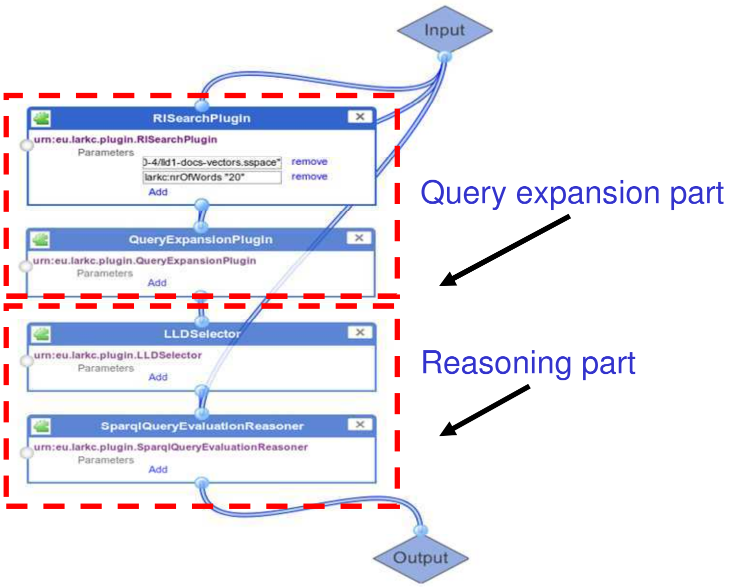

Input
RISearchPlugin
urn:eu.larkc.plugin.RISearchPlugin
Parameters
0-4/lld1-docs-vectors.sspace"
larkc:nrOfWords "20"
remove
remove
Add
QueryExpansionPlugin
urn:eu.larkc.plugin.QueryExpansionPlugin
Parameters
Add
LLDSelector
urn:eu.larkc.plugin.LLDSelector
Parameters
Add
SparqlQueryEvaluationReasoner
urn:eu.larkc.plugin.SparqlQueryEvaluationReasoner
Parameters
Add
Output
Query expansion part
Reasoning part

Fig. 7. Realization of query expansion in the Linked Life Data reasoning workflow.

5 Conclusions

We proposed a technology that allows a resource discovery process to be enhanced by integration with the reasoning. The technology is based on the Large Knowledge Collider (LarKC). LarKC is very promising platform for creation of new-generation semantic reasoning applications. The LarKC's main value is twofold. On the one hand, it enables a new approach for large-scale reasoning based on the technique for interleaving the identification, the selection, and the reasoning phases. On the other hand, through over the project's life time (2008-2011), LarKC has evolved in an outstanding, service-oriented platform for creating very flexible but extremely powerful applications, based on the plug-in's realization concept. The LarKC plug-in marketplace has already comprised several tens of freely available plug-ins, which implement new know-how solutions or wrap existing software components to offer their functionality to a much wider range of applications as even originally envisioned by their developers. Moreover, LarKC offers several additional features to improve the performance and scalability of the applications, facilitated through the parallelization, distributed execution, and monitoring platform. LarKC is an open source development, which encourages collaborative application development for Semantic Web. Despite being quite a young solution, LarKC has already established itself as a very promising technology in the Semantic Web world. Some evidence of its value was a series of Europe- and world-wide Semantic Web challenges won by the LarKC applications. It is important to note that the creation of LarKC applications, including the ones discussed in the paper, was also possible and without LarKC, but would have required much more (in order of magnitude) development efforts and financial investments.

We believe that the availability of such platform as LarKC will make a lot of developers to rethink their current approaches for resource discovery as well as semantic reasoning towards their tighter coupling and wider adoption of the service-oriented paradigm.

6 Acknowledgment

This work has been performed for the LarKC project (http://www.larkc.eu), partly funded by the European Commission's IST activity of the 7th Framework Programme (ICT-FP7-215535).

7 References

1. McCarthy, P.: Introduction to Jena. IBM Developer Works, http://www.ibm.com/developerworks/xml/library/j-jena/
2. Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A., Katz, Y.: Pellet: a practical owl-dl reasoner. Journal of Web Semantics, http://www.mindswap.org/papers/PelletJWS.pdf
3. Fensel, D., van Harmelen, F.: Unifying Reasoning and Search to Web Scale. IEEE Internet Computing, 11(2), 96--95 (2007).
4. Broekstra, J., Klein, M., Decker, S., Fensel, D., van Harmelen, F., Horrocks, I.: Enabling knowledge representation on the Web by extending RDF schema. Proceedings of the 10th international conference on World Wide Web (WWW '01), ACM, 467--478 (2001).
5. Donovang-Kuhlisch, M.: Smart City Process Support and Applications as a Service – from the Future Internet. Future Internet Assembly 2010, http://fi-ghent.fi-week.eu/files/2010/12/1430-Margarete-Donovang-Kuhlisch.pdf (2010)
6. Guo, Y., Pan, Z., Heflin, J.: LUBM: A Benchmark for OWL Knowledge Base Systems. Web Semantics, 3(2), 158--182 (2005)
7. High Level Expert EU Group: Riding the wave - How Europe can gain from the rising tide of scientific data. Final report, October 2010, http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707
8. Thompson, B., Personick, M.: Large-scale mashups using RDF and bigdata. Semantic Technology Conference (2009)
9. Hustadt, U., Motik, B., Sattler, U.: Data Complexity of Reasoning in Very Expressive Description Logics. Proc. IJCAI 2005, Edinburgh, UK, July 30–August 5 2005. Morgan Kaufmann Publishers, 466--471 (2005)

10. McKendrick, J.: Size of the data universe: 1.2 zettabytes and growing fast, ZDNet.
11. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a streaming world! Reasoning upon rapidly changing information. IEEE Intelligent Systems, 24(6), 83--89 (2009)
12. Fensel, D., van Harmelen, F.: Unifying Reasoning and Search to Web Scale. IEEE Internet Computing. 11(2), 96--95 (2007)
13. Cheptsov, A., Assel, M.: Towards High Performance Semantic Web – Experience of the LarKC Project. inSiDE - Journal of Innovatives Supercomputing in Deutschland, 9(1), 72--75 (2011)
14. Huang, Z., van Harmelen, F., Teije, A.: Reasoning with inconsistent ontologies. Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'05, 454--459 (2005)
15. Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Staab, S., Stojanovic, L., Stojanovic, N., Studer, R., Stumme, G., Sure, Y., Tane, J., Volz, R., Zacharias, V.: KAON - Towards a Large Scale Semantic Web. Tjoa, Proceedings of the Third international Conference on E-Commerce and Web Technologies, 304--313 (2002)
16. Huang, Z.: Interleaving Reasoning and Selection with Semantic Data. Proceedings of the 4th International Workshop on Ontology Dynamics (IWOD-10), ISWC2010 Workshop (2010)
17. Deelman, E., Gannon, D., Shields, M., Taylor I.: Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25(5) (2009)
18. Gil, Y., Ratnakar, V., Fritz, C.: Assisting Scientists with Complex Data Analysis Tasks through Semantic Workflows. Proceedings of the AAAI Fall Symposium on Proactive Assistant Agents, Arlington, VA.
19. IRIS - Integrated Rule Inference System - API and User Guide, http://iris-reasoner.org/pages/user_guide.pdf
20. Fensel, D., van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E., Fischer, F., Huang, Z., Kiryakov, A., Lee, T., Schooler, L., Tresp, V., Wesner, S., Witbrock, M., Zhong, N.: Towards LarKC: A Platform for Web-Scale Reasoning. Proceedings of the 2008 IEEE international Conference on Semantic Computing ICSC, 524--529 (2008)
21. Assel, M., Cheptsov, A., Gallizo, G., Celino, I., Dell'Aglio, D., Bradeško, L., Witbrock, M., Della Valle, E.: Large knowledge collider: a service-oriented platform for large-scale semantic reasoning. Proceedings of the International Conference on Web Intelligence, Mining and Semantics (2011)
22. Assel, M., Cheptsov, A., Gallizo, G., Benkert, K., Tenschert, A.: Applying High Performance Computing Techniques for Advanced Semantic Reasoning. eChallenges e-2010 Conference Proceedings. Paul Cunningham and Miriam Cunningham (Eds). IIMC International Information Management Corporation (2010)
23. Roman, D., Bishop, B., Toma, I., Gallizo, G., Fortuna, B.: LarKC Plug-in Annotation Language. Proceedings of The First International Conferences on Advanced Service Computing – Service Computation 2009 (2009)
24. Della Valle, E., Celino, I., Dell'Aglio, D.: The Experience of Realizing a Semantic Web Urban Computing Application. T. GIS, vol. 14, iss. 2, 163--181 (2010)
25. Sahlgren, M.: An introduction to random indexing. Proceedings of Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering TKE 2005, 1--9 (2005)

26. Jurgens, D., Stevens, K.: The S-Space Package: An Open Source Package for Word Space Models. Proceedings of the ACL 2010 System Demonstrations, 30--35 (2010)
27. Assel, M., Cheptsov, A., Czink, B., Damljanovic, D., Quesada, J.: MPI Realization of High Performance Search for Querying Large RDF Graphs using Statistical Semantics. Proceedings of the 1st Workshop on High-Performance Computing for the Semantic Web (HPCSW2011), co-located with the 8th Extended Semantic Web Conference, ESWC2011, Heraklion, Greece, May 29 (2011)
28. Le Phuoc, D., Polleres, A., Morbidoni, C., Hauswirth, M., Tummarello, G.: Rapid semantic web mashup development through semantic web pipes. Proceedings of WWW2009 Research Track (2009)
29. Ruta, M.: If objects could talk: novel resource discovery approaches in pervasive environments.
http://www.iaria.org/conferences2010/filesUBICOMM10/MicheleRuta_NexTech2010_Keynote_Speech-2.pdf
30. Efthimiadis, E.: Query Expansion. Martha E. Williams (ed.), Annual Review of Information Systems and Technology (ARIST), v31, 121--187 (1996)
31. Quilitz, B., Leser, U.: Querying Distributed RDF Data Sources with SPARQL. Proceedings of the 5th European Semantic Web Conference (ESWC2008)

Should I quit using my resource? Modeling Resource Usage through Game Theory

Paraskevas V. Lekeas

Department of Applied Mathematics,
University of Crete, Crete, Greece
plekeas@gmail.com

Abstract. Existing web infrastructures have supported the publication of a tremendous amount of resources, and over the past few years Data Resource Usage has become an everyday task for millions of users all over the world. In this work we model Resource Usage as a Cooperative Cournot Game in which a resource user and the various resource services are engaged. We give quantified answers as to when it is of interest for the user to stop using part of a resource and to switch to a different one. Moreover, we do the same from the perspective of a resource's provider.

Keywords: Resource Usage, Cournot Competition, Game, Core

1 Introduction

Data Resource[1] Usage is an everyday task for millions of users all over the world. Exchanging information, communicating, working and various other aspects of our life have been inevitably affected by data repositories which can be accessed through various channels, such as the Web and the Internet via different technologies, interfaces and infrastructures [1]. Usually once someone has identified an appropriate resource, he interacts with it by exchanging information. This sort of interaction is heavily commercialized, and a huge industry[2] has been established, which invests a great amount of money in marketing web services and products that provide access to resources quite often freely. This is why from now on we will use the word "provider" to refer to the underlying structure responsible for a resource. These providers most of the time are extremely interested in developing integrated resource services[3] in order to attract users, and, more importantly, to convince them to keep using these. This is because users are valuable: They provide information to the resource by interacting with

[1] When we refer to a resource we have in mind that in the background there exists a set of electronic mechanisms or internet infrastructures that created this resource for the purpose of value generation (either a profit when there is an underlying company or some other social gain, such as [2]). See also [3] for a related taxonomy.

[2] End-user spending for IT services worldwide estimated to be $763 billion in 2009 [4].

[3] We prefer this term instead of the term "web service" since many other alternative channels exist like satellite and cellphone grids, ad hoc networks, etc.

it, they bring money through the adds or subscription fees, and of course they bring new users to broaden the profit cycle.

A living example is Google which proposes a web resource experience through the integration of different technologies in order for users to continue using its services. Opening a Google account is a fairly easy one-minute process and instantaneously the new user has access to different cutting-edge services like Android OS apps, adaptive web search through the Google search engine, cloud services, access to landline phone calls, teleconferencing services and much more. A "perfect" user of Google would be the one who uses all these services explicitly through the Google APIs, sharing no data with any other competitive resource (e.g. AWS [5] or Ubuntu One [6]) and thus enriching only Google's resource knowledge repositories. However, many times it is the case that not all services or technologies of a resource are welcomed by users and sometimes users tend to accept only specific services from a resource ignoring some others. Also a situation that is not so good for a provider is the case where users decide to quit its resource and switch to a different one that provides similar or better services [7].

In this work we investigate the following problem. When do users tend to partially[4] abandon a specific resource? What can resource providers do about that? Is there a way to formulate the above trends in order to be evaluated and measured? In order to approach the above questions we model the various user - service interactions within a resource with different plays of the user, which are engaged either in a cooperative or in a non-cooperative manner. Each of these plays generates a value, which is to be conceived as a measure of the user's satisfaction for the appropriate service.

In the rest of the paper we proceed as follows: Section 2 gives a motivating example and formulates Resource Usage as a cooperative Cournot game. Section 3 studies the cases of partial rejection of a resource and also the possible reactions of the provider to prevent that. Section 4 concludes with a discussion and future work.

2 Modeling Resource Usage

Before describing our model let us give a stimulating example.

2.1 The case of Zoogle+ resource

Imagine the following scenario[5]: Zoogle Inc. decides to offer a new web integrated data resource, named Zoogle+ that will provide its users with the following set of

[4] Partially means that the user is unsatisfied only with some of the services and wants to switch but likes the rest and wants to keep them.

[5] Any explicit or implicit references to facts or persons is accidental and imaginary. Beware also that in Greek, Zoogle is pronounced almost the same as the Greek word "" which means Jungle referring to the chaotic and controversial informational nature of the WWW.

services: $N = \{email, cloud, voip\}$. A user u decides to try Zoogle+ for a certain period of time, and for this reason he signs up creating an account. Since u wants to be accurate in his calculations he uses a worth function $v(\cdot)$ to rate how good his experience is. Since it is Zoogle+'s policy to prohibit the exclusive use of only one service, ignoring the rest, the single use of a service for u is worth 0, i.e. $v(\{i\}) = 0$, $\forall\, i \in N$. Moreover, when u uses any two of the services he rates his experience with a value of 2, i.e. $v(\{i, j\}) = 2$, $\forall\, i, j \in N$ with $i \neq j$. Finally, when using all three of the services he calculates a value of 2.5, i.e. $v(N) = 2.5$. What should the user do? Should he be totally loyal to Zoogle+ or not?

On the one hand, when u uses all the services he gets on average 0.83 for each one, but, on the other hand, when he selects only two of the services he gets 0 for the one not selected and 1 for each of the rest. Therefore, u decides to maximize his satisfaction and thus selects to use only two services of Zoogle+ and to seek the third in an external resource. So the answer in this case is that u will partially leave. What would happen if $v(N)$ is worth 3 to u? Clearly then he should stay loyal to Zoogle+ because he would maximize his satisfaction in this case, since there is no combination of services that would give him more than 1 (on average).

Underlying the intuition of the above example is the idea of the core in cooperative game theory [8]. Briefly speaking, a cooperative game is a situation in which a group of N players, by making decisions that take into account each other's actions and responses, decides to act together to generate some value. This value (or what is left if we put aside the costs of playing the game) must later be split according to certain rules agreed upon by these players. The various ways this split can be made are defined as solution concepts. One of the most used solution concepts is that of the core. According to this, the split is in the core if no subset of players can benefit (earn more) by breaking away from the whole set N. Under this light, user u decided to partially leave Zoogle+ when $v(N) = 2.5$ because the core of the game was empty, while when $v(N) = 3$ the core became non-empty and thus u used all the services. In what follows we formulate Resource Usage as a cooperative Cournot game.

2.2 The model

Suppose that a user u decides to sign up for a resource R in order to *use* its i, $i \in \{1, \cdots, n\}$, different services. The phrase "use a service" refers to the interaction of u with the service in order for some desired tasks to be completed. For example, using the GUI of a service, entering data by typing, downloading files, writing scripts and compiling them online can be perceived as parts of such an interaction. Let p_i denote the interaction of u with service i. Call each p_i a *play* that u does with the service. Every play[6] generates some value for u. Since u needs to make an effort to generate this value we assume that the *per unit cost* to u for playing p_i is c. For example, if a play generates 2 units of value for u, then the cost of the play would be $2c$. In general the cost function for q_i units of

[6] from now on the terms "play" and "interaction" will be used interchangeably.

value produced in a play would be cq_i. We assume for clarity of the exposition that c is the same for every p_i, $i = 1, \cdots, n$. If we now take the value generated from a play and subtract the cost spent to produce it, then we will find how much the play is *worth* to the user, or in other words we will find the profit of the user from using the specific service, which intuitively represents a measure of how satisfied the user is with the service. So each play contributes some worth to the user and if we add all these contributions from all the plays we will have the total worth to the user u from using all the services of R.

As is generally accepted, maximizing user satisfaction constitutes the key issue to every service. This forces every play p_i to seek to maximize its contribution to the total worth earned by u. But this happens under the following restriction. The user u has a limited time to spend interacting with the services and thus he must split this into his needs wisely in order to acomplish his different tasks through R. This means that no play can monopolize all the available time of u. Moreover, u's multitasking abilities are limited by nature. So spending more time on service i might, on the one hand raise satisfaction from i but on the other might lower satisfaction from service j ($j \neq i$). A mathematical model that describes such interactions among p_i's is the Cournot competition [9]. Under this model it is assumed that the plays do not cooperate with each other but instead decide independently and at the same time about how much value they should produce for u. Thus if with π_i we declare the profit of each p_i we will have that:

$$\pi_i = \max_i(\textit{value generated from i-th play} - \textit{cost spent}) \tag{1}$$

Assume that u in play i creates q_i, $i = 1, \cdots, n$ units of value. In economics usually the function that describes the total value generated from the i-th play in its simplest form is given by $(a - \sum_{j=1}^{n} q_j)q_i$. Here a is a positive constant that represents the size of the environment into which interactions happen. For our model, a could e.g. represent the total size of data held by the resource available for use or the total time that u wants to dedicate to R. If, for example, a represents the time available, then the demand function intuitively says that the more value is generated by all the plays the less time remains to be used and vice versa. Since now the cost spent for i-th play is cq_i using the above in (1) we will have:

$$\pi_i = \max_i(a - \sum_{j=1}^{n} q_j - c)q_i \ , i = 1, \cdots, n \tag{2}$$

The solutions of (2) are the Nash equilibria of the plays of u in the case that these plays do not cooperate. These equilibria will help us find the worth of u and reason about his loyalty to R. It is easy to prove (the proof can be found in the Appendix, or for a more general case, in [9]) that the solutions of (2) give the following worth for each play p_i in the case that these act independently (non-cooperatively):

$$\pi_i = \left(\frac{a-c}{n+1}\right)^2 \tag{3}$$

Assume now that all the p_i's decide to act in a cooperative manner. In this case due to symmetry we can imagine all the plays combined together into a unique play, so (2) collapses into a single equation the maximization of which gives:

$$v(N) \equiv \frac{(a-c)^2}{4} \tag{4}$$

where with $v(N)$ we denote the total worth produced when all plays cooperate. Before applying our model to quantify u's loyalty to R let us discuss what cooperation and non-cooperation means for the plays p_i. When we say that a set of plays cooperate, we mean that they do not harm each other but instead act as a unity to produce a common value. On the other hand, when there is non-cooperation each play does not care about the other plays but wants only to maximize its own value. Since this idea might sound a bit subtle we give the following example: In the case of Zoogle+ if the 3 services acted in a cooperative manner then their total worth according to (4) would be $\frac{(a-c)^2}{4}$. If, on the other hand, they acted in a non-cooperative manner, from (3) each would be worth $\left(\frac{a-c}{3+1}\right)^2$ and their total worth would be $3\left(\frac{a-c}{4}\right)^2 < \frac{(a-c)^2}{4}$. So it is clear that when under cooperation the plays produce more.

3 User Loyalty

Let us now turn our attention to the loyalty of u to R. As said earlier p_i's can act in a cooperative or non-cooperative manner in order to achieve their goals. Apart from total cooperation and non-cooperation there exist intermediate situations in which some of the plays might decide to cooperate and some others will decide to deviate (non-cooperate). It is exactly these cases that will shed some light on u's loyalty to R. Consider the following scenario: User u is thinking of abandoning a non-empty set $S \subset N$ of services because he discovered that on a different resource R' he earns $v(S)$ from these. In order to make his final decisions he first reasons as follows: "If I stay loyal to R, my total worth from (4) is $v(N)$ and on average I earn $\frac{v(N)}{n}$. On the other hand, if I partially switch to R' I would earn on average $\frac{v(S)}{s}$, where $|S| = s$. So I partially switch to R' if $\frac{v(S)}{s} > \frac{v(N)}{n}$." The idea just described describes the notion of the core in the cooperative game played by p_i, $i = 1, \cdots, n$. We say that the core is non-empty when for any non-empty set of services chosen, u on average earns less than if all the services were chosen. In other words, the core is non-empty when:

$$\frac{v(S)}{s} \leq \frac{v(N)}{n}, \forall S \subset N \tag{5}$$

So finally the user partially deviates from R if he discovers at least one set of services S^* that violates (5). For this set we will have that $\frac{v(S^*)}{s^*} > \frac{v(N)}{n} \Rightarrow v(S^*) > \frac{s^*(a-c)^2}{4n}$. In this case u would have an incentive to partially leave R.

3.1 Resource provider's view

Our approach is not only valuable to users but also to a resource provider. This is because under our model the provider can adopt some strategies to fight a potential rejection of a user. Using equation (5) we see that the provider must try to keep the core of the game non-empty. So first of all the provider must decide to provide its services in a cooperative way. We can observe this trend in many resource providers nowadays. For example, Google recently (March 2012) unified its services in order to act cooperatively. In this way there is the trend that any two services will cooperate in order to produce a common value. Thus through Gmail you can view or process your attachments through GoogleDocs or use the Dashboard to synchronize all your e-mail contacts with your Android device. In this manner Google tries to keep its core non-empty and thus give incentives to users for more satisfaction. Of course there is always the case of a user using the services in a non-cooperative manner, but then from (3) he would earn less. Moreover, the provider should try to increase the ratio $\frac{(a-c)^2}{4n}$. This can be done in the following ways: First, the provider can help u to reduce his cost c. This can be achieved in many ways, e.g. by upgrading its hardware, by hiring a qualified service [10], by adopting process completeness strategies [11] or by improving the service tutorials and introducing online help desks [12, 13]. Second, it might consider increasing the factor a, which as we said might represent the size of data or the time available. For example, a social network might strive to attract more users thus increasing a so that the current users will belong to a bigger society and become more satisfied, resulting in more options for interaction. The same idea also applies to the critical mass of Service Overlay Networks [14]. Finally, it might consider reducing the number of services it provides (reducing the denominator of (5) increases the fraction), for example by obsoleting outdated services or not so popular ones.

Another important factor for a provider is to collect user rating information data for its services so as to compute its own estimations of how satisfied the users are. According to (5) the closer the provider's estimation $v_{provider}(S)$ is to the user's one $v_{user}(S)$, the more an effective strategy can be adopted to maintain its customers, and this is because in this way the provider has a clear image of what its users like. Also providers should follow user trends to estimate the potential S^*'s that make its resource vulnerable either by asking for feedback from the users or by outsourcing this task to experts [15, 16]. Moreover, more complex scenarios can be adopted from the provider's point of view in order to further refine its strategies. For this the provider can design more complicated games into which users would engage themselves. For example, the provider will not only consider how to satisfy the user but also how to earn more money, so instead of having the players compete in order to find the equilibria between value produced and cost spent we could have players competing between satisfaction offered and money earned and cost spent.

4 Discussion and future work

As stated previously, the value generated by the players and the cost spent are related to the user's loyalty to the resource. But can the above be calculated by the user? After all, there are many controversial metrics that can be used for rating. For example, user u in the Zoogle+ example might have various concerns when using the resource: how user friendly are Zoogle+'s interfaces? How much storage space am I allowed? What are the privacy policies of Zoogle+? Do most of my friends belong to Zoogle+ too? And there are even more. An approach to these concerns is the following: On the one hand, the value generated through a user - service interaction must be perceived as a combination A) of the amount of information in the form of data created, exchanged, stored, or retrieved by the user, and B) of the user's personal metrics. For example one such metric is the one we adopted as a cost factor, i.e. the amount of the user's time invested to produce the value through interaction (programming, typing, asking queries, etc.). And this is something natural, but other functions can be used too, such as money spent by the user, bandwidth or CPU resources used, or a combination of the above. One can consult [17] and most of the references therein for a recent treatment of QoS properties and measurement metrics of services.

In our cooperative game we used as a notion of fairness one in which the value generated by the players must split equally among them. This is called a game with transferable utilities since user u makes the decision based on the average value calculated by all the plays. This means that the transfering of profit between any two services x, y is allowed, so as to maintain the same average. But since there exist many different notions of fairness such as the Shapley Value, it is of particular interest to extend the analysis to these notions as well.

Finally, since under our model we assumed that each service is somewhat of the same nature as every other service, in a more realistic scenario in which the services differ, we could have used a differentiation parameter γ and the demand from service i would change to $(a - q_i - \gamma \sum_{j=1, j \neq i}^{n} q_j)$, thus resulting in a more complex worth function. This case is a subject of ongoing work.

References

1. M. Hilbert and P. Lopez. *The World's Technological Capacity to Store, Communicate, and Compute Information.* Science, Vol. 332 no. 6025, 60-65, American Association for the Advancement of Science, 2011.

2. https://www.findthemissing.org/en
3. K. Vanthournout, G. Deconinck, and R. Belmans. *A taxonomy for resource discovery.* Personal Ubiquitous Computing, vol 9-2 p. 81-89, Springer-Verlag, 2005.
4. D. Blackmore, K. Hale, J. Harris, F. Ng, and C. Morikawa. *Market Share Analysis: IT Services Rankings, Worldwide, 2009.* Gartner, Inc, 29 April 2010.
5. http://aws.amazon.com/
6. https://one.ubuntu.com/
7. M. Torkjazi, R. Rejaie and W. Willinger. *Hot Today, Gone Tomorrow: On the Migration of MySpace Users.* Proceedings of the 2nd ACM workshop on Online social networks (WOSN), p. 43-48, Barcelona, Spain 2009.
8. M. J. Osborne and A. Rubinstein. *A course in game theory.* MIT Press, 1994.
9. A. A. Cournot. *Researches into the Mathematical Principles of the Theory of Wealth* (English translation of the original). Macmillan, New York, 1897.
10. D. Wanchun, Q. Lianyong, Z. Xuyun, C. Jinjun. *An evaluation method of outsourcing services for developing an elastic cloud platform.* The Journal of Supercomputing, Springer Netherlands, 2010 published online. doi: 10.1007/s11227-010-0491-2
11. G. Piccolia, M. K. Brohmanb, R. T. Watsonc, and A. Parasuramand. *Process completeness: Strategies for aligning service systems with customers service needs.* Computers in Industry, 41, 2, p. 129 - 145, Elsevier, 2000 Business Horizons, 52, 4, p. 367 - 376, Elsevier, 2009.
12. F. Schubert, C. Siu, H. Cheung, L. Peng Chor, and L. Shigong. *An integrated help desk support for customer services over the World Wide Web - a case study.* Computers in Industry, 41, 2, p. 129 - 145, Elsevier, 2000.
13. F. Schubert, C. Siu, H. Cheung, and L. Peng Chor. *Web-based intelligent helpdesk-support environment.* International Journal of Systems Science, 33:6, 389-402, Taylor & Francis, 2002.
14. N. Lam. *Capacity Allocation in Service Overlay Networks.* Ph.D. Dissertation, McGill University (Canada), 2011.
15. C. Benko. *OUTSOURCING EVALUATION. A profitable process.* Information Systems Management, 10, 2, Taylor & Francis, 1993.
16. R. McIvor. *How the transaction cost and resource-based theories of the firm inform outsourcing evaluation.* Journal of Operations Management, 27, 1, p. 45-63, Elsevier 2009.
17. D. A. D'Mello and V. S. Ananthanarayana. *Dynamic selection mechanism for quality of service aware web services.* Enterprise Information Systems, 4:1, 23-60, Taylor & Francis, 2010.

Appendix

We prove below the Nash equilibria for the game described in section 2.2. The first order derivative conditions of (2) $\forall i, i = 1, \cdots, n$ give:

$$\frac{\partial}{\partial q_i}[(a - \sum_{j=1}^{n} q_j - c)q_i] = a - \sum_{j=1,j\neq i}^{n} q_j - c - 2q_i = 0 \Rightarrow q_i = \frac{a - c - \sum\limits_{j=1,j\neq i}^{n} q_j}{2}$$

The system gives: $\tilde{q} \equiv q_1 = \cdots = q_n$, thus $\tilde{q} = \frac{a-c-(n-1)\tilde{q}}{2}$ or $\tilde{q} = \frac{a-c}{n+1}$. Now plugging $\tilde{q}$ in (2) gives the result.

FaCETa: Backward and Forward Recovery for Execution of Transactional Composite WS*

Rafael Angarita[1], Yudith Cardinale[1], and Marta Rukoz[2]

[1] Departamento de Computación, Universidad Simón Bolívar,
Caracas, Venezuela 1080
[2] LAMSADE, Université Paris Dauphine
Université Paris Ouest Nanterre La Défense
Paris, France
{yudith,rangarita}@ldc.usb.ve, marta.rukoz@lamsade.dauphine.fr

Abstract. In distributed software contexts, Web Services (WSs) that provide transactional properties are useful to guarantee reliable Transactional Composite WSs (TCWSs) execution and to ensure the whole system consistent state even in presence of failures. Fails during the execution of a TCWS can be repaired by forward or backward recovery processes, according to the component WSs transactional properties. In this paper, we present the architecture and an implementation of a framework, called FaCETa, for efficient, fault tolerant, and correct distributed execution of TCWSs. FaCETa relies on WSs replacement, on a compensation protocol, and on unrolling processes of Colored Petri-Nets to support fails. We implemented FaCETa in a Message Passing Interface (MPI) cluster of PCs in order to analyze and compare the behavior of the recovery techniques and the intrusiveness of the framework.

1 Introduction

Large computing infrastructures, like Internet increase the capacity to share information and services across organizations. For this purpose, Web Services (WSs) have gained popularity in both research and commercial sectors. Semantic WS technology [20] aims to provide for rich semantic specifications of WSs through several specification languages such as OWL for Services (OWL-S), the Web Services Modeling Ontology (WSMO), WSDL-S, and Semantic Annotations for WSDL and XML Schema (SAWSDL) [15]. That technology supports WS composition and execution allowing a user request be satisfied by a Composite WS, in which several WSs and/or Composite WSs work together to respond the user query.

WS Composition and the related execution issues have been extensively treated in the literature by guaranteeing user QoS requirements and fault tolerant execution [7, 11, 16, 18, 21]. WSs that provide transactional properties are useful to guarantee reliable Transactional Composite WSs (TCWSs) execution, in order to ensure that the whole system remains in a consistent state

* This work was supported by the Franco-Venezuelan CNRS-FONACIT project N°22782

even in presence of failures. TCWS becomes a key mechanism to cope with challenges of open-world software. Indeed, TCWS have to adapt to the open, dynamically changing environment, and unpredictable conditions of distributed applications, where remote services may be affected by failures and availability of resources [27].

Generally, the control flow and the order of WSs execution is represented with a structure, such as workflows, graphs, or Petri-Nets [3, 6, 7, 14]. The actual execution of such TCWS, carried out by an EXECUTER, could be deployed with centralized or distributed control. The EXECUTER is in charge of *(i)* invoking actually WSs for their execution, *(ii)* controlling the execution flow, according to data flow structure representing the TCWS, and *(iii)* applying recovery actions in case of failures in order to ensure the whole system consistence; fails during the execution of a TCWS can be repaired by forward or backward recovery processes, according to the component WSs transactional properties.

In previous works [9, 10] we formalized a fault tolerant execution control mechanism based on Colored-Petri Nets (CPN), which represent the TCWS and the compensation process. In [9] unrolling algorithms of CPNs to control the execution and backward recovery were presented. This work was extended in [10] to consider forward recovery based on WS replacement; formal definitions for WSs substitution process, in case of failures, were presented. In [10], we also proposed an EXECUTER architecture, independent of its implementation, to execute a TCWS following our proposed fault tolerant execution approach.

In this paper, we present an implementation of our EXECUTER framework, called FACETA (FAult tolerant Cws Execution based on Transactional properties), for efficient, fault tolerant, and correct distributed execution of TCWSs. We implemented FACETA in a Message Passing Interface (MPI) cluster of PCs in order to analyze and compare the efficiency and performance of the recovery techniques and the intrusiveness of the framework. The results show that FACETA efficiently implements fault tolerant strategies for the execution of TCWSs with small overhead.

2 TCWS Fault-Tolerant Execution

This Section recall some important issues related to transactional properties and backward and forward recovery, presented in our previous works [7, 9, 10]. We consider that the Registry, in which all WSs are registered with their corresponding *WSDL and OWLS documents*, is modeled as a Colored Petri-Net (CPN), where WS inputs and outputs are represented by places and WSs, with their transactional properties, are represented by colored transitions - colors distinguish WS transactional properties [7]. The CPN representing the Registry describes the data flow relation among all WSs.

We define a query in terms of functional conditions, expressed as input and output attributes; *QoS* constraints, expressed as weights over criteria; and the required global transactional property as follows. A Query Q is a 4-tuple (I_Q, O_Q, W_Q, T_Q), where:

- I_Q is a set of input attributes whose values are provided by the user,

- O_Q is a set of output attributes whose values have to be produced by the system,
- $W_Q = \{(w_i, q_i) \mid w_i \in [0,1]$ with $\sum_i w_i = 1$ and q_i is a QoS criterion$\}$, and
- T_Q is the required transactional property; in any case, if the execution is not successful, nothing is changed on the system and its state is consistent.

A TCWS, which answers and satisfies a Query Q, is modeled as an acyclic marked `CPN`, called CPN-$TCWS_Q$, and it is a sub-net of the Registry `CPN`[1]. The *Initial Marking* of CPN-$TCWS_Q$ is dictated by the user inputs. In this way, the execution control is guided by a unrolling algorithm.

2.1 Transactional Properties

The transactional property (TP) of a WS allows to recover the system in case of failures during the execution. The most used definition of individual WS transactional properties ($TP(ws)$) is as follows [8, 13]. Let s be a WS: s is **pivot** (p), if once s successfully completes, its effects remains forever and cannot be semantically undone (compensated), if it fails, it has no effect at all; s is **compensatable** (c), if it exists another WS s', which can semantically undo the execution of s, even after s successfully completes; s is **retriable** (r), if s guarantees a successfully termination after a finite number of invocations; the **retriable** property can be combined with properties p and c defining **pivot retriable** (pr) and **compensatable retriable** (cr) WSs.

In [11] the following TP of TCWS have been derived from the TP of its component WSs and their execution order(sequential or parallel). Let tcs be a TCWS: tcs is **atomic** ($\boldsymbol{a}$), if once all its component WSs complete successfully, they cannot be semantically undone, if one component WS does not complete successfully, all previously successful component WSs have to be compensated; tcs is **compensatable** (c), if all its component WSs are compensatable; tcs is **retriable** (r), if all its component WSs are retriable; the retriable property can be combined with properties $\boldsymbol{a}$ and c defining **atomic retriable** ($\boldsymbol{a}r$) and **compensatable retriable** (cr) TCWSs.

According to these transactional properties, we can establish two possible recovery techniques in case of failures:

- *Backward* recovery: it consists in restoring the state (or a semantically closed state) that the system had at the beginning of the TCWS execution; i.e., all the successfully executed WSs, before the fail, must be compensated to undo their produced effects. All transactional properties (p, $\boldsymbol{a}$, c, pr, $\boldsymbol{a}r$, and cr) allow backward recovery;
- *Forward* recovery: it consists in repairing the failure to allow the failed WS to continue its execution. Transactional properties pr, $\boldsymbol{a}r$, and cr allow forward recovery.

[1] A marked `CPN` is a `CPN` having tokens in its places, where tokens represent that the values of attributes (inputs or outputs) have been provided by the user or produced by a WS execution.

2.2 Backward Recovery Process: unrolling a Colored Petri-Net

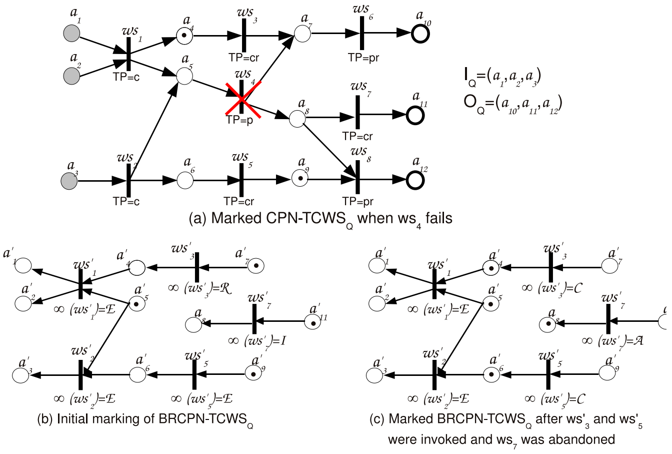
The global TP of CPN-$TCWS_Q$ ensures that if a component WS, whose TP does not allow forward recovery fails, then all previous executed WSs can be compensated by a backward recovery process. For modeling TCWS backward recovery, we have defined a backward recovery `CPN`, called BRCPN-$TCWS_Q$, associated to a CPN-$TCWS_Q$ [9]. The component WSs of BRCPN-$TCWS_Q$ are the compensation WSs, s', corresponding to all c and cr WSs in CPN-$TCWS_Q$. The BRCPN-$TCWS_Q$ represents the compensation flow, which is the inverse of the execution order flow. In BRCPN-$TCWS_Q$ a color of a transition s' represents the execution state of its associated transition s in the CPN-$TCWS_Q$ and is updated during CPN-$TCWS_Q$ execution. Color(s') $\in$ {I='initial', R='running', E='executed', C='compesated', A='Abandonned'}thus, if color(s')='E' means that its corresponding WS s is being executed. In [7, 9] we propose techniques to automatically generate both `CPNs`, CPN-$TCWS_Q$ and BRCPN-$TCWS_Q$.

The execution control of a TCWS is guided by a unrolling algorithm of its corresponding CPN-$TCWS_Q$. A WS is executed if all its inputs have been provided or produced, i.e., each input place has as many tokens as WSs produce them or one token if the user provide them. Once a WS is executed, its input places are unmarked and its output places (if any) are marked.

The compensation control of a TCWS is also guided by a unrolling algorithm. When a WS represented by a transition s fails, the unrolling process over CPN-$TCWS_Q$ is halted,an *Initial Marking* on the corresponding BRCPN-$TCWS_Q$ is set (tokens are added to places associated to input places of the faulty WS s, and to places representing inputs of BRCPN-$TCWS_Q$, i.e., places without predecessors) and a backward recovery is initiated with the unrolling process over BRCPN-$TCWS_Q$. We illustrate a backward recovery in Figure 1. The marked CPN-$TCWS_Q$ depicted in Figure 1(a) is the state when ws_4 fails, the unrolling of CPN-$TCWS_Q$ is halted, and the *Initial Marking* on the corresponding BRCPN-$TCWS_Q$ is set to start its unrolling process (see Figure 1(b)), after ws'_3 and ws'_5 are executed and ws_7 is abandoned before its invocation, a new *Marking* is produced (see Figure 1(c)), in which ws'_1 and ws'_2 are both ready to be executed and can be invoked in parallel. Note that only **compensatable** transitions have their corresponding compensation transitions in BRCPN-$TCWS_Q$.

2.3 Forward Recovery Process: Execution with WS Substitution

During the execution of TCWSs, if a failure occurs in an advanced execution point, a backward recovery may incur in high wasted resources. On the other hand, it is hard to provide a **retriable** TCWS, in which all its components are **retriable** to guaranty forward recovery. We proposed an approach based on WS substitution in order to try forward recovery [10]. TCWS composition and execution processes deal with *service classes* [1], which group WSs with the same semantic functionality, i.e., WSs providing the same operations but having different WSDL interfaces (input and output attributes), transactional support, and *QoS*. When a WS fails, if it is not **retriable**, instead of backward recovery, an substitute WS is searched to be executed on behalf of the faulty WS.



(a) Marked CPN-TCWS$_Q$ when ws$_4$ fails

(b) Initial marking of BRCPN-TCWS$_Q$

(c) Marked BRCPN-TCWS$_Q$ after ws'$_3$ and ws'$_5$ were invoked and ws$_7$ was abandoned

Fig. 1. Example of Backward Recovery

In a *service class*, the functional equivalence is defined according the WSs input and output attributes. A WS s is a functional substitute, denoted by $\equiv_F$, to another WS s^*, if s^* can be invoked with at most the input attributes of s and s^* produces at least the same output attributes produced by s. s is an Exact Functional substitute of s^*, denoted by $\equiv_{EF}$, if they have the same input and output attributes. Figure 2 illustrates several examples: $ws_1 \equiv_F ws_2$, however $ws_2 \not\equiv_F ws_1$, because ws_1 does not produce output a_5 as ws_2 does. $ws_1 \equiv_F ws_3$, $ws_3 \equiv_F ws_1$, and also $ws_1 \equiv_{EF} ws_3$.

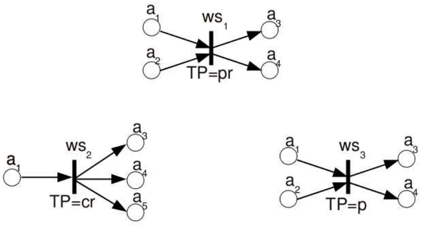


Fig. 2. Example of functional substitute WSs

In order to guarantee the TCWS global TP, a WS s can be replaced by another WS s^*, if s^* can behave as s in the recovery process. Hence, if $TP(s)$=p, in which case s only allows backward recovery, it can be replaced by any other

WS because all transactional properties allow backward recovery. A WS with $TP(s) = pr$ can be replaced by any other **retriable** WS (pr,$\boldsymbol{ar}$,cr), because all of them allow forward recovery. An **atomic** WS allows only backward recovery, then it can be replaced by any other WS which provides backward recovery. A **compensatable** WS can be replaced by a WS that also provides compensation as c and cr WSs. A cr WS can be only replaced by another cr WS because it is the only one allowing forward and backward recovery. Thus, a WS s is Transactional substitute of another WS s^*, denoted by $\equiv_T$, if s is a Functional substitute of s^* and their transactional properties allow the replacement.

In Figure 2, $ws_1 \equiv_T ws_2$, because $ws_1 \equiv_F ws_2$ and $TP(ws_2) = cr$, then ws_2 can behave as a pr WS; however $ws_1 \not\equiv_T ws_3$, even $ws_1 \equiv_F ws_3$, because as $TP(ws_3) = p$, w_3 cannot behave as a pr WS. Transactional substitution definition allows WSs substitution in case of failures.

When a substitution occurs, the faulty WS s is removed from the CPN-$TCWS_Q$, the new s^* is added, but we keep the original sequential relation defined by the input and output attributes of s. In that way, the CPN-$TCWS_Q$ structure, in terms of sequential and parallel WSs, is not changed. For **compensatable** WSs, it is necessary Exact Functional Substitute to do not change the compensation control flow in the respective BRCPN-$TCWS_Q$. In fact, when a **compensatable** WS is replaced, the corresponding compensate WS must be also replaced by the new corresponding one in the BRCPN-$TCWS_Q$. The idea is to try to finish the TCWS execution with the same properties of the original one.

2.4 Protocol in case of Failures

In case of failure of a WS s, depending on the $TP(s)$, the following actions could be executed:

- if $TP(s)$ is **retriable** (pr, $\boldsymbol{ar}$, cr), s is re-invoked until it successfully finish (forward recovery);
- otherwise, another Transactional substitute WS, s^*, is selected to replace s and the unrolling algorithm goes on (trying a forward recovery);
- if there not exist any substitute s^*, a backward recovery is needed, i.e., all executed WSs must be compensated in the inverse order they were executed; for parallel executed WSs, the order does not matter.

When in a *service class* there exist several WSs candidates for replacing a faulty s, it is selected the one with the best quality measure. The quality of a transition depends on the user query Q and on its QoS values. WSs Substitution is done such that the substitute WS locally optimize the QoS. If several transitions have the same value of quality, they can be randomly selected to be the substitute. A similar quality measure is used in [7] to guide the composition process. Then, during the execution, we keep the same heuristic to select substitutes.

3 FaCETa: An TCWS Executer with Backward and Forward Recovery Support

In this Section we present the overall architecture of FaCETa, our execution framework. The execution of a TCWS in FaCETa is managed by an Execu-

TION ENGINE and a collection of software components called ENGINE THREADS, organized in a three levels architecture. In the first level the EXECUTION ENGINE receives the TCWS (represented by a `CPN`). It is in charge of initiating, controlling, and monitoring the execution of the TCWS. To do so, it launches, in the second layer, an ENGINE THREAD for each WS in TCWS. Each ENGINE THREAD is responsible for the execution control of its WS. They receive WS inputs, invoke the respective WS, and forward its results to its peers to continue the execution flow. In case of failure, all of them participate in the backward or forward recovery process. Actual WSS are in the third layer. Figure 3 roughly depicts the overall architecture.

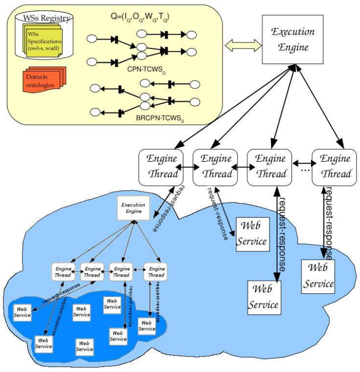
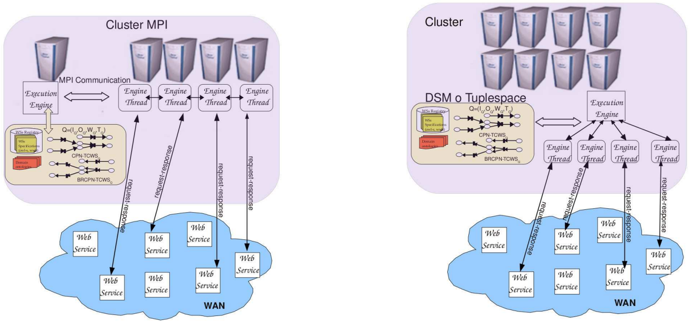


Fig. 3. FACETA Architecture

By distributing the responsibility of executing a TCWS across several ENGINE THREADS, the logical model of our EXECUTER allows distributed execution of a TCWS and is independent of its implementation, i.e., this model can be implemented in a distributed memory environment supported by message passing (see Figure 4(a)) or in a shared memory platform, e.g., supported by a distributed shared memory [22] or tuplespace [19] systems (see Figure 4(b)). The idea is to place the EXECUTER in different physical nodes (e.g., a high available and reliable cluster computing) from those where actual WSS are placed. The EXECUTION ENGINE needs to have access to the WSS Registry, which contains the *WSDL and OWLS* documents. The knowledge required at runtime by each ENGINE THREAD (e.g., WS semantic and ontological descriptions, WSS predecessors and successors,transactional property, and execution control flow) can be directly extracted from the `CPNs` in a shared memory implementation or sent by the EXECUTION ENGINE in a distributed memory implementation. In this paper, we have implemented a prototype of FACETA in a distributed memory platform using MPI.

Typically, a component of a TCWS can be a simple transactional WS or TCWS. Thus, we consider that transitions in the `CPN`, representing the TCWS,



(a) Distributed memory system (b) Distributed shared memory system

Fig. 4. Implementation of FaCETa.

could be WSs or TCWSs. WSs have its corresponding *WSDL and OWLS* documents. TCWSs can be encapsulated into an Executer; in this case the Execution Engine has its corresponding *WSDL and OWLS* documents. Hence, TCWSs may themselves become a WS, making TCWS execution a recursive operation (see Figure 3).

3.1 Distributed Memory Implementation of FaCETa

We implemented FaCETa in a MPI Cluster of PCs (i.e., a distributed memory platform) following a Master/Slaves-SPDM (Single Process Multiple Data) parallel model. The Execution Engine run in the front-end of the Cluster waiting user execution requests. To manage multiple client requests, the Execution Engine is multithreading. The deployment of a TCWS implies several steps: *Initial,* WS *Invocation*, and *Final* phases. In case of failures, recovery phases could be executed: *Replacing* phase, allowing forward recovery or *Compensation* phase for backward recovery.

Whenever the Execution Engine (master) receives a CPN-$TCWS_Q$ and its corresponding BRCPN-$TCWS_Q$, it performs the *Initial* phase: *(i)* start, in different nodes of the cluster, an Engine Thread (peer slaves) responsible for each transition in CPN-$TCWS_Q$, sending to each one its predecessor and successor transitions as CPN-$TCWS_Q$ indicates (for BRCPN-$TCWS_Q$ the relation is inverse) and the corresponding *WSDL and OWLS* documents (they describe the WS in terms of its inputs and outputs, its functional substitute WSs, and who is the compensation WS, if it is necessary); and *(ii)* send values of attributes in I_Q to Engine Threads representing WSs who receive them. Then the master wait for a successfully execution or for a message *compensate* in case of a backward recovery is needed.

Once Engine Threads are started, they receive the part of CPN-$TCWS_Q$ and BRCPN-$TCWS_Q$ that each Engine Thread concerns on, sent by the Execution Engine in the *Initial* phase. Then, they wait for the input values needed to invoke its corresponding WS. When an Engine Thread receives all input values (sent by the master or by other peers) and all its predecessor peers have finished, it executes the WS *Invocation* phase, in which the actual WS is remotely invoked. If the WS finishes successfully, the Engine Thread sends WS output values to Engine Threads representing its successors and wait for a *finish* or *compensate* message. If the WS fails during the execution, the Engine Thread tries a forward recovery: if TP(WS) is **retriable**, the WS is re-invoked until it successfully finish; otherwise the Engine Thread executes the *Replacing* phase: the Engine Thread has to determine the best substitute among the set of functional substitute WSs; it calculates the quality of all candidates according their *QoS* criteria values and the preferences provided in the user query; the one with the best quality is selected to replace the faulty WS; this phase can be executed for a maximum number of times ($MAXTries$). If replacing is not possible, the *Compensation* phase has to be executed: the Engine Thread responsible of the faulty WS sends the message *compensate* to Execution Engine and control tokens to successor peers of the compensation WS, in order to inform about this failure and start the unrolling process over BRCPN-$TCWS_Q$; once an Engine Thread receives all control tokens, it invokes the compensation WS; the unrolling of BRCPN-$TCWS_Q$ ensure the invocation of compensation WSs, s', in the inverse order in which their corresponding WS, s, were executed. Note that forward recovery is executed only by the Engine Thread responsible of the faulty WS, without intervention of the master neither other peers; while backward recovery need the intervention of all of them.

If the TCWS was successfully executed, in the *Final* phase the master receives all values of attributes of O_Q, in which case it broadcasts a *finish* message to all slaves to terminate them, and returns the answer to user; otherwise it receive a *compensate* message indicating that a backward recovery has to be executed, as it was explained above, and return an error message to user.

Assumptions: In order to guarantee the correct execution of our algorithms, the following assumptions are made: *(i)* the network ensures that all packages are sent and received correctly; *(ii)* the Execution Engine and Engine Threads run in a reliable cluster, they do not fail; *(iii)* the Engine Threads receive all WS outputs when its corresponding WS finishes, they cannot receive partial outputs from its WS; and *(iv)* the component WSs can suffer silent or stop failures (WSs do not response because they are not available or a crash occurred in the platform); we do not consider runtime failures caused by error in inputs attributes (e.g., bad format or out of valid range) and byzantine faults (the WS still responds to invocation but in a wrong way).

4 Results

We developed a prototype of FaCETa, using Java 6 and MPJ Express 0.38 library to allow the execution in distributed memory environments. We deployed FaCETa in a cluster of PCs: one node for the Execution Engine and one node for each Engine Thread needed to execute the TCWS. All PCs have the same configuration: Intel Pentium 3.4GHz CPU, 1GB RAM, Debian GNU/Linux 6.0, and Java 6. They are connected through a 100Mbps Ethernet interface.

We generated 10 **compensatable** TCWSs. All those TCWSs were automatically generated by a composition process [7], from synthetic datasets comprised by 800 WSs with 7 replicas each, for a total of 6400 WSs. Each WS is annotated with a transactional property and a set of *QoS* parameters, however for our experiments we only consider the response time as the *QoS* criteria. Replicas of WSs have different response times.

The OWLS-API 3.0 was used to parse the WS definitions and to deal with the OWL classification process.

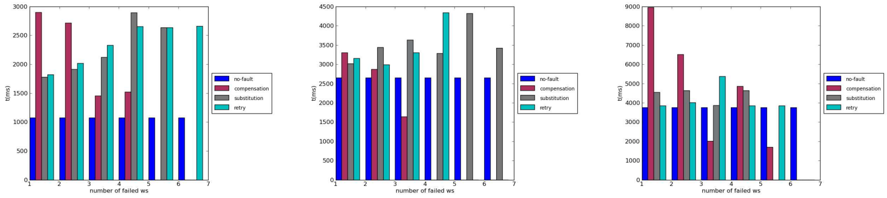
The first group of experiments were focussed on a comparative analysis of the recovery techniques. The second group of experiments evaluates the overhead incurred by our framework in control operations to perform the execution of a TCWS and to execute the fault tolerant mechanisms.

To simulate unreliable environments, we define five different conditions wherein all WSs have the same probability of failure: 0.2, 0.15, 0.1, 0.005, and 0.001. The executions on these unreliable environments were done in three scenarios to support the fails: *(i)* backward recovery (compensation, red bars in Figure 5), *(ii)* forward recovery because all WSs are retriable (retry, light blue bars in Figure 5), and *(iii)* forward recovery (substitution, gray bars in Figure 5). On each scenario all TCWSs were executed 10 times.

Each TCWS was also executed 10 times in a reliable environment, in which all WSs have 0 as probability of failures (no-faulty, blue bars in Figure 5) in order to classify them according their average total execution time in three groups: less than 1500ms (Figure 5(a)), (ii) between 1500ms and 3500ms (Figure 5(b), and (more than 3500ms (Figure 5(c)).

In Figure 5 we plot the average of the total execution time according the number of failed WSs, in order to compare all recovery techniques. The results show that when the number of failed WSs is small (i.e., the probability of failures is less than 20%) backward recovery (compensation) is the worst strategy because almost all component WSs had been executed and have to be compensated. Moreover, when the average of the total execution time of TCWSs is high (greater than 1500ms) forward recovery with retry strategy is better than forward recovery with substitution due to the substitute normally has a bigger response time than the faulty WS. By the other side, in cases in which the probability of failure is greater than 30%, backward recovery whit compensation behaves better than the other ones (even the final results is not produced) because there are many faulty services and only few have to be compensated.

Another issue that can be observed it is the number of outputs received before the backward recovery mechanism has to be executed. In this experiment, the average percentage of outputs received before compensation was 37%. All these outputs are lost or delivered as a set of incomplete (and possibly meaningless and useless) outputs to the user. This percentage is related to the average percentage of compensated services, which is 80%, confirming the overhead, the possible unfulfillment of *QoS* requirements, and the lost outputs. Definitely, backward



(a) Total Exec. Time less than 1500ms

(b) Total Exec. Time between 1500ms and 3500ms

(c) Total Exec. Time more than 3500ms

Fig. 5. Executions on the unreliable environments

recovery should be executed only in absence of another alternative, at early stages of execution of a TCWS, or high faulty environments.

To measure the intrusiveness of FaCETa incurred by control activities, we execute the same set of experiments describe above, but we set to 0 the response time of all WSs. Table 4 shows the average overhead under all different scenarios.

	Average Overhead (ms)	% overhead increased
No Fault	611.7	
Compensation	622.38	2%
Substitution	612.82	0.2%
Retry	612.01	0.05%

Table 1. Average overhead incurred by FaCETa

The average overhead of FaCETa only depends on the number of components WSs in a TCWS. It does not depend on the total response time of TCWS. It means that while the total response time is higher the overhead % will decrease. It is clear that the reason behind the backward recovery strategy overhead (increased by 2%) is the amount of coordination required to start the compensation and the fact that a new WS execution (the compensation WS execution) has to be performed for each successfully executed WS, in order to restore the consistent system state. Additionally, the compensation process has to be done following the unrolling algorithm of the respective BRCPN-$TCWS_Q$. We do not consider to wait before the retry of a failed WS execution; therefore, the increased overhead of retry a WS is almost imperceptible.

As the *service class* for each WS is sent by the Execution Engine in the *Initial* phase, each Engine Thread has the list of the functional substitute WSs sorted according their quality, then there is virtually no overhead when searching for a functional substitute WS to replace a faulty one.

Based on the results presented above, we can conclude that FaCETa efficiently implements fault tolerant strategies for the execution of TCWSs with admissible small overhead.

5 Related Work

Regarding fault tolerant execution of Composite WSs (CWSs), there exist centralized and distributed approaches. Generally centralized approaches [17, 23, 25] consider, besides compensation process, alternative WSs in case of failures or absent WSs, however they extend the classical 2PC protocol, which is time consuming, and they are not transparent to users and developers.

In distributed approaches, the execution process proceeds with collaboration of several entities. We can distinct two kinds of distributed coordination approach. In the first one, nodes interact directly based on a peer-to-peer application architecture and collaborate, in order to execute a CWS with every node executing a part of it [2, 5, 16, 18, 28]. In the second one, they use a shared space for coordination [4, 12, 19].

FENECIA framework [16] provides an approach for managing fault tolerance and *QoS* in the specification and execution of CWSs. FENECIA introduces WS-SAGAS, a transaction model based on arbitrary nesting, state, vitality degree, and compensation concepts to specify fault tolerant CWS as a hierarchy of recursively nested transactions. To ensure a correct execution order, the execution control of the resulting CWS is hierarchically delegated to distributed engines that communicate in a peer-to-peer fashion. A correct execution order is guaranteed in FENECIA by keeping track of the execution progress of a CWS and by enforcing forward and backward recovery. To manage failures during the run-time it allows the execution retrial with alternative candidates. FACTS [18], is another framework for fault tolerant composition of transactional WSs based on FENECIA transactional model. It combines exception handling strategies and a service transfer based termination protocol. When a fault occurs at run-time, it first employs appropriate exception handling strategies to repair it. If the fault cannot be fixed, it brings the TCWS back to a consistent termination state according to the termination protocol (by considering alternative services, replacements, and compensation). In [28] a fault handling and recovery process based on continuation-passing messaging, is presented. Nodes interpret such messages and conduct the execution of services without consulting a centralized engine. However, this coordination mechanism implies a tight coupling of services in terms of spatial and temporal composition. Nodes need to know explicitly which other nodes they will potentially interact with, and when, to be active at the same time. In [2] all replicas of a WS are simultaneously invoked. Only results of the first replica finished are accepted, other executions are halted or ignored. As our work, in [5] a rollback workflow is automatically created considering the service dependencies. Those frameworks support users and developers to construct CWS based on WS-BPEL technologies, then they are not transparent to users and developers.

Another series of works rely on a shared space to exchange information between nodes of a decentralized architecture, more specifically called a tuple space [12, 19]. The notion of a tuplespace is a piece of memory shared by all interacting parties. Using tuplespace for coordination, the execution of a (part of a) workflow within each node is triggered when tuples, matching the templates registered by the respective nodes, are present in the tuplespace. Thus, the templates a component uses to consume tuples, together with the tuples it produces, represent its coordination logic. In [19] approach to replace a centralized BPEL engine by a set of distributed, loosely coupled, cooperating nodes, is presented. This approach presents a coordination mechanism where the data is managed using a tuplespace and the control is driven by asynchronous messages exchanged between nodes. This message exchange pattern for the control is derived from a Petri Net model of the workflow. In [19], the workflow definition is transformed into a set of activities, that are distributed by passing tokens in the Petri Net. In [12] an alternative approach is presented, based on the chemical analogy. Molecules (data) are floating in a chemical solution, and react according to reaction rules (program) to produce new molecules (resulting data). The proposed architecture is composed by nodes communicating through a shared space containing both control and data flows, called the multiset. Through a shared multiset, containing the information on both data and control dependencies needed for coordination, chemical WSs are co-responsible for carrying out the execution of a workflow in the CWS in which they appear. Their coordination is decentralized and distributed among individual WS chemical engine executing a part of the workflow. As this approach, in our approach the coordination mechanism stores both control and data information independent of its implementation (distributed or shared memory). However, none of these works manage failures during the execution.

Facing our approach against all these works, we overcome them because the execution control is distributed and independent of the implementation (it can be implemented in distributed or shared memory platforms), it efficiently executes TCWSs by invoking parallel WSs according the execution order specified by the CPN, and it is totally transparent to users and WS developers, i.e., user only provides its TCWS, that could be automatically generated by the composition process [7] and no instrumentation/modification/specification is needed for WSs participating in the TCWS; while most of these works are based on WS-BPEL and/or some control is sitting closely to WSs and have to be managed by programmers.

There exist some recent works related to compensation mechanism of CWSs based on Petri-Net formalism [21, 24, 26]. The compensation process is represented by Paired Petri-Nets demanding that all component WSs have to be compensatable. Our approach considers other transactional properties (e.g., *pr*, *cr*, ***ar***) that also allows forward recovery and the compensation Petri-Net can model only the part of the TCWS that is compensable. Besides, in those works, the Petri-Nets are manually generated and need to be verified, while in our approach they are automatically generated.

6 Conclusions and Future Work

In this paper we have presented FaCETa, a framework for ensuring *correct and fault tolerant execution order* of TCWSs. The execution model is distributed, can be implemented in distributed or share memory systems, is independent of implementation of WS providers, and is transparent to users and developers. To support failures, FaCETa implements forward recovery by replacing the faulty WS and backward recovery based on a unrolling process over a CPN representing the compensation flow. We have presented a distributed memory implementation of FaCETa in order to compare the behavior of both recovery techniques. The results show that FaCETa efficiently implements fault tolerant strategies for the execution of TCWSs with small overhead.

We are currently working on implementing FaCETa in a distributed shared memory platform in order to test the performance of the framework in centralized and decentralized platforms. Our intention is to compare both implementations under different scenarios (different characterizations of CPNs) and measure the impact of compensation and substitution on QoS.

References

1. Valdino Azevedo, Marta Mattoso, and Paulo Pires, *Handling dissimilarities of autonomous and equivalent web services*, Proc. of Caise-WES, 2003.
2. Johannes Behl, Tobias Distler, Florian Heisig, Rudiger Kapitza, and Matthias Schunter, *Providing fault-tolerant execution of web-service–based workflows within clouds*, Proc. of the 2nd Int. Workshop on Cloud Computing Platforms (CloudCP), 2012.
3. Antonio Brogi, Sara Corfini, and Razvan Popescu, *Semantics-based composition-oriented discovery of web services*, ACM Trans. Internet Techn. **8** (2008), no. 4, 1–39.
4. Paul Buhler and José M. Vidal, *Enacting BPEL4WS specified workflows with multiagent systems*, The Workshop on Web Services and Agent-Based Eng., 2004.
5. Omid Bushehrian, Salman Zare, and Navid Keihani Rad, *A workflow-based failure recovery in web services composition*, Journal of Software Engineering and Applications **5** (2012), 89–95.
6. Yudith Cardinale, Joyce El Haddad, Maude Manouvrier, and Marta Rukoz, *Web service selection for transactional composition*, Elsevier Science-Procedia Computer Science Series (Int. Conf. on Computational Science(ICCS) **1** (2010), no. 1, 2689–2698.
7. Yudith Cardinale, Joyce El Haddad, Maude Manouvrier, and Marta Rukoz, *CPN-TWS: A colored petri-net approach for transactional-qos driven web service composition*, Int. Journal of Web and Grid Services **7** (2011), no. 1, 91–115.
8. Yudith Cardinale, Joyce El Haddad, Maude Manouvrier, and Marta Rukoz, *Transactional-aware Web Service Composition: A Survey*, IGI Global - Advances in Knowledge Management (AKM) Book Series, 2011.
9. Yudith Cardinale and Marta Rukoz, *Fault tolerant execution of transactional composite web services: An approach*, Proc. of The Fifth Int. Conf. on Mobile Ubiquitous Computing, Systems, Services qnd Technologies (UBICOMM), 2011.
10. Yudith Cardinale and Marta Rukoz, *A framework for reliable execution of transactional composite web services*, Proc. of The Int. Conf. on Management of Emergent Digital EcoSystems (MEDES), 2011.
11. Joyce El Haddad, Maude Manouvrier, and Marta Rukoz, *TQoS: Transactional and QoS-aware selection algorithm for automatic Web service composition*, IEEE Trans. on Services Computing **3** (2010), no. 1, 73–85.

12. Hector Fernandez, Thierry Priol, and Cédric Tedeschi, *Decentralized approach for execution of composite web services using the chemical paradigm*, IEEE Int. Conf. on Web Services, 2010, pp. 139–146.
13. Walid Gaaloul, Sami Bhiri, and Mohsen Rouached, *Event-based design and runtime verification of composite service transactional behavior*, IEEE Trans. on Services Computing **3** (2010), no. 1, 32–45.
14. Chad Hogg, Ugur Kuter, and Hector Munoz-Avila, *Learning Hierarchical Task Networks for Nondeterministic Planning Domains*, The 21st Int. Joint Conf. on Artificial Intelligence (IJCAI-09), 2009.
15. Farrell J. and H Lausen, *Semantic annotations for wsdl and xml schema*, January 2007, W3C Candidate Recommendation. http://www.w3.org/TR/sawsdl/.
16. Neila Ben Lakhal, Takashi Kobayashi, and Haruo Yokota, *FENECIA: failure endurable nested-transaction based execution of compo site Web services with incorporated state analysis*, VLDB Journal **18** (2009), no. 1, 1–56.
17. An Liu, Liusheng Huang, Qing Li, and Mingjun Xiao, *Fault-tolerant orchestration of transactional web services*, Web Information Systems – WISE 2006 (Karl Aberer, Zhiyong Peng, Elke Rundensteiner, Y anchun Zhang, and Xuhui Li, eds.), Lecture Notes in Computer Science, vol. 4255, Springer Berlin-Heidelberg, 2006, pp. 90–101.
18. An Liu, Qing Li, Liusheng Huang, and Mingjun Xiao, *FACTS: A Framework for Fault Tolerant Composition of Transactional Web Services*, IEEE Trans. on Services Computing **3** (2010), no. 1, 46–59.
19. Daniel Martin, Daniel Wutke, and Frank Leymann, *Tuplespace middleware for petri net-based workflow execution*, Int. J. Web Grid Serv. **6** (2010), 35–57.
20. S. McIlraith, T.C. Son, and H. H. Zeng, *Semantic web services*, IEEE Intelligent Systems **16** (2001), no. 2, 46–53.
21. Xiaoyong Mei, Aijun Jiang, Shixian Li, Changqin Huang, Xiaolin Zheng, and Yiyan Fan, *A compensation paired net-based refinement method for web services composition*, Advances in Information Sciences and Service Sciences **3** (2011), no. 4.
22. Jesús De Oliveira, Yudith Cardinale, Jesús Federico, Rafael Chacón, and David Zaragoza, *Efficient distributed shared memory on a single system image operating system*, Latin-American Conf. on High Performance Computing, 2010, pp. 1–7.
23. Jonghun Park, *A high performance backoff protocol for fast execution of composite web services*, Computers and Industrial Engineering **51** (2006), 14–25.
24. Fazle Rabbi, Hao Wang, and Wendy MacCaull, *Compensable workflow nets*, Formal Methods and Software Engineering - 12th Int. Conf. on Formal Engineering Methods, LNCS, vol. 6447, 2010, pp. 122–137.
25. M. Schafer, P. Dolog, and W. Nejdl, *An environment for flexible advanced compensations of web service transactions*, ACM Transactions on the Web **2** (2008).
26. Yonglin Wang, Yiyan Fan, and Aijun Jiang;, *A paired-net based compensation mechanism for verifying Web composition transactions*, The 4th Int. Conf. on New Trends in Information Science and Service Science, 2010.
27. Qi Yu, Xumin Liu, Athman Bouguettaya, and Brahim Medjahed, *Deploying and managing web services: issues, solutions, and directions*, The VLDB Journal **17** (2008), 537–572.
28. Weihai Yu, *Fault handling and recovery in decentralized services orchestration*, The 12th Int. Conf. on Information Integration and Web-based Applications #38; Services, iiWAS, ACM, 2010, pp. 98–105.

Discovering Semantic Equivalence of People behind Online Profiles

Keith Cortis, Simon Scerri, Ismael Rivera, and Siegfried Handschuh

Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland
firstname.surname@deri.org

Abstract. Users are currently required to create and separately manage duplicated personal data in heterogeneous online accounts. Our approach targets the crawling, retrieval and integration of this data, based on a comprehensive ontology framework which serves as a standard format. The motivation for this integration is to enable single point management of the user's personal information. The main challenge faced by this approach is the discovery of semantic equivalence between contacts described in online profiles, their attributes and shared posts. Contacts found to be semantically equivalent to persons that are already represented within the user's personal information model are linked together. In this paper we outline our part-syntactic, part-semantic approach to online profile integration, the current status and future plans for research and development concerned with this challenge.

Keywords: semantic equivalence, online profile, personal information model, ontologies, social networks, semantic lifting, semantic web

1 Introduction

At present, the typical computer literate user is forced to create a personal profile for each online account they would like to use. Since the recent shift towards the usage of remote data management and sharing services, this necessity has become even more pressing. Popular online accounts now vary from general social networking platforms to specific email, instant messaging, calendaring, task management and file-sharing services as well as business-oriented customer management services. Personal data in these accounts ranges from the more static identity-related information, to more dynamic information about one's social network as well as physical and online presence. In the context of this paper, we refer to all these kinds of personal data, stored on one of many distinguishable online accounts, as a user's 'online profile'.

The current situation results in personal data being unnecessarily duplicated over different platforms, without the possibility to merge or port any part of it [2], thus forcing users to also manage this data separately and manually. This is reflected in a survey[1] that we conducted, where 16% *always*, 20% *frequently* and 38% *sometimes* use the same personal information within their 'business' (e.g. professional networks) and 'administrative' (e.g. e-commerce) profiles. On the other hand, the 'social/private'

[1] http://smile.deri.ie/node/517

profiles of 12% *always*, 6% *frequently* and 40% *sometimes* contain the same personal information as their business/administrative profiles. Our aim is to enable the user to create, aggregate and merge multiple online profiles into one digital identity, through the di.me[2] userware - a single access point to the user's personal information sphere [25]. The latter also refers to personal data on a user's multiple devices (e.g. laptops, tablets, smartphones). This makes the di.me userware sophisticated and novel since it does not only 'attack' the distributed/duplicated online profile management problem, but targets the integration of distributed/duplicated personal information found across multiple local and remote sources. The already integrated data is stored in the user's personal information sphere which holds all the valuable information of the user on a personal server. The advantage of creating a digital identity within the di.me userware is that of automatically integrating several identities into one with no, or minimal, user effort. This would then enable easier management of the multiple identities, without expecting existing systems to adopt our model or the user to do the integration manually, both of which aren't practical. Results from our survey also outline that 32.7% would *extremely* favour a system that synchronises and shows you personal information collected from different personal online sites, 26.5% favour the idea *quite a bit*, whilst 20.4% are *moderately* in favour. Additionally, 30.6% would *extremely* favour a system that enables you to centrally modify and update your information in different personal sites from one location, 30.6% favour the idea *quite a bit*, whilst 14.3% are *moderately* in favour. These statistics motivate the development of the di.me userware.

In this paper we will only focus on the integration of heterogeneous online user profiles, a task which is not straightforward for two main reasons. First, no common standards exists for modelling profile data in online accounts [20], making the retrieval and integration of federated heterogeneous personal data instantly a hard task. A second problem is that the nature of some of the personal data on digital community ecosystems [13], such as known contacts (resources) and presence information, is dynamic. To address these difficulties, we propose the use of a standard format that is able to handle both the more static as well as dynamic profile data. This comes in the form of an integrated ontology framework consisting of a set of re-used, extended as well as new vocabularies provided by the OSCA Foundation (OSCAF)[3] (only the most relevant ontologies will be mentioned in this paper). Our approach is to map and integrate various online user profiles onto this one standard representation format. The first stage of this approach discovers semi/unstructured information by crawling attributes that are available through online account APIs, resulting in a separate representation for each respective online profile. These representations maintain links to the source account as well as to the external identifiers of the specific online profile attributes. Additionally, all crawled attributes, in our case the profile information, are aggregated into what we refer to as the user's 'super profile'. The second stage of our approach targets the mapping of attributes for each of the represented online profiles with equivalent attributes for the super profile. The use of ontologies and Resource Description Framework (RDF)[4] as the main data representation means that the mapping we pursue considers both syn-

[2] http://www.dime-project.eu/

[3] http://www.oscaf.org/

[4] http://www.w3.org/RDF/

tactic as well as semantic similarities in between online profile data. Our approach is performing semantic lifting and not traditional ontology matching since we are discovering resources from a user's profile (schema of particular online account) which are then mapped to our ontology framework. We then attempt to discover semantic equivalence between persons (this includes both the user and their contacts) that are known in multiple online accounts, based on the results of individual attribute matching. An appropriate semantic equivalence metric is one of the requirements for aspiring self-integrating system [21], such as the di.me userware.

Several techniques may be required in-order to discover if two or more online persons are semantically equivalent. The most popular techniques are syntactic based, i.e. a string/value comparison is performed on the various person profile attributes. Our ontology-based approach allows us to extend the matching capabilities 'semantically', ensuring more accurate results based on clearly-specified meanings of profile attribute types, as well as through an exploration of their semantic (in addition to syntactic) relatedness. The discovery of semantically equivalent person representations results in their semantic integration at the Personal Information Model (PIM) level of the user's data. The PIM handles unique personal data that is of interest to the user such as the user's singular digital identity, files, task lists and emails. It is an abstraction of the possibly multiple occurrences of the same data as available on multiple online accounts and devices. The users have complete control over their accessed personal data, since our approach does not target sharing of personal information. In the remainder of the paper we start by discussing and comparing related work in Section 2. Details on our approach are then provided in Section 3. An update of the current status and prototype implementation is then provided in Section 4, before a list of our targeted future aspirations and a few concluding remarks in Section 5.

2 Related Work

The process of *matching* takes two schemas or ontologies (each of which is made up of a set of discrete entities such as tables, XML[5] elements, classes, properties, etc.) as an input, producing relationships (such as equivalence) found between these entities as output [26]. COMA++ [3] is one of the most relevant schema and ontology matching tools that finds out the semantic correspondences among meta-data structures or models. Given that these matching problems are overcome, it would benefit service interoperability and data integration in multiple application domains. Several techniques and prototypes were implemented in-order to solve the matching problem in a semi-automatic manner such as Falcon-AO (ontology matching) [14], thus reducing manual intervention. Our approach is different to the mentioned traditional approach since we aren't concerned with matching two conceptualisations (schemas or ontologies), but a schema of an online account e.g. a social network to an ontology or set of ontologies. We refer to this process as semantic lifting, since we are lifting semi/unstructured information (the user's profile attributes) from a schema as discussed in Section 3.1, which is manually mapped to an interoperable standard (ontology framework) as discussed in Section 3.2.

[5] http://www.w3.org/XML/

Findings in [15] suggest that provided enough information is available, multiple user profiles can be linked at a relatively low cost. In fact, their technique produces very good results by considering a user's friends list on multiple online accounts. Earlier approaches rely on just a specific Inverse Functional Property (IFP) value e.g. email address or name [17],[12]. However, as pointed out in [5], IFP assumptions are rather shallow, since users are able to create multiple accounts even within the same social network (e.g. a business-related profile, social profile, etc.) each time using different identification, e.g. email addresses.

A number of approaches rely on formal semantic definitions, through the use of ontologies and RDF, to enable portability of online profiles. The work by [22] presents an online application that transforms a user's identity on a social network (Facebook) into a portable format, based on the Friend of a Friend (FOAF) ontology [6]. The approach described in [20] goes on step forward, attempting to integrate multiple online profiles that have been converted to FOAF. As opposed to IFP approaches, this approach takes into consideration all (FOAF) profile attributes, assigning different importance levels and similarity measures to each. Although FOAF enables a much richer means for profile attribute comparison, we use a more comprehensive conceptualisation through the Nepomuk Contact Ontology (NCO) [19], which is integrated into a comprehensive ontology framework. This integration enables attributes in multiple profiles to be semantically related to unique, abstract representations in the user's PIM. Once the technique in [20] sees the profiles transformed to a FOAF representation, a number of techniques are used for syntactic matching between short strings and entire sentences. In addition, the syntactic-based aspect of our matching will also perform a Linguistic Analysis to yield further information about the typed profile attributes. Named Entity Recognition (NER) can discover more specific types than the ones known (e.g. identifying city and country in a postal address) and recognise abbreviations or acronyms in attribute labels.

Many approaches enhance the otherwise syntactic-based profile matching techniques with a semantic-based extension. In particular, the above-cited work by Raad et. al. is supplemented with an Explicit Semantic Analysis [11], whose aim is to detect semantic similarity between profile attributes through the computation of semantic relatedness between text, based on Wikipedia. A similar approach [27] uses snippets returned from an online encyclopedia to measure the semantic similarity between words through five web-based metrics. Our approach will consider semantic relatedness to determine similarity between entities not only based on their labels or values, but also on a semantic distance to other relevant concepts. For example, although an address in one profile might consist of just the city, and another address might refer to only the country, the fact that the city in the first profile is known to be in the country defined for the second profile will be considered as a partial match.

The calculation of such measures within different systems or domains is a very important task, given the increase in access to heterogeneous and independent data repositories [9]. Research efforts conducted by [28] identify three common approaches for calculating the semantic distance between two concepts, namely i) the knowledge-based approach which uses remote Knowledge Bases (KBs) such as WordNet[6] (count edge distance) [7], ii) lexico-syntactic patterns (makes binary decisions), and iii) statisti-

[6] http://wordnet.princeton.edu/

cal measures (uses contextual distributions or concept co-occurrences). The mentioned techniques are not relevant for certain cases, as the concept distances cannot be calculated. This means that such a process is not straightforward, especially if a personal KB is used, where a good distance metric needs personal adjustments in-order to cater for a particular user's needs. Normally for a personal ontology (can be domain specific), in our case the PIM, several concepts are not available within remote KBs. Therefore, it's impossible to calculate the semantic distance between two concepts, if remote KBs are used alone. Hierarchical semantic KBs such as the ones constructed on an "is a" structure, can be fundamental for finding the semantic distance between concepts [16].

There is one major distinction between our approach and the semantic-based approaches described above. Although remote KBs such as DBpedia[7] are to be considered as a backup, the KB on which we initially perform a similarity measure is the user's own PIM. The PIM is populated partly automatically - by crawling data on the user's devices, applications and online accounts, and partly by enabling the user to manually extend the representations of their own mental models. The advantage here is that the PIM contains information items that are of direct interest to the user, and is thus more relevant to the user than external structured or partly structured KBs. Therefore, the semantic matching of profiles is bound to yield more accurate results, based on a KB that is more personal and significantly smaller.

3 Approach

Our online profile (instance) matching approach will involve four successive processes (A-D), as outlined by Fig. 1 and discussed below.

3.1 Retrieval of User Profile Data from Online Accounts

The first step is to retrieve personal information from various online accounts such as Facebook, Twitter and LinkedIn, and is fairly straightforward once the required API calls are known. We target several categories of online profile data such as the user's own identity-related information, their online posts, as well as information about the user's social network, including the identities and posts shared by their contacts.

3.2 Mapping User Profile Data to the Ontology Framework

Once online profile data has been retrieved from an online account, it is mapped to two particular ontologies in our ontology Framework. Identity-related online profile information is stored as an instance of the NCO Ontology, which represents information that is related to a particular contact. The term 'Contact' is quite broad, since it reflects every bit of data that identifies an entity or provides a way to communicate with it. In this context, the contact can also refer to the user's own contact information. Therefore, both the user and their contacts as defined in an online profile are represented as instances of *nco:Contact*. Presence and online post data for the user is stored as instances of the LivePost Ontology (DLPO)[8], a new ontology for the representation of

[7] http://dbpedia.org/

[8] http://www.semanticdesktop.org/ontologies/dlpo/—currently a candidate OSCAF submission

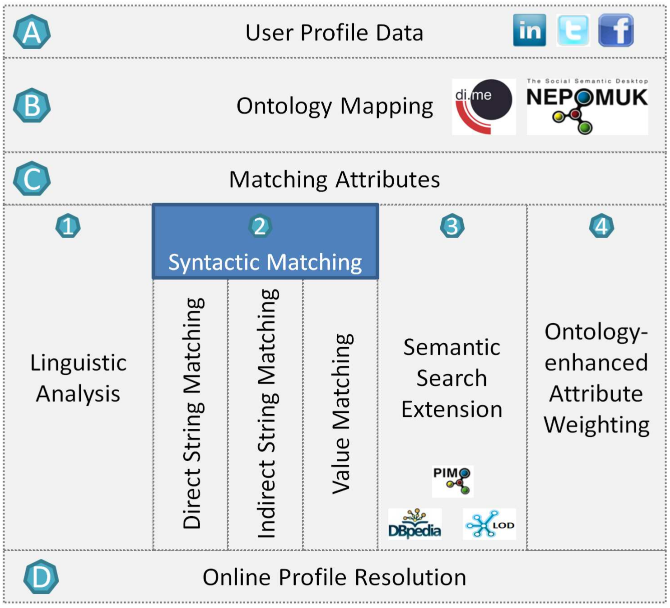


Fig. 1. Approach Process

dynamic personal information that is popularly shared in online accounts, such as multimedia posts (video/audio/image), presence posts (availability/activity/event/checkin), messages (status messages/comments) and web document posts (note/blog posts).

Fig. 2 demonstrates how the above ontologies can be used to store online profile data from an online account (OnlineAccountX). The figure also shows the user's super profile (di.meAccount). An explanation of how the other ontologies in the framework can be used to effectively integrate the two profiles once semantic equivalence is discovered, is provided later on. The upper part of the figure refers to the T-box, i.e. the ontological classes and attributes, whereas the lower part represents the A-box, containing examples of how the ontologies can be used in practice (straight lines between the A- and T-box denote an instance-of relationship).

The attributes of the online user profiles will be mapped to their corresponding properties within our ontology framework. The example shows five identity-related profile attributes that have been mapped to the NCO (affiliation, person name, organisation, phone number, postal address). Presence-related profile information is also available in the form of a complex-type 'livepost', consisting of a concurrent status message - "Having a beer with Anna @ESWC12 in Iraklion", a check-in (referring to the *pimo:City* representation for Heraklion through *dlpo:definingResource*) and an event post (referring to the *pimo:Event* instance representing the conference through *dlpo:definingResource*). *dlpo:definingResource* defines a direct relationship between a 'livepost' subtype and a PIM item. A person, "Anna" is also tagged in this post, as referred by *dlpo:relatedResource*. This property creates a semantic link between a 'livepost' and the relevant PIM items.

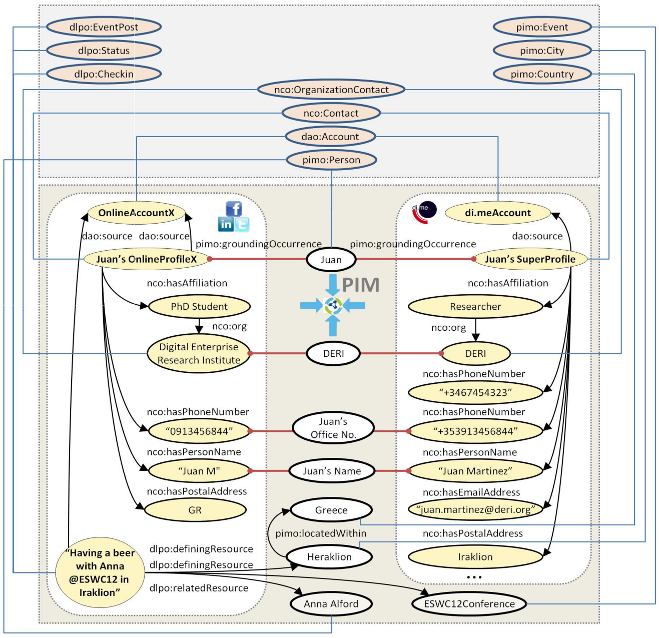


Fig. 2. Approach Scenario

3.3 Matching User Profile Attributes

Our approach towards matching the user profile attributes (metadata matching), considers the data both at a semantic and syntactic level. It involves four successive processes as outlined within the third level (C) of Fig. 1.

Linguistic Analysis Once the transformation and mapping of the user's profile data to its RDF representation has been performed, a matching process is initiated against the user's PIM in order to find similar attributes or links and relations between them. In the case that the profile attribute is known to contain an atomic value (e.g. a person's name, phone number, etc.), no further linguistic analysis is performed. However, profiles attributes may contain more complex and unstructured information such as a postal address (e.g. "42 Upper Newcastle Road, Lower Dangan, Galway, Ireland"). For such attributes, a deeper linguistic analysis is required to discover further knowledge from their values; concretely, a decomposition into different entities or concepts is the goal pursued. In the postal address example, the aim is to find out that '42 Upper Newcastle Road' refers to the most specific part of the address information, 'Lower Dangan' to an area or district, 'Galway' to a city and 'Ireland' to a country. The techniques applied to extract or decompose the attribute values are regular expressions and gazetteer lookups. Typically both techniques work well when the domain or structure is known. Therefore, the algorithm distinguishes profile attributes by type or nature, which is known at this stage, to apply different regular expressions and use different gazetteers. Abbreviations and acronyms are also covered in this analysis by including entries for them in the gazetteers (e.g. a gazetteer for countries also includes the ISO 3166 codes).

Finally, there are special profile attributes which let the user describe themselves, or even include hyperlinks to their personal websites. The text in these attributes is also analysed by a Natural Language Processing pipeline in order to extract named-entities and perform the proper lookups in the gazetteers. However, users normally just provide a hyperlink to their personal website[9], due to size limitations for the description attribute, or the reluctance of the users to enter such information within all their online profiles. In such cases, the hyperlink is resolved and its content is extracted for further analysis, mainly to discover any named-entities (e.g. city) that will enrich the description attribute. Despite of not having these meaningful links between the entities and the profile, such information is used in order to re-balance the weights of certain attributes as described in the Ontology-enhanced Attribute Weighting sub-section below. For example, a Twitter user who only provides a username in her profile would not be able to match to a richer profile which contains her name, postal address, phone number, email, etc., if the username is unknown. However, if this information is found in the form of entities within her personal website, the likelihood that two profiles match increases considerably.

Syntactic Matching Straightforward **value matching** is applied on attributes that have a non-string literal type (e.g. birth date or geographical position), since these have a strict, predefined structure. For attributes of type string (*xsd:string*), if their ontology type (e.g. person name, postal address) is either known beforehand or discovered through NER, **direct string matching** is applied. In both cases, the matching takes as input the attribute in consideration against PIM instances of a similar type. For example, in Fig. 2., the label of the organisation (*nco:OrganizationContact* instance) specified within the *nco:org* property for the user's online account profile (i.e. 'Digital Enterprise Research Institute') is matched against other organisation instances within the PIM. The super profile instance 'DERI' is one example of other PIM instances having the same type. The fact that in this case one of these two equivalent profile organisation attributes is an acronym for the other one will be taken into consideration by the employed string matching technique. In the event that the attribute entity remains unknown even after NER is performed, **indirect string matching** is applied over all PIM instances, regardless of their type.

A string matching metric is used for syntactically matching user profile attribute values that are obtained from an online account to attribute values that are stored in the PIM KB. The recursive field matching algorithm proposed by Monge and Elkan [18] is applied for matching string values. A degree of 1.0 signifies that string 'A' and string 'B' fully match or one string abbreviates the other. On the other hand, a degree of 0.0 signifies that there is no match between two strings. All sub-fields of string 'A' and string 'B' are also compared against each other. The Monge-Elkan string matching metric (1), considered as one of the most accurate [8], is defined as follows:

[9] In an analysis of the 'description' field for 125 Twitter users, we found that 54% were linking to a web page that contains personal information about them.

$$match\left(A,B\right)=\frac{1}{|A|}\sum_{i=1}^{|A|}\max\{sim'\left(A_i,B_j\right)\}_{j=1}^{|B|} \tag{1}$$

—where *sim'* is a particular secondary distance function that is able to calculate the similarity between two tokens 'A' and 'B'. The major reason for choosing this metric is that it holds for matching an attribute value to its abbreviation or acronym, unlike other metrics considered such as Levenshtein distance, Jaro and Jaro-Winkler. This technique considers the abbreviation that is either: i) a prefix, ii) a combination of a prefix and suffix, or iii) an acronym, of the expanded string, or else a concatenation of prefixes from the expanded string. Our plan is to extend this metric to match non-trivial variations of an expanded string e.g. username 'ramauj' to the full name 'Juan Martinez'.

Semantic Search Extension Once the syntactic matching is complete, a semantic search extension process follows. Referring again to our example, the user's address known for the super profile (di.meAccount) is listed as 'Iraklion', and is related to an instance of a *pimo:City*, 'Heraklion'. The one just retrieved from the online account profile (OnlineAccountX) refers to 'GR', which is found to be related to a particular instance of *pimo:Country*, 'Greece'. Although the two address attributes do not match syntactically, they are semantically related. Given that the profile in question is the user's, it is highly likely that through some other data which is either automatically crawled or enriched by the user, the PIM contains references to both these locations, and that semantic relationships exist in between. In the example, through the PIM KB, the system already knows that the city and country instances related to both addresses are in fact related through *pimo:locatedWithin*. This constitutes a partial semantic match, to be taken into consideration when assigning semantic-based attribute weights. If such data didn't exist within the PIM KB (main KB for matching), remote KBs such as DPBedia or any other dataset that is part of the Linked Open Data cloud[10], will be accessed to determine any possible semantic relationship. Another example centres around Juan's two roles, listed as a 'Researcher' for 'DERI' within his super profile, and as a 'PhD Student' for 'Digital Enterprise Research Institute' on his online account. Although less straightforward, a semantic search here would largely support the syntactic search in determining that there is a high match between these two profile attributes, after finding that DERI is a research institute which employs several Researchers and PhD students.

Ontology-enhanced Attribute Weighting To discover semantic equivalence between persons in online profiles or otherwise, an appropriate metric is required for weighting the attributes which were syntactically and/or semantically matched. Factors that will be taken into account by the metric are the total number of attributes that were mapped to our ontology framework, the number of syntactically matched attributes, the number of attributes matched based on the semantic search extension, and the importance of attributes depending on the target domain of the specific online account. In addition, ontology-enhanced attribute weights are an added benefit of our ontology framework

[10] http://lod-cloud.net/

over other ontologies such as FOAF. Attribute constraints defined in the NCO ontology, such as cardinality and inverse functional properties, enable the assignment of different predefined weights to the attributes. Thus, the properties that have a maximum or an exact cardinality of 1 have a higher impact on the likelihood that two particular profiles are semantically equivalent. Carrying even a higher predefined weight are inverse functional properties, which uniquely identify one user. Examples of attributes having cardinality constraints are first name, last name and date of birth, whereas an example of a inverse functional property is a private email address or a cell phone number. Profile attributes such as affiliation, organisation, city and country have no such cardinality constraints defined in the ontology, and as a result they have a lower weight.

3.4 Online Profile Matching

Based on the attribute weighting metric, we define a threshold for discovering semantic equivalence between elements of a user's online profiles, i.e. personal identity and information that is already known and represented at the PIM level. A user can then be suggested to merge all kinds of profile information, e.g. their 'organisation' from various online profiles into their super profile, depending on this threshold. This includes marking contacts for the same unique person as 'known' over multiple online accounts.

The actual integration of semantically-equivalent personal information across distributed locations is realised through the 'lifting' of duplicated data representations onto a more abstract but unique representation in the PIM. The Personal Information Model Ontology (PIMO) [23] provides a framework for representing a user's entire PIM, modelling data that is of direct interest to the user. By definition, PIMO representations are independent of the way the user accesses the data, as well as their source, format, and author. Initially, the PIM will be populated with any personal information that is crawled from a user's particular online account or device. Therefore, if there is no match of a particular entity, a new instance is created. In the example shown in Fig. 2, Juan's PIM (grey area) 'glues' together all the things he works with uniquely, irrespective of their multiple 'occurrences' on different devices and/or online accounts. First and foremost, the PIM includes a representation for the user himself, as a *pimo:Person* instance. This instance refers to the two shown profiles through the *pimo:groundingOccurrence* property, which relates an 'abstract' but unique subject to one or more of its occurrences. For example, the unique *pimo:City* instance has multiple occurrences in multiple accounts, and is related to both Juan's postal address and his check-in as defined on his online account. The advantage of using ontologies is evident here - resources can be linked at the semantic level, rather than the syntactic or format level. For example, although the user's name or organisation differ syntactically, the discovery that they are semantically equivalent is registered within the PIM.

4 Implementation

This section describes the development progress so far. The current prototype employs the Scribe OAuth Java library[11] to retrieve data from a LinkedIn profile. Scribe supports

[11] https://github.com/fernandezpablo85/scribe-java

Table 1. Ontology Mapping of LinkedIn Attributes

Query	Attribute	Ontology Mapping
http://api.linkedin.com/v1/people/∼:(id,first-name,last-name,location:(name),picture-url,summary,positions,phone-numbers,im-accounts,date-of-birth)	id	$\xrightarrow{nao:externalIdentifier}$ <value>
	first-name	$\xrightarrow{nco:hasPersonName}$ nco:PersonName $\xrightarrow{nco:nameGiven}$ <value>
	last-name	$\xrightarrow{nco:hasPersonName}$ nco:PersonName $\xrightarrow{nco:nameFamily}$ <value>
	location:(name)	$\xrightarrow{nco:hasLocation}$ geo:Point $\xrightarrow{nao:prefLabel}$ <value>
	picture-url	$\xrightarrow{nco:photo}$ <value>
http://api.linkedin.com/v1/people/∼/connections:(id,first-name,last-name,location:(name),picture-url,summary,positions,phone-numbers,im-accounts,date-of-birth)	summary	$\xrightarrow{nao:description}$ <value>
	positions	$\xrightarrow{nco:hasAffiliation}$ nco:Affiliation $\xrightarrow{nco:title/role/department/org}$ <value>
	phone-numbers	$\xrightarrow{nco:hasPhoneNumber}$ nco:PhoneNumber $\xrightarrow{nco:phonenumber}$ <value>
	im-accounts	$\xrightarrow{nco:hasIMAccount}$ nco:IMAccount $\xrightarrow{nco:imAccountType/imID}$ <value>
	date-of-birth	$\xrightarrow{nco:hasBirthDate}$ nco:BirthDate $\xrightarrow{nco:birthdate}$ <value>
http://api.linkedin.com/v1/people/∼:(current-share)	id	$dlpo : LivePost$ $\xrightarrow{nao:externalIdentifier}$ <value>
	timestamp	$dlpo : LivePost$ $\xrightarrow{dlpo:timestamp}$ <value>
	comment	$dlpo : LivePost$ $\xrightarrow{dlpo:textualContent}$ <value>
	source name	$dlpo : LivePost$ $\xrightarrow{dao:source}$ dao:Account $\xrightarrow{nao:prefLabel}$ <value>

major 1.0a and 2.0 OAuth APIs such as Google, Facebook, LinkedIn, Foursquare and Twitter, and can thus be used to extend future profiles.

Table 1 shows different types of LinkedIn service calls that our prototype supports (column one). The first retrieves a user's profile data, whereas the second retrieves a user's contact profile data. The third query retrieves status updates from the user. The calls return a set of LinkedIn profile data for the user or their connections, of which we currently map the shown list (column two) to the specific concepts and properties in our ontology framework (column three). The first set of ontology properties in the third column are attached to the *nco:Contact* instance representing the user or one of their contacts (omitted from the Table), whereas the second set of ontology properties are attached to the respect *dlpo:LivePost* instance. Both instances are linked to the online account from which they where retrieved via *dao:source*, this case being a representation of the LinkedIn online account.

Since the LinkedIn API[12] data is returned in XML, we required a transformation of this data into an RDF representation, for mapping to our ontologies. The translation between XML to RDF is quite a tedious and error-prone task, despite the available tools and languages. Although an existing approach is to rely on Extensible Stylesheet Language Transformations (XSLT)[13], the latter was designed to handle XML data, which in contrast to RDF possesses a simple and known hierarchical structure. Therefore, we use the XSPARQL [1] query language. XSPARQL (W3C member submission) provides for a more natural approach based on merging XQuery[14] and SPARQL[15] (both W3C Recommendations). The transformation between the XML LinkedIn data into our RDF representation (using Turtle[16] as the serialization format) is declaratively expressed in a

[12] https://developer.linkedin.com/rest

[13] http://www.w3.org/TR/xslt

[14] http://www.w3.org/TR/xquery/

[15] http://www.w3.org/TR/sparql11-query/

[16] http://www.w3.org/TR/turtle/

XSPARQL query, which also covers the transformation of profile data from any social network adherent to the OpenSocial standard[17].

For the linguistic analysis and NER process, presented in Section 3.3, we have selected the General Architecture for Text Engineering (GATE)[18] platform, which allows decomposing complex processes—or *'pipeline'*—into several smaller tasks or modules with a specific purpose or using a specific (even third-party) algorithm. GATE is is distributed with the ANNIE information extraction system [10], which includes a variety of algorithms for sentence analysis and pre-defined gazetteers for common entity types (e.g. countries, organizations, etc.), which we extended with acronyms or abbreviations where necessary. We employ the GATE *Large KB Gazetteer* module in order to make use of the information stored within the user's PIM, since it can get populated dynamically from RDF data.

Listing 1.1 shows an example of online profile data retrieved from the LinkedIn account for user "Juan Martinez". The RDF representation (in Turtle syntax) shows how the data is mapped to our ontology framework, through the XSPARQL transformer. The LinkedIn account representation (_:acct1 as an instance of *dao:Account*) contains references to two contacts known within (_:c1, _:c2 as instances of *nco:Contact*), one of which (_:c1) is the Juan's own contact representation. Shown attached to Juan's contact instance is a series of identity-related information as well as one status message post (instance of *dlpo:Status*). This example highlights the comprehensiveness of our integrated ontology framework in dealing with various types of online profile data, when compared to other integrated ontology approaches such as the use of FOAF and Semantically-Interlinked Online Communities [4]. More importantly, it also illustrates how integration of online profile data is achieved at the semantic level. Once the two contacts in the online profile (including the one for the user) are discovered to be semantically equivalent to persons that are already represented in the PIM, a link is created between them through *pimo:groundingOccurrence*. The PIM Metadata at the bottom of Listing 1.1 demonstrates how the same unique person representations at the level of the PIM can point to multiple occurrences for that person, e.g. contacts for that person as discovered in online accounts, including the ones just retrieved from LinkedIn.

5 Conclusions and Future Work

In this paper, we discuss the possibility of eliminating the need for the user to separately manage multiple digital identities in unrelated online accounts. Our approach targets the crawling and retrieval of user profile data from these accounts, and their mapping onto our comprehensive ontology framework, which serves as a standard format for the representation of profile data originating from heterogeneous distributed sources. Our main target is the discovery of semantic equivalence between contacts described in online profiles, through a metric which computes a weighted semantic similarity of their individual attributes. Aggregated profile data is lifted onto a unique PIM representation and integrated in a super profile. The integrated data in the PIM is the main KB for matching since it's personalised and thus contains the most valuable information about

[17] http://code.google.com/apis/opensocial/

[18] http://gate.ac.uk/

```
#LinkedIn Profile Metadata
_:acct1 a dao:Account .
_:acct1 nao:prefLabel "LINKEDINAccount" .
_:c1 a nco:Contact .
_:c1 dao:source _:acct1 .
_:c1 nao:externalIdentifer "J7qb-67bTP" .
_:c1 nco:hasPersonName _:cn12 .
_:cn12 a nco:PersonName .
_:cn12 nco:nameGiven "Juan" .
_:cn12 nco:nameFamily "Martinez" .
_:c1 nco:hasAffiliation _:pos8 .
_:pos8 a nco:Affiliation .
_:pos8 nao:externalID 224093780 .
_:pos8 nco:role "Strategy Manager" .
_:pos8 nco:start "2003-1-1T00:00:00Z" .
_:pos8 nco:org _:org16 .
_:org16 a nco:OrganizationContact .
_:org16 nie:title "Ingeneria Ltd." .
...
_:stms644819790 a dlpo:Status .
_:stms644819790 dao:source _:acct1 .
_:stms644819790 nao:externalIdentifier "s6448190" .
_:stms644819790 dlpo:timestamp "2011-10-26T21:32:52" .
_:stms644819790 dlpo:textualContent "Seeking Job" .
...
_:c2 a nco:Contact .
_:c2 dao:source _:acct1 .
_:c2 nco:hasPersonName _:cn22 .
_:cn22 a nco:PersonName .
_:cn22 nco:nameGiven "Anna" .
_:cn22 nco:nameFamily "Alford" .
...

#PIM Metadata
_:PIM a pimo:PIM .
_:PIM pimo:creator _:user .
_:user a pimo:Person .
_:user pimo:groundingOccurrence _:c1 .
_:user pimo:groundingOccurrence _:c23 .
_:user pimo:groundingOccurrence _:c18 .
...
_:user foaf:knows _:person35 .
_:person35 pimo:groundingOccurrence _:c2 .
_:person35 pimo:groundingOccurrence _:c53 .
...
```

Listing 1.1. User Profile Transformer Output and PIM Integration

the user. One of the motivations for the di.me userware (currently in development), is to enable the user a single entry-point to distributed personal information management. This would enable easier management for the user, where no or minimal user effort would be required for the integration of such personal information.

The current prototype is able to retrieve a user's profile data from LinkedIn, but more online accounts are being targeted. The technology for syntactic matching will be further improved through linguistic analysis. Our most challenging future enhancement is the envisaged semantic extension to the current syntactic-based profile attribute matching. Research contributions will on the other hand focus on defining an appropriate semantic-based attribute weighting for each matched attribute, together with the definition of a metric which takes into account all the resulting weighted matches and the identification of a threshold that determines whether two or more online profile refer to the same person. Online posts are also taken into consideration [24]. An analysis of posts from multiple accounts can help us discover whether two or more online profiles are semantically equivalent.

Finally, a comprehensive evaluation of our system would be performed on three levels — i) syntactic matching, ii) semantic matching, and iii) a combination of. This would help determine whether our part-syntactic, part-semantic approach actually yields better results.

Acknowledgements. This work is supported in part by the European Commission under the Seventh Framework Program FP7/2007-2013 (*digital.me* – ICT-257787) and in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (*Líon-2*).

References

1. W. Akhtar, J. Kopecky, T. Krennwallner, and A. Polleres. Xsparql: Traveling between the xml and rdf worlds and avoiding the xslt pilgrimage. In *Proc. 5th European Semantic Web Conference (ESWC2008)*, pages 432–447, Berlin, Heidelberg, 2008.
2. D. Appelquist, D. Brickley, M. Carvahlo, R. Iannella, A. Passant, C. Perey, and H. Story. A standards-based, open and privacy-aware social web. W3c incubator group report, W3C, december 2010.
3. D. Aumueller, H. Do, S. Massmann, and E. Rahm. Schema and ontology matching with coma++. In *Proc. ACM SIGMOD international conference on Management of data*, pages 906–908, New York, NY, USA, 2005.
4. D. Berrueta, D. Brickley, S. Decker, S. Fernndez, C. Grn, A. Harth, T. Heath, K. Idehen, K. Kjernsmo, A. Miles, A. Passant, A. Pollares, and L. Polo. Sioc core ontology specification. Technical report, 2010.
5. S. Bortoli, H. Stoermer, P. Bouquet, and H. Wache. Foaf-o-matic - solving the identity problem in the foaf network. In *Proc. Fourth Italian Semantic Web Workshop (SWAP2007)*, 2007.
6. D. Brickley and L. Miller. Foaf vocabulary specification 0.98. Technical report, 2010.
7. A. Budanitsky and G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proc. Workshop on Wordnet and other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.
8. W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proc. IJCAI-03 Workshop on Information Integration (IIWeb-03)*, pages 73–78, Acapulco, Mexico, Aug.9-10 2003.
9. V. Cross. Fuzzy semantic distance measures between ontological concepts. In *Proc. Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the*, volume 2, pages 635–640, june 2004.
10. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.
11. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. Twentieth International Joint Conference for Artificial Intelligence (IJCAI-07)*, pages 1606–1611, Hyderabad, India, Jan.6-12 2007.

12. J. Golbeck and M. Rothstein. Linking social networks on the web with foaf: A semantic web case study. In *Proc. Twenty-Third Conference on Artificial Intelligence (AAAI'08)*, pages 1138–1143, Chicago, Illinois, USA, 13-17 2008.
13. M. Ion, L. Telesca, F. Botto, and H. Koshutanski. An open distributed identity and trust management approach for digital community ecosystems. In *Proc. International Workshop on ICT for Business Clusters in Emerging Markets, June 2007. Michigan State University*, 2007.
14. N. Jian, W. Hu, G. Cheng, and Y. Qu. Falcon-ao: Aligning ontologies with falcon. In *Proc. K-Cap 2005 Workshop on Integrating Ontologies. (2005)*, pages 87–93, 2005.
15. S. Labitzke, I. Taranu, and H. Hartenstein. What your friends tell others about you: Low cost linkability of social network profiles. In *Proc. 5th International ACM Workshop on Social Network Mining and Analysis*, San Diego, CA, USA, Aug.20 2011.
16. Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15:871–882, 2003.
17. P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3(2-3):211–223, 2005.
18. A. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *Proc. Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270, 1996.
19. A. Mylka, L. Sauermann, M. Sintek, and L. van Elst. Nepomuk contact ontology. Technical report, 2007.
20. E. Raad, R. Chbeir, and A. Dipanda. User profile matching in social networks. In *Proc. 13th International Conference on Network-Based Information Systems, 2010*, pages 297–304, Takayama, Gifu Japan, 2010.
21. S. R. Ray. Interoperability standards in the semantic web. *Journal of Computing and Information Science in Engineering, ASME*, 2:65–69, 2002.
22. M. Rowe and F. Ciravegna. Getting to me: Exporting semantic social network from facebook. In *Proc. Social Data on the Web Workshop, International Semantic Web Conference*, 2008.
23. L. Sauermann, L. van Elst, and K. Mller. Personal information model (pimo). Oscaf recommendation, OSCAF, february 2009.
24. S. Scerri, K. Cortis, I. Rivera, and S. Handschuh. Knowledge discovery in distributed social web sharing activities. In *Making Sense of Microposts (#MSM2012)*, pages 26–33, 2012.
25. S. Scerri, R. Gimenez, F. Herman, M. Bourimi, and S. Thiel. digital.me - towards an integrated personal information sphere. In *Proc. Federated Social Web Europe Conference (FSW 2011)*, Berlin, Germany, 2011.
26. P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics IV*, 3730:146–171, 2005.
27. S. A. Takale and S. S. Nandgaonkar. Measuring semantic similarity between words using web documents. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 1(4), 2010.
28. H. Yang and J. Callan. Learning the distance metric in a personal ontology. In *Proc. 2nd international workshop on Ontologies and information systems for the semantic web*, pages 17–24, New York, NY, USA, 2008.

Nenad Stojanvoic, Ljiljana Stojanovic and Opher Etzion
(Editors)

Proceedings

of the

7th International Workshop on Semantic Business Process Management (SBPM2012)

In conjunction with the

9th Extended Semantic Web Conference (ESWC2012)

May 27, 2012, Heraklion, Greece

Organizing Committee:

Nenad Stojanovic (FZI Research Center for Information Technologies, Karlsruhe)
Ljiljana Stojanovic (FZI Research Center for Information Technologies, Karlsruhe)
Opher Etzion (IBM Research, Haifa)

Workshop URI: http://sbpm2012.fzi.de

Preface

The 7th International workshop on Semantic Business Process Management continues the tradition of previous workshops from this serial: collecting and discussing new semantic-based approaches and ideas for a more efficient BPM. This year the most important issue is the semantic-based dynamicity in BPM systems, which very nice correlates with the predictions that Gartner, Inc. for the next period:

By 2012, 20 per cent of customer-facing processes will be knowledge-adaptable and assembled just in time to meet the demands and preferences of each customer, assisted by BPM technologies.

> Today's capability to proactively change processes is merely an interim step for process improvement. The next evolution will be processes that self-adjust based on the sensing of patterns in user preferences, consumer demand, predictive capabilities, trending, competitive analysis and social connections.

By 2013, dynamic BPM will be an imperative for companies seeking process efficiencies in increasingly chaotic environments.

> IT organisations are striving to become better aligned with the demands placed on the business. Pressure to reduce the latency of change in business processes is driving a need for more dynamic and systematised measures. Adopting a more dynamic form of BPM, which focuses on enabling process changes to occur when and as needed will enable organisations to better respond to unanticipated change requirements in business processes, and to handle process changes more effectively.

More than a half of papers are dealing with the different approaches for managing these dynamics in business processed by using semantic technologies. Rest of the papers is targeting the dynamicity from the modeling point of view (design time). Therefore, this workshop is a nice opportunity for exchanging ideas and starting new cooperation in this domain, as it has been in the past.

Table of Contents

Enabling Semantic Search in Large Open Source Communities

Gregor Leban, Lorand Dali, Inna Novalija

Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana
{gregor.leban, lorand.dali, inna.koval}@ijs.si

Abstract. This paper describes methodology used for building a domain specific ontology. Methods that allow automatic concept and relation extraction using domain-related glossaries are presented in this research. The constructed ontology contains vocabulary related to computer science and software development. It is used for supporting different functionalities in the ALERT project, which aims to improve software development process in large open source communities. One of the uses of the ontology is to provide semantic search functionality, which is a considerable improvement over the keyword search that is commonly supported.

1 Introduction

Open source communities and software development organizations in general are often using several different communication channels for exchanging information among developers and users. Beside a source code management system (SCMS), these software developing communities also frequently use an issue tracking system (ITS), a forum, one or more mailing lists and a wiki. Each of these channels typically serves a different purpose. Issue tracking system allows the users of the software to report to the developers issues they encountered or to suggest new features. Forums and mailing lists have a similar purpose which is to allow open discussions between the members of the community. Wikis are commonly used as platforms for providing software documentation, user guides and tutorials for the users.

A problem that is common for the present day and that is becoming more and more troubling for large and medium open source communities is the information overload. Users are generating large amounts of information on different communication channels and it is difficult to stay up-to-date. For illustration, consider the KDE community [1], which is offering a wide range of open source products. In April 2012 KDE had approximately 290.000 bug reports in their ITS, 126.000 posts on their forum and 163 active mailing lists where according to one of the KDE developers between 30-80 emails are exchanged per day.

Providing help in processing and managing such large amounts of information is one of the main goals of the ALERT[1] project [2]. ALERT is a European project that

[1] ALERT is acronym for Active support and reaL-time coordination based on Event pRocessing in open source software development

aims to develop a system, which will be able to help users (especially developers) in large open source communities. The system, once finished, will be able to collect and process all the posts (emails, issues, forum posts, etc.) that are generated in different communication channels used by the community. The information will be processed and stored in a way so that it will provide support for different features of the system.

One of the features that were specified as very important by the use-case partners of the project was advanced search functionality. Search that is supported on communication channels such as ITS and forums is only a simple Boolean keyword search. A fundamental problem with keyword search is that different people use different words (synonyms) to refer to the same concept. Therefore not all posts that discuss the same concept can be retrieved by a simple keyword-based search.

As an improvement to keyword search we would like to provide in the ALERT system *semantic* search. What we mean by semantic is that the search is performed using the actual concept that the search term represents. Consider, for example, that the user performs a search for “dialog”. The concept of the dialog can be represented also with other terms, such as “window” or “form”. Instead of returning the results that directly mention “dialog” we therefore also want to return results that contain any of the term synonyms. Additionally, since search is based on the actual concepts we can also exploit the fact that concepts can be related to each other. When searching for one concept we can therefore also consider including results about some closely related concepts. In KDE domain, for example, searching for “email client” should also return posts containing concept “KMail” which is the KDE’s email application.

In order provide semantic search we have to use an ontology. Each class in the ontology should represent a concept that can have one or more labels (synonyms). When a new post is created in one of the communication channels, the ALERT system annotates or tags it with the concepts that are mentioned in the post. These annotations are stored in a knowledge base, which allows us then to quickly find all posts tagged with a particular concept.

The main question that needs answering is what ontology should be used for annotating the posts. Since we are providing support for software developing communities the important concepts that should be annotated are the ones related to computer science and software development. Since we were not able to find any such existing ontology we had to construct it ourselves. The process that we used to construct such an ontology is the main contribution of the paper. The steps in the process are general and can be reused also for constructing other domain specific ontologies.

The remainder of this paper is organized as follows. Section 2 provides details of the methodology used for building the Annotation ontology. The process consists of two main steps – (a) identifying the computer science specific terminology that we wish to represent in the ontology, and (b) constructing the relations between the concepts. In Section 3 we describe how the ontology can be used in ALERT to provide the semantic search functionality. Section 4 describes related work and Section 5 provides the conclusions.

2 Building the Annotation ontology

2.1 Creating ontology concepts

As stated before, the concepts that we wish to have in the Annotation ontology are related to computer science and software development. In order to obtain a relevant set of terms we searched online for glossaries related to computer science. The two web sites that we found to be most up-to-date and relevant for our purpose were *webopedia.com* [5] and *whatis.techtarget.com* [6]. For each of the terms we were also able to obtain a description of the term which in most cases contained links to several related terms. To identify terms especially related to software development we used the *stackoverflow* website [7], which is a Q-A system with more than 2.5 million questions related to software development. *Stackoverflow* contains an up-to-date list of tags that are used to tag the questions. Most popular tags together with their descriptions were also included in the starting set of concepts.

After obtaining the set of terms, our first goal was to merge synonyms. Merging of terms was performed in two ways. First way was using a synonym list that we were able to obtain from the *stackoverflow* website and which contained around 1,400 synonym pairs. The second way was by using the term descriptions and searching in them for patterns such as “X also known as Y” or “X (abbreviated Y)”. In cases when such patterns were identified, terms X and Y can be considered as synonyms and be represented as the same concept. In this way we obtained a set of concepts where each concept has one or more labels that represent the concept.

In the next step we wanted to link the concepts to corresponding Wikipedia articles. This allows us to obtain more information about the concepts and potentially also extend the ontology with new related concepts. By using a semi-automatic approach, we make the repetition of the process relatively easy to do, such that future updates of the ontology are not too costly. By identifying a corresponding Wikipedia article we are also able to implicitly create links to well-known knowledge bases which are extracted from Wikipedia, such as DBpedia, Yago and Freebase.

Our approach for mapping concepts to Wikipedia articles has several steps. First, we link the concept labels to Wikipedia articles. We do this by automatically matching the labels to the titles of articles to see which article corresponds to each label. In this process, the following two challenges were identified:

a) The article with the matching title is a disambiguation page i.e. a page containing links to pages which each describe one of the meanings of the concept. For example *TTL* is mapped to a page which contains links to *Time to Live*, *Transistor Transistor Logic*, *Taiwan Tobacco and Liquor*, etc.

b) Some of the computer science concepts are so frequently used in common language that they are not considered ambiguous. In this case a computer science concept can be mapped on Wikipedia to something completely unrelated to computer science. An example of such a concept is *ant* which in computer science refers to *Apache ant*, a software tool for automatic build processes, but is mapped to the Wikipedia article about the ant insect.

The disambiguation pages were not difficult to identify since they typically contain phrases such as ‘*X may mean…*’ or ‘*X may refer to…*’. We have defined rules to automatically match these patterns and exclude disambiguation pages from further analysis.

After mapping labels to the corresponding Wikipedia pages we used the content of these pages to identify new terms which were not covered by the glossaries. To do this, we only used the first paragraph of each article, which usually gives a short definition of the term. Often it also contains links to articles describing closely related concepts. We used the articles linked in the first paragraph as candidates for new terms and sorted them by their frequency. We expect that if an article was linked to by many articles that we know are about computer science, then this article is very likely about a computer science concept as well. Based on this assumption, titles of the frequently appearing articles were added to the ontology as new concepts.

After obtaining the final set of concepts we also wanted to organize the concepts into a hierarchy of categories. For this purpose we used text mining techniques and in particular the OntoGen [8] toolbox which interactively uses k-means clustering [9] to group the concepts into a hierarchy and extracts keywords to help the user in assigning a name to each category. In this way we were able to semi-automatically define 31 categories such as “operating systems”, “programming languages” and “companies”.

2.2 Creating relations between the ontology concepts

An important part of the ontology are also relations between the concepts. With regard to our task of semantic search, the relations allow us to expand the search to also include closely related concepts.

To create the relations between the concepts we can use the information that was available on the online glossaries. As we mentioned, the descriptions of the terms usually contained several links to other related terms. These links can be used to automatically create relations between the corresponding ontology concepts. Since hyperlinks don’t contain any additional semantic information about the type of relation we can only create some general kind of relation between the concepts. In our ontology we represented them using a *linksTo* relation.

In order to obtain more specific and usable relations we decided to apply natural language processing (NLP) techniques on term descriptions with the goal of identifying semantic relations between the concepts. Consider, for example, the following sentence from the “C#” concept description:

> “C# is a high level, general-purpose object-oriented programming language created by Microsoft. “

If “C#” and “Microsoft” are concepts in the ontology then it is possible using NLP techniques to identify that the verb connecting the two concepts is “created by”. The task of creating relations between concepts is in this way reduced to simply defining a mapping from verbs to appropriate relations.

A detailed list of steps involved in creating the relations is as follows. The input to the procedure was the list of ontology concepts and all the descriptions of the concepts. First we identify in the descriptions sentences that mention two or more concepts. Next we use Stanford parser [10] to generate a dependency parse of the sentence. The dependencies provide a representation of grammatical relations between words in a sentence. Using the dependency parse and the co-occurring ontology concepts, we can extract the path from one ontology concept to another one. As a next step, we used Stanford Part-Of-Speech (POS) tagger [11] to tag the words in the sentence. Of all the tags we are only interested in the verb (with or without preposition) that connects the two concepts. As a result we can obtain triples, such as:

XSLT, used by, XML schema
WSDL, describes, Web Service
Microsoft, created, Windows
Apple Inc., designed, Macintosh

In the next step we use WordNet [12] and group the obtained verbs into synsets (synonym sets). From all the sentences we obtained verbs that can belong to around 750 different WordNet synsets. Of all these synsets we only considered those that can be mapped to relations *isPartOf, hasPart, creator* and *typeOf.* We decided to include these relations because they are mostly hierarchical and can be used to expand the search conditions. WordNet synsets that were used to obtain these relations were:

- *isPartOf* and *hasPart* relations were obtained from “include” and “receive have” synsets
- *creator* relations were obtained from “make create”, “form constitute make”, “implement”, “construct build make”, “produce bring forth”, “introduce present acquaint”, “make do” and “plan project contrive design” synsets
- *typeOf* relations were obtained from “establish base ground”, “include”, “exist be” and “integrate incorporate” synsets

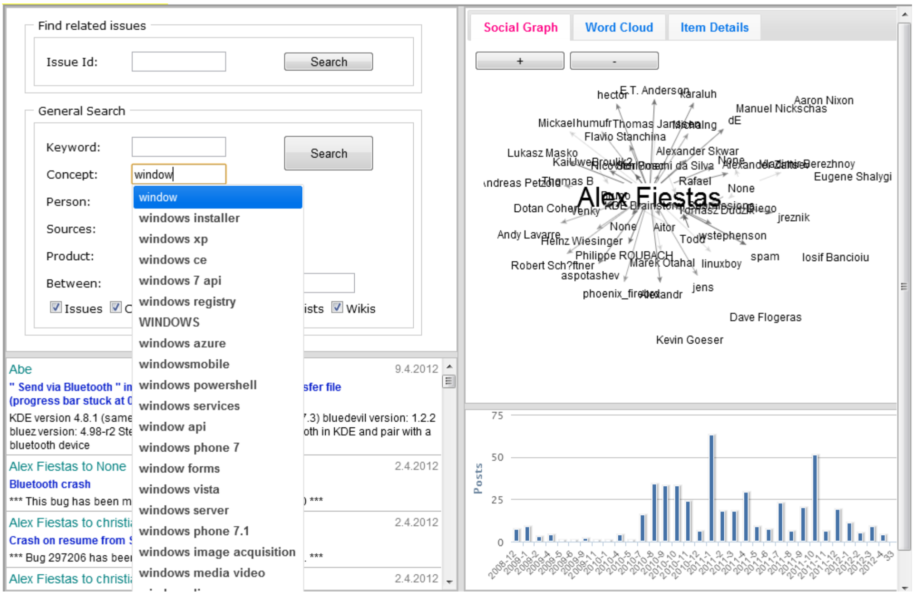
In addition to these relations we also included a few other types of relations:

- *subclass* and s*uperclass* relationships have been obtained by using the OntoGen [8] text mining tool
- *sameAs* relationships provide links to the identical DBpedia resources.
- *linksTo* relations were used for all relations that we extracted but were not mapped to some more specific type of relation (like *isPartOf, creator*, etc.).

2.3 Filtering and publishing the ontology as RDF

Before the ontology was finished we wanted to make sure that it doesn’t contain any unnecessary concepts. By checking the terms on the online glossaries we noticed that some of them are obsolete and therefore irrelevant for our ontology. To determine if a concept is relevant or not we decided to again use the stackoverflow website. For each concept we searched in how many questions the concept is mentioned. If the concept was mentioned in less than 10 questions we decided to treat it as irrelevant and we removed the concept and its relations from the ontology. The value 10 was chosen experimentally by observing which concepts would be removed at different thresholds. An example of a concept that was removed by this procedure is HAL/S (High-order Assembly Language/Shuttle) which was found only in one question on the *stackoverflow* website.

The final version of the generated ontology contains 6,196 concepts and 91,122 relationships and is published in the Resource Description Framework (RDF) format.



Abe 9.4.2012
" Send via Bluetooth " in … sfer file (progress bar stuck at 0
KDE version 4.8.1 (same … 7.3) bluedevil version: 1.2.2 bluez version: 4.98-r2 Ste … oth in KDE and pair with a bluetooth device

Alex Fiestas to None 2.4.2012
Bluetooth crash
*** This bug has been m …) ***

Alex Fiestas to christi… 2.4.2012
Crash on resume from S
*** Bug 297206 has been … . ***

Alex Fiestas to christi… 2.4.2012

Figure 1. Search interface provided by the ALERT system

3. Using the Annotation ontology for semantic search

The created ontology can be used to annotate all the posts that are generated in the communication channels monitored by the ALERT system. When a new post is created we annotate it with concepts that are mentioned in the text. We do this by checking the labels of the concepts and determining if any of them appears in the text. The post with its annotations is then stored in the Knowledge base and can be used for searching.

Since the ALERT project is still in progress we currently only have a preliminary version of the search interface. A screenshot of the interface is shown in Figure 1. The search form in the top left corner allows the user to specify a rich set of search conditions. Beside the keyword search the conditions can also include:

- Concepts. The user can specify a concept from the Annotation ontology in order to find posts that are annotated with this concept.
- Authors of the posts. All posts from the communication channels have authors and they can be specified as a condition.
- Source code (files, classes, methods). By monitoring source code management systems used by the community we are aware of all the files, classes and methods developed in the project. The information is stored in the knowledge base and can be used to find all the commits where a particular file/class/method was modified.
- Time constraints. All posts have an associated time stamp. The search interface allows us to set a particular time period that we are interested in.

- Filters by post type. The user can specify what type of posts (issues, emails, forum posts, etc.) he would like to see in the list of results.

After performing the search, the list of posts that match the query is displayed below the search form. Along with the list of results, the system also provides two visualizations of the results. Social graph of the people involved in the resulting posts is displayed on the right side of the screen. It shows who is corresponding with whom and highlights the most active people. Below the social graph is the timeline visualization that shows the distribution of results over time. It is an important aggregated view of the results since it can uncover interesting patterns.

4. Related work

The automatic and semi-automatic ontology learning methods usually include a number of phases. Most approaches define the set of the relevant ontology extension sources, preprocess the input material, build ontology according to the specified methodology, evaluate and reuse the composed ontology.

As Reinberger and Spyns [13] state, the following steps can be found in the majority of methods for ontology learning from text: collecting, selecting and preprocessing of an appropriate corpus, discovering sets of equivalent words and expressions, establishing concepts with the help of the domain experts, discovering sets of semantic relations and extending the sets of equivalent words and expressions, validating the relations and extended concept definition with help of the domain experts and creating a formal representation.

As suggested in [14], ontology learning from text is just one phase in the methodology for semi-automatic ontology construction preceded by domain understanding, data understanding and task definition and followed by ontology evaluation and ontology refinement.

In our approach we have utilized the traditional steps for ontology development, like terms extraction, synonyms extraction, concepts definition, establishment of concept hierarchies, relations identification [15].

Fortuna et al. [8] developed an approach to semi-automatic data-driven ontology construction focused on topic ontology. The domain of interest is described by keywords or a document collection and used to guide the ontology construction. OntoGen [8] uses the vector-space model for document representation. In current work, the tool has been utilized for defining the hierarchical relationships between concepts.

Learning relations in the ontology was addressed by a number of researchers. Taxonomic relations have been extracted by Cimiano et al. [16]. Moreover, Maedche and Staab [17] contributed to the approach, which allowed discovering conceptual relations from text.

5. Conclusions

In this paper we have proposed an approach for building a domain specific ontology related. The methods for concept and relation extraction have been suggested and applied in order to build an ontology related to computer science and software development. The generated ontology is used in the ALERT project among other things to provide semantic search functionality. The advantages of the semantic search over keyword search are (a) the avoidance of issues with synonyms, and (b) the ability for expanding the search by including related concepts in the search. The current version of the ALERT system provides a preliminary interface for performing the semantic search by entering the concept name. In future we plan to improve the interface to allow the user also to extend the search to related concepts.

Acknowledgements

This work was supported by the Slovenian Research Agency, European Social Fund and ALERT (ICT-2009.1.2)

References

[1] KDE, http://www.kde.org.

[2] ALERT project, http://www.alert-project.eu.

[3] OW2, http://www.ow2.org.

[4] Morfeo project, http://www.morfeo-project.org.

[5] Webopedia, http://www.webopedia.com.

[6] Computer Glossary, Computer Terms, http://whatis.techtarget.com.

[7] Stack Overflow, http://www.stackoverflow.com.

[8] B. Fortuna, M. Grobelnik, D. Mladenic, OntoGen: Semi-automatic Ontology Editor, HCI, 9 (2007), 309-318.

[9] J.B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

[10] M-C. de Marneffe, B. MacCartney and C. D. Manning, Generating Typed Dependency Parses from Phrase Structure Parses, in: LREC 2006, 2006.

[11] K. Toutanova, C.D. Manning, Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, in: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000, pp. 63-70.

[12] WordNet, http://wordnet.princeton.edu.

[13] M. L. Reinberger, P. Spyns, Unsupervised Text Mining for the Learning of DOGMA-Inspired Ontologies, in: Buitelaar P.; Handschuh S.; Magnini B. (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.

[14] M. Grobelnik, D. Mladenic, Knowledge Discovery for Ontology Construction, in: Davies, J.; Studer R.; Warren P. (Eds.), Semantic Web Technologies: Trends and Research in Ontology-Based Systems, John Wiley & Sons, 2006, 9–27.

[15] P. Buitelaar, P. Cimiano, B. Magnini (Eds.), Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.

[16] P. Cimiano, A. Hotho, S. Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, Journal of Artificial Intelligence Research (JAIR), 24 (2005), 305-339.

[17] A. Maedche, S. Staab, Discovering conceptual relations from text, in: W. Horn (Ed.), ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, August 21-25, 2000, IOS Press, Amsterdam, 2000, 321-324.

[16] P. Cimiano, A. Hotho, S. Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, Journal of Artificial Intelligence Research (JAIR), 24 (2005), 305-339.

[17] A. Maedche, S. Staab, Discovering conceptual relations from text, in: W. Horn (Ed.), ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, August 21-25, 2000, IOS Press, Amsterdam, 2000, 321-324.

Definition of a nuclear crisis use-case management to s(t)imulate an event management platform

Anne-Marie Barthe-Delanoë[1], Sebastien Truptil[1],
Roland Stühmer[2], Frederick Benaben[1]

[1] Université de Toulouse – Mines Albi, 81000 Albi, France
[2] FZI Forschungszentrum Informatik, Karlsruhe, Germany
{sebastien.truptil, anne-marie.barthe, frederick.benaben}@mines-albi.fr
roland.stuehmer@fzi.de

Abstract. The European PLAY project aims at providing platform for event management. This platform should be tested and stimulated through several use-cases. Obviously, these use-cases should be relevant on the business point of view, but to make them relevant, it could be interesting to be able to redesign them as often as required, in order to improve their business context. This article presents a specific crisis use-case for the PLAY platform evaluation and also a technical framework dedicated to make this use-case as agile as possible. The general principle is to fill the gap between business level (process models) and technical level (workflows definition and web-services implementation). The way the use-case will be simulated (to stimulate the PLAY platform) and the way the use-case will be designed and potentially re-designed (to be simulated) are described in this article.

Keywords: events, web-services, use-case, nuclear crisis, SOA, business processes, workflows.

Introduction

The European PLAY project (Grant number: FP7-ICT-2009-5) aims at designing an event management platform. Any event provider, such as electronic devices, information systems, etc., would be able to send its events to the PLAY platform through a cloud infrastructure. The PLAY platform provides an event market place containing (i) the events received from event providers and (ii) new events generated by the Complex Event Processing tool (CEP layer) from the combination of the previous ones (the deduction is rule-based). Any event consumer would then receive the events from the topic it has subscribed for. Event consumers could finally use these events to act on a better way, according to the way the situation evolves.

In this article, we aim at presenting the way one specific use-case could be implemented to stimulate the PLAY platform and demonstrate its features. This use-case concerns a nuclear crisis. The global objective is to define the workflows and the web-services used to simulate the crisis management and using the PLAY platform to

adfa, p. 1, 2011.

run and adapt the overall behavior of the crisis cell. Furthermore, in order to match with the business level definition of the crisis management, this article introduces the mechanisms in charge of ensuring the direct generation of these workflows and web-services from the process models and activities.

In the first section of this article, we introduce the global architecture of the demonstration platform and the way workflows and web-services will be used in a SOA context. The second section presents the use-case scenarios. The remainder of the article describes the automatic transformation of business models (process cartography) into technical components (workflows and web-services) that will be implemented in the demonstration platform of the first section.

1 Overview of the global architecture

In our use-case of a nuclear crisis management, the scenarios are very complex and a lot of sub-processes are involved. One of the objectives of our current research work is to simulate this use-case through a demonstration platform.

This platform will be based on the Service Oriented Architecture (SOA) principles [1] and will be able to run the three levels of processes (decisional, operational, support). Technically, we will use the Distributed Service Bus (ESB) PETALS developed by the French open-source software editor PETALS Link [2]. Such a technical infrastructure requires the description of the processes as workflows in a runnable language (Business Process Execution Language (BPEL) [3] for instance). In order to make this task easier, all the sub-processes will be described with Business Process Modeling Notation (BPMN) [4].

Furthermore, considering the fact that we have to describe the event exchanges between actors during the crisis response, BPMN is a good choice: this language is not only strongly aligned with computer implementation of workflows, but also structurally event-oriented (events are represented via circles and can be typed). BPMN is at the intersection between our need to represent events and the technical requirements of our demonstration platform (proximity between BPMN and workflow language).

1.1 The PLAY Platform

Figure 1 shows the conceptual architecture of the PLAY Platform. The Distributed Service Bus (DSB) provides the Service-oriented Architecture and Event-driven Architecture (EDA) [5] infrastructure to connect components, devices and end user services. The DSB enables the federation of separate SOAs through the formation of domains, which can be allowed to exchange events. Thus, distributed sources of events can be combined in the platform.

The Event Cloud provides storage and forwarding of events so that interested parties can be notified of events according to content-based subscription. The storage operates as an event history to fulfil queries for older events, which do not need real-time results e.g., when generating statistics. The Event Cloud is comprised of a peer-to-peer system of storage nodes organised in a CAN network [6].

The Distributed Complex Event Processing (DCEP) component has to detect complex events and do reasoning over events in real-time. Events per se might not be meaningful, but meaningful events can be derived from available, simpler events. The platform can readily detect such derived events, because it has knowledge of all events and applies event patterns, as described in [7], to the input events.

Finally, the Service Adaptation Recommender suggests changes (adaptations) of services' configurations, composition or workflows, in order to overcome problems or achieve higher performance.

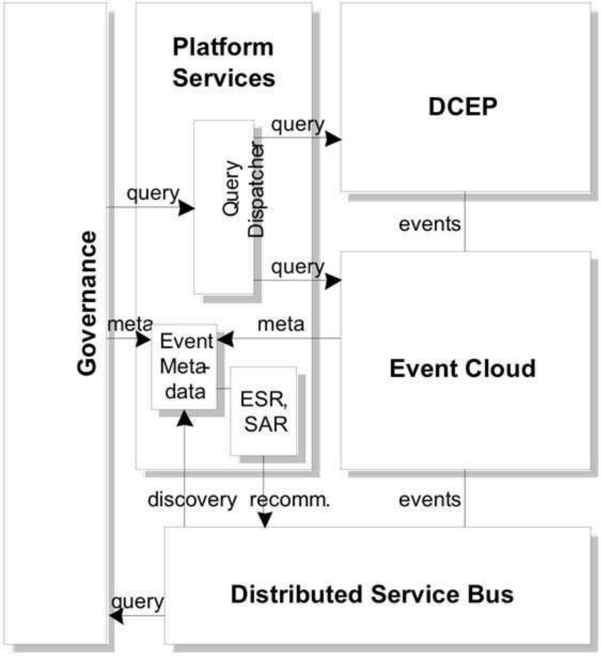


Fig. 1. PLAY Platform conceptual architecture

1.2 The Simulation Platform

Basically, the structure of the demonstration platform will be the following: several ESBs will run (now, the prototype uses only one), thanks to their workflow engine, several BPEL workflows (representing decisional, operational or support processes) among several web-services (representing activities of actors that might be invoked in a crisis management context).

Each web-service is able to generate events (such as status but also business events like radiation measures or requested resources). These events are formatted using the WS-Notification [8] standard: they embedded at least a timestamp and an ID number. Any generated event follows one of the eight event types (Figure 2) we have defined for the crisis response domain.

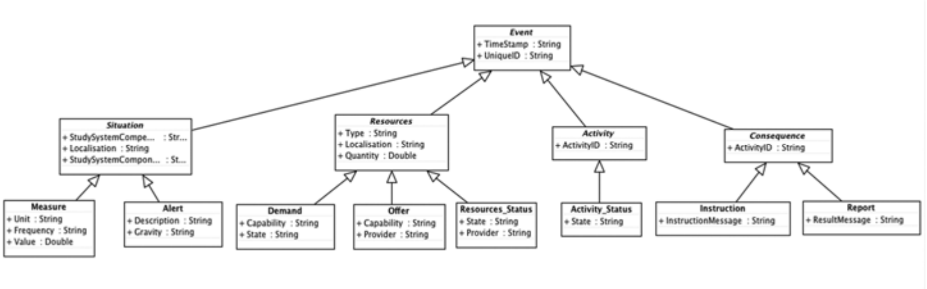


Fig. 2. The eight event types for crisis response domain.

Then, these events are sent to a special service of ESB. This special service (event manager or event proxy) is in charge of gathering events, translating them into an appropriate format for further processing (in the case of PLAY this is an RDF Schema) and sending them to the cloud platform of PLAY (see Figure 3). The PLAY platform can use these events to generate new events and enrich the event market place. The event manager is also in charge of receiving new events from the cloud PLAY platform in order to send them to the web-services that are subscribers for that type of event.

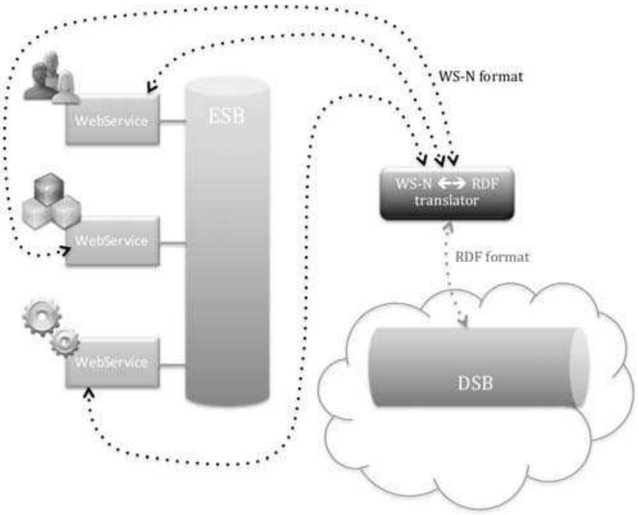
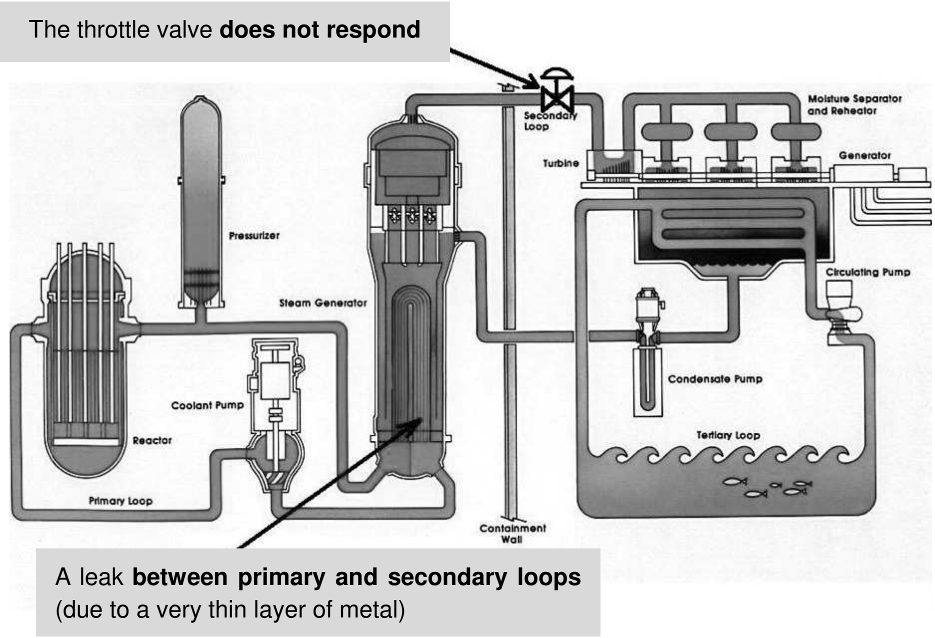


Fig. 3. Technical architecture (with event sources and exchanges) of the simulation platform

2 Description of the use-case through a BPM approach

The considered crisis situation use-case takes place in a French nuclear plant which reactor is water-pressurized (water-pressurized type is used for all nuclear reactors in

France, exception made for a single reactor [9]). The radiation leak in our scenario results of the combination of two issues (as presented in Figure 4).



The throttle valve **does not respond**

A leak **between primary and secondary loops** (due to a very thin layer of metal)

Fig. 4. Nuclear Crisis Use-Case: the triggering event.

i. The metal of the steam generator is very thin. Due to the wearing effect of time, a leak appeared in the steam generator. As a consequence, the water within the primary loop, which is contaminated, spreads through the secondary loop.
   Consequences: The steam (and the water) of the secondary loop are contaminated and the pressure within the secondary loop increases.
ii. The throttle valve, a safety device of the secondary loop, opens due to the increased pressure inside the secondary loop. Unfortunately, it does not respond to the manual bypass of the safety loop, requiring its closure.
   Consequences: The steam of the secondary loop, contaminated, escapes from the secondary loop to the atmosphere.

To solve, or at least reduce, this crisis situation, several stakeholders are involved. They are grouped into an organization called “crisis cell”, which is in charge of the crisis response. The representative of the French national authority (the prefect), outside the nuclear plant, pilots this crisis cell. Delegates of each actor are present in the crisis cell. Firemen, policemen, weather survey network, scientists, emergency medical service, and any other actor involved in response process has one representative in the crisis cell. The delegates validate the feasibility of the decisions, ensure link with actors on the field and ensure communication between actors.

Each actor involved in the crisis response has its own abilities, the events it is listening to and the events it is able to generate and to send to the cloud (i.e. the technological platform that manages the events).

The crisis cell is structured according to three kinds of process, regarding the standards of business process cartography (as defined by the European standard NF EN ISO 9001 version 2000 [10]): decisional process, operational process and support process. These three processes may be detailed through seven sub-processes, as shown on Figure 5.

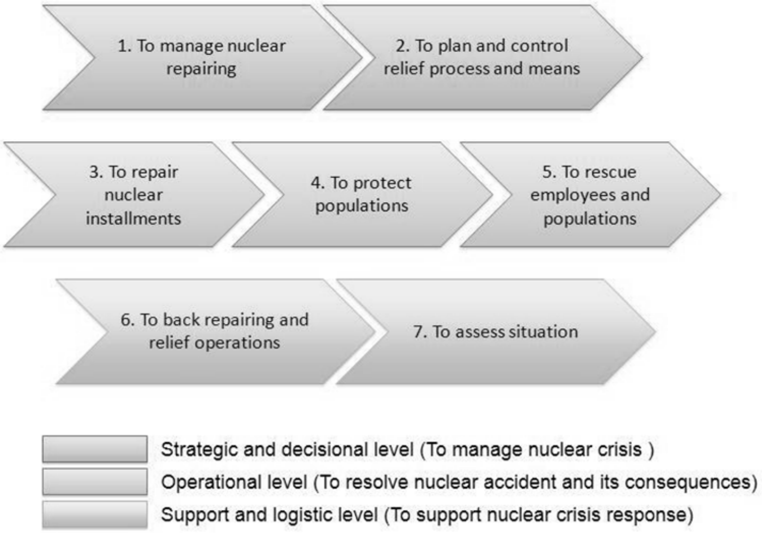

1. To manage nuclear repairing
2. To plan and control relief process and means
3. To repair nuclear installments
4. To protect populations
5. To rescue employees and populations
6. To back repairing and relief operations
7. To assess situation
Strategic and decisional level (To manage nuclear crisis)
Operational level (To resolve nuclear accident and its consequences)
Support and logistic level (To support nuclear crisis response)

Fig. 5. Nuclear Crisis Overall Treatment/ Management

We have designed eighty-three BPMN processes to cover the whole cartography of the interactions between the crisis cell stakeholders. As an example for this article, we focus on a little part of that process cartography we have already implemented in our prototype. The following Figure 6 presents several swim lanes (horizontal containers) that represent the involved actors for this process (here the French weather forecast network Meteo France and the Radiation Survey Network) and the PLAY system. Each pool embeds its own activities and flows, while exchanges between pools are represented through flows generating events.

The matching workflow of the previously presented BPMN process now runs on our prototype of simulation platform (we have designed the matching BPEL files and also the web services called by the processes).

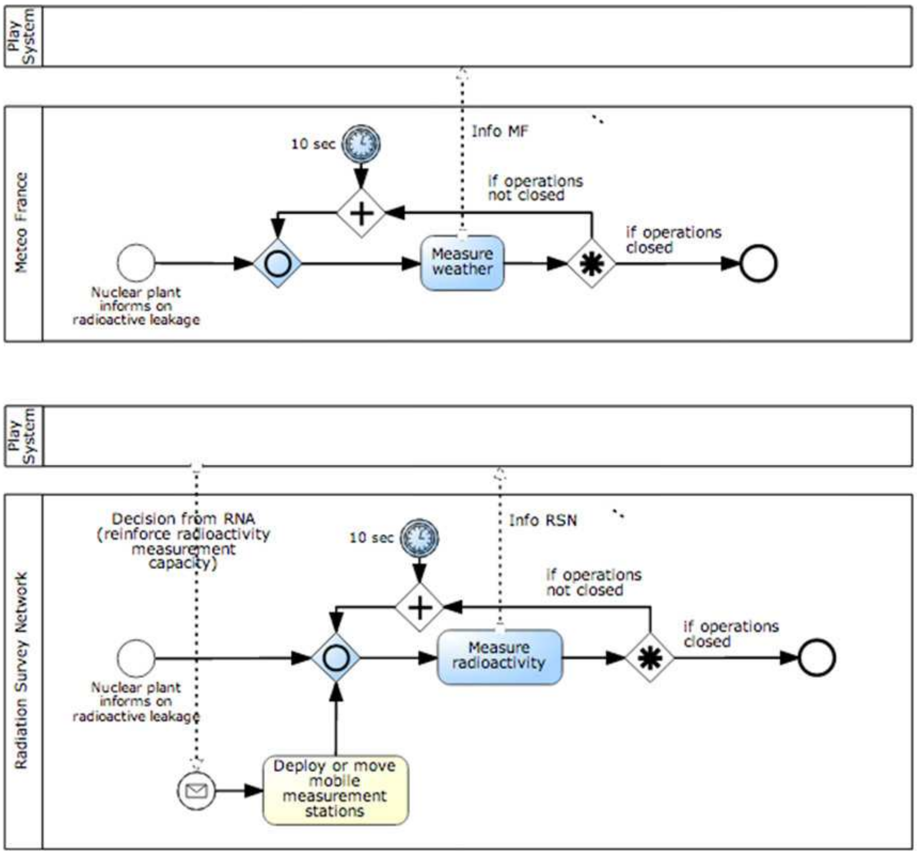


Fig. 6. Focus on a part of the crisis response in BPMN.

3 Use of the s(t)imulation platform

The input of this step is a set of ordered business activities, each under the responsibility of an actor. Nevertheless the business activities could not be directly used by the platform, it is necessary to match these business activities with technical services (as an operation of web-services). Research works try to define semantics service search engine, like [11] or [12], in charge of realizing automatically this kind of matching.

Since the aim of our research work in the PLAY project is to configure a technical platform, which has to simulate a business description, we decided at first to manually realize this matching before improving our work with any semantics service search engine.

Based on this choice, the automatic configuration of the technical platform, represented by the Figure 7, is divided in four main steps for each business process:

Step 1: In practice, once the knowledge about the crisis situation is gathered, the crisis response is represented by a set of BPMN models. In all these models, a pool represents each partner and a pool represents the cloud.

Our set of generic tools, as a result of our internal research work, allow to retrieve the business services used in the BPMN files and to match them with « real » services, i.e. technical services provided by the collaboration partners. The tools extract the whole business services information from the BPMN process description: they produce a list of services (name of the service and its responsible partner's name).

Step 2: The next step is to match these business services with technical services. For the moment, we suppose that a business service matches with a single technical operation (i.e. a single operation of a WebService). This matching is manually done with the help of our tools.

For each business service, the user has to choose an operation from a WebService (thanks to WSDL file). If the concerned partner doesn't provide a relevant WSDL file, our tools allow creating such a file and the associated WebService.

Once the WSDL file is chosen, the tools parse it and propose the operations contained into it (and sort it by port types) to the user. The step is repeated as many times as needed to match all the business services with technical operation. Through a Model Driven Engineering (MDE) (model transformation of the BPMN files) and with this service matching, we will obtain the workflow, expressed in Business Process Execution Language (BPEL), and a set of configuration files (that are linked to the server used to run the workflow)

Step 3: In PLAY, the workflow is run on an Enterprise Service Bus (ESB), which is PETALS, developed by the French editor PETALS Link. This ESB is compliant with the JBI standard (also based on JSR208 standard).

The JBI standard implies the use of Services Unit (SU) and Services Assembly (SA) that compose the configuration of the services on the ESB. The several SUs and SAs are automatically generated for all the services of the collaborative process. A SU is composed of the WSDL of the service and a JBI file that defines, in a unique way for the ESB, the web-service. A SA makes the link between a protocol (SOAP, HTTP, …) and the web-services through the SU [13]. They are necessary to use both partners' services (which can be real web services or just interfaces that make the link between the SOA architecture and a manual/technical operation) and the mediation service (the BPEL orchestrator which runs the collaborative process described into the BPEL file).

Step 4: Finally, all the artifacts on the ESB are automatically deployed. These artifacts are composed of, on the one hand all the SAs and SUs created during the previous step, and, on the other hand all the binding component (BC) needed to communicate with the web-services (one BC per protocol) and the potentially requires service engine (for instance a workflow engine).

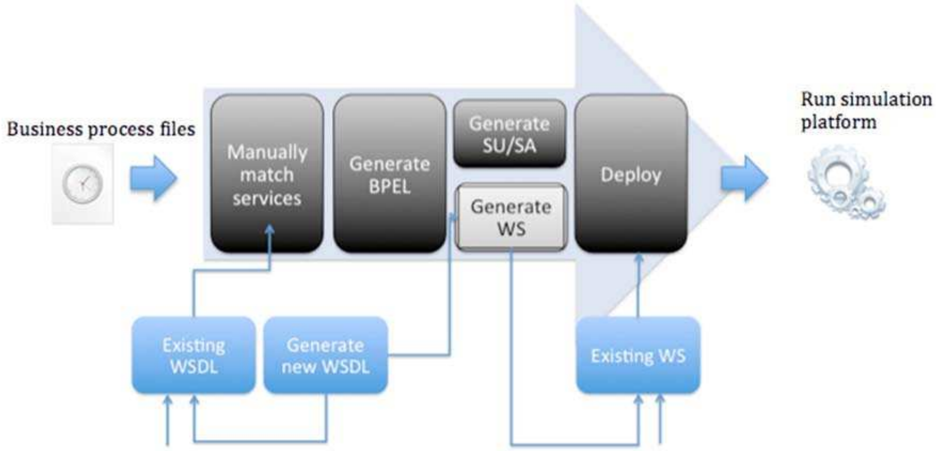


Fig. 7. The (almost) automated settings chain

Step 5: In PLAY, our research work consists in simulating the execution of the defined crisis response. Consequently, we need graphical user interface for each service, in order to simulate the response for the invocation of any operation by the workflows. We use a tool name EasiestDemo, also developped by PEtALS Link [14]. It allows the creation of a graphical interface for each operation of a WebService. A graphical interface is composed of TextBox for each input and output elements of the operation and the colors of the interface are defined for each actor in a XML file.

Conclusion

To demonstrate the powerful capabilities of the PLAY platform, and especially the way interactions and interconnections of events could be handled, complex workflows have been be designed. These workflows must also be directly connected to the cloud infrastructure of PLAY. There are two main issues in this objective: (i) the quality of the considered use-case (and of the associated workflows) and (ii) the way these use-cases can be easily executed to stimulate the PLAY platform.

This article tried to deal with both these issues. The presented use-case is a very complete one that could be easily made simpler or more complex. The main objective of this approach is to show how a platform like PLAY could support and hide the complexity of a concrete business situation. Choreography and orchestration of heterogeneous business processes in a real size imply a high-level of complexity. In crisis management context, crucial decision tasks (that require human vision) could be embedded in that complexity. Consequently, these crucial tasks could be partially hidden and decision makers could miss the fact that these tasks are so important (among the mass of other tasks). By assuming orchestration and choreography, an event-driven platform like PLAY could deal with the quantity of computable tasks in order to highlight critical ones (the ones that requires decision makers).

To deal with both the main issues of the overall objective (business quality of the use-case and the ability of workflow simulation to stimulate the PLAY platform), the whole set of services; workflows and events are currently being implemented. Simultaneously, the whole structure of the PLAY platform (presented in figure 1) is also being deployed. The concrete and complete confrontation between both these worlds (even if there are already partial confrontations, component by component) will be a great step of the PLAY project and should demonstrate the functional capabilities of the PLAY platform. However, there are still a lot of questions concerning non-functional aspects. For instance, scalability of the simulation platform will be a crucial issue in order to evaluate scalability of the PLAY platform. Some other points like quality of service or security, even if less crucial, are also to be considered.

Acknowledgments. The authors would like to thank very gratefully the members of the PLAY consortium (FZI, Karlsruhe; ARMINES/CGI, Albi; PetalsLink, Toulouse; ICCS, Athens; CIM, Nis; INRIA, Sofia-Antipolis; Orange-Labs, Sophia-Antipolis) and European commission for their support.

References

1. Vernadat, F.: Interoperable enterprise systems: Principles, concepts, and methods. In: Annual Reviews in Control, vol. 31, no. 1, pp. 137—145 (2007)
2. PEtALS Link, http://www.petalslink.com
3. OASIS, Web Services Business Process Execution Language Version 2.0, http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html (2007)
4. Object Management Group, Business Process Model And Notation (BPMN), Version 1.2, http://www.omg.org/spec/BPMN/1.2/PDF/ (2009)
5. Luckham, D., Schulte, R.: Event Processing Glossary – Version 1.1, Event Processing Technical Society (2008)
6. Filali, I., Pellegrino, L., Bongiovanni, F., Huet, F., Baude, F.: Modular P2P-based Approach for RDF Data Storage and Retrieval. In: Proceedings of The Third International Conference on Advances in P2P Systems. (2011)
7. Etzion O., Niblett P.: Event Processing in Action. Manning Publications Co. (2010)
8. OASIS, Web Services Base Notification 1.3 OASIS Standard (2006).
9. Electricité de France (EDF): Panorama de l'électricité : les différents types de réacteurs n nucléaires, http://www.edf.com/html/panorama/production/industriels/nucleaire/types_reacteurs.html
10. Norme européenne NF EN ISO 9001 version 2000, Systèmes de management de la qualité – Exigences. AFNOR (2000)
11. Bénaben, F., Boissel-Dallier, N., Lorré, J. P. et Pingaud, H.: Semantic Reconciliation in Interoperability Management through Model-Driven Approach. In: Proceedings of PRO-VE 2010, pp.705—712, (2010)
12. Dong, H., Hussain, F.K., Chang, E.: A service search engine for the industrial digital Ecosystems. In: IEEE Trans. on Industrial Electronics, vol 99, (2009)
13. JSR 208: The Java Community Process (SM) Program - JSRs: *Java Specification Requests - detail JSR# 208*, (2005).
14. EasiestDemo, http://research.petalslink.org/display/easiestdemo/EasiestDemo+-+Open+source+BPEL+to+Java+generator

Context Management in Event Marketplaces

Yiannis Verginadis[1], Ioannis Patiniotakis[1], Nikos Papageorgiou[1], Dimitris Apostolou[1], Gregoris Mentzas[1], Nenad Stojanovic[2]
[1] Institute of Communications and Computer Systems,
National Technical University of Athens,
{jverg, ipatini, npapag, dapost, gmentzas}@mail.ntua.gr
[2] FZI Forschungszentrum Informatik, Karlsruhe, Germany
nstojano@fzi.de

Abstract. This paper refers to methods and tools for enabling context detection and management based on events. We propose a context model that builds on top of previous efforts and we give details about the mechanisms developed for context detection in event marketplaces. In addition, we show how simple or complex events can be used in combination with external services in order to derive higher level context with the use of Situation-Action-Networks (SANs). Specifically, we present two different approaches, one for detecting low level context and another one for deriving higher-level contextual information using SANs. We present an illustrative scenario for demonstrating the process of specialization of our generic context model and its instantiation based on real-time events.

Keywords: Context, Event Marketplace, Detecting Context, Deriving Context

1 Introduction

Context is "any information that can be used to characterize the situation of an entity, i.e., a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." [1]. Context detection is considered important in the so-called event marketplaces [2] (see e.g., http://pachube.com, a platform offering a service based architecture, a range of graphing and visualization tools, event detection via triggers, along with cost-effective data storage) for enhancing the user's experience when interacting with the event marketplace.

Events from event marketplaces are an important source of context for service-based applications that consume them because they may convey important information, which is relevant for service execution and adaptation. To achieve the goal of injecting event processing results to context, an event-based context model is needed along with context detection and derivation mechanisms. In previous work [3] Situation-Action-Networks have been proposed as a hierarchical goal-directed modeling approach comprising nodes with specific semantics used to model goal decompositions, enriched with flow control capabilities. SANs provide means to decompose goals into subgoals and capabilities for seeking and achieving the high-level goals, involving situations (i.e. complex event patterns), context conditions and actions.

The simplest SAN possible is a two level tree with a parent (root) node and three child nodes, each of them having specific semantics. A parent node models the Goal sought. The leftmost child node describes a situation that must occur, in order to start goal seeking. The middle child node corresponds to context update and requires that a specific contextual condition is true before continuing with the SAN traversal. The rightmost child node specifies the action to be taken in order to fulfil the goal. Rightmost node can also be a sub-goal node with its own three child nodes, or it can even be a construct joining several sub-goals in sequence or in parallel. As the SAN becomes more complex, involving several subgoals (Figure 1), it deepens and reveals its hierarchical and goal-directed characteristics. In this work, SANs are extended so that they can be used for detecting and deriving context from events.

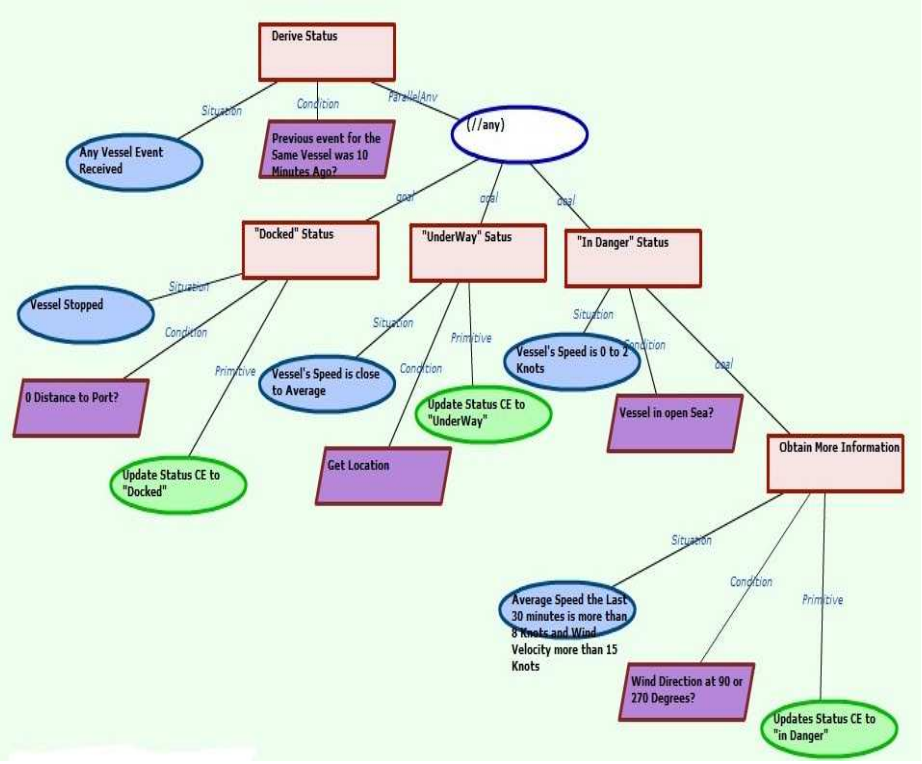

Derive Status
Situation
Condition
Parallel/Any
(//any)
Any Vessel Event Received
Previous event for the Same Vessel was 10 Minutes Ago?
goal
goal
goal
"Docked" Status
"UnderWay" Satus
"In Danger" Status
Vessel Stopped
Situation
Condition
Primitive
0 Distance to Port?
Update Status CE to "Docked"
Situation
Condition
Primitive
Vessel's Speed is close to Average
Update Status CE to "UnderWay"
Get Location
Situation
Condition
goal
Vessel's Speed is 0 to 2 Knots
Vessel in open Sea?
Obtain More Information
Situation
Condition
Primitive
Average Speed the Last 30 minutes is more than 8 Knots and Wind Velocity more than 15 Knots
Wind Direction at 90 or 270 Degrees?
Updates Status CE to "in Danger"

Fig. 1. SAN Illustration based on the marine vessel traffic scenario: Pink nodes denote Goals, Blue nodes denote Situations, Magenta nodes denote Context Condition and Green nodes denote Actions

This paper continues with a discussion about related work in the domain of event-based context management, while in section 3 it presents a generic context model that is considered appropriate for the needs of event marketplaces. In section 4, we consider two different approaches, one for detecting low level context and another one for deriving higher-level contextual information using SANs. In section 5, we show how the generic context model can be specialized so that it can be instantiated to support an example scenario. We conclude in section 6 with a summary of our event-based context management approach.

2 Related Work

Context-awareness in service-oriented systems refers to the capability of a service or service-based application to be aware of its physical environment or situation and to respond proactively and intelligently based on such awareness; see e.g. [4]. Through the use of context, a new generation of service-based applications is expected to arise for the benefit of coping with the dynamic nature of the Internet; see e.g. [5], [6]. To reflect the varying nature of context and to ensure a universal applicability of context-aware systems, context is typically represented at different levels of abstraction [7]. At the first level of raw context sources there are context data coming from sensor devices, or user applications. At the next levels, context is represented using abstraction approaches of varying complexity. The work in [8] reviews models of context that range from key-value models, to mark-up schemes, graphical models, object-oriented models, logic-based models and ontology-based models. In [9] an ontological model of the W4H classification for context was proposed. The W4H ontology provides a set of general classes, properties, and relations exploiting the five semantic dimensions: identity (who), location (where), time (when), activity (what) and device profiles (how). The five dimensions of context have been also pointed out earlier in [10] where it was stated that context should include the 'five W': Who, What, Where, When, and Why.

Our work focuses on detecting context changes which correspond to either atomic or complex events and use complex event processing to model and identify them. Similarly to [11], we focus on events as a source of context because they are snippets of the past activities; therefore event processing may be viewed as a context detecting technology. Event processing results may be transferred to other applications, injecting context related information into services and processes. Based on the context definition of Dey and Abowd [1] and the associated five dimensions of context expressed in ontological model of the W4H [9], we define a high-level context model following an object-based modelling approach which can be easily specialized for different applications. We use semantic querying to extract contextual information from event payloads. Moreover, we exploit the reasoning capabilities of Situation-Action- Networks to enable dynamic derivation of context from multiple event streams and external services.

3 Context Model

We propose a context model as a stepping stone for facilitating event-based context detection and derivation functionality, in order to better understand situations in dynamic service oriented environments that demand for new additional information sources or/and lead to a number of service adaptations as means for successfully coping with dynamic environmental changes. In order to achieve the goal of extracting contextual information, analyzing them and then deriving higher level context, we follow an event-based context modelling approach. In this section, we present such a Context Model (Figure 2), expressed in UML 2.0 class diagram. This model is based on the W4H model [9] that describes the five main elements associated within a context; the five elements are arranged into a quintuple (When, What, Where, Who, How).

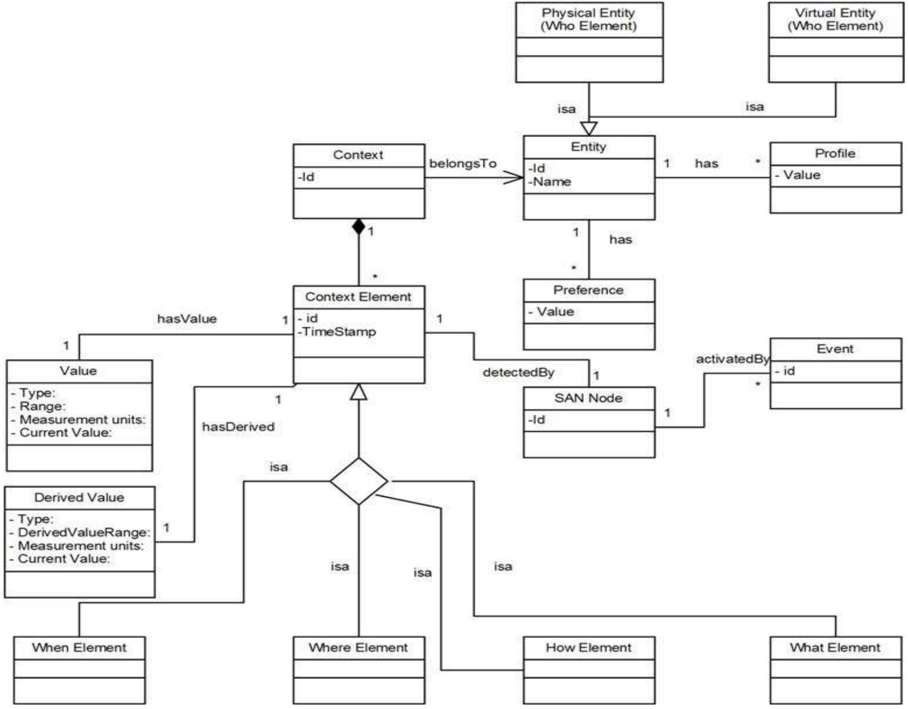


Fig. 2. Context Model

This Context Model expresses the temporal (i.e. When), spatial (i.e. Where), declarative (i.e. Who, What) and explanatory (i.e. How) dimensions of context having as central point of focus the notion of Entity. We refer to either physical or virtual entities with specific profiles and preferences that characterise them (e.g. vessel, port authority information system etc.). This way context obtains substance around the notion of an entity which can be a customer of an event marketplace system. The context class in our model constitutes the aggregation of several different context elements that may refer to five dimensions of context. Each Context element can have a value that can be acquired from the situation node of a SAN and/or a derived value that arises from any kind of reasoning process or call of external services. All context related information can be captured as objects which can store either a single scalar value or multiple values such as vectors, sets, lists etc. As any of the available context models [8], our model needs to become domain or application specific in order to be useful. Next, we show how SAN Editor can be used to specialize and instantiate the generic context model.

4 Event-based Context Management

In our context modelling approach and implementation, we consider entities as being able to own SAN trees. The scope of context elements is distinguished into three levels: “Local”: Context elements can be updated and used only by a specific SAN instance. “Entity”: Context elements can be updated and used by any of the SANs owned by the same entity. “Global”: Context elements can be updated and used by all SANs independently to which entity they belong to.

Using the SAN Editor, we can perform context model specializations based on the application scenario and can formulate the necessary queries to events for extracting contextual information. We provide two approaches for acquiring context from simple or complex events and instantiating our context model. Both approaches use the SAN Editor for:

1. defining SPARQL queries to specific RDF event payload information that can update the values of an entity’s context elements; and
2. defining SANs that can use information from several event streams, analyse them and/or combine them with external services, in order to update the derived values of context elements. In this way, we succeed in acquiring higher level context compared to the lower level information that events carry.

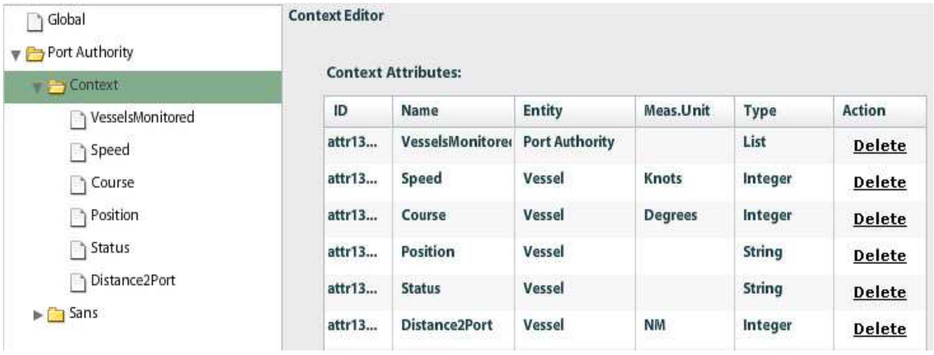
The application of both approaches is presented in the following section through the marine vessel traffic illustrative scenario, which uses events related to marine traffic control that can be used to detect potentially dangerous vessel movements informing a controller when two vessels are approaching each other.

5 Illustrative Scenario

A vast amount of real time events are available from portals connected to automatic identification systems (AIS) that contain important vessel information worldwide (e.g., speed, course, vessel type, wind conditions etc.) and the several different users/authorities that might be interested in them. In order to exploit efficiently all these information in an automated way we use our context model and present how it can be specialized for the specific application domain while we give a glimpse to its possible run time instantiations.

Context Model Specialization: Our context model needs to become application specific in order to be useful. In this section, we focus on context model specialisation which pertains the definition of entities along with their context elements necessary for capturing the context in terms of a specific application scenario. We use the marine vessel traffic scenario which is related to vessel and marine traffic control observing systems.

In this scenario, we consider the entity Port Authority as the owner of all SANs discussed below while the entity of interest is the Vessel. In order to capture contextual information related to Vessels' context, we have defined the following Context Elements that shape the specialization of our context model: Speed, Course, Position, Status, Distance2Port.



- Global
- Port Authority
  - Context
    - VesselsMonitored
    - Speed
    - Course
    - Position
    - Status
    - Distance2Port
  - Sans

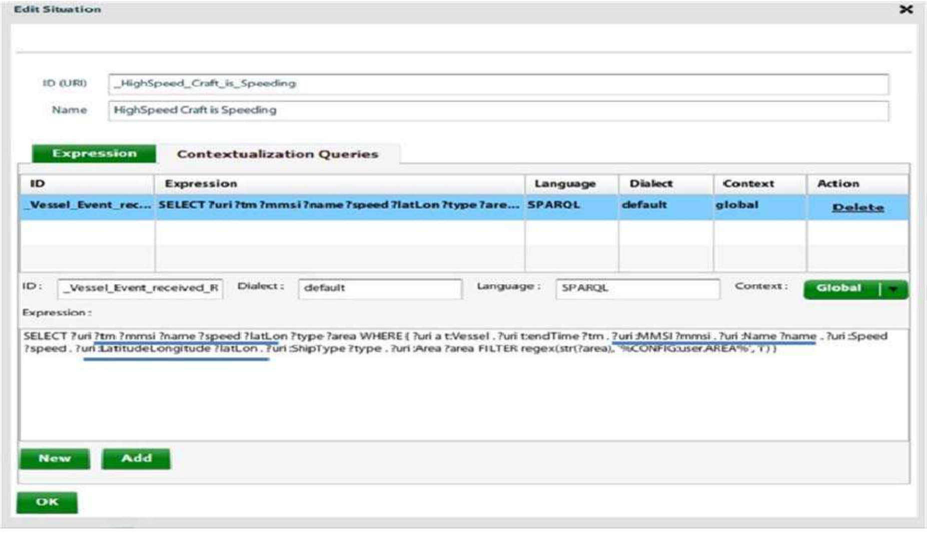
Context Editor

Context Attributes:

ID	Name	Entity	Meas.Unit	Type	Action
attr13...	VesselsMonitore	Port Authority		List	Delete
attr13...	Speed	Vessel	Knots	Integer	Delete
attr13...	Course	Vessel	Degrees	Integer	Delete
attr13...	Position	Vessel		String	Delete
attr13...	Status	Vessel		String	Delete
attr13...	Distance2Port	Vessel	NM	Integer	Delete

Fig. 3. Context Model Specialisation for the marine vessel traffic scenario

In Figure 3, the reader can find the complete list of the five context elements associated with the Vessel entity, specialising the context model for the marine vessel traffic scenario, using the SAN editor. This model specialisation will be instantiated at run time through the context detection and derivation approaches that are presented in the following sections.



Edit Situation

ID (URI): _HighSpeed_Craft_is_Speeding

Name: HighSpeed Craft is Speeding

ID	Expression	Language	Dialect	Context	Action
_Vessel_Event_rec...	SELECT ?uri ?tm ?mmsi ?name ?speed ?latLon ?type ?are...	SPARQL	default	global	Delete

ID: _Vessel_Event_received_R

Dialect: default

Language: SPARQL

Context: Global

Expression:

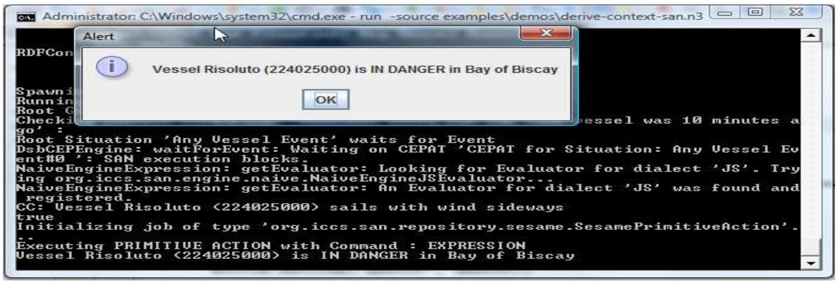
```
SELECT ?uri ?tm ?mmsi ?name ?speed ?latLon ?type ?area WHERE { ?uri a t:Vessel . ?uri t:endTime ?tm . ?uri :MMSI ?mmsi . ?uri :Name ?name . ?uri :Speed
?speed . ?uri :LatitudeLongitude ?latLon . ?uri :ShipType ?type . ?uri :Area ?area FILTER regex(str(?area), '%CONFIG:user.AREA%', 'i') }
```

Fig. 4. SAN Editor Screenshot – Performing SPARQL Queries (about Position)

Detecting Context: In this section, we discuss our first approach for acquiring context from simple or complex events and instantiating our context model. Using SAN Editor, we are able to define SPARQL queries to specific event payload information that update the values of an entity's context elements. As we show in the following figure 3 during our experiment we received events regarding a specific vessel called “Risoluto”. Details regarding the entity such as profile information automatically update the context of this entity based on the detected events in the situation node of a SAN.

Figure 4 depicts a screenshot of SAN editor with the required SPARQL queries for instantiating the “Position” context element of the vessel entity (Latitude/Longitude). Specifically, we query the vessel entity event payload with respect to the “LatLon” information. Similarly, other queries are used in the editor regarding the “Speed” and “Course” context elements and refer to event=based detection of low level context.

Deriving Context using SANs: Our second approach that we apply for extracting context from simple or complex events and instantiating our context model using SANs. We define a number of SANs that can use information from several event streams and combine them with external services in order to update the derived value class of context elements. In this way, we succeed in acquiring higher level context compared to the lower level information that events carry.



Administrator: C:\Windows\system32\cmd.exe - run -source examples\demos\derive-context-san.n3

Alert

Vessel Risoluto (224025000) is IN DANGER in Bay of Biscay

```
RDFCon

Spawni
Runnin
Root C
Checki                                           essel was 10 minutes a
go' :
Root Situation 'Any Vessel Event' waits for Event
DsbCEPEngine: waitForEvent: Waiting on CEPAT 'CEPAT for Situation: Any Vessel Ev
ent#0 ': SAN execution blocks.
NaiveEngineExpression: getEvaluator: Looking for Evaluator for dialect 'JS'. Try
ing org.iccs.san.engine.naive.NaiveEngineJSEvaluator...
NaiveEngineExpression: getEvaluator: An Evaluator for dialect 'JS' was found and
 registered.
CC: Vessel Risoluto (224025000) sails with wind sideways
true
Initializing job of type 'org.iccs.san.repository.sesame.SesamePrimitiveAction'.
..
Executing PRIMITIVE ACTION with Command : EXPRESSION
Vessel Risoluto (224025000) is IN DANGER in Bay of Biscay
```

Fig. 5. SAN Engine Screenshot for Derived Status

This context derivation can be complex and may involve multi-level SANs. Figure 1 shows a SAN that upon traversal will be able to update the derived value class of the Status context element. Specifically, the status of the vessel becomes “Docked” whenever we detect a vessel that has been stopped and its distance from any port is close to zero or “UnderWay” whenever vessel’s speed is close to average and “In Danger” when the system realizes that the vessel has almost stopped (away from any port) and strong winds are blowing from the side. Figure 5 is a screenshot of the run-time execution of the specific SAN for deriving the vessel’s status. A pop up alert has been added in order to better demonstrate the context derivation regarding the Status context element.

6 Conclusions

In this paper we presented methods and tools for enhancing context detection and management based on events. This proposed context management approach presented here is considered appropriate for the needs of event marketplaces. We described a Context Model that was used by the developed mechanisms for performing event-based context detection and presented two different approaches for detecting low level context (using SAN Editor) and deriving higher-level contextual information using Situation-Action-Networks (SANs). We provided with a meaningful context model specialization and demonstrated how simple or complex events coming from an event marketplace can be used and combined with external services, in order to derive higher level context with the use of SANs.

Acknowledgment

This work has been partially funded by the European Commission under project PLAY (Grant FP7-258659). The authors would like to thank the project partners for their advices and comments regarding this work.

References

1. Dey, AK & Abowd, GD: Towards a Better Understanding of Context and Context-Awareness, In Proceedings of the PrCHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness, pp. 304-307 (2000)
2. Stühmer, S. and Stojanovic, N.: Large-scale, situation-driven and quality-aware event marketplace: the concept, challenges and opportunities. In Proceedings of the 5th ACM international conference on Distributed event-based system, NY, USA, 403-40 (2011)
3. Verginadis, Y., Patiniotakis, I., Papageorgiou, N., Stuehmer, R.: Service Adaptation Recommender in the Event Marketplace: Conceptual View. R. Garcia-Castro et al. (Eds.): ESWC 2011 Workshops, LNCS 7117, pp. 194--201. Springer, Heidelberg (2011).
4. Abowd, GD, Ebling, M, Hunt, G, Lei, H & Gellersen, HW: Context-Aware Computing, IEEE Pervasive Computing, Vol 1, Issue 3, pp. 22–23 (2002)
5. Sheng, QZ, Nambiar, U, Sheth, AP, Srivastava, B, Maamar, Z & Elnaffar S: WS3: international workshop on context-enabled source and service selection, integration and adaptation, In Proceedings of the International Workshop on Context enabled Source and Service Selection, Integration and Adaptation, Beijing, China, pp 1263-1264 (2008)
6. Sheng, QZ, Yu, J & Dustdar, S: Enabling Context-Aware Web Services: Methods, Architectures, and Technologies, 1st ed., Chapman and Hall/CRC (2010)
7. Luther, M., Fukazawa, Y., Wagner, M., Kurakake, S.: Situational reasoning for task-oriented mobile service recommendation. The Knowledge Engineering Review 23(01), 7–19 (2008)
8. Bettini, C, Brdiczka, O, Henricksen, K, Indulska, J, Nicklas, D, Ranganathan, A & Riboni, D: A survey of context modelling and reasoning techniques, Pervasive and Mobile Computing, Vol. 6, Issue 2, pp.161–180 (2010)
9. Truong, HL, Manzoor, A & Dustdar, S: On modeling, collecting and utilizing context information for disaster responses in pervasive environments, In Proceedings of the 1st int. workshop on Context-aware software technology and applications, pp. 25–28 (2009)
10. Abowd, GD & Mynatt E.D.: Charting past, present, and future research in ubiquitous computing, ACM Transactions on Computer-Human Interaction, Vol. 7, Issue 1, pp. 29–58 (2000)
11. Etzion, O, Skarbovsky, I, Magid, Y, Zolotorevsky, N, & Rabinovich, E: Context Aware Computing and its utilization in event-based systems, Tutorial presented in DEBS, Cambridge, UK (2010)

Dealing with Vagueness in Semantic Business Process Management through Fuzzy Ontologies

Panos Alexopoulos and José-Manuel Gómez-Pérez

iSOCO, Avda del Partenon 16-18, 28042, Madrid, Spain,
{palexopoulos,jmgomez}@isoco.com

Abstract. One of the primary focuses of Semantic Business Process Management is the application of ontology-based semantics for the machine processable representation of business processes and the automation of their management lifecycle. Towards that direction, various ontologies have been proposed, each covering one or more aspects of the knowledge required to describe a business process. Yet, one major limitation of these ontologies is their inability to express knowledge that is vague as they are based on bivalent ontological formalisms. In this context we argue in this paper, through concrete examples and use cases, in favor of using fuzzy ontologies for the effective capture, representation and exploitation of the vagueness that may characterize business processes and we provide initial directions of how this may be practically achieved.

1 Introduction

The use of ontology-based semantics for the modeling of business processes is an emerging research area that aims at creating process descriptions with explicit and shareable meaning, thus achieving better management of organizational knowledge and higher level of process automation [5] [9]. In this context, a number of approaches and ontological schemas have been proposed, covering not only the strict notion of business process [5] [8] but also wider related aspects like organizational structures or business functions [7] [12].

Yet, a dimension of business process knowledge that has so far been inadequately considered within the relevant community is that of vagueness. Vagueness, typically manifested by terms and concepts like Tall, Strong, Expert etc., is a quite common phenomenon in human knowledge and it is related to our inability to precisely determine the extensions of such concepts in certain domains and contexts. That is because vague concepts have typically fuzzy boundaries which do not allow for a sharp distinction between the entities that fall within the extension of these concepts and those which do not. This is not usually a problem in individual human reasoning but it can become one when multiple people need to agree on the exact meaning of such terms and when machines need to reason with them. For example, a system could never use the statement *"This process requires many people to execute"* in order to determine the number of people actually needed for the process.

In this paper we argue over the need for a systematic way of capturing, representing and ultimately using vague knowledge in business process management and we propose for that the utilization of techniques and methods from the area of fuzzy ontologies [2]. The latter are extensions of classical ontologies that based on principles of Fuzzy Set Theory [11] allow the assignment of truth degrees to vague ontological elements in an effort to quantify their vagueness. As such, and given that the prevailing approaches for semantic business process management are ontology-based, fuzzy ontologies are a natural candidate for dealing with vagueness in this area. Of course, the idea of applying fuzzy representation techniques in business process modeling has been the subject of some works so far [6] [14], the most recent being that of [13] where the authors focus on the vagueness that may characterize decision situations in event-driven business processes and propose for that the utilization of fuzzy linguistic variables and fuzzy rules. However, the primary limitation of the above approaches in terms of vagueness treatment is the lack of ontology-based semantics for describing vague business process knowledge in an explicit, formal and shareable way.

A basic argument for this is that the modeling inconsistencies that may arise due to the freedom that business analysts have to name and describe process knowledge can also occur (and in fact even more severely) when the analysts need to describe vague concepts and terms. As argued in [2] vagueness is a phenomenon with a high level of subjectivity and context-dependence and therefore it is very important that the various interpretations a piece of vague knowledge may have are explicitly defined and shared among those who are intended to use this knowledge. A second argument has to do with the semantic querying and reasoning capabilities that fuzzy ontological formalisms can provide. The ability to automatically infer, when querying a business process model, implicit facts, is equally important in the presence of vagueness as it is in the classical case.

The focus of this paper is on the illustration of the need and applicability of fuzzy ontologies in vague business process management so that the foundations for future research on this area may be set. Therefore in section 2 we draw and analyze examples from state of the art business process ontologies and related application scenarios in order to highlight the forms in which vagueness may be present in semantic business process information and the need for a formal treatment of it. In section 3 we describe the notion of fuzzy ontologies and we show how these may be developed and used for business process modeling by means of relevant state of the art methods and techniques. Finally, in the last section conclusions and directions for future work are provided.

2 Vagueness in Business Processes

2.1 Vagueness in Semantic Modelling

Vagueness as a semantic phenomenon is typically manifested through predicates that admit borderline cases [10], i.e. cases where it is unclear whether or not the predicate applies. For example, some people are borderline tall: not clearly tall and not clearly not tall. In the relevant literature two basic kinds of vagueness are identified: *degree-vagueness* and *combinatory vagueness* [10]. A predicate has degree-vagueness if the existence of borderline cases stems from the apparent lack of crisp boundaries between application and non-application of the predicate along some dimension. For example, *Bald* fails to draw any sharp boundaries along the dimension of hair quantity while *Red* can be vague along the dimensions of brightness and saturation. On the other hand, a predicate has combinatory vagueness if there is a variety of conditions all of which have something to do with the application of the predicate, yet it is not possible to make any sharp discrimination between those combinations which are sufficient and/or necessary for application and those which are not. An example of this type is *Religion* as there are certain features that all religions share (e.g. beliefs in supernatural beings, ritual acts etc.), yet it is not clear which of these features are able to classify something as a religion.

It should be noticed that vagueness is different from inexactness or uncertainty. For example, stating that someone is between 170 and 180 cm is an inexact statement but it is not vague as its limits of application are precise. Similarly, the truth of an uncertain statement, such as *"Today it might rain"*, cannot be determined due to lack of adequate information about it and not because the phenomenon of rain lacks sharp boundaries.

In an ontology the elements that can be vague are typically concepts, relations, attributes and datatypes [2]. A concept is vague if, in the given domain, context or application scenario, it admits borderline cases, namely if there are (or could be) individuals for which it is indeterminate whether they instantiate the concept. Primary candidates for being vague are concepts that denote some phase or state (e.g Adult, Child) as well as attributions, namely concepts that reflect qualitative states of entities (e.g. Red, Big, Broken etc.). Similarly, a relation is vague if there are (or could be) pairs of individuals for which it is indeterminate whether they stand in the relation. The same applies for attributes and pairs of individuals and literal values. Finally, a vague datatype consists of a set of vague terms which may be used within the ontology as attribute values. For example, the attribute *performance*, which normally takes as values integer numbers, may also take as values terms like *very poor, poor, mediocre, good* and *excellent.* Thus vague datatypes are identified by considering the ontology's attributes and assessing whether their potential values can be expressed through vague terms.

2.2 Vagueness in Business Process Knowledge

The term "process knowledge" refers to the information describing the control flow of a process as well as its content, namely all artifacts that its definition may refer to. These artifacts are typically derived from and express the business environment and the organizational context of the process. Vague pieces of information and knowledge may appear in all three dimensions of process knowledge, namely structure, domain and organizational context.

To illustrate this point, we have considered and analyzed, with the help of our company's consultants, two different cases of business process knowledge. The first case involved a set of generic business process related ontologies, developed in project SUPER [1], which may serve as reusable knowledge schemas in practical semantic business process modeling scenarios. The analysis of these ontologies, which included among others the Business Process Modeling Ontology (BPMO), the Business Goals Ontology (BGO), the Business Roles Ontology (BROnt) and the Business Motivation Ontology (BGO), involved the identification within them of elements that can be interpreted as vague, according to the definitions of the previous paragraph. Our criterion for classifying an element as vague or not was merely the potential existence of borderline cases, not the number of them. That meant that even if an element could potentially have only one borderline case, it was considered vague.

The outcome of this analysis is summarized in table 1 where a sample of the elements we managed to identify as vague, along with a brief explanation of their vagueness, is presented. As one can easily see, the identified as vague elements are quite central to their respective ontologies (e.g. the hasBusinessGoal relation) and as such they are expected to be found in many relevant application scenarios. Furthermore, the use of vague terms like "desired" in the definition of elements (e.g "Desired Result") indicates that in practice there could be an almost infinite number of vague ontological elements in these ontologies that would be the result of the combination of such terms with non-fuzzy elements (e.g. "Loyal Customer", "Expert Analyst" etc.).

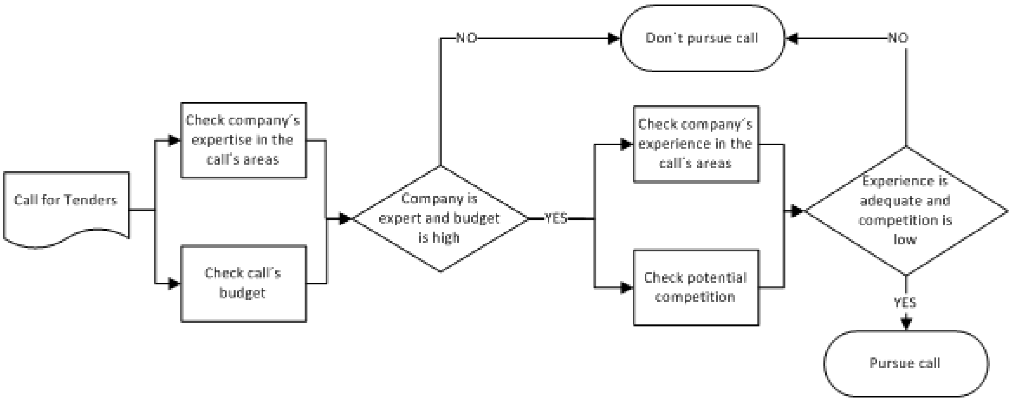
Table 1. Exemplary vague elements from business process ontologies

Element	**Ontology**	**Vagueness**
Managerial Role	BROnt	Combinatory vagueness due to the lack of sharp discrimination between those conditions that are necessary for someone to be considered as having a managerial role
CompetitorRole	BROnt	Degree-vagueness along the dimensions of the number of competitor's business areas and target markets that constitute someone as a competitor
hasBusinessDomain	BPMO	Combinatory vagueness due to the lack of sharp discrimination between those conditions that are necessary for something to belong to a given domain
Strategic Goal	BGO	Combinatory vagueness due to the lack of sharp discrimination between those conditions that are necessary for a goal to be strategic
Desired Result	BMO	Combinatory vagueness when criteria for desirability have not been set or are vague, degree-vagueness when these criteria are arithmetic

[1] http://www.ip-super.org/

The second case of business process knowledge we considered involved a specific application scenario, derived from [1], where the process of tender call evaluation had to be modelled as part of a decision support system. A tender call is an open request made by some organization for a written offer concerning the procurement of goods or services at a specified cost or rate. The evaluation of a tender call by a company refers to the process of deciding whether it should devote resources for preparing a competitive tender in order to be awarded the bid. A diagram describing this business process is depicted at figure 1.

Fig. 1. Tender Call Evaluation Process


Call for Tenders
Check company's expertise in the call's areas
Check call's budget
Company is expert and budget is high
NO
YES
Don't pursue call
Check company's experience in the call's areas
Check potential competition
Experience is adequate and competition is low
NO
YES
Pursue call

Our analysis of this process involved identifying which aspects of it (structure, domain knowledge etc) had vague characteristics. Our findings can be summarized as follows: First, some of the process's various decision conditions according to which a specific action is decided are vague. For example, in order to take the decision about pursuing the call, two criteria that need to be satisfied are i) the budget of the project to be **high** and ii) the company's experience to be **adequate**. In both cases there could be borderline cases as it is indeterminate what is the exact threshold over which the budget is considered high (degree vagueness) or how many years and how many projects are required exactly for the company to be considered experienced in a given area (degree vagueness in two dimensions). Second, many of the underlying organizational and domain pieces of knowledge that are needed for performing various steps of the overall process are also vague. For example, the assessment of the potential competition for the call requires knowledge about the company's competitors. Yet, the existence of other companies that are borderline competitors is possible, mainly due to the lack of clear criteria about what constitutes a competitor and what not (combinatory vagueness). A similar argument can be made for the knowledge about the company's areas of expertise.

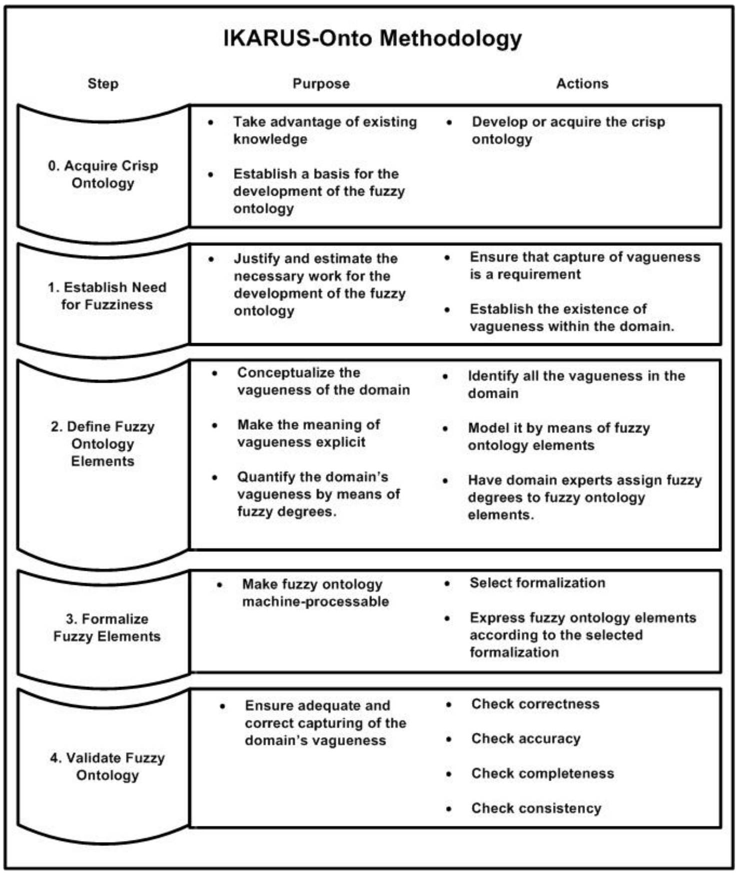
This second case illustrates, apart from the existence of vagueness in a common business process, the potential problems that may be caused during the latter's execution when this vagueness is not formally considered. Different people who will perform the same process will most likely produce different results, exactly because they will interpret various pieces of knowledge in a different manner (e.g. what budget is considered "high" or which companies are competitors). And it should be noted that this is not merely a problem of inadequate measurement or lack of concrete business rules but an inherent problem caused by the vagueness of human knowledge. For example, even if there is a business rule suggesting that competitors are those who have clients in the same industries and services in the same areas the question remains: what is the minimum number of similar clients or services that a given company needs to have in order to be considered a competitor? In the next section we describe how fuzzy ontologies may be practically applied for dealing with questions like that.

3 Modeling and Using Vague Business Processes Knowledge through Fuzzy Ontologies

A fuzzy ontology utilizes notions from Fuzzy Set Theory in order to formally represent the vague ontological elements described in paragraph 2.1. The basic elements it provides include i) **Fuzzy Concepts**, namely concepts to whose instances may belong to them to certain degrees (e.g. *Goal X is an instance of StrategicGoal at a degree of 0.8*), ii) **Fuzzy Relations/Attributes**, namely relations and attributes that link concept instances to other instances or literal values to certain degrees (e.g. *John is expert at Knowledge Management to a degree of 0.5*) and iii) **Fuzzy Datatypes**, namely sets of vague terms which may be used within the ontology as attribute values (e.g. the attribute *experience* mentioned above). In a fuzzy datatype each term is mapped to a fuzzy set that assigns to each of the datatype's potential exact values a fuzzy degree indicating the extent to which the exact value and the vague term express the same thing (e.g. *a consultant with 5 years of experiences is considered junior to a degree of 0.4*)

As with classical ontologies, using and applying fuzzy ontologies in practical scenarios requires corresponding methods and tools for developing them, formally representing them and performing reasoning and querying over them. Two recent developments that may cover a great part of these requirements is the IKARUS-Onto [2] methodology and the Fuzzy OWL 2 framework [4]. The first, depicted at figure 2, provides concrete steps and guidelines for identifying vague knowledge and conceptually modelling it by means of fuzzy ontology elements, placing particular emphasis into the explicitness and shareability of the vagueness's meaning. The second enables the formalization of fuzzy ontologies through the OWL 2 language and provides querying and reasoning services over them through a corresponding reasoner [3]. The adoption and application of the above frameworks in semantic business process modeling is quite straightforward as they are based on and extend already established methods and tech-

Fig. 2. The IKARUS-Onto Methodology



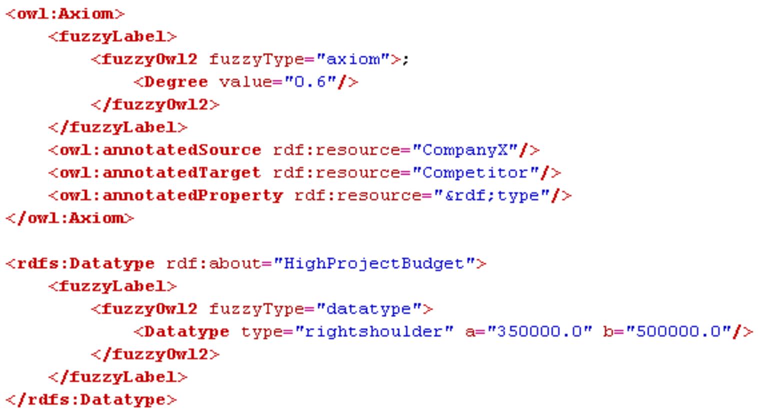
IKARUS-Onto Methodology

Step	Purpose	Actions
0. Acquire Crisp Ontology	• Take advantage of existing knowledge • Establish a basis for the development of the fuzzy ontology	• Develop or acquire the crisp ontology
1. Establish Need for Fuzziness	• Justify and estimate the necessary work for the development of the fuzzy ontology	• Ensure that capture of vagueness is a requirement • Establish the existence of vagueness within the domain.
2. Define Fuzzy Ontology Elements	• Conceptualize the vagueness of the domain • Make the meaning of vagueness explicit • Quantify the domain's vagueness by means of fuzzy degrees.	• Identify all the vagueness in the domain • Model it by means of fuzzy ontology elements • Have domain experts assign fuzzy degrees to fuzzy ontology elements.
3. Formalize Fuzzy Elements	• Make fuzzy ontology machine-processable	• Select formalization • Express fuzzy ontology elements according to the selected formalization
4. Validate Fuzzy Ontology	• Ensure adequate and correct capturing of the domain's vagueness	• Check correctness • Check accuracy • Check completeness • Check consistency

niques from traditional ontology modeling. Thus, the semantic treatment of a business process whose related knowledge includes vagueness would involve i) the building of a process ontology without considering vagueness ii) the fuzzification of this ontology through IKARUS-Onto and Fuzzy OWL 2 and iii) the querying/reasoning over the fuzzy process ontology through Fuzzy OWL 2.

To assess the feasibility and potential value of our approach we applied it in developing a fuzzy ontology for the aforementioned tender call evaluation process. Due to space limitations we cannot describe this ontology in detail, nevertheless it suffices to say that pieces of knowledge with degree-vagueness were typically modelled as fuzzy datatypes (e.g. budget and experience) while those with combinatory vagueness (e.g. competitors) were modelled as fuzzy relations or concepts. Figure 3 depicts two sample definitions in Fuzzy OWL 2.

Fig. 3. Sample of Fuzzy Tender Process Ontology



```
<owl:Axiom>
    <fuzzyLabel>
        <fuzzyOwl2 fuzzyType="axiom">;
            <Degree value="0.6"/>
        </fuzzyOwl2>
    </fuzzyLabel>
    <owl:annotatedSource rdf:resource="CompanyX"/>
    <owl:annotatedTarget rdf:resource="Competitor"/>
    <owl:annotatedProperty rdf:resource="&rdf;type"/>
</owl:Axiom>

<rdfs:Datatype rdf:about="HighProjectBudget">
    <fuzzyLabel>
        <fuzzyOwl2 fuzzyType="datatype">
            <Datatype type="rightshoulder" a="350000.0" b="500000.0"/>
        </fuzzyOwl2>
    </fuzzyLabel>
</rdfs:Datatype>
```

In any case, the developed fuzzy ontology formed the basis of a simple decision support system which queried the ontology in order to evaluate the tender call evaluation criteria and provide a suggestion to its users on whether they should pursue a given call. We had this system used by some of our consultants and we asked from them to provide some informal feedback. The positive remarks we received regarded the explicitness of the fuzzy elements's meaning (a result of using IKARUS-Onto) as well as the automation achieved in the retrieval of vague knowledge that made easier the evaluation of the fuzzy decision criteria. On the other hand, the task of defining fuzzy degrees and membership functions was deemed as quite difficult and time-consuming, indicating thus the need for more automated methods for vague knowledge acquisition.

4 Conclusions and Future Work

In this paper we utilized concrete examples and use cases from the literature in order to highlight the omnipresence of vagueness in business process knowledge and to argue in favour of using fuzzy ontologies for dealing with it in a semantically rich manner. Furthermore, we provided a small but representative example of how some of the latest fuzzy ontology engineering methods and tools may be practically used for dealing with vagueness in business process modeling. In the future we intend to further substantiate and support the use of fuzzy ontologies in semantic business process modeling by providing dedicated methods and tools for the capture, management and exploitation of vague process knowledge and by applying and evaluating these methods in real application scenarios.

Acknowledgements

The research leading to this work has received partial funding from the ECs FP7 programme: (FP7/2011-14) under grant agreement no. 287615 (PARLANCE).

References

1. Alexopoulos P, Wallace M, Kafentzis K, Thomopoulos (2009) A Fuzzy Knowledge-based Decision Support System for Tender Call Evaluation. 5th IFIP Conference on Artificial Intelligence Applications & Innovations.
2. Alexopoulos P., Wallace M.,Kafentzis K. and Askounis D. (2011) A Methodology for Developing Fuzzy Ontologies, Knowledge and Information Systems, pp. 1-29, Springer.
3. Bobillo F, Delgado M, Gomez-Romero J (2008) DeLorean: A Reasoner for Fuzzy OWL 1.1. 4th International Workshop on Uncertainty Reasoning for the Semantic Web, October 2008.
4. Bobillo F., Straccia U., Fuzzy Ontology Representation using OWL 2. International Journal of Approximate Reasoning 52(7):1073-1094, 2011.
5. Cabral, Liliana; Norton, Barry and Domingue, John (2009). The business process modelling ontology. In: 4th International Workshop on Semantic Business Process Management, Crete, Greece, 2009.
6. Cox, E. (2002). Knowledge-Based Business Process Modeling: Complex Systems Design Through A Fusion of Computational Intelligence And Object-Oriented Models. PC AI, 16(2), 1523
7. Filipowska, A., Kaczmarek, M. and Markovic, I., 2008. Organizational Ontologies to Support Semantic Business Process Management. In Information Systems Journal.
8. Chiara Di Francescomarino, Chiara Ghidini, Marco Rospocher, Luciano Serafini, and Paolo Tonella. Semantically-aided business process modeling. In 8th International Semantic Web Conference, pages 114 129. Springer, 2009.
9. M. Hepp, F Leymann, J. Domingue, A. Wahler, and D. Fensel. Semantic business process management: A vision towards using semantic web services for business process management. In IEEE International Conference on e-Business Engineering, pages 535540, 2005.
10. Hyde D. (2008) Vagueness, Logic and Ontology. Ashgate New Critical Thinking in Philosophy.
11. Klir G, Yuan B (1995) Fuzzy Sets and Fuzzy Logic, Theory and Applications. Prentice Hall.
12. Markovic, I. and Kowalkiewicz, M., 2008. Linking Business Goals to Process Models in Semantic Business Process Modeling. In 12th International IEEE Enterprise Distributed Object Computing Conference. IEEE, pp. 332-338.
13. Thomas, O.; Dollmann, T.; Loos, P. (2008): Rules Integration in Business Process Models A Fuzzy Oriented Approach. In: Enterprise Modelling and Information Systems Architectures 3, Nr. 1, S. 1830
14. Völkner, P., Werners, B. (2002). A simulation-based decision support system for business process planning. Fuzzy Sets and Systems, 125(3), 275288.