Proceedings of the

KWEPSY2007

Knowledge Web PhD Symposium 2007

co-located with the 4th Annual European Semantic Web Conference [ESWC2007**]**

June 6, 2007

Innsbruck, Austria

Edited by

Elena Simperl, University of Innsbruck, Austria
Joerg Diederich, L3S Research Center, Hannover, Germany
Guus Schreiber, Vrije Universiteit Amsterdam, Netherlands

Table of Contents

The technical program included the following full papers:

The program was complemented by a poster session as follows:

The workshop Website is available online at http://ontoworld.org/wiki/KWEPSY2007

The Knowledge Web PhD Symposium Series

The Knowledge Web PhD Symposium KWEPSY aims at bringing together doctoral students within the Semantic Web community to open their work up to discussion in a European forum, and to obtain valuable feedback from leading scientists in the field. Participants to the symposium receive constructive comments with respect to topic-specific research issues and are assisted in formulating a coherent research narrative for their doctoral work. Though organized under the umbrella of **Knowledge Web**, the symposium is open to all PhD students carrying out research on topics related to the Semantic Web, whilst priority is given to 1st/2nd year PhD students, because they are still in the process of defining the scope of their research.

Students submit an extended abstract, structured in accordance to a pre-defined template, which has been designed to highlight the key methodological components required for a sound research narrative. The submission needs to address the following aspects:

- Describe the research problem of the PhD thesis and argument its relevant for the Semantic Web area.
- Describe the state of the art, emphasizing the need for improvement and the feasibility of your approach.
- Summarize the expected contributions and outline the real-world use cases (applications, target audience) which will mainly benefit from your work.
- Sketch the research methodology that you have adopted (or you are planning to adopt), in particular your approach to evaluating/validating the results.
- Describe your proposed approach, clearly differentiating between the results achieved so far, the remaining work and the (planned) evaluation.
- Compare and contrast your approach with other existing approaches, in particular highlighting the shortcomings of other approaches, which your approach is planning to tackle.
- Conclude your summary with an outline of the planned future work.

Papers should not exceed 5 pages and are peer-reviewed by at least two members of the scientific advisory board. The submissions are reviewed against the following criteria (in this order):

- Conformance of the submitted abstract to the given template.
- Novelty and originality of the research work.
- Rigorousness and scientific soundness of the overall approach and of the results so far.
- Clarity of the presentation.
- Relevance of the work with respect to the Semantic Web field.

The selected participants are given the opportunity to open their work up to discussion in front of other students and an expert audience (either in a regular presentation session or in a poster session). Each accepted contribution is assigned to a scientific advisor who provides extended feedback to the presented research achievements and to the accuracy of the applied methodology. In addition to full papers, a limited number of papers is accepted as posters, for which the authors are required to submit a 2 page version of the original submission.

The symposium is scheduled as a full-day event, consisting of full paper presentations and a poster session. Each full paper is presented in a talk of 25 minutes (15 minutes presentation, 10 minutes discussion). As a general template for the presentations the speakers are recommended to use the following structure:

- Problem statement (1-2 slides)
- Research questions and expected contributions (1 slide)
- Proposed solution (4-6 slides)
- Evaluation or evaluation plans (1-2 slides)
- Future work (1 slide)

Best papers are selected by the scientific advisors and awarded during the symposium.

The Knowledge Web PhD Symposium has been held in 2006 and 2007 co-located with the European Semantic Web Conference ESWC.

KWEPSY 2007

The KWEPSY 2007 PhD Symposium brought together 40 researchers, be that doctoral students or senior researchers. The symposium program included 12 full paper presentations, a poster session and extensive discussions. In particular scientific advisors of the accepted papers were kind to give in-depth feedback of the work surveyed.

The symposium organizers, Elena Simperl, Jörg Diederich and Guus Schreiber, received 55 submissions of papers in response to the call for contributions. As a result of the peer reviewing process, 12 full papers and 17 posters of these were selected for publication in these proceedings. The program committee consisted of the following Semantic Web experts from industry or academia: **Alain Leger**, France Telecom, France, **Asuncion Gomez Perez**, Universidad Politecnica Madrid, Spain, **Anupriya Ankolekar**, University of Karlsruhe, Germany, **Axel Polleres**, University Rey Juan Carlos, Spain, **Daniel Olmedilla**, L3S Research Center, Hannover, Germany, **Carole Goble**, University of Manchester, UK, **Diana Maynard**, University of Sheffield, UK, **Elena Simperl**, University of Innsbruck, Austria, **Enrico Franconi**, Free University of Bozen-Bolzano, Italy, **Enrico Motta**, The Open University, UK, **Fabien Gandon**, INRIA Sophia-Antipolis, France, **Frank van Harmelen**, Vrije Universiteit Amsterdam, Netherlands, **Guus Schreiber**, Vrije Universiteit Amsterdam, Netherlands, **Heiner Stuckenschmidt**, University of Mannheim, Germany, **Holger Wache**, Vrije Universiteit Amsterdam, Netherlands, **Ilya Zaihrayeu**, University of Trento, Italy, **Jeff Z. Pan**, University of Aberdeen, UK, **Jerome Euzenat**, INRIA Rhone-Alpes, France, **Jörg Diederich**, L3S Research Center, Hannover, Germany, **Sebastian Schaffert**, Salzburg Research, Austria, **John Breslin**, NUI Galway, Ireland, **Lora Aroyo**, University of Eindhoven, Netherlands, **Lyndon Nixon**, Free University of Berlin, Germany, **Marco Ronchetti**, University of Trento, Italy, **Martin Dzbor**, The Open University, UK, **Michal Zaremba**, University of Innsbruck, Austria, **Pavel Shvaiko**, University of Trento, Italy, **Peter Haase**, University of Karlsruhe, Germany, **Philipp Cimiano**, University of Karlsruhe, Germany, **Richard Benjamins**, Isoco, Spain, **Robert Tolksdorf**, Free University of Berlin, Germany, **Rose Dieng**, INRIA, France, **Sergio Tessaris**, Free University of Bozen-Bolzano, Italy, **Stefan Decker**, NUI Galway, Ireland, **Tomas Vitvar**, NUI Galway, Ireland, **Valentina Tamma**, University of Liverpool, UK, **Walter Binder**, Lugano University, CH, **Wolfgang Nejdl**, L3S Research Center, Hannover, Germany, **Yiannis Kompatsiaris**, Centre for Research and Technology Hellas, Greece, **York Sure**, University of Karlsruhe, Germany

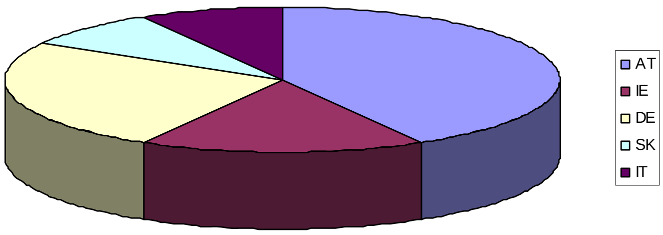
Several members of the program committee agreed to become a scientific advisor of at least one of the student who were accepted to give a presentation in the symposium. They were **Diana Maynard**, University of Sheffield, UK, **Elena Simperl**, University of Innsbruck, Austria, **Enrico Franconi**, Free University of Bozen-Bolzano, Italy, **Heiner Stuckenschmidt**, University of Mannheim, Germany, **Holger Wache**, Vrije Universiteit Amsterdam, Netherlands, **Jerome Euzenat**, INRIA Rhone-Alpes, France, **Jörg Diederich**, L3S Research Center, Hannover, Germany, **Sebastian Schaffert**, Salzburg Research, Austria, **Peter Haase**, University of Karlsruhe, Germany.

The organization committee would like to thank all PC members and in particular the scientific advisors for their thorough and substantial reviews, which were crucial for the success of the actual event.

At the beginning of the symposium Elena Simperl gave a short introductory talk, which outlined the motivation and objectives of the symposium, introduced the technical program and the best paper award. The technical program consisted of three sessions of paper presentations, a poster session, and a wrap-up session in which the attendees participated in a lively discussion on the general objectives, the format and the organization of the symposium.

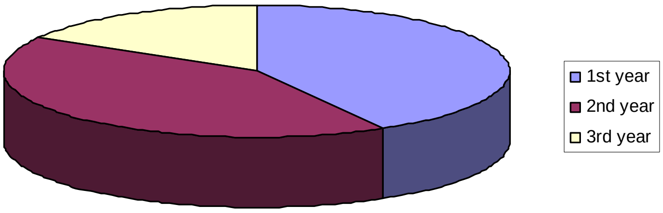
The presentations held on this workshop covered the areas Semantic Web Services, languages and reasoning, ontology engineering, information extraction and ontology learning. The authors were affiliated to institutions widely spread across Europe, with a slight majority of Austrian institutions which can be traced back to the location of the event (Figure 1).

Figure 1: Distribution of papers per country



Further on, the work reported in the accepted papers was in various stages, from early to very advanced, while most of the papers did not describe completed PhD research (Figure 2).

Figure 2: Stage of the work reported in the papers



The technical program included the following full papers:

- *Caching for Semantic Web Services*, Michael Stollberg, Digital Enterprise Research Institute, University of Innsbruck
- *Towards Novel Techniques for Reasoning in Expressive Description Logics based on Binary Decision Diagrams*, Uwe Keller, Digital Enterprise Research Institute, University of Innsbruck
- *Process Mediation in Semantic Web Services*, Emilia Cimpian, Digital Enterprise Research Institute, University of Innsbruck
- *Scalable Web Service Composition with Partial Matches*, Adina Sirbu, Jörg Hoffmann, Digital Enterprise Research Institute, University of Innsbruck
- *Improving Email Conversation Efficiency by Enhancing Email with Semantics*, Simon Scerri, Digital Enterprise Research Institute, National University of Ireland Galway
- *Towards Distributed Ontologies with Description Logics*, Martin Homola, Comenius University
- *A Framework for Distributed Reasoning on the Semantic Web Based on Open Answer Set Programming*, Cristina Feier, Digital Enterprise Research Institute, University of Innsbruck
- *Logic as a power tool to model negotiation mechanisms in the Semantic Web Era*, Azzura Ragone, SisInfLab, Politecnico di Bari, Italy
- *Improving the Usability of Large Ontologies by Modularization*, Anne Schlicht, University of Mannheim
- *Inferential Ontology Learning*, Vit Novacek, Digital Enterprise Research Institute, National University of Ireland Galway
- *Imprecise SPARQL: Towards a Unified Framework for Similarity-Based Semantic Web Tasks*, Christoph Kiefer, Department of Informatics, University of Zurich
- *Semiautomatic Creation of Semantic Networks*, Lars Bröcker, Fraunhofer IAIS

The program was complemented by a poster session as follows:

- *Towards Cross-Media Document Annotation*, Ajay Chakravarthy, Department of Computer Science, Sheffield
- *Semantic Business Process Modeling*, Yan Zhixian, Digital Enterprise Research Institute, University of Innsbruck
- *Towards Open Ontology Engineering*, Katharina Siorpaes, Digital Enterprise Research Institute, University of Innsbruck
- *Ontology-based Virtual Data Integration for E-Science Grids*, Andreas Langegger, Institute of Applied Knowledge Processing, Johannes Keppler University Linz
- *Research on collaborative information sharing systems*, Davide Eynard, Dipartimento di Elettronica e Informazione, Politecnico di Milano
- *On the communication and coordination issues of Semantic Web Services using Triple Space Compunting*, Omair Shafiq, Digital Enterprise Research Institute, University of Innsbruck
- *Reasoning with Large Data Sets*, Darko Anicic, Digital Enterprise Research Institute, University of Innsbruck
- *Towards a Semantic Wiki for Science*, Christoph Lange, Computer Science, Jacobs University Bremen
- *Ontology-Driven Management of Space Middleware*, Reto Krummenacher, Digital Enterprise Research Institute, University of Innsbruck
- *Intelligent Search ina Collection of Video Lectures*, Angela Fogarolli, Dept. of Information and Communication Tech., University of Trento
- *A Collaborative Semantic Space for Enterprise,* Alexandre Passant, Laboratoire LaLICC, Universite Paris IV Sorbonne
- *A Directed Hypergraph Model for RDF*, Amadis Antonio Martinez Morales, Universidad de Carabobo, Venezuela
- *Applying Semantic Technologies to the Design of Open Service-oriented Architectures for Geospatial Applications*, Thomas Usländer, Fraunhofer IITB
- *Pattern-based Ontology Construction*, Eva Blomqvist, Jönköping University
- *Semantic Group Formation*, Asma Ounnas, School of Electronics and Computer Science, University of Southhampton
- *Combining HTN-DL Planning and CBR to compound Semantic Web Services*, Antonio Sanchez-Ruiz, Dep. Ingenieria del Software e Inteligencia Artificial, Universidad Complutense de Madrid
- *Ontology Mapping Specification Language*, Francois Scharffe, Digital Enterprise Research Institute, University of Innsbruck

From the accepted full papers mentioned above the following three have been selected as best paper candidates based on the recommendations of the reviewers:

- *Caching for Semantic Web Services*, Michael Stollberg, Digital Enterprise Research Institute, University of Innsbruck
- *Towards Distributed Ontologies with Description Logics*, Martin Homola, Comenius University, Slovakia
- *Improving the Usability of Large Ontologies by Modularization*, Anne Schlicht, University of Mannheim, Germany

The best paper award was sponsored by the **BIT Joint School for Information Technology** and was won by Anne Schlicht from the University of Mannheim.

Conclusions and outlook

The feedback we received in the wrap-up session was similar to the previous edition of the event (especially mentioning this time the focus on 1st/2nd year students), even though the time for presentation and discussion was considered to be too short and also the poster session should have been longer. An introductory session for the scientific advisors was considered useful and will be re-introduced next year. Furthermore, some participants complained that the feedback they received from the reviewing process was very high level. This was because of the overwhelming number of submissions, which is could not been foreseen and which increased the workload for the reviewers considerably. A nice idea (though difficult to implement) was to actually make the mentors present the work of the student to deepen the discussion on the topic (though this will exclude the ability of the students to train presentations about their theses).

In summary, the Knowledge Web PhD symposium KWEPSY 2007 has been positively evaluated by both the mentors and the attending PhD students. The setup of the symposium was very successful and we plan only minor modification, such as the introduction of the mentors' introductory presentations or a higher number of scientific advisors to be able to accommodate a larger number of submissions expected for 2008.

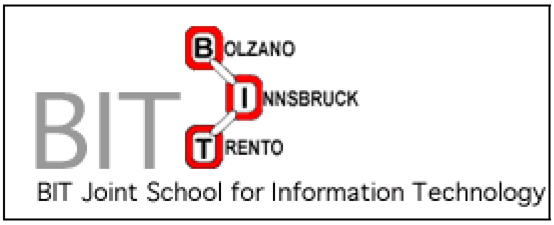
Starting from 2007, the symposium has become an integral part of the European Semantic Web Conference (ESWC) which ensures that the symposium will continue to exist even beyond the end of the Knowledge Web project. This is also underlined by the fact that several institutions in the Semantic Web field such as the European Association for Semantic Web Education EASE and the Semantic Technologies Institute **STI International** decided to officially endorse the PhD symposium starting from next year.

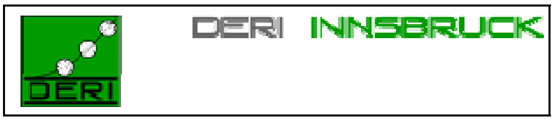
Innsbruck, Hannover and Amsterdam
August, 2007

Elena Simperl
Jörg Diederich
Guus Schreiber

Sponsors

The PhD Symposium is supported by





Caching for Semantic Web Service Discovery

Michael Stollberg

Digital Enterprise Research Institute (DERI),
University of Innsbruck, Austria
michael.stollberg@deri.org

Abstract. This document is an extended abstract on a PhD work that develops an efficient, scalable, and stable Web service discovery engine. These qualities become important for discovery engines that serve as a software component in automated SOA technologies. Based on a profound formal specification, the approach is to capture design time discovery results and then use this knowledge for efficient runtime discovery. The work is evaluated by a statistical time efficiency comparison with other Web service discovery engines, and by a applicability study in real-world SOA applications.

Keywords: Semantic Web Services, Goals, Functional Descriptions, Discovery, Efficiency, Scalability, Stability

1 Introduction

Discovery is one of the central reasoning tasks in SOA systems, concerned with the detection of usable Web services for a specific request or application context. Aiming at the automation of this task, most existing works on semantically enabled Web service discovery focus on the quality of the applied matchmaking techniques. However, the following qualities become important for using an automated Web service discovery engine as a reliable software component in a SOA system: *efficiency* as the time required for finding a usable Web service, *scalability* as the ability to deal with large numbers of available Web services, and *stability* as the behavioral constancy among several invocations.

My PhD work addresses this challenge by applying the concept of caching to Web service discovery. For this, I extend the goal-driven approach that is promoted by the WSMO framework (`www.wsmo.org`). I distinguish *goal templates* as generic objective descriptions and *goal instances* as instantiations of a goal template that denotes concrete client requests. At design, Web service discovery for goal templates is performed. The result is stored in a graph that organizes goal templates by their semantic similarity and captures the minimal knowledge on the usable Web services for each goal template. This knowledge is utilized for efficient runtime discovery, i.e. the detection of a usable Web service for solving a goal instance that is defined by a client. In particular, this is achieved by:

1. *pre-filtering* as only the Web services that are usable for the corresponding goal template are potential candidates for the goal instance, and
2. *minimal use of a reasoner* for matchmaking because in certain situations the usability of a Web service for a goal instance can be directly inferred.

2 Solution Overview

My work extends the approach for Web service discovery promoted by the WSMO framework with a refined goal model and a rigid formalization for the functional aspects of Web service discovery. On this basis, the so-called *Semantic Discovery Caching* technique (short: SDC) caches the minimal knowledge in order to optimize the computational qualities of Web service discovery.

2.1 Web Service Discovery Framework

Figure 1 shows the conceptual model as a dataflow diagram. It deals with three entities: *Web services* that have a formal description and are accessible via a WSDL interface, *goal templates* as formalized, generic objective descriptions that are stored in the system, and *goal instances* that formally describe a concrete request by instantiating a goal template with concrete inputs. At design time, Web services for goal templates are discovered. The result is cached in the SDC graph, the knowledge structure for optimizing the Web service discovery process. At runtime, a concrete client request is formulated as a goal instance. The runtime discovery finds one usable Web service for solving this. It uses the cached knowledge for optimization, in particular for pre-filtering and minimizing the number of necessary matchmaking operations.

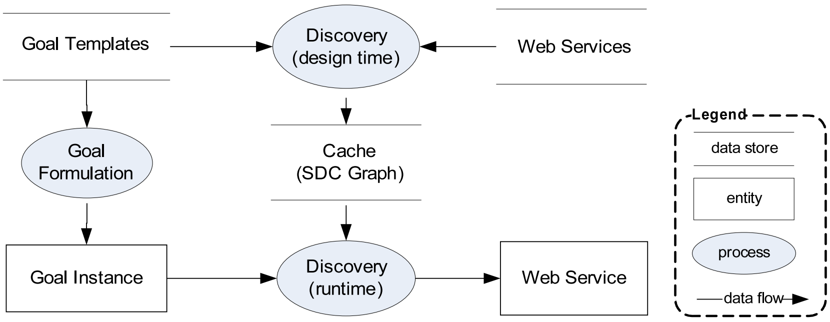


Fig. 1. Overview of Web Service Discovery Framework

In contrast to an invocation request for a Web service, a goal formally describes a client objective of getting from the current state of the world into a state wherein the objective is satisfied. This provides an abstraction layer for facilitating problem-oriented Web service usage: the client merely specifies the objective to be achieved as a goal, and the system discovers, composes, and executes suitable Web services for solving this. The distinction of goal templates and goal instances allows to better support the goal formulation by clients (e.g. by form-based instantiation through a graphical user interface), and – more importantly – provides the foundation for the two-phase Web service discovery outlined above.

I consider functional aspects as the primary aspect for discovery: if a Web service does not provide the functionality for solving a goal, then it is not usable and other, non-functional aspects are irrelevant. For this, the possible solution for goals and possible executions of Web services are formally described by functional descriptions $\mathcal{D} = (\Sigma, \Omega, IN, \phi^{pre}, \phi^{\mathit{eff}})$; Σ is the signature, Ω are domain ontologies, IN are the input variables, the precondition ϕ^{pre} and the effect ϕ^{eff} constraint the start- and end states. As the design time discovery result, the usability of a Web service W for a goal template $\mathcal{G}$ is expressed in terms of matching degrees (*exact, plugin, subsume, intersect, disjoint*). A goal instance is defined as a pair $GI(\mathcal{G}) = (\mathcal{G}, \beta)$ with the corresponding goal template $\mathcal{G}$ and an input binding β that is used to invoke a Web service W for solving $GI(\mathcal{G})$. If W is usable for $\mathcal{G}$ under the degrees *exact* or *plugin*, then W is also usable for any $GI(\mathcal{G})$; under the degrees *subsume* and *intersect*, additional matchmaking is required at runtime; if W is not usable for $\mathcal{G}$ it is also not usable for $GI(\mathcal{G})$.

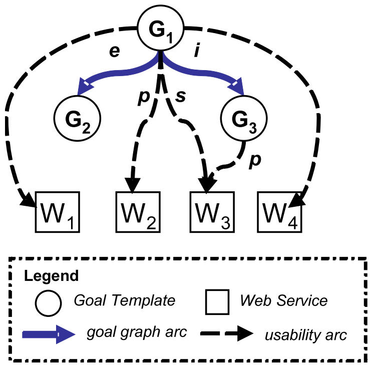
2.2 Semantic Discovery Caching

The main contribution of my work is the SDC technique as the solution for enabling efficient, scalable, and stable Web service discovery. Its purpose is to improve the computational quality of the runtime discovery process by exploiting the relationships between goal templates, goal instances, and Web services.

The central element is the SDC Graph that provides an index structure for efficient search of goal templates and usable Web services. It organizes goal templates with respect to their semantic similarity, and keeps the minimal knowledge on the usability of the available Web services. Two goal templates $\mathcal{G}_i$ and $\mathcal{G}_j$ are considered to be similar if they have at least one common solution; if this is given, then mostly the same Web services are usable for them. In consequence, the upper layer of a SDC graph is the *goal graph* that organizes goal templates in a subsumption hierarchy, and the lower layer is the *usability cache* that captures the minimal knowledge on the usability of the available Web services. Upon this cache structure, the discovery operations make use of inference rules between the similarity degree of goal templates and the usability degree of Web services.

For illustration, Figure 2 shows an example of an SDC graph along with the most relevant inference rules. This considers three goal templates: $\mathcal{G}_1$ for package shipment within Europe, $\mathcal{G}_2$ for Switzerland, and $\mathcal{G}_3$ for Germany. As each solution for $\mathcal{G}_2$ is also a solution of $\mathcal{G}_1$, their similarity degree is *subsume*; the same holds between $\mathcal{G}_3$ and $\mathcal{G}_1$. These relationships are expressed by directed arcs in goal graph. Besides the goal templates, let there be some Web services, among them e.g. W_1 that provides package shipment within Europe, W_2 throughout the whole world, W_3 within the European Union, and W_4 within the Commonwealth. Their usability degree for each goal template is explicated by directed arcs in the usability cache. This knowledge is efficiently used for runtime discovery. Consider a goal instance for shipping a package from Munich to Berlin: its corresponding goal instance is $\mathcal{G}_3$; because W_1, W_2, and W_3 are usable for $\mathcal{G}_3$ under the *plugin* degree, we know that each of them is usable for solving the goal instance without the need of a matchmaker during runtime discovery.

Structure of an SDC Graph



inference rules for $subsume(\mathcal{G}_1, \mathcal{G}_2)$

(1) $exact(\mathcal{G}_1, W) \Rightarrow plugin(\mathcal{G}_2, W)$.
(2) $plugin(\mathcal{G}_1, W) \Rightarrow plugin(\mathcal{G}_2, W)$.
(3) $subsume(\mathcal{G}_1, W) \Rightarrow exact(\mathcal{G}_2, W)$ or
(4) $subsume(\mathcal{G}_1, W) \Rightarrow plugin(\mathcal{G}_2, W)$ or
(5) $subsume(\mathcal{G}_1, W) \Rightarrow subsume(\mathcal{G}_2, W)$ or
(6) $subsume(\mathcal{G}_1, W) \Rightarrow intersect(\mathcal{G}_2, W)$ or
(7) $subsume(\mathcal{G}_1, W) \Rightarrow disjoint(\mathcal{G}_2, W)$.
(8) $intersect(\mathcal{G}_1, W) \Rightarrow plugin(\mathcal{G}_2, W)$ or
(9) $intersect(\mathcal{G}_1, W) \Rightarrow intersect(\mathcal{G}_2, W)$ or
(10) $intersect(\mathcal{G}_1, W) \Rightarrow disjoint(\mathcal{G}_2, W)$.
(11) $disjoint(\mathcal{G}_1, W) \Rightarrow disjoint(\mathcal{G}_2, W)$.

Fig. 2. Example of a SDC Graph and Inference Rules

The SDC graph during its life time are maintained by algorithms that handle the addition, removal, and modification of goal templates and Web services. Two refinements ensure that the SDC graph exposes sophisticated search properties: (1) the only similarity degree that occurs in the goal graph is *subsume*, and (2) the minimization of the usability cache in order to avoid redundancy. The SDC technique is implemented as a discovery component of the WSMX system, available at the SDC homepage: `members.deri.at/~michaels/software/sdc/`.

3 Evaluation

To demonstrate the achievable quality increase for Web service discovery, I have run several comparison test between the SDC-enabled runtime discovery and an engine that applies the same matchmaking techniques but does not make use of the cached knowledge. Table 1 shows a snapshot of the statistical prepared test results; details and the original test data are available from SDC homepage. This clearly shows that the SDC discovery is **efficient** (the average time is always lower), **scalable** (the time for the SDC discovery remains the same for increasing numbers of Web services), and **stable** (the standard deviation is significantly smaller than the one of the comparison engine).

Another relevant aspect is the appropriateness of the assumptions that underly the conceptual model. For this, I have examined the applicability in real-world settings – e.g. in one of the world's largest SOA systems at telecommunication provider *Verizon*. In summary, there are many Web services that provide similar functionalities but differ in the detailed usage conditions. Also, the usage requests posted by the consuming applications can be expressed in terms of goals; these can be organized in a fine-grained subsumption hierarchy in the SDC graph so that its benefits for efficient runtime discovery can be exploited. Besides, the distinction of goal templates and goal instances has been regarded by practioneers as suitable way for realizing problem-oriented Web service usage.

Table 1. Comparison Test Statistics (all values in seconds)

No. of WS	engine	mean μ	median $\bar{x}$	standard deviation σ
10	SDC	0.28	0.27	0.03
	non-SDC	0.41	0.39	0.21
100	SDC	0.29	0.28	0.03
	non-SDC	3.96	3.68	2.55
2000	SDC	0.31	0.29	0.05
	non-SDC	72.96	65.55	52.13

4 Related Work and Publications

Very few existing works address the computational quality of Web service discovery techniques. I am not aware of any other approach that addresses this problem in a similar way. The following outlines the relationship to related research fields; details are discussed in the publications listed below.

Semantic Web Service Discovery. Most works are only concerned with the matchmaking techniques. As a contribution to this end, my work is based on a formal model that describes requested and provided functionalities on the level of executions of Web services and solutions for goals (*cf.* Section 2).

Web Service Repository Indexing. Other approaches reduce the search space for discovery by indexing Web service repositories. Keyword-based categorization as already supported by UDDI is imprecise in comparison to the SDC graph. More sophisticated solutions create a search tree based on formal descriptions; this can achieve logarithmic search time, but – in contrast to SDC – still requires several matchmaking operations for each request.

Caching. Caching techniques are a well-established means for performance increase in several areas of computing. Respective studies show that caching can achieve the highest efficiency increase if there are many similar requests. The SDC graph can be understood as a cache structure for Web service discovery.

Scalable Ontology Repositories. Works on scalable ontology reasoning infrastructures minimize the reasoning effort at runtime, e.g. by materalization and organization of the available knowledge at design time. However, such techniques can not replace the SDC technique because it defines a specific knowledge structure and algorithms for Web service discovery.

Publications (most relevant)

Stollberg, M. and Norton, B.: *A Refined Goal Model for Semantic Web Services.* In Proc. of the 2nd International Conference on Internet and Web Applications and Services (ICIW 2007), Mauritius, 2007.

Stollberg, M.; Keller, U.; Lausen, H. and Heymans, S.: *Two-phase Web Service Discovery based on Rich Functional Descriptions.* In Proc. of the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, 2007.

Stollberg, M.; Hepp, M., Hoffmann, J.: *Efficient and Scalable Web Service Discovery with Caching.* Submitted to 6th International Semantic Web Conference (ISWC 2007).

Towards Novel Techniques for Reasoning in Expressive Description Logics based on Binary Decision Diagrams *

Uwe Keller

Digital Enterprise Research Institute (DERI), University of Innsbruck, Austria
uwe.keller@deri.org

Abstract. We propose to design and study new techniques for description logic (DL) reasoning based on a prominent data structure that has been applied very successfully in various domains in computer science where one has to face the efficient representation and processing of large scale problems: *Binary Decision Diagrams* (BDDs). BDDs have been used very successfully for reasoning in propositional logics, and have been lifted to the level of first-order logics, too. In both cases, they provide a rich semantic structure to guide proof search. Therefore, we believe that (i) BDDs are interesting to study in the context of reasoning for a logic of intermediate expressivity (such as DLs) and (ii) that they provide a fertile ground for the design of novel efficient methods for reasoning in particular expressive DLs. The project will help to enrich the available machinery of DL reasoning techniques.

1 Introduction

Description Logics (DLs) [1] are a family of class-based knowledge representation formalisms characterised by the use of various constructors to build complex classes from simpler ones, and by an emphasis on the provision of sound, complete and (empirically) tractable reasoning services. They have a wide range of applications, but are most widely known as the basis for ontology languages such as OWL. Recently, [14] pointed out that the increasing use of DL-based ontologies in areas such as e-Science and the Semantic Web however is already stretching the capabilities of existing DL systems, and brings with it a range of challenges for future research on reasoning methods for DL. Key issues here are the provision of efficient algorithms that allow (advanced) applications (i) to scale up to knowledge bases of practical relevance and (ii) to leverage expressive languages for capturing domain knowledge. However, expressiveness of DLs comes at a price: the theoretically high (worst-case) complexity of relevant reasoning tasks. Hence, it is unlikely, that there is a *single* method, that performs well in all possible cases. Rather, one can expect that specific techniques perform well one particular classes of problems.

So far, research in practical DL reasoning methods has centered around structural subsumption algorithms [2] and tableau methods [13], and have recently been extended by the application of the resolution principle [16,18] (and optimized evaluation techniques from the area of deductive databases) to expressive DLs. Automata-based approaches (e.g.(e.g. [24]) (although possible in theory) have had nearly no impact on the development of practical reasoning algorithms for DLs.

Based on the observation of recent trends in the area of DL reasoning and the research challenge identified in [14] for this field, we propose to design and research novel techniques for Description Logic reasoning that are based on a well-known principles of reasoning that (a) has been studied for other (especially more expressive) logics, (b) proved itself to be a successful method of reasoning for these logics and (c) work significantly different from current state-of-the-art techniques in DL reasoning. More specifically, we propose to investigate the use of *Binary Decision Diagrams* (BDDs) [3] and their manifold variants and extensions as a fundamental framework for realizing well-known reasoning tasks, in particular for expressive DLs.

2 Binary Decision Diagrams and their Variants

A Binary Decision Diagram (BDD) [3] is a simple data structure for representing an (n-ary) boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$. A boolean function $f(x_1, \ldots, x_n)$ can be represented as a rooted, directed, acyclic graph, which consists of decision nodes and two terminal nodes called 0-terminal and 1-terminal. Each decision node is labeled by a Boolean variable x_i and has two child nodes called low child and high child. The edge from a node to a low (high) child represents an assignment of the variable to 0 (1). Such a BDD is called *ordered* if different variables appear in the same order on all paths from the root. It is called *reduced* if the graph is reduced according to two rules: (i) merge any isomorphic subgraphs, and (ii) eliminate any node whose two children are isomorphic. Consequently, reduced BDDs reuse structures in the BDD representation to a maximum extent and therefore shrink the size of the representation. Most

* This work has been funded by the Austrian Federal Ministry for Transport, Innovation, and Technology under the project *Semantic Engineering Support Environment* (SEnSE, FIT-IT contract FFG 810807).

often, the term BDD refers actually to Reduced Ordered Binary Decision Diagram (ROBDD), i.e. BDDs that are reduced in regard of a specific (given) order. The advantage of an ROBDD is that it is canonical (unique) for a particular boolean function: Although the boolean function might have various (equivalent) descriptions, the respective ROBDD is unique. A path from the root node to the 1-terminal represents a (possibly partial) variable assignment for which the represented Boolean function has the value true. As the path descends to a low child (high child) from a node, then that node's variable is assigned to 0 (1).

The fundamental and most important characteristic of ROBDDs hereby is (i) an extreme compression in many practical cases (after the application of reduction rules to remove eliminate redundancies) and (ii) very fast implementations of standard operations on boolean functions. Although the (naive) representation of a boolean function in a BDD might be very large (and require exponential space wrt. the number of boolean input variables), given some fixed ordering on input variables, BDDs can most often be reduced to an OBDDs (representing the same function) such that the resulting representation actually is comparably small (e.g. polynomial wrt. the BDD representation). In particular, for a given ordering the reduced form is unique, and the OBDD for the $n - ary$ boolean function which returns always 0 consists only of the 0-terminal node. The achievable compression crucially depends on the chosen variable ordering, i.e for many boolean functions there exists an ordering such that the corresponding reduced OBDD has a minimal size. On the other hand, there are functions (e.g. the multiplication function) that are inherently difficult, i.e. no variable ordering exists such that the reduced OBDD has small size. Practical experience shows, that such functions are rare in many industrial applications. Finding an optimal variable ordering is known to be intractable [26], however heuristics often work well in practice [23].

BDDs have been applied in various domains, most prominently hardware verification [17], in Computer Aided Design (CAD), and in software to synthesize circuits (logic synthesis). Very often, they superseded previously known methods. Various variations and generalizations of BDDs have been developed over time to overcome limitations for particular domains, e.g. Zero Suppressed Decision Diagrams (ZDDs), Binary Moment Diagrams (BMDs), Free Binary Decision Diagrams (FBDDs), (reduced ordered) Multi-valued Decision Diagrams ((RO)MDDs).

3 How to Reason with BDDs

BDDs can be used for reasoning in propositional logics straightforwardly: A propositional formula $\phi(x_1, \ldots, x_n)$ containing n propositional variables x_i can be seen as an n-ary boolean function. To construct a BDD for $\phi(x_1, \ldots, x_n)$ one can apply *Shannon's decomposition principle*: for all boolean variable assignments $x_1, \ldots, x_n \in \{0, 1\}$ it holds that $\phi \Leftrightarrow (x_i \Rightarrow \phi\{x_i/1\}) \wedge (\neg x_i \Rightarrow \phi\{x_i/0\})$, where $\phi\{x/\phi'\}$ denotes the formulae which is constructed from ϕ by replacing all occurrences of x by ϕ'. The principle can be recursively applied, potentially in regard of a given ordering $\prec$ on the propositional variables x_i in ϕ. Reduction and simplification operations can be applied after each step to construct a ROBDD. Since the ROBDD for a boolean function is unique, one can read off immediately from the constructed BDD, if the respective input formula is unsatisfiable (or valid): ϕ is unsatisfiable (valid) if the respective ROBDD contains only the node 0 (1). This shows immediately that the construction of an ROBDD in the worst-case is expensive (unless $P = NP$). However, in practice (especially when applying a suitable variable ordering $\prec$) the construction of ROBDDs can be done efficiently. Alternatively, BDDs can be constructed bottom up, too, starting from atomic subformulae stepwise to increasingly complex sub-formulae of ϕ, since the application of logical operators (e.g. $\wedge$,$\vee$,$\neg$,$\Rightarrow$, $\Leftrightarrow$) to combine formulae to more complex ones can be implemented very efficiently (i.e. in linear or quadratic time in the size of the BDDs to be combined) as standard BDD graph operations on the the corresponding BDDs.

Interestingly, BDDs are very rich structures for storing semantic information about the input formulae ϕ: in the propositional case, paths from the root to the 1 leaf node compactly represent all *models* of ϕ (wrt. to the given propositional signature). Analogously, all pathes from the root to the 0 leaf node capture compactly all *counter models* for ϕ, i.e. interpretation for which ϕ is not satisfied. Further, if one considers a 1-path as a conjunction of literals and the set of 1-pathes disjunctively combined, then the BDD contains a *disjunctive normal form* of ϕ. At the same time, one can directly interpret the set of 0-pathes in the BDD as a *conjunctive normal form* for ϕ. This is promising since proof search strategies can be implemented on top of BDDs that use either normal form representation. Since tableau methods can be seen as processes that derive a disjunctive normal form for an input formula ϕ and resolution methods as processes that iteratively extend conjunctive normal forms of ϕ, we expect that techniques from both fields can be considered for the design of efficient proof search strategies. Further, we believe that BDDs are able to provide a uniform structure that can be used to realize a variety of reasoning tasks for logics (beyond satisfiability), because they are inherently encode compactly a lot of semantic (i.e. syntax-independent) information about ϕ: in the propositional case this is the whole truth table of ϕ.

The link to First-order Logics (FOLs) is as well rather straightforward: Let Σ be a first-order signature (including two 0-ary predicates 1 and 0 denoting the respective truth values and a set $\mathcal{V}$ of variable names). The set of terms $Term(\Sigma)$ is defined as usual as the smallest set containing all variables $x \in \mathcal{V}$ and is closed under the application of n-ary function symbols $f \in \Sigma$ to any combination of n terms. The set of atomic formulae $\mathcal{L}_0(\Sigma)$ is defined as the set of expressions that can be generated from terms by applying any n-ary predicate symbol $p \in \Sigma$ to any combination of n terms. The First-order Logic $\mathcal{L}(\Sigma)$ over signature Σ is then defined as the smallest set of expression that contains all atomic formulae

$\phi \in \mathcal{L}_0(\Sigma)$ and is closed under the application of the usual logical junctors $\neg, \wedge, \vee, \Rightarrow, \Leftrightarrow$ as well as the application of any quantor $Q \in \{\exists, \forall\}$ to any pair of variable $x \in \mathcal{V}$ and formula $\phi \in \mathcal{L}(\Sigma)$.

For the sake of simplicity (and without loss of generality[1]), we consider here only formulae $\phi \in \mathcal{L}(\Sigma)$ in universal prenex form, i.e. have the form $\phi = \forall x_1, \ldots, x_n.M_\phi$ with M_ϕ a quantifier free formulae. The described techniques can be extended to the full language $\mathcal{L}(\Sigma)$ as shown e.g. in [22,9]. For such a ϕ we construct the Binary Decision Diagram $BDD_\phi = bdd(M_\phi)$, i.e. a graph (V, E) with vertices $v \in V$ and l-labeled edges $e = (v, l, v') \in E$ recursively as follows:

$$bdd(\psi) = \begin{cases} (\{l\}, \emptyset) & \textit{if } \psi = l \in \{0,1\} \\ (\{a, 0, 1\}, \{(a, +, 1), (a, -, 0)\}) & \textit{if } \psi = a \in \mathcal{L}_0(\Sigma) \setminus \{0,1\} \\ negbdd(bdd(\psi')) & \textit{if } \psi = \neg\psi' \\ conjunctbdd(bdd(\psi'), bdd(\psi'')) & \textit{if } \psi = \psi' \wedge \psi'' \\ disjunctbdd(bdd(\psi'), bdd(\psi'')) & \textit{if } \psi = \psi' \vee \psi'' \\ implbdd(bdd(\psi'), bdd(\psi'')) & \textit{if } \psi = \psi' \Rightarrow \psi'' \\ biimplbdd(bdd(\psi'), bdd(\psi'')) & \textit{if } \psi = \psi' \Leftrightarrow \psi'' \end{cases}$$

whereby $negbdd, conjunctbdd, disjunctbdd, implbdd, biimplbdd$ represent standard operations to negate BDDs and to combine BDDs conjunctively, disjunctively, by implication, and biimplication. The resulting BDD_ϕ therefore contains only nodes that represent atomic subformulae occurring in ϕ, 1, or 0; atomic subformulae are considered as (unstructured) propositional letters. For any given ϕ, BDD_ϕ can be constructed in finite time (and usually very fast). Reduction (wrt. a fixed order $\prec$ on atomic formulae in $\mathcal{L}_0(\Sigma)$ can be applied as in the propositional case. If BDD_ϕ is considered as a formula (in the so-called *if-then-else* or Shannon normal form), then BDD_ϕ is logically equivalent to M_ϕ.

Clearly, since BDD_ϕ in general contains atomic formulaes with variables, we can not determine the unsatisfiability of ϕ directly from the graph structure. However, it is a compact, logically equivalent representation of M_ϕ that allows to check (in many practical cases) efficiently for unsatisfiability if no variable were present in ϕ, or if M_ϕ contains variables but is already propositionally unsatisfiable. Hence, the question is how to deal with the variables (and therefore the quantifiers) in $\phi = \forall x_1, \ldots, x_n.M_\phi$ which is equivalent to $\forall x_1, \ldots, x_n.BDD_\phi$. Here, Herbrand's theorem [4] provides the theoretical means to identify the missing piece to devise a proof procedure for FOL, since it allows to reduce FOL unsatisfiability to unsatisfiability on propositional logic: A formulae of the form $\phi = \forall x_1, \ldots, x_n.M_\phi$ is unsatisfiable iff. there exists a $k \in \mathbb{N}$ and a substitution σ such that $(M_\phi^1 \wedge M_\phi^2 \wedge \ldots \wedge M_\phi^k)\sigma$ is a propositionally unsatisfiable formulae, whereby M_ϕ^i denotes a „new" copy of M_ϕ where the variables $x_1, \ldots, x_n$ in M_ϕ have been renamed uniquely to $x_1^i, \ldots, x_n^i$ such that they do not occur by any other copy M_ϕ^j and $x_k^i \neq x_l^i$ if $x_k \neq x_l$.

Therefore, we can devise a FOL proof procedure by enriching the data structure BDD_ϕ with a search procedure that attempts to find suitable number of extension step k and ground substitution σ, such that $\Pi_{k,\sigma}(\phi) = (M_\phi^1 \wedge M_\phi^2 \wedge \ldots \wedge M_\phi^k)\sigma$ can be demonstrated as being unsatisfiable. Since $\Pi_{k,\sigma}(\phi)$ is propositional and can be efficiently constructed (for any k) from BDD_ϕ (essentially by application of the standard $conjunctbdd$ operation), the compact representation of M_ϕ and the „built-in" unsatisfiability check are promising features of BDDs as the basis of a FOL proof procedure. For a given k (starting with $k = 1$), the search procedure can try to find a suitable substitution σ that „falsifies" (or refutes) the BDD and iteratively increase the number of required copies k if all possibilities have been explored but turned out to be unsuccessful. Clearly, blind guessing of candidates σ is absolutely undesirable. As in Semantic Tableau and Resolution, the proof procedure should take the formulae (and its structure) itself into account and use well-known tools such as unification and the computation of most general unifiers. Here, BDDs provide again a rich structure and various options for this specific purpose (even if orderings are not used), such as analysis and elimination of 1-paths or strategies that work with 0-paths instead. Clearly, the proof procedure is only guaranteed to terminate in the case of an unsatisfiable formula. For FOL, this can not be changed since the set of satisfiable formulas in FOL is not recursively enumerable.

A straightforward way to apply BDDs to DL reasoning could then be as follows: many DLs can be considered as very restricted subsets of FOLs, where the syntactic restrictions lead to decidability of fundamental reasoning tasks such as unsatisfiability of a knowledge base. This even works for very expressive DLs as long as they can be „embedded" to FOL. The main question here is how to achieve the termination of the FOL proof procedure in these cases. Clearly, there are two parameters to play with: (a) the translation function that embeds a given DL knowledge base into a set of first-order formulae, and (b) suitable refinements (or restrictions) of the proof search process based on the specific characteristics of the underlying DL (such as the finite tree model property) or the syntactic structure of the generated set of FOL formulae. In regard of (b), we are very optimistic, since BDDs can be used to generate some tableau-like as well as some resolution-like (micro) inference steps (when simplifying the BDD that represent the current state of the proof search), we expect that certain well-studied techniques can be rebuilt in the BDD framework. At the same time we can exploit in the BDD framework that it possible to do both at the same time: checking unsatisfiability of a formula (non-existence of consistent and deductively complete 1-paths) as well as its satisfiability (the presence of a consistent deductively complete 1-path). Therefore, novel techniques for reasoning (even for very expressive DLs), potentially interweaving both processes can be designed and investigated thoroughly.

[1] Every formulae $\phi \in \mathcal{L}(\Sigma)$ can be transformed into an equi-satisfiable formula in universal prenex form in polynomial time [19]

In the past, a few approaches [21,8,22,12] on how to generalize the principles underlying BDDs and OBDDs from the propositional level to the first-order level have been studied. Each of them could serve as a distinct starting point for our purposes and will be investigated closely. A brief overview of essential underlying principles is given in [10]. Interestingly, [11] shows theoretically that BDDs and Resolution are fundamentally different techniques for Propositional Logic, whereby the argument carries over to FOL. Further, [22] discusses the relation of their specific approach to First-order Semantic Tableaux, and points out the specifically important advantage of the BDD-based method over Semantic Tableaux which is the property of very compact representations during proof search.
All these approaches target at First-order Logics and therefore at *checking for unsatisfiability* of an input formula. For reasoning in DLs, a possible and slightly different approach would be the following: one exploits the rich structure of BDDs to search for models of an input formula, i.e. to build a *model generation procedure* based on BDDs. This is essentially the basic idea underlying the DL tableau procedures (that work on a different representation than BDDs) and can be expected to simplify termination proofs for all input formulas (e.g. for DLs with the finite model property).
The resulting model generation algorithm will be different from the proposed unsatisfiability checking algorithms for FOL, since it uses the information that is represented in the BDD in a different way and applies different modifications. Still, both algorithms can be represented and performed on top of the BDD representation of the input formulae (or knowledge base). This suggests to study the possibility and efficiency of deductive process that interweave both activities (i.e. theorem proving and disproving). In consequence, a resulting model generation procedure would be applicable (when dropping certain DL-specific assumptions) to First-order Logics, too, and potentially result in novel techniques for model generation for FOLs.

4 Related Work

In the following we briefly discuss approaches that are relevant for DL reasoning and make major use of Binary Decision Diagrams during the proof search.

The Knowledge Cartographer Approach. The work reported in [5,6,7] takes a purely set-theoretic perspective on DL reasoning, especially on TBoxes. The underlying idea is simple, yet elegant: Given a DL signature Σ for any interpretation over Σ the universe under consideration is partitioned into a number of non-overlapping sets (so-called *atomic regions*). Given any interpretation, the extension (or interpretation) of any concept expression over Σ can be composed by atomic regions (via set-theoretic union) only. If we consider a given TBox $\mathcal{T}$ (as it is common in practical applications), then the possible partitions of the universe of models of $\mathcal{T}$ are often restricted severely (in comparison to the partitions for arbitrary interpretations) and the number of atomic regions decreases drastically. Hence, if n is the number of atomic regions, then any concept expression can be identified with an n-dimensional bit vector in $\mathbb{B}^n$ (the so-called *signature*). The base vectors of the canonical basis of $\mathbb{B}^n$ represent the atomic regions themselves. Since concept are constructed essentially by means of set-theoretic operations, the most important (yet simple) concept constructors (such as $\sqcap, \sqcup, \neg$) can be very efficiently mapped (and implemented) by means of bit-operations on signature. Important semantic tests between concept expressions can be check by simple comparison of the bit vectors (that are linear in the size n of the bit vectors), e.g. concept C_1 is subsumed by concept C_2 if for the corresponding signatures (or bit vectors) it holds that $sig(C_1) \leq sig(C_n)$. However, in the worst-case the required length n of the bit vectors is exponential in the size of the signature. The key problem in this approach is to determine needed atomic regions for a given signature Σ and TBox $\mathcal{T}$. Ordered BDDs are taken as an efficient means for computing the signatures that need to be considered for a given TBox $\mathcal{T}$. In this sense, $\mathcal{T}$ is compiled in a preprocessing step into a semantic data structure that later on simplifies particular semantic checks, such as concept subsumption. The approach is defined for a restricted subset of the DL $\mathcal{ALC}$ and can not deal with arbitrary concept descriptions for ABox queries. The reported evaluation results seem to indicate superior performance over state-of-the-art systems, especially in the presence of ABoxes of significant size. One has to keep in mind here, that the a system for a rather limited DL is compared against more general DL system. Further, the performed result are not well documented and do not give a clear indication of scientific significance of measured experiment.

A BDD-based calculus for Reasoning in the Modal Logic K. Very recently [20] proposed a novel satisfiability checking procedure for formulae in the basic modal logic $\mathbf{K}$. It has been reported that a corresponding implementation is competitive or even superior to existing highly-optimized modal reasoning systems for certain knowledge bases (KB). In particular, for formulae that require extensive modal reasoning, the method seems to perform very well. The authors note, that the method can be extended to the multi-modal logic $\mathbf{K_{(m)}}$. Since $\mathbf{K_{(m)}}$ can be considered as a syntactic variant of the description logic $\mathcal{ALC}$ [25] (where m corresponds to the number of role names in the underlying DL signature), the proof procedure is suitable for reasoning with the non-trivial DL $\mathcal{ALC}$, too. It is not clear to what extent it is possible to transfer the approach to other more expressive DLs than $\mathcal{ALC}$. Further, testing satisfiability of a formula wrt. to background knowledge is not covered in [20].

In contrast, to these specific related BDD-based techniques the approach that we propose addresses DL reasoning by refinement and tailoring of BDD-based calculi for a logic that is more expressive than many expressive DLs, namely First-order Logic. It has therefore inherently the advantage to be applicable to a wide range of expressive DLs, that go beyond $\mathcal{ALC}$. Further, it has the potential to support expressive extensions of DLs that result in undecidable yet empirically still tractable knowledge representation frameworks. We expect that major elements of the Knowledge Cartographer approach naturally arise in our framework and allow to related both approaches on a more detailed technical level.

5 Conclusions & Future Work

We have proposed to investigate the use of BDDs and their manifold variants and extensions as a fundamental framework for realizing well-known reasoning tasks, in particular for expressive DLs. BDDs are very rich structures capturing a variety of useful information that can be exploited by inference calculi during proof search. We are especially interested in investigating and refining BDD-based calculi that have been proposed for First-order Logics a couple of years ago. Since there are various different ways of using BDDs to design new inference calculi and numerous variants of the BDD data structure, we expect that BDDs give us sufficiently many options for our investigation and gives enough room to come up with useful novel inference techniques. Using the approach of refining FOL techniques naturally enables us to cope with various expressive DLs (such as the ones underlying the Web Ontology Language (OWL) in version 1.0 ($\mathcal{SHOIN}(\mathbf{D})$) and version 1.1. ($\mathcal{SROIQ}$ [15], with extensions for metamodeling and n-ary datatypes). Further, the investigation of the behavior of the developed methods for certain tractable subsets of such DLs (e.g. the ones discussed in OWL 1.1 language proposal[2]) is certainly desirable.

We expect that the proposed research agenda helps to evolve the state-the-art in the field of Description Logic reasoning by (i) extending the available machinery of tools for reasoning in expressive DLs by distinctively novel methods and (ii) by provision of a deep understanding of the strengths and potential weaknesses of the developed novel methods. Our ultimate aim is to design techniques to help to increase the possible range of applications of DLs for knowledge-based and intelligent systems. As far as possible, we aim to identify potential extensions to existing DL-based knowledge representation frameworks to make them more expressive for applications while still staying in an empirically tractable framework. This is well in line with the needs of advanced applications of DLs as is has been discussed in [14]. Technical details of a BDD-based inference calculus for $\mathcal{ALC}$ are subject to an upcoming paper. In all approaches to BDD-based FOL proof procedures that we are aware of, specific means of equality reasoning or reasoning with of concrete data-types have not been studied yet. However, this is needed to deal with popular features in expressive DLs such as cardinality restrictions or concrete domains and it therefore subject of investigation after we are able to deal with $\mathcal{ALC}$. Eventually, investigations on how to deal with large ABoxes (e.g. provided and managed by a relational database system) and the integration of rules and rule-based reasoning could complete the outlined research project along dimensions that are currently observable as main lines of research in the DL field.

References

1. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
2. Alexander Borgida and Peter F. Patel-Schneider. A semantics and complete algorithm for subsumption in the classic description logic. *J. Artif. Intell. Res. (JAIR)*, 1:277–308, 1994.
3. Randal E. Bryant. Symbolic boolean manipulation with ordered binary-decision diagrams. *ACM Comput. Surv.*, 24(3):293–318, 1992.
4. Melvin Fitting. *First-Order Logic and Automated Theorem Proving*. Springer-Verlag, second edition edition, 1996.
5. Krzysztof Goczyla, Teresa Grabowska, Wojciech Waloszek, and Michal Zawadzki. The cartographer algorithm for processing and querying description logics ontologies. In *Advances in Web Intelligence Third International Atlantic Web Intelligence Conference (AWIC 2005), Lodz, Poland*, pages 163–169, 2005.
6. Krzysztof Goczyla, Teresa Grabowska, Wojciech Waloszek, and Michal Zawadzki. Cartographic approach to knowledge representation and management in kasea. In *Proceedings of International Workshop on Description Logics (DL 2005), Edinburgh, Scotland, UK*, 2005.
7. Krzysztof Goczyla, Teresa Grabowska, Wojciech Waloszek, and Michal Zawadzki. The knowledge cartography - a new approach to reasoning over description logics ontologies. In *Theory and Practice of Computer Science, 32nd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2006), Merín, Czech Republic*, pages 293–302, 2006.
8. Jean Goubault. Proving with bdds and control of information. In *In Proceedings of the 12th International Conference on Automated Deduction (CADE) Nancy, France 1994*, pages 499–513, 1994.
9. Jean Goubault. A BDD-Based Simplification and Skolemization Procedure. *Logic Jnl IGPL*, 3(6):827–855, 1995.
10. Jean Goubault and Joachim Posegga. BDDs and automated deduction. In *International Syposium on Methodologies for Intelligent Systems*, pages 541–550, 1994.

[2] `http://owl1_1.cs.manchester.ac.uk/tractable.html`

11. J. F. Groote and H. Zantema. Resolution and binary decision diagrams cannot simulate each other polynomially. *Discrete Applied Mathematics*, 130(2):157–171, August 2003.
12. Jan Friso Groote and Olga Tveretina. Binary decision diagrams for first-order predicate logic. *J. Log. Algebr. Program.*, 57(1-2):1–22, 2003.
13. Ian Horrocks. *Optimising Tableaux Decision Procedures for Description Logics*. PhD thesis, University of Manchester, 1997.
14. Ian Horrocks. Applications of description logics: State of the art and research challenges. In Frithjof Dau, Marie-Laure Mugnier, and Gerd Stumme, editors, *Proc. of the 13th Int. Conf. on Conceptual Structures (ICCS'05)*, number 3596 in Lecture Notes in Artificial Intelligence, pages 78–90. Springer, 2005.
15. Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible $\mathcal{SROIQ}$. In *Proc. of the 10th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2006)*, pages 57–67. AAAI Press, 2006.
16. Ullrich Hustadt. *Resolution-Based Decision Procedures for Subclasses of First-Order Logic*. PhD thesis, Universität des Saarlandes, Saarbrücken, Germany, November 1999.
17. J.R. Burch, E.M. Clarke, K.L. McMillan, D.L. Dill, and L.J. Hwang. Symbolic Model Checking: 10^{20} States and Beyond. In *Proceedings of the Fifth Annual IEEE Symposium on Logic in Computer Science*, pages 1–33, Washington, D.C., 1990. IEEE Computer Society Press.
18. Boris Motik. *Reasoning in Description Logics using Resolution and Deductive Databases*. PhD thesis, Univesität Karlsruhe (TH), Karlsruhe, Germany, January 2006.
19. A. Nonnengart and C. Weidenbach. Computing small clause normal forms. In A. Robinson and A. Voronkov, editors, *Handbook of Automated Reasoning*, volume I, chapter 6, pages 335–367. Elsevier Science, 2001.
20. Guoqiang Pan, Ulrike Sattler, and Moshe Y. Vardi. Bdd-based decision procedures for the modal logic k. *Journal of Applied Non-Classical Logics*, 16(1-2):169–208, 2006.
21. Joachim Posegga. *Deduktion mit Shannongraphen für Prädikatenlogik erster Stufe.*, volume 51 of *DISKI*. Infix Verlag, St. Augustin, Germany, 1993.
22. Joachim Posegga and Peter H. Schmitt. Automated deduction with shannon graphs. *Journal of Logic and Computation*, 5(6):697–729, 1995.
23. Richard Rudell. Dynamic variable ordering for ordered binary decision diagrams. In *ICCAD '93: Proceedings of the 1993 IEEE/ACM international conference on Computer-aided design*, pages 42–47, Los Alamitos, CA, USA, 1993. IEEE Computer Society Press.
24. U. Sattler and M. Y. Vardi. The hybrid mu-calculus. In R. Goré, A. Leitsch, and T. Nipkow, editors, *Proceedings of the International Joint Conference on Automated Reasoning*, volume 2083 of *LNAI*, pages 76–91. Springer Verlag, 2001.
25. Klaus Schild. A correspondence theory for terminological logics: Preliminary report. In *In Proceedings of the International Joint Conference of Artificial Intelligence (IJCAI 1991)*, pages 466–471, 1991.
26. Seiichiro Tani, Kiyoharu Hamaguchi, and Shuzo Yajima. The complexity of the optimal variable ordering problems of shared binary decision diagrams. In *ISAAC '93: Proceedings of the 4th International Symposium on Algorithms and Computation*, pages 389–398, London, UK, 1993. Springer-Verlag.

Process Mediation in Semantic Web Services*

Emilia Cimpian

Digital Enterprise Research Institute,
Institute for Computer Science, University of Innsbruck,
Technikerstrasse 21a, A-6020 Innsbruck, Austria
emilia.cimpian@deri.org

Abstract. The Semantic Web Services initiatives are aiming to develop automatic and dynamic solutions for the semantically described Web services discovery, invocation and execution. The automation of all this activities is possible only if both the requestor and the provider of a service are semantically describing the requested and the provided functionalities, as well as the behavior they are going to have during the service's invocation. However, several mismatches may occur, on several levels: data, process or functionality. This paper is focusing on overcoming the process heterogeneity problems, from the processes compatibility point of view.

1 Introduction

An intense research activity regarding Semantic Web services has been going on during the last years. But only the semantic descriptions attached to data or to the Web services deployed using todays technologies does not solve the heterogeneity problems that may come up due to the distributed nature of the Web itself. As such, the heterogeneity existing in representing data and processes or in the multitude of choices in representing the requested and the provided functionalities, and in the various forms of the communication patterns (public processes) are problems that have to be solved before being able to fully benefit of the semantic enabled Web and Web services. Considering that these problems can not be avoided, dynamic mediation solutions that fully exploit the semantic descriptions of data and services are required.

As mediation is a rather broad and well-studies field at both semantic ([8], [2]) and non-semantic level ([6], [5]), this paper focuses further on only a subset, namely on process mediation in the context of Semantic Web services.

The discussion is held in the context of Web Service Modeling Ontology (WSMO)[1] [4], [3], a framework that offers all the necessary instruments to semantically describe the Web services and all the related aspects. One of the main reasons in choosing WSMO as the semantic framework for Web services is that it realizes the importance of

[*] The work is funded by the European Commission under the projects ASG, DIP, enIRaF, InfraWebs, Knowledge Web, Musing, Salero, SEKT, Seemp, SemanticGOV, Super, SWING and TripCom; by Science Foundation Ireland under the DERI-Lìon Grant No.SFI/02/CE1/I13; by the FFG (Österreichische Forschungsförderungsgesellschaft mbH) under the projects Grisino, RW^2, SemBiz, SemNetMan, SeNSE, TSC, OnTourism.

[1] The author has been an active member of the WSMO working group since 2004

mediators and treats them as first class citizens. WSMO offers specific means to semantically describe concrete mediation solutions and to directly refer to them when needed (e.g. from ontologies or Web services).

This paper is further structured as follows: Section 2 presents the addressed problem, while Section 3 provides an overview of the current state of the art in the field, illustrating how the approach further described in the paper is different, and what are its advantages. The expected contribution and the research methodology followed are presented in Section 4 and Section 5 respectively. Section 6 concludes the paper.

2 Problem Definition

By process mediation we understand the action of overcoming the heterogeneity problems between two processes involved in a collaborative task (that is, one process is generating information needed by the other process). What this thesis is focusing on is finding technologies and developing tools that would allow two processes to interact, even if this interactions is not a straight-forward one.

Consider for example that one of the processes expects (from the environment) certain information in order to continue its execution. On the other hand, the other process is going to generate the needed information, but in different format, order, or in terms of a different ontology. As all the information needed by the first process exists, the process mediator will have to ensure that the data, as generated by the second process, is transformed in order to match the first process' needs.

3 State of the Art Overview

Process mediation is still a poorly explored research field, in the context of Semantic Web Services. The existing work represents only visions of mediator systems able to resolve in a (semi-) automatic manner the processes heterogeneity problems, without presenting sufficient details about their architectural elements. Still, these visions represent the starting points and valuable references for the future concrete implementations.

Two integration tools, Contivo[2] and CrossWorlds[3] seemed to be the most advanced ones in this field.

Contivo is an integration framework which uses metadata representing messages organized by semantically defined relationships. One of its functionalities is that it is able to generate transform code based on the semantic of the relationships between data elements, and to use this code for transforming the exchange messages. However, Contivo is limited by the use of a purpose-built vocabulary and of pre-configured data models and formats.

CrossWorlds is an IBM integration tool, meant to facilitate the B2B collaboration through business processes integration. It may be used to implement various e-business models, including enhanced intranets (improving operational efficiency within a business enterprize), extranets (facilitating electronic trading between a business and its

[2] http://www.contivo.com/

[3] http://www.sars.ws/hl4/ibm-crossworlds.html

suppliers) and virtual enterprizes (allowing enterprizes to link to outsourced parts). The draw-backs of this approach is that different applications need to implement different collaboration and connection modules, in order to interact. As a consequence, the integration of a new application can be done only with additional effort.

Important results are expected from a newly started European project, SUPER[4] which aims to enhance widely accepted business processes industrial standards with semantic, and to provide a comprehensive tools stack in order to support the entire life-cycle of semantically described business processes. It is exactly the outputs of this type of initiatives the process mediator designed in this thesis is able to act upon.

Through our approach we aim to provide dynamic mediation between various parties using WSMO for describing goals and Web Services. As described in this paper this is possible without relying on any hard-coded transformations.

4 Expected Contribution

The main expected results of the research carried out are as follows:

- formalization of a set of solvable mismatches - the process mediation can not aim at solving any type of mismatches that can occur during the inter-operation of two processes independently designed, but only of a sub-set; that is, some restriction have to be imposed, for example that all the needed information is provided;
- formalization of a set of unsolvable mismatches - this set is useful for determining in a timely manner whether two processes can not inter-operate;
- implementation of a prototype able to overcome the solvable mismatches.

The achievement of these goals will represent a step-forward for the process mediation from two perspectives: firstly, from the process representation perspective, none of the current approaches is addressing the semantically described process mediation, which considering the emergence of semantic technologies is nowadays an important aspect; secondly, this type of process mediation will boost the semantic Web service invocation technologies, as the automatic invocation of such a service is due to fail as soon as an inconsistency between the service's and the requestor's behavior occurs.

5 Research Methodology

The following steps need to be taken in solving the addressed problem: a)identification of solvable and unsolvable mismatches, b)formalization and resolution of the mismatches and c) prototype implementation. The focus of the research was so far in identifying an initial set of solvable and unsolvable mismatches, and in the development of a prototype able to cope with the solvable mismatches. On the other hand, formalizing the mismatches was not address yet, being still an important open issue.

[4] http://www.ip-super.org/

The first step in achieving the goals is the identification of solvable and unsolvable mismatches. Although in the beginning this identification was based strictly on theoretical assumptions and toy use-cases, the set has been extended based on the real use-cases obtained from several European and Austrian research projects [5].

A list containing the initial set of resolvable mismatches that the process mediator intends to address is provided below.

Stopping an unexpected message (Figure 1. a)): in case one process generates some information that the other one does not want to receive, the mediator should just retain and store it. This information can be send later, if needed, or it will just be deleted after the communication ends.

Inversing the order of messages (Figure 1. b)): in case one of the processes generates the information in a different order than the one the other process wants to receive. The messages that are not yet expected will be stored and sent when needed.

Splitting a message (Figure 1. c)): in case one of the processes sends in a single message multiple information and the other one expects to receive it in different messages.

Combining messages (Figure 1. d)): in case one of the processes expects a single message, containing information sent by the other one in multiple messages.

Sending a dummy acknowledgement (Figure 1. e)): in case one of the processes expects an acknowledgement for a certain message, and the other partner does not intend to send it, even if it receives the message.

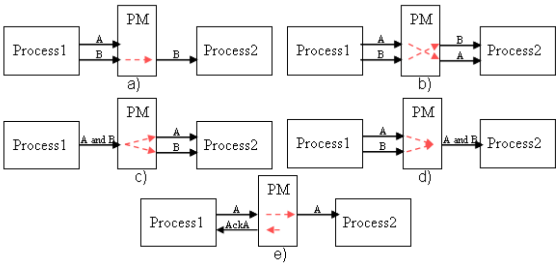


Fig. 1: Addresses Mismatches

Similarly, a set of unsolvable mismatches has been determined; due to space limitations they are not presented in this paper. For a detailed description of these unsolvable mismatches please see [1].

By combining several types of solvable mismatches previously presented more complex mismatches can be successfully solved. However, a combination of solvable mismatches with one or more unsolvable ones leads to more complex unsolvable mismatches.

A process mediation prototype able to cope with the solvable mismatches has been already developed. It is able o parse processes semantically described using Web Service Modeling Language (WSML[6]) and to deal with the heterogeneity problems previously presented. In case the two processes use different underlying ontologies the

[5] Two of the most illustrative projects from this point of view are SUPER and SemBiz

[6] http://www.wsmo.org/wsml

services of a data mediator for mapping between the ontologies (like the one described in [7]) are needed. A complete description of the algorithm implemented by this prototype, as well as its architecture is presented in [1].

6 Conclusions

This thesis is addressing the process mediation in a semantic environment. This is currently an important aspect, as the semantic description of services and service requests does not consist only of data expressed using ontologies, but also of semantically described processes. This thesis aims to develop a set of general methodologies for identifying and solving heterogeneity problems that may appear between semantically described processes.

A direct application is solving the heterogeneity problems that may occur during the invocation of a service, considering that both the invoker and the service have well defined interfaces defining their behaviors, and that they are not going to adjust these behaviors according to their conversation partner.

The prototype that is going to be delivered with this thesis implements dynamic techniques able to detect and overcome on the fly (during run-time) the mismatches existing between given semantically described processes.

References

1. E. Cimpian and A. Mocan. WSMX Process Mediation Based on Choreographies. In *Proceedings of the 1st International Workshop on Web Service Choreography and Orchestration for Business Process Management at the BPM 2005*, Nancy, France, 2005.
2. E. Cimpian, A. Mocan, and M. Stollberg. Mediation enabled semantic web services usage. *Proceedings of the First Asian Semantic Web Conference*, 09 2006.
3. J. B. Domingue, D. Roman, and M. Stollberg (eds.). Web Service Modeling Ontology (WSMO) - An Ontology for Semantic Web Services. Position Paper at the W3C Workshop on Frameworks for Semantics in Web Services, June 9-10, 2005, Innsbruck, Austria, 2005.
4. C. Feier, A. Polleres, R. Dumitru, J. Domingue, M. Stollberg, and D. Fensel. Towards intelligent web services: The web service modeling ontology (WSMO). *International Conference on Intelligent Computing (ICIC)*, 2005.
5. J.Madhavan, P. A. Bernstein, P. Domingos, and A. Y. Halevy. Representing and reasoning about mappings between domain models. *Proc. of Eighteenth National Conference on Artificial intelligence*, pages p.80–86, July 2002.
6. A. Maedche, B. Motik, N. Silva, and R. Volz. Mafra - a mapping framework for distributed ontologies. *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW)*, September 2002.
7. A. Mocan, E. Cimpian, and M. Kerrigan. Formal Model for Ontology Mapping Creation. In *Proceedings of the 5th Intl. Semantic Web Conference (ISWC 2006)*, November 2006.
8. M. Paolucci, N. Srinivasan, and K. Sycara. Expressing WSMO Mediators in OWL-S. Hiroshima, Japan, 2004.

Scalable Web Service Composition with Partial Matches

Adina Sirbu and Jörg Hoffmann*

Digital Enterprise Research Institute (DERI)
University of Innsbruck, Austria
firstname.lastname@deri.org

Abstract. We will investigate scalable algorithms for automated Semantic Web Service Composition (WSC). Our notion of WSC is very general: it allows the generation of new constants by Web service outputs; the composition semantics includes powerful background ontologies; and we use the most general notion of matching, *partial matches*, where several services can cooperate each covering only a part of a requirement. Herein, we define a first formalism. We show that automatic composition is very hard: even testing solutions is Π_p^2-complete. We identify a special case that covers many relevant WSC scenarios, and where solution testing is only **coNP**-complete. While **coNP** is still hard, in the area of planning under uncertainty, scalable tools have been developed that deal with the same complexity. In our next step, we will adapt the techniques underlying one of these tools to develop a scalable tool for WSC. In the long term, we will investigate richer formalisms, and accordingly adapt our algorithms.

1 Introduction

In any task that involves automatic processing of Web services, one needs support for background ontologies. In the case of WSC, almost all existing solutions compile the problem into AI Planning formalisms. The motivation is that planning tools have become many times more scalable in recent years, through the use of heuristic functions and other search techniques, e.g. [8]. The problem here is that those tools cannot handle background ontologies. The following example illustrates the importance of those:

Example 1. A service shall be composed that inputs a constant of concept A, and outputs one of concept C. (E.g., A may be a trip request and C a ticket.) The ontology defines the concepts A, B, $B_1, \dots, B_n$, C, and states that each $B_i \subseteq B$, and $\bigcup_1^n B_i \supseteq B$. (E.g., B may be a geographical region and the B_i its parts.) An available service ws_{AB} transforms A into B, and for each i an available service ws_{B_iC} transforms B_i into C. A solution first applies ws_{AB}, and then applies the ws_{B_iC} in conjunction.

In Example 1, reasoning over the background ontology is necessary to (1) understand which services can be used, and to (2) test whether a given composition is actually a solution. Such reasoning can be modelled through the *background theories* explored in some planning works, e.g., [4, 6]. However, incorporating this notion into the modern scalable planning tools poses serious challenges, and has not yet even been tried. Due to the background theory, even computing a state transition – which is now a form of belief revision – is a computationally very hard task. The existing planning tools dealing

[*] Ph.D. Supervisor

with background theories, e.g., [4, 6], map the problem into generic deduction, which is well known for its lack of scalability. The existing planning tools dealing with WSC, e.g., [9, 2], ignore the ontology and assume *exact matches*. In Example 1, this would require $B = B_i$ instead of $B \cap B_i \neq \emptyset$. Obviously, this renders the example unsolvable.

We identify an interesting special case of WSC. We exploit the fact that Web services may output new constants.[1] Now, if all ramifications of a Web service concern only propositions involving at least one new constant, then a belief revision is not necessary. We term this special case WSC with *forward effects*: the effects are "forward" in the sense that no backwards-directed belief revision is necessary. E.g., the services in Example 1 have forward effects. The same holds for many WSC scenarios from the literature, and from real case studies. A simple example is the wide-spread "virtual travel agency", where Web services must be linked that book travel and accommodation, generating new constants corresponding to tickets and reservations.

We introduce a framework for planning with background theories.[2] We show that testing whether an action sequence is a solution – *solution testing* – is Π_2^p-complete in general, but only **coNP**-complete with forward effects. Of course, **coNP** is still hard. However, *planning under uncertainty* has the same complexity of solution testing, and scalable tools for this case have already been developed. Adapting them for WSC with forward effects will be our next step. In particular, the Conformant-FF tool [7] is based on CNF reasoning, which can be naturally extended to our setting.

Section 2 introduces our formalism, Section 3 discusses forward effects. Sections 4 provides some details regarding our future developments, Section 5 concludes.

2 WSC with Partial Matches

The (planning) terminology of our formalism corresponds to WSC is as follows. Web services are planning "operators"; their input/output behaviour maps to input/output parameters on which preconditions and effects are specified [1, 3, 5]. The background ontology is the background "theory". The precondition in the goal is equivalent to sets of "initial literals" and "initial constants". The effect in the goal is the "goal condition".

We assume supplies of logical predicates p, q, variable names x, y and constant names a, b, c, d, e; *(ground) literals* are defined as usual. For variables X, $\mathcal{L}^X$ is the set of literals using only variables from X. We write $l[X]$ for a literal l with variable arguments X. For a tuple C of constants substituting X, we write $l[C/X]$. In the same way, we use the substitution notation for any construct involving variables. Positive ground literals are *propositions*. A *clause* is a disjunction of literals with universal quantification on the outside, e.g. $\forall x.(\neg p(x) \vee q(x))$. A *theory* is a conjunction of clauses. An *operator* o is a tuple $(X_o, \text{pre}_o, Y_o, \text{eff}_o)$, where X_o, Y_o are sets of variables, pre_o is a conjunction of literals from $\mathcal{L}^{X_o}$, and eff_o is a conjunction of literals from $\mathcal{L}^{X_o \cup Y_o}$. The intended meaning is that X_o are the inputs and Y_o the outputs, i.e., the new constants created by the operator. For an operator o, an *action* a is given by $(\text{pre}_a, \text{eff}_a) \equiv (\text{pre}_o, \text{eff}_o)[C_a/X_o, E_a/Y_o]$ where C_a and E_a are vectors

[1] Namely, the outputs model the generated data.

[2] [10] defines a similar WSC formalism, but considering plug-in matches and restricting the background theory, instead of the service effects, to obtain efficiency.

of constants; for E_a we require that the constants are pairwise different. In this way "operators" are Web services and "actions" are service calls. *$\mathcal{WSC}$ tasks* are tuples $(\mathcal{P}, \mathcal{T}, \mathcal{O}, C_0, \phi_0, \phi_G)$. Here, $\mathcal{P}$ are predicates; $\mathcal{T}$ is the theory; $\mathcal{O}$ is a set of operators; C_0 is a set of constants, the initial constants supply; ϕ_0 is a conjunction of ground literals, describing the possible initial states; ϕ_G is a conjunction of literals with existential quantification on the outside, describing the goal states, e.g., $\exists x, y.(p(x) \wedge q(y))$.[3] All predicates are from $\mathcal{P}$, all constants are from C_0, all constructs are finite.

Assume we are given a task $(\mathcal{P}, \mathcal{T}, \mathcal{O}, C_0, \phi_0, \phi_G)$. *States* in our formalism are pairs (C_s, I_s) where C_s is a set of constants, and I_s is a *C_s-interpretation*, i.e., an interpretation of all propositions formed from the predicates $\mathcal{P}$ and the constants C_s. We need to define the outcome of applying actions in states. Given a state s and an action a, a is *applicable in s* if $I_s \models \text{pre}_a$, $C_a \subseteq C_s$, and $E_a \cap C_s = \emptyset$. We allow *parallel actions*. These are sets of actions which are applied at the same point in time. The result of applying a parallel action A in a state s is $res(s, A) :=$

$$\{(C', I') \mid C' = C_s \cup \bigcup_{a \in A, appl(s,a)} E_a, I' \in min(s, C', \mathcal{T} \wedge \bigwedge_{a \in A, appl(s,a)} \text{eff}_a)\}$$

Here, $min(s, C', \phi)$ is the set of all C'-interpretations that satisfy ϕ and that are minimal with respect to the partial order defined by $I_1 \leq I_2$:iff for all propositions p over C_s, if $I_2(p) = I_s(p)$ then $I_1(p) = I_s(p)$. This is a standard semantics where the ramification problem is addressed by requiring minimal changes to the predecessor state s [11]. A is *inconsistent* with s iff $res(s, A) = \emptyset$; this can happen in case of conflicts. Note that $res(s, A)$ allows non-applicable actions. This realizes partial matches: a Web service can be applied as soon as it matches at least one possible situation.

We refer to the set of states possible at a given time as a *belief*. The *initial belief* is $b_0 := \{s \mid C_s = C_0, s \models \mathcal{T} \wedge \phi_0\}$. A parallel action A is inconsistent with a belief b if it is inconsistent with at least one $s \in b$. In the latter case, $res(b, A)$ is undefined; else, it is $\bigcup_{s \in b} res(s, A)$. This is extended to action sequences in the obvious way. A *solution* is a sequence $\langle A_1, \dots, A_n \rangle$ s.t. for all $s \in res(b_0, \langle A_1, \dots, A_n \rangle) : s \models \phi_G$.

When assuming *fixed arity* – a constant upper bound on the arity of all variable vectors (e.g., used in predicates) – transformation to a propositional representation is polynomial. Even in this case, solution testing is Π^p_2-complete in $\mathcal{WSC}$. Further, we have proved that polynomially bounded solution existence is Σ^p_3-complete.

Theorem 1 (Solution testing in $\mathcal{WSC}$). *Assume a $\mathcal{WSC}$ task with fixed arity, and a sequence $\langle A_1, \dots, A_n \rangle$ of parallel actions. It is Π^p_2-complete to decide whether $\langle A_1, \dots, A_n \rangle$ is a solution.*

3 Forward Effects

The high complexity of $\mathcal{WSC}$ motivates the search for interesting special cases. As stated, here we define a special case where every change an action makes to the state involves a new constant. A $\mathcal{WSC}$ task $(\mathcal{P}, \mathcal{T}, \mathcal{O}, C_0, \phi_0, \phi_G)$ has *forward effects* iff:

[3] The existential quantification is needed to give meaning to the creation of new constants.

- For all $o \in \mathcal{O}$, and for all $l[X] \in \text{eff}_o$, we have $X \cap Y_o \neq \emptyset$. In words, the variables of every effect literal contain at least one output variable.
- For all clauses $cl[X] \in \mathcal{T}$, where $cl[X] = \forall X.(l_1[X_1] \vee \cdots \vee l_n[X_n])$, we have $X = X_1 = \cdots = X_n$. In words, in every clause all literals share the same arguments.

The set of all such tasks is denoted with $\mathcal{WSC}|_{fwd}$. The second condition implies that effects involving new constants can only affect literals involving new constants. Given a state s and a parallel action A, define $res|_{fwd}(s, A) :=$

$$\{(C', I') \mid C' = C_s \cup \bigcup_{a \in A, exec(s,a)} E_a, I'|_{C_s} = I_s, I' \models \mathcal{T} \wedge \bigwedge_{a \in A, exec(s,a)} \text{eff}_a\}$$

Here, $I'|_{C_s}$ denotes the restriction of I' to the propositions over C_s.

Proposition 1 (Semantics of $\mathcal{WSC}|_{fwd}$). *Assume a $\mathcal{WSC}|_{fwd}$ task, a state s, and a parallel action A. Then $res(s, A) = res|_{fwd}(s, A)$.*

Hence, the action semantics becomes a lot simpler with forward effects, no longer needing the notion of minimal changes with respect to the previous state. We get:

Proposition 2 (Solution testing in $\mathcal{WSC}|_{fwd}$). *Assume a $\mathcal{WSC}|_{fwd}$ task with fixed arity, and a sequence $\langle A_1, \dots, A_n \rangle$ of parallel actions. It is* **coNP***-complete to decide whether $\langle A_1, \dots, A_n \rangle$ is a solution.*

It is currently an open problem what the complexity of deciding polynomially bounded solution existence is in $\mathcal{WSC}|_{fwd}$. With Proposition 2, membership in Σ_p^2 is easy to see. However, we suspect that the problem is actually **coNP**-complete. We have one half of a proof, but some tricky issues must still be resolved regarding the generation of exponentially many constants.

4 Tool and Language Developments

Our next step will be to develop a tool for $\mathcal{WSC}|_{fwd}$. We focus on achieving scalability: we expect practical SWS scenarios to involve large sets of Web services, involving huge search spaces (of partial compositions). We will try to overcome this by designing heuristic solution distance and search node filtering techniques. Specifically, we will start from the ideas underlying the Conformant-FF (CFF) [7] planning tool.

CFF represents beliefs in terms of propositional CNF formulas, and uses SAT reasoning to test solutions. We will adapt this to our purposes, simply by adapting the generation of the CNFs. Note that this is possible only in $\mathcal{WSC}|_{fwd}$, not in $\mathcal{WSC}$, for complexity reasons. CFF also introduces heuristic and filtering techniques, based on an abstraction of the planning problem. Namely, for each belief CFF computes an abstract solution using approximate SAT reasoning, and then uses the abstract solution to inform the search. We will apply the same principle for $\mathcal{WSC}|_{fwd}$. This will differ from CFF in the different structure of our CNFs, necessitating different approximate reasoning, and in that we will explore typical forms of background theories (e.g., subsumption hierarchies) to obtain better distance estimates. Further, in difference to us, CFF (and indeed every existing planning tool) does not allow the generation of new constants. We will devise new heuristics for dealing with this. Note that this is critical:

as already indicated, exponentially many constants may be generated in general, so one needs heuristics identifying which are important. Those heuristics need to be clever: if they remove too many constants, then the solutions may be cut out; if they remove too few constants, then the search space may explode.

In the long term, our line of research will be to incrementally enrich the language our tool accepts, accordingly adapting the algorithms. For a start, some generalisations of $\mathcal{WSC}|_{fwd}$ are possible without losing Proposition 1. Most importantly, instead of requiring that *every* effect literal involves a new constant, one can postulate this only for literals that may actually be affected by the background theory. This may be important, e.g., , for dealing with updates on attributes of existing constants. If such a language turns out to not be enough for many practical examples, then we will look into how feasible it is to drop the forward effects restriction and deal with full $\mathcal{WSC}$. Note that, due to Theorem 1, this would require QBF solving for reasoning about beliefs. We further plan to investigate into allowing non-deterministic outcomes of Web services. These would be modelled as operators with lists of alternative effects, each of which may occur. We expect that we can handle this by appropriately extending the CNF formulas underlying beliefs, as well as the associated machinery.

5 Conclusion

We have introduced a formalism for WSC, and identified a special case for which no belief revision is necessary. We will solve some open complexity problems, and develop a tool inspired from CFF. We will incrementally extend the tool to richer languages.

References

1. A. Ankolenkar et al. DAML-S: Web service description for the semantic web. In *ISWC*, 2002.
2. R. Akkiraju, B. Srivastava, I. Anca-Andreea, R. Goodwin, and T. Syeda. Semaplan: Combining planning with semantic matching to achieve web service composition. In *ICWS*, 2006.
3. D. Martin et al. OWL-S: Semantic markup for web services. In *SWSWPC*, 2004.
4. T. Eiter, W. Faber, N. Leone, G. Pfeifer, and A. Polleres. A logic programming approach to knowledge-state planning, II: The DLVK system. *AI*, 144(1-2):157–211, 2003.
5. D. Fensel, H. Lausen, A. Polleres, J. de Bruijn, M. Stollberg, D. Roman, and J. Domingue. *Enabling Semantic Web Services: The Web Service Modeling Ontology*. Springer-Verlag, 2006.
6. E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, and H. Turner. Nonmonotonic causal theories. *AI*, 153(1-2):49–104, 2004.
7. J. Hoffmann and R. Brafman. Conformant planning via heuristic forward search: A new approach. *AI*, 170(6–7):507–541, May 2006.
8. J. Hoffmann and B. Nebel. The FF planning system: Fast plan generation through heuristic search. *JAIR*, 14:253–302, 2001.
9. S. Ponnekanti and A. Fox. SWORD: A developer toolkit for web services composition. In *WWW*, 2002.
10. J. Scicluna and J. Hoffmann. Semantic web service composition: Background theories, parallelisation, and business policies. In *ESWC PhD Symposium*, 2007. Submitted.
11. M. Winslett. Reasoning about actions using a possible models approach. In *AAAI*, 1988.

Aiding the Workflow of Email Conversations by Enhancing Email with Semantics

Simon Scerri[1]

[1] Digital Enterprise Research Institute,
National University of Ireland Galway
IDA Business Park, Galway, Ireland.
{Simon.Scerri}@deri.org

Abstract. Despite persisting in popularity, email is still plagued with information overload, hindering the workflow of data handled by the user. Just as Semantic Web technologies promise to revolutionize the Web, we aspire to use the same technology to enhance electronic mail with useful semantics. Thus we will tackle one of the largest flaws of the email communication genre - the lack of shared expectations about the form and content of the interaction, which can be attributed to the lack of explicit semantics covering context and content of exchanged messages. Earlier research showed that email content can be captured by applying speech act theory. We will refine and extend this work to develop an email speech act ontology and outline a non-deterministic model that predicts the user's best course of action upon sending or receiving an email.

Keywords: Email, Speech Act Theory, Metadata Extraction, Semantic Web.

1 Introduction

Email persists as on of the top features of the internet. Studies on lexical densities of email discourse showed [1] that despite being a written form of communication, email texts are closer to spoken rather than written discourse. Email has been regarded as a new genre [2], where a genre is a patterning of communication which structures communication by creating shared expectations about the form and content of the interaction, thus easing the burden of production and interpretation [3]. Email workflow is very inefficient because it lacks these shared expectations on how and when the exchanged information is to be acted upon. Processing incoming messages is frequently postponed, sometimes indefinitely, due to different priorities [4] or because the mental effort required would lead to distraction from other tasks. Whereas it should be the email sender's interest to make any expectations explicit to the recipient, the latter frequently ends up having to invest more time to extract and act upon implicit expectations. Apart from being subject to misinterpretation, this process puts off the recipient from immediately trying to act upon a message. In a nutshell, the lesser the effort required out of the recipient, the greater the chance that the sender's expectations are fulfilled in a timely manner.

2 Background and Related Work

Speech Act Theory [5] states that in saying something one is doing something, and is mainly concerned with the difference between the three meanings of utterances or written text: the Locutionary, or literal meaning; the Illocutionary, or the social function the speaker is performing; and the Perlocutionary, or the result or effect produced in the given context. For the speech act ‘Could you please close the door’, the Illocutionary force is that the speaker is requesting an action, the Perlocutionary force on the hearer means they are expected to close the door, rather than answering a question with a yes or no which would be the Locutionary meaning. The theory was applied to Email a number of times, in particular for email classification based on the sender’s intent [1][2][6], focus detection of threaded email conversations [7], predicting actions on email messages [8] and easing task management arising through email [9] amongst others. Although these provided promising results, they had a serious limitation since the expectations accompanying messages were only guessed on arrival, and thus never confirmed by the sender. An email message is frequently multi-purpose, realizing several purposes at the same time. Therefore our approach goes beyond simple email classification, since we consider specific segments within an email and not the email as a whole. Other relevant research work involved the introduction and formalization of Semantic Email processes [10]. Based on the Semantic Web paradigm this involved exchanging messages having predefined intents. One drawback is that users have to resort to predefined templates and this lack of flexibility limits the practicality of the approach. Also, average users are not willing to migrate from an email system that works to a different email system, even if the latter provides less ambiguous dialogue and more efficient results.

3 Semantically Enhanced Email

Although email has many weaknesses, it also provides a fundamentally right model for a communication system [11]; the major advantages of the model being asynchronosity, threading and the fact that it is a command central system. Therefore we would like to retain the basic email model, but extend its functionalities by adding a semantic layer to the model. In particular, this will clearly state the otherwise implicit intents and expectations associated with speech acts in a message. We believe that by making this information explicit, the user is aided with the exchange of information. As a result email’s disconnected workflow becomes more efficient. By:

- Fine-tuning existing email speech act taxonomies presented in earlier work [6] and creating our own email speech act ontology;
- Outlining a predictive model for illocutionary and perlocutionary reactions attributed to speech acts in email messages;
- Applying the results within extensions to popular email clients capable of capturing and embedding semantic information in exchanged messages ;

we aspire to achieve this scenario and thus substantially reduce the occurrence and consequences of the given problems. In this paper, we are mainly concerned with the first two steps and we will elaborate on our ideas in the coming sections.

3.1 Email Speech Act Ontology

Our ontology is a refinement and extension to an existing taxonomy [6], which regarded speech acts as conjunctions of various *Verbs* and *Nouns* as a pair (*v-n*). By including further parameters in our speech act model, we believe that our ontology is much more powerful than any of its predecessors. In particular, we directly addressed our main concerns: the intents and expectations accompanying speech acts – by including specific parameters in the model. This is reflected in Our verb hierarchy in Fig. 1, which differs between the two most basic verb roles at the highest level: *Initiative*, initiating a conversational thread; or *Continuative*, continuing an earlier conversation. The roles are then refined into *Requestive*, when something is being requested out of the recipient e.g. 'Can you go to the meeting?'; *Informative*, when the act is not in response to any request and requires no further dialogue e.g. 'I'm going to the meeting'; and *Responsive*, when satisfying a former request e.g. 'Yes I will go to the meeting'. The *Imperative* role is both a requestive and an informative since its behavior corresponds to both definitions above, e.g. 'Go to the meeting'. The four end verbs can manifest particular roles in particular situations. Whereas *Request* and *Decline* perform a requestive and responsive role respectively, *Deliver* can double for two roles: 'Here is the requested file' is Responsive whereas if the file wasn't requested it is Informative. *Commit* is yet more versatile and can manifest all roles.

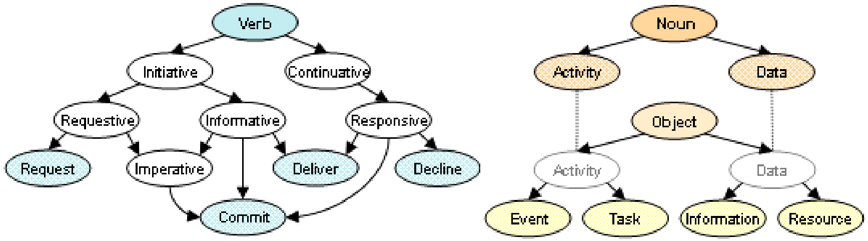
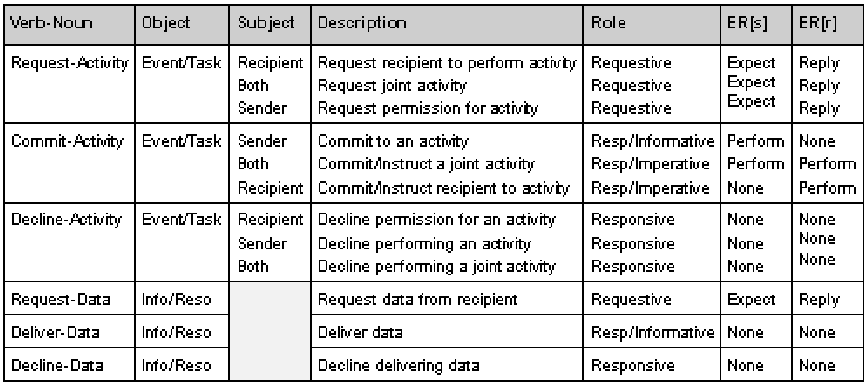


Fig. 1. Speech Act Verb, Noun and Object

In our ontology, we categorize the nouns in two major concepts: *Data*, representing something which occurs strictly within the boundary of email and *Activity*, representing something occurring outside the world of email. We extended our speech act definition to include a *Speech Act Object* representing instances of nouns rather than subclasses. Modeling the workflow and predictions for multiple verb-object pairs can be done by considering the abstract verb-noun pair. *Event* and *Task* are Activity instances and *Information* and *Resource* are Data instances. Previous work differed between a speech act requesting permission to attend an event and another requesting someone to attend. We think that these speech acts are fundamentally similar, with the only difference being whether the recipient or the sender is tied to the activity in the request. Speech acts can also have both sender and recipient tied to the activity. We therefore extended our speech act definition to also include a *Speech Act Subject*, applicable only to speech acts with Activity nouns, where the subject can be the *Sender*, *Recipient*, or *Both*. Given these new parameters we define a *Realized Speech Act* (v-(o)[s]); where o denotes possible noun instances and s denotes the subject of activity noun instances if applicable.

3.2 Predicting Reactions on a Speech Act

Table 1. Realized Speech Acts Combinations and Expected Reactions



Verb-Noun	Object	Subject	Description	Role	ER[s]	ER[r]
Request-Activity	Event/Task	Recipient Both Sender	Request recipient to perform activity Request joint activity Request permission for activity	Requestive Requestive Requestive	Expect Expect Expect	Reply Reply Reply
Commit-Activity	Event/Task	Sender Both Recipient	Commit to an activity Commit/Instruct a joint activity Commit/Instruct recipient to activity	Resp/Informative Resp/Imperative Resp/Imperative	Perform Perform None	None Perform Perform
Decline-Activity	Event/Task	Recipient Sender Both	Decline permission for an activity Decline performing an activity Decline performing a joint activity	Responsive Responsive Responsive	None None None	None None None
Request-Data	Info/Reso		Request data from recipient	Requestive	Expect	Reply
Deliver-Data	Info/Reso		Deliver data	Resp/Informative	None	None
Decline-Data	Info/Reso		Decline delivering data	Responsive	None	None

The intents and expectations around which our ontology is designed correspond to the illocutionary and perlocutionary forces of the speech acts respectively. We now outline a non-deterministic predictive model to address them. We define the *Illocutionary Expected Reaction* [*ERs*] as the course of action expected out of the speech act sender on sending, and the *Perlocutionary Expected Reaction* [*ERr*] as the course of action expected out of the recipient on acknowledgment. We categorize reactions into *Passive* and *Active* reactions. Passive reactions are *Expect*, where the sender expects a response on sending a speech act; and *None*, where the sender or recipient is expected to do nothing on issuing or receiving the speech act. Active reactions are *Reply*, when the recipient is expected to reply on getting a speech act; and *Perform*, for speech acts which demand an Activity, e.g. Task, from the sender or recipient on sending or getting a speech act. We apply this predictive model to our realized speech act definition as (v-(o)[s]) {ERs}→{ERr}, denoting that on sending a speech act specific expected reactions for both sender and recipient are generated.

Not all combinations of the verb-noun pairs in the ontology are relevant. Whereas committing to an event makes sense, committing to a resource does not. Table 1 is an exhaustive table presenting all relevant speech acts given as the verb-noun pairs, their respective noun instances, and their activity subjects if applicable. A brief description for each realized speech act is given along the verb role and the expected reactions generated for the sender on sending and recipient on acknowledgment. The table highlights the fact that one speech-act can serve more than one role and can thus have more than one predictive force. If a person A requests another person B to attend to an event (Request-Event[Recipient]), then A's speech act has a requestive role. On sending, A expects a response, whereas when reading the email B is expected to reply. On the other hand, if A *instructed* B to go to the event in the first place (Commit-Event[Recipient]), the role of the speech act is imperative and therefore both informative and requestive. On sending A is expected to do nothing whereas on acknowledging the speech act B is expected to perform.

4 Future Directions and Conclusion

We are currently evaluating how well our speech act model fits a real corpus of threaded email messages, and how it compares to previous work in [6], by using the Kappa statistic to measure human annotator agreement for both models. After considering the results, we plan to extend popular Email Clients to enable semantic email by providing: semi-automatic content metadata extraction through text analytics; context metadata retention through threaded-based email handling; and invisible semantic annotation of email (based on our ontology) with such metadata through a MIME extension that allows for an RDF content-type in the email headers.

We believe that the presented models can be a sound basis for achieving our goal: improving the data workflow efficiency for the user. Although they are generic enough to be applied to other communication media, the problem addressed here mostly concerns disconnected workflows – where implicit expectations have a larger impact on workflow efficiency. We want to achieve a scenario where email users are aided by smarter email clients that predict their actions on the basis of the semantics accompanying speech acts in email. The semantic email-aware email client will aid the user by autonomously aiding the workflow of personal information generated by email. The client will suggest the most appropriate action for speech acts the user is creating or acting upon. Rather than going through unread mails, a user will be able to periodically check or even be reminded of speech acts they were expected to act upon and never did. If supported by personal information management tools, the email client might suggest saving a task in a task list once the user commits to it. This scenario would improve the overall efficiency of email conversations.

References

1. Khosravi, H., Wilks, Y.: Routing email automatically by purpose not topic. Natural Language Engineering, Vol. 5 (1999) 237–250.
2. Goldstein, J., Sabin, R.E.: Using Speech Acts to Categorize Email and Identify Email Genres. System Sciences, HICSS2006 (2006).
3. Erikson, T.: Making Sense of Computer-Mediated Communication (CMC): Conversations as Genres. Hawaiian International Conference on System Services, HICSS2000 (2000).
4. Whittaker, S., Bellotti, V., Gwizdka, J.: Email and PIM: Problems and Possibilities. CACM.7 (2007).
5. Searle, J.: Speech Acts. Cambridge University Press (1969).
6. Carvalho, V., Cohen, W.: Improving Email Speech Act Analysis via N-gram Selection. HLT/NAACL, ACTS Workshop, New York City, NY (2006).
7. Feng, D., Shaw, E., Kim, J., Hovy, E.: Learning to detect conversation focus of threaded discussions. Human Language Technology Conference, New York (2006) 208-215.
8. Dabbish, L., Kraut, R., Fussell, S., Kiesler, S.,: Understanding email use: predicting action on a message. SIGCHI conference on Human factors in computing systems (2005) 691-700.
9. Khoussainov, R., Kushmerick, N.: Email task management: An iterative relational learning approach. Conference on Email and Anti-Spam, CEAS'2005 (2005).
10. McDowell, L., Etzioni, O., Halevey, A., Levy, H.:Semantic email. In Proceedings of the 13th World Wide Web Conference, ACM Press, New York (2004) 244-254
11. Armstrong, E.: What's wrong with email? Bootstrap Alliance (2000)

Towards Distributed Ontologies with Description Logics

Martin Homola

Comenius University, Bratislava, Slovakia
homola@fmph.uniba.sk

1 Introduction

Mainstream research on Description Logics (DLs) usually treats DL knowledge bases as monolithic structures (cf. [1]). It has been pointed out, however, that in environments such as Semantic Web, distribution of ontological knowledge across various sources is expected and accepted [2]. Several use-cases can be provided for the envisioned distributed ontology environment:

- Two distinct applications may use two different ontologies to refer to the same concept. Thanks to mapping between these ontologies, association of these concepts can be derived.
- One particular application (e.g., a semantic annotation of a document) can use different ontologies to refer to two distinct concepts. These concepts, may be related by mapping, and so further semantic consequences can be derived.
- Developers may choose to map from a foreign ontology instead of repeating a complex description of some concept. Reuse of concepts can be facilitated.

As Kalfoglou et al. conclude in [3], "...ontology mapping nowadays faces some of the challenges we were facing ten years ago when the ontology field was at its infancy. We still do not understand completely the issues involved." This suggests that accomplishing these scenarios is a long-term and incremental task. In our research, we focus on DLs, a formalism with precise, logical semantics, with encouraging computational properties and practical reasoner implementations [1]. Also, the most prominent ontology language suggested by W3C, OWL Web Ontology Language [4], is derived form DLs.

From the DL point of view, the actual line of research suggests itself:

1. Describe useful syntactic constructs and intuitions behind them.
2. Provide formal model-theoretic semantics for these constructs, yielding a logic with reasoning tasks such as satisfiability of concepts and entailment of concept subsumption.
3. Provide reasoning algorithms for the decision tasks.
4. Develop and optimize implementations of reasoners.

Moreover, each time different set of constructs is put together in Step 1, the following steps need to be repeated, in order to investigate properties and practical usability of thus constructed DL.

In Step 1, for a start, the existing research on ontology mapping (see [3] for a survey) can provide us with intuitions and suggest syntactic constructs for ontology combination. In Step 2, we shall craft formal frameworks that allow combinations of DL knowledge bases using the syntactic constructs selected in Step 1. Reasoning algorithms and implementations follow in Steps 3 and 4.

There are several existing approaches. Distributed Description Logics (DDLs) of Borgida, Serafini and Tamilin [5–7], a framework in which concepts of DL knowledge bases are associated by so called bridge rules, thus allowing for inter-ontology subsumption. Grau et. al in [8] combine DL knowledge bases with $\mathcal{E}$-connections [9], thus allowing for links – inter-ontology role relationships.

It is also noted in [8] that DDLs expose unintuitive behaviour in some situations as demonstrated therein (see below). We find the idea of inter-ontology subsumption appealing. We have analyzed the problem mentioned in [8] and found that it can be "fixed". We summarize our results below in this paper.

2 Distributed Description Logics

A DDL knowledge bases consist of a distributed TBox – a set of local TBoxes, each over its own DL language $\mathcal{L}_i$ – and a set of bridge-rules $\mathfrak{B}$ between these local TBoxes. Each of the local TBoxes $\mathcal{T}_i$ is a collection of GCIs of the form: $i : C \sqsubseteq D$, where the prefix i identifies the TBox the GCI belongs to. Bridge-rules are of two forms, *into* bridge-rules and *onto* bridge-rules:

$$i : A \overset{\sqsubseteq}{\rightarrow} j : G \ , \qquad\qquad i : B \overset{\sqsupseteq}{\rightarrow} j : H \ .$$

For precise formal semantics of DDLs, please refer to [7]. A particularly attractive feature of DDLs is that they capture the reuse of concepts between several ontologies. This combines well with the basic assumption of Semantic web that no central ontology but many ontologies with redundant knowledge will exist.

In [8], it is noted that certain properties of subsumption relations are not modeled properly by DDL. This problem is demonstrated by the following example that we borrow from [8]. Consider the ontology O:

$$\begin{aligned} \mathrm{NonFlying} &\equiv \neg\mathrm{Flying} \ , &\qquad \mathrm{Penguin} &\sqsubseteq \mathrm{Bird} \ , \\ \mathrm{Bird} &\sqsubseteq \mathrm{Flying} \ , & \mathrm{Penguin} &\sqsubseteq \mathrm{NonFlying} \ . \end{aligned}$$

And the distributed counterpart of O, divided into two ontologies A (on the left) and B (on the right):

$$\begin{aligned} \mathrm{NonFlying_A} &\equiv \neg\mathrm{Flying_A} \ , &\qquad A : \mathrm{Bird_A} &\overset{\sqsupseteq}{\rightarrow} B : \mathrm{Penguin_B} \ , \\ \mathrm{Bird_A} &\sqsubseteq \mathrm{Flying_A} \ . & A : \mathrm{NonFlying_A} &\overset{\sqsupseteq}{\rightarrow} B : \mathrm{Penguin_B} \ . \end{aligned}$$

While Penguin in O is not satisfiable, the corresponding $\mathrm{Penguin_B}$ in B is. The problem is that the DDL framework allows an interpretation to associate each instance x of $\mathrm{Penguin_B}$ with two distinct elements, say y_1 and y_2, one instance of $\mathrm{Bird_A}$ and the other one of $\mathrm{NonFlying_A}$, even if $\mathrm{Bird_A}$ and $\mathrm{NonFlying_A}$ are disjoint. We agree with [8] that it is intuitive to expect that certain relations among concepts of one ontology propagate along bridge rules. So, we would expect $\mathrm{Penguin_B}$ to be unsatisfiable, as it is a "subconcept of two imported concepts" $\mathrm{Bird_A}$ and $\mathrm{NonFlying_A}$ which in their original ontology are disjoint.

3 Our Contribution So Far

We address the problem outlined above by introducing so called conjunctive bridge-rules. We use the following syntax (into and onto form respectively):

$$i : C \overset{\sqsubseteq}{\twoheadrightarrow} j : G \ , \qquad\qquad i : D \overset{\sqsupseteq}{\twoheadrightarrow} j : H \ .$$

Recall from [7], that distributed interpretation is a tuple $\mathfrak{I} = ((\Delta^{\mathcal{I}_1}, \cdot^{\mathcal{I}_1}), \dots, (\Delta^{\mathcal{I}_n}, \cdot^{\mathcal{I}_n}), r)$ comprising of local interpretations $(\Delta^{\mathcal{I}_i}, \cdot^{\mathcal{I}_i})$ for each local TBox $\mathcal{T}_i$ and $r(\cdot) = \bigcup_{i \neq j} r_{ij}(\cdot)$ interprets the mapping. In addition to the clauses that formally define the semantics of DDL (please refer to [7]), the semantics of conjunctive bridge-rules is given by the following two clauses:

1. $\mathfrak{I} \models_{\mathrm{d}} i : C \overset{\sqsubseteq}{\twoheadrightarrow} j : G$ if for each $i : D \overset{\sqsubseteq}{\twoheadrightarrow} j : H$, $r_{ij}\big(C^{\mathcal{I}_i} \cap D^{\mathcal{I}_i}\big) \subseteq G^{\mathcal{I}_j} \cap H^{\mathcal{I}_j}$,
2. $\mathfrak{I} \models_{\mathrm{d}} i : C \overset{\sqsupseteq}{\twoheadrightarrow} j : G$ if for each $i : D \overset{\sqsupseteq}{\twoheadrightarrow} j : H$, $r_{ij}\big(C^{\mathcal{I}_i} \cap D^{\mathcal{I}_i}\big) \supseteq G^{\mathcal{I}_j} \cap H^{\mathcal{I}_j}$,

where $\mathfrak{I} \models_{\mathrm{d}} R$ is to be read $\mathfrak{I}$ satisfies the bridge rule R.

Our choice of adding new kind of bridge-rules instead of replacing the old semantics by the new one is to underline that both kinds can co-exist and be used according to the intentions of the ontology editor. We continue with characterization of conjunctive bridge-rules. First, they are stronger than the original form in a sense: the semantic condition imposed by a bridge-rule of the original form is also imposed by the corresponding conjunctive form.

Theorem 1. *Given a distributed TBox $\mathfrak{T}$ with a set of bridge-rules $\mathfrak{B}$ and some local TBoxes $\mathcal{T}_i$ and $\mathcal{T}_j$ such that $i \neq j$ and $i : C \overset{\sqsubseteq}{\twoheadrightarrow} j : G \in \mathfrak{B}$ ($i : C \overset{\sqsupseteq}{\twoheadrightarrow} j : G \in \mathfrak{B}$), for each distributed interpretation $\mathfrak{I}$ such that $\mathfrak{I} \models_{\mathrm{d}} \mathfrak{T}$ it holds that $r_{ij}\big(C^{\mathcal{I}_i}\big) \subseteq G^{\mathcal{I}_j}$ ($r_{ij}\big(C^{\mathcal{I}_i}\big) \supseteq G^{\mathcal{I}_j}$) respectively.*

Next theorem shows that choosing conjunctive bridge-rules solves the problem outlined by the example above.

Theorem 2. *Given a distributed TBox $\mathfrak{T}$ with a set of bridge-rules $\mathfrak{B}$ and some local TBoxes $\mathcal{T}_i$ and $\mathcal{T}_j$ such that $i \neq j$, if for some $n > 0$ the bridge-rules $i : C_1 \overset{\sqsubseteq}{\twoheadrightarrow} j : G_1, \dots, i : C_n \overset{\sqsubseteq}{\twoheadrightarrow} j : G_n$ are all part of $\mathfrak{B}$ then for every distributed interpretation $\mathfrak{I}$ such that $\mathfrak{I} \models_{\mathrm{d}} \mathfrak{T}$ it holds that $r_{ij}\Big((C_1 \sqcap \dots \sqcap C_n)^{\mathcal{I}_i}\Big) \subseteq (G_1 \sqcap \dots \sqcap G_n)^{\mathcal{I}_j}$.*

Likewise, if for some $n > 0$ the bridge-rules $i : C_1 \overset{\sqsupseteq}{\twoheadrightarrow} j : G_1, \dots, i : C_n \overset{\sqsupseteq}{\twoheadrightarrow} j : G_n$ are all part of $\mathfrak{B}$ then for every distributed interpretation $\mathfrak{I}$ such that $\mathfrak{I} \models_{\mathrm{d}} \mathfrak{T}$ it holds that $r_{ij}\Big((C_1 \sqcap \dots \sqcap C_n)^{\mathcal{I}_i}\Big) \supseteq (G_1 \sqcap \dots \sqcap G_n)^{\mathcal{I}_j}$.

And so, for concepts involved in conjunctive bridge-rules, the intersection is always “properly related” (subset/superset, w.r.t. the kind of the bridge-rules) to the image of the intersection of their mapped counterparts. In the example above, the concept $\mathrm{Penguin_B}$ is mapped to both $\mathrm{Bird_A}$ and $\mathrm{NonFlying_A}$ which are disjoint. If we would use conjunctive bridge rules in this example, we would have ${\mathrm{Penguin_B}}^{\mathcal{I}_\mathrm{B}} \cap {\mathrm{Penguin_B}}^{\mathcal{I}_\mathrm{B}} \subseteq r_{ij}\left({\mathrm{Bird_A}}^{\mathcal{I}_\mathrm{A}} \cap {\mathrm{NonFlying_A}}^{\mathcal{I}_\mathrm{A}}\right)$, which yields ${\mathrm{Penguin_B}}^{\mathcal{I}_\mathrm{B}} \subseteq r_{ij}(\emptyset)$ and so ${\mathrm{Penguin_B}}^{\mathcal{I}_\mathrm{B}} \subseteq \emptyset$. That is, in the distributed knowledge base, the concept $\mathrm{Penguin_B}$ is unsatisfiable, according to our intuition.

In [5–7] various desiderata for DDLs are stated. These include:

Monotonicity. Bridge-rules do not delete local subsumptions.

Simple subsumption propagation. Combination of into and onto bridge-rules allows for propagation of subsumption across ontologies. Formally, if $i : C \overset{\sqsupseteq}{\rightarrow} j : G \in \mathfrak{B}$ and $i : D \overset{\sqsubseteq}{\rightarrow} j : H \in \mathfrak{B}$ then $\mathfrak{T} \models_\mathrm{d} i : C \sqsubseteq D \implies \mathfrak{T} \models_\mathrm{d} j : G \sqsubseteq H$.

Generalized subsumption propagation. If $i : C \overset{\sqsupseteq}{\rightarrow} j : G \in \mathfrak{B}$ and $i : D_k \overset{\sqsubseteq}{\rightarrow} j : H_k \in \mathfrak{B}$, for $1 \leq k \leq n$ then $\mathfrak{T} \models_\mathrm{d} i : C \sqsubseteq \bigsqcup_{k=1}^{n} D_k \implies \mathfrak{T} \models_\mathrm{d} j : G \sqsubseteq \bigsqcup_{k=1}^{n} H_k$.

These desiderata are satisfied even in the presence of conjunctive bridge-rules. Moreover, subsumption propagates over conjunctive bridge-rules even for intersection of concepts, as follows.

Theorem 3. *If $i : C \overset{\sqsupseteq}{\twoheadrightarrow} j : G \in \mathfrak{B}$ and $i : D_k \overset{\sqsubseteq}{\twoheadrightarrow} j : H_k \in \mathfrak{B}$, $1 \leq k \leq n$ then $\mathfrak{T} \models_\mathrm{d} i : C \sqsubseteq \bigsqcap_{k=1}^{n} D_k \implies \mathfrak{T} \models_\mathrm{d} j : G \sqsubseteq \bigsqcap_{k=1}^{n} H_k$.*

These results have not been published yet. We have submitted a technical paper to DL-2007. Other interesting desiderata for DDLs published in [7] include requirements that local inconsistency does not pollute the entire distributed system and that the information flow along the bridge-rules respects the direction of these bridge-rules (i.e., no information flows the other way). Evaluating conjunctive bridge-rules with respect to these desiderata is subject to our ongoing research.

4 Conclusion and Future Work

We decided to pursue research for our Ph.D. dissertation in the area of Distributed Ontologies based on Description Logics. Several approaches exist in this field, most notably those of [8], exploiting the $\mathcal{E}$-connections framework, and those of [5–7] introducing Distributed Description Logics (DDLs). We find the latter one interesting, since it allows for inter-ontology subsumption – a notion that combines well with the vision of Semantic Web. However, an unintuitive behaviour of DDLs has been demonstrated in [8]. We have found it natural to start our research, concentrating on this problem. We have figured out that the semantics of DDLs can be amended to solve this problem. There are several possibilities to continue with our research:

- Continue with the evaluation of conjunctive bridge-rules (e.g., with respect to the remaining desiderata for DDLs.)
- Devising a reasoning algorithm, studying its computational complexity.
- Implementation and practical evaluation. (Depends on the previous item.)
- Exploiting the possibility to combine DDLs with links used in $\mathcal{E}$-connections.

We intend to address these issues in our Ph.D. research. We will continue with the evaluation of conjunctive bridge-rules first. Also, there is a tableaux algorithm known for the original DDL framework (see [6, 7]), we will start by checking whether it can be adopted for conjunctive bridge-rules effectively. We would like to continue with the complexity analysis next. The last item of the list are very interesting as well.

Acknowledgement. The author is a Ph.D. student in third year, however, he only switched to Distributed Ontologies a year ago. Support from Slovak Agency for Promotion of Research and Development projects, contract no. APVV-20-P04805 and APVV-99-PO5005; and project VEGA of Slovak Ministry of Education and Slovak Academy of Sciences no. 1/0173/03 is acknowledged.

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The Description Logic Handbook. Cambridge University Press (2003)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284**(5) (2001) 34–43
3. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: The state of the art. In: Semantic Interoperability and Integration. Number 04391 in Dagstuhl Seminar Proceedings (2005)
4. : OWL web ontology language overview. A W3C Recommendation. Available online, at `http://www.w3.org/TR/owl-features/`
5. Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. Journal of Data Semantics **1** (2003) 153–184
6. Serafini, L., Tamilin, A.: Local tableaux for reasoning in distributed description logics. In: Procs. of DL'04. CEUR-WS (2004)
7. Serafini, L., Borgida, A., Tamilin, A.: Aspects of distributed and modular ontology reasoning. In: Procs. of IJCAI'05. (2005) 570–575
8. Cuenca Grau, B., Parsia, B., Sirin, E.: Combining OWL ontologies using $\mathcal{E}$-connections. Journal of Web Semantics **4**(1) (2006) 40–59
9. Kutz, O., Lutz, C., Wolter, F., Zakharyaschev, M.: $\mathcal{E}$-connections of abstract description systems. Artificial Intelligence **156**(1) (2004) 1–73

A Framework for Distributed Reasoning on the Semantic Web Based on Open Answer Set Programming

Cristina Feier

Digital Enterprise Research Institute, Universität Innsbruck,AT
{firstname.lastname}@deri.org

Abstract. The Semantic Web envisions a Web where information is represented by means of formal vocabularies called ontologies for enabling automatic processing and retrieval of information. It is expected that ontologies will be interconnected by mappings forming a network topology. Reasoning with such a network of ontologies is a big challenge due to scalability issues. The local model semantics seems to be a promising approach in this direction that has served as the basis of several frameworks/distributed languages. The intent of the work described in this paper is to define a new framework for representing and reasoning with interconnected ontologies that will be based on a new language for representing ontologies and mappings called Distributed Open Answer Set Programming (DOASP). DOASP is a syntactical extension of OASP in the direction of distributedness and its semantics is a combination between the local model semantics and the OASP semantics. The reason for choosing OASP is the emerging interest in hybrid formalisms for the Semantic Web.

1 Research Context and Problem Statement

While the current Web is only usable by humans, the Semantic Web [1] envisions to make information processable and services on the Web usable by machines as well. The main technology for establishing the Semantic Web is the creation of so-called *ontologies*, which can be used to represent information and services on the Web. Ontologies are commonly defined as *formal specifications of shared conceptualizations* [2, 3] where the sharing aspect is important: ontologies have to be *reusable*.

This reusability does not mean that on the Web there will be a single ontology for capturing all the knowledge (not even all the knowledge corresponding to a given domain); there will be different ontologies that describe the same domain at different levels of granularity or from different perspectives, or that describe partially overlapping domains. In order to effectively use such overlapping ontologies one needs a means to describe the way they are interconnected. In particular, one uses logical axioms called *mappings* to relate elements of one ontology to elements of others.

In order to represent such ontologies and mappings, logical languages such as Description Logics (DLs) [4] or Logic Programming(LP) [5] can be used. The usage of formal languages allows in general for well-defined formal reasoning and thus the necessary automation of tasks such as checking consistency of ontologies. Some popular languages for representing ontologies anchored in one or both of the mentioned formalisms are WSML [6], OWL [7], SWRL [8].

Description Logics are attractive as the basis for ontology representation formalisms due to their many different decidable expressive fragments and their suitability for conceptual modeling as they adopt the Open World Assumption. However standard DLs lack support for non-monotonicity and have rather rigid constructs for expressing knowledge. Logic Programming (LP) on the other hand is a knowledge representation formalism with support for non-monotonic reasoning through the *negation as failure* operator. Moreover, the rule-based presentation of LP makes representing knowledge rather intuitive. A disadvantage of most LP approaches is their Closed World Assumption which is not realistic in an open like the Web, where knowledge is notoriously incomplete. One approach which tries to reconcile the use of negation of failure with the Open World Assumption which is characteristic to the Semantic Web is to restrict the use of negation in the form of so-called negation-as-failure [**?**]. Other approaches which try to combine advantages of the DL and LP paradigms are the so-called hybrid representation formalisms. Among them is also the language *Open Answer Set Programming* (OASP) [9–11] which keeps the rule-based presentation and the nonmonotonic capability from LP and drops the closed-domain assumption to allow for open domains as is the case in Description Logics (and first-order logic).

In the context of interconnected ontologies with mappings, an important factor, besides the syntax and semantics of the specific language(s) used for representing ontologies, is the choice for a semantics and algorithms for reasoning with the whole network of ontologies. The simplest approach, also called *global semantics* is to assume that all information relevant to a query (i.e., all the relevant ontologies and mappings) is put together and reasoning is performed as with a local ontology. Such an approach is considered for example in [12]. While this is a simple approach to deal with interconnected ontologies, it has some inconveniences: language specificity is constrained (ontologies and mappings have all to be expressed into formalism(s) which the particular reasoning engine supports), and local inconsistency propagates to the whole reasoning space. Also, computing a global model for the whole reasoning space imposes a heavy computational burden for the inferencing task.

Some existing work from the AI area of contextual reasoning seems to come handy for reasoning with networks of ontologies. In [13] a semantics called *local model semantics*, that captures the main intuitions behind context modeling, which are *locality* and *compatibility*, has been introduced. The advantages of this semantics over the simple approach described above are the possibility of using different local reasoners, thus also different formalisms for representing ontologies, the possibility to isolate local inconsistencies, and better scalability. It is common for the systems that implement this semantics to be based on distributed reasoning procedures.

Current approaches for representing and reasoning with interconnected ontologies using the local model semantics are based on Propositional Logic [14–17], Default Logic [18], First Order Logic [19, 20] and Description Logics [21, 22]. As concerns the reasoning support, propositional multi-context systems were put in correspondence with bounded modal logic K_n and shown that any contextual satisfiability problem can be reduced to that of satisfying some formula in K_n whose modal depth is at most equal to one [15]. Also contextual satisfiability problems were tractably encoded into purely propositional ones, which enables SAT-based implementations of such systems [14, 16].

A special case is that of Propositional Logic theories connected by bridge rules in which negation as failure can appear, the so-called information chain theory [17]. In this case a solution based on a fix-point operator that computes at each step a chain/anti-chain pair was devised. An algorithm based on a variation of the Well-Founded Semantics for Default Logic, WFS2 [23], was devised for reasoning with contextual default theories [18]. A similar algorithm (in the sense that it isolates local inconsistencies) based on the stable model semantics was devised for reasoning with ground ASP. A sound and complete calculus for DFOL based on ML [24] is presented in [19]. A distributed tableaux reasoning algorithm for DDL that works only for acyclic TBoxes is described in [25] and [21].

The main goal of the project is, given the advantages of OASP compared to DLs, and the benefits of the local model semantics, similarly to the extension of the DL semantics to the interconnected case in [21], to extend OASP to its distributed variant *Distributed Open Answer Set Programming* (DOASP). Moreover, I will investigate decidability of this extended framework (fragments) together with associated reasoning procedures.

2 Objectives and Approach

The main objectives of this work are:

- to define a language based both on LP and DL for representing ontology spaces, more specifically a distributed version of Open Answer Set Programming, called Distributed Open Answer Set Programming.
- to provide a declarative semantics for this distributed language in the style of Local Model Semantics described in [20].
- to design algorithms for reasoning with this language according to the proposed semantics.

In the following I give a short overview of how each of the precedent issues is/will be addressed together with corresponding evaluation criteria.

DOASP. Distributed Open Answer Set Programming is a language that extends Open Answer Set Programming in order to accommodate the representation of context spaces in accordance with the two principles underlying the notion of context: locality and compatibility. A context is a set of OASP rules. Every literal has attached the identifier of the context it is part of. The context of the head of the rule is always the context the rule makes part from. Rules that contain only literals belonging to the current context are local rules, while those who contain foreign literals in the body are bridge rules. Thus, it is possible for a bridge rule to connect more than two ontologies, depending on the number of foreign literals in the body.

Semantics. The semantics for DOASP is a blend between the local model semantics [13] and the OASP semantics. The most straightforward way to define it would be to simply extend the stable model based semantics of OASP to the distributed case similarly to the extension of the stable model semantics in the information chain approach. However, defining the semantics in such a way leads to the propagation of local inconsistencies, and does not lend itself to distributed reasoning. The possibility to define the semantics of DOASP similar with the well-founded based semantics defined for the information chain approach (the one based on a fix point which iterates over chain, anti-chain pairs) will be investigated, or furthermore, a para-consistent semantics like in the case of Con-DL might be considered.

The evaluation will consist in considering several concrete scenarios (DOASP encodings of real world situations) in order to justify the intuition behind the new semantics and the utility of the approach.

Algorithms. Most of the DL-based approaches which adopt the local model semantics are able to deal only with very restricted ontology spaces, while the non-monotonic approaches based on this kind of semantics rely on propositionalization. This, together with the fact that so far only decidable fragments of OASP were identified [11] without associated effective reasoning algorithms, points to the fact that the the identification of such algorithms is a non-trivial task. Note that due to the presence of open domains, propositionalization cannot be applied (at least, not straightforwardly) for reasoning with OASP as is commonly done with ASP.

In order to identify practical reasoning algorithms for reasoning with integrating languages, it is promising to investigate the relation with pure database languages. For example, in [26] the relation of Open Answer Set Programming with the database language Datalog LITE was established. Given the close relation of Datalog LITE with the relational algebra, one could build upon existing database techniques for reasoning with DOASP. Another direction of investigation is about the possibility of employing modularity results from the Stable Model Semantics community like splitting sets [27], Rosati's weak safeness condition [28], in order to identify what has to be evaluated together and what can be evaluated separately.

As full-fledged OASP is undecidable, and thus also DOASP, I will start with considering subsets of the language with reduced expressivity and then incrementally I will consider more expressive subsets for devising algorithms. As an evaluation criteria, the algorithms has to be proven as sound and complete.

An orthogonal direction for reasoning is the use of heuristics in order to maintain soundness but not necessarily preserving completeness. Performing an exhaustive search on the Web seems not desirable in many concrete scenarios, so pruning the context space or simplifying it might be an option.

3 Status, Future work

I consider myself to be at the end of the phase of defining the research problem. So far, the syntax and a declarative semantics for DOASP has been defined. I have been interested in this topic for around one and a half years. As future work, I plan to follow the steps mentioned in the previous section.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284** (2001) 34–43
2. Gruber, T.: An ontology for engineering mathematics. Technical report, KSL-94-18, Stanford, Ca (1994)

3. Borst, W.: Construction of engineering ontologies. Technical report, University of Tweenty. Enschede, NL-Centre for Telematica and Information Technology (1997)
4. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook. Cambridge University Press (2003)
5. Gelfond, M., Lifschitz, V.: The Stable Model Semantics for Logic Programming. In: Proc. of ICLP'88, Cambridge, Massachusetts, MIT Press (1988) 1070–1080
6. de Bruijn, J., eds.: The WSML Family of Representation Languages. Deliverable D16.1v0.2, WSML (2005) Available from http://www.wsmo.org/TR/d16/d16.1/v0.2/.
7. Dean, M., Schreiber, G., eds.: OWL Web Ontology Language Reference. (2004) W3C Recommendation 10 February 2004.
8. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A semantic web rule language combining OWL and RuleML. 21 may 2004, W3C (2004)
9. Heymans, S., Nieuwenborgh, D.V., Vermeir, D.: Open answer set programming with guarded programs. ACM Transactions on Computational Logic (TOCL) (2006)
10. Heymans, S., Nieuwenborgh, D.V., Vermeir, D.: Conceptual logic programs. Annals of Mathematics and Artificial Intelligence (Special Issue on ASP) (2006)
11. Heymans, S., Van Nieuwenborgh, D., Vermeir, D.: Guarded Open Answer Set Programming. In: LPNMR 2005. Number 3662 in LNAI, Springer (2005) 92–104
12. Haase, P., Wang, Y.: A decentralized infrastructure for query answering over distributed ontologies. In: Proc. of the 22nd Annual ACM Symposium on Applied Computing. (2007)
13. Ghidini, C., Giunchiglia, F.: Local model semantics, or contextual reasoning = locality+compatibility. Artificial Intelligence **127** (2001) 221–259
14. F. Roelofsen, L.S., Cimatti, A.: Many hands make light work: Localized satisfiability for multi-context systems. In: ECAI-04, Valencia, Spain (2004)
15. Serafini, L., Roelofsen, F.: Complexity of contextual reasoning. In: Nineteenth National Conference on Artificial Intelligence (AAAI-04), San Jose, California, USA (2004)
16. Serafini, L., Roelofsen, F.: Satisfiability for propositional contexts. In: KR' 04). (2004) 369
17. Roelofsen, F., Serafini, L.: Minimal and absent information in contexts. In: (IJCAI-05), Edinburgh, Scotland (2005)
18. Brewka, G., Roelofsen, F., Serafini, L.: Contextual default reasoning. In: Proc. IJCAI-07, Hyderabad, India (2007)
19. Ghidini, C., Giunchiglia, F.: Distributed first order logic. Frontiers of Combining Systems 2, Research Studies Press Ltd. (2000) 121–139
20. Ghidini, C., Serafini, L.: Distributed first order logic. Technical report, ITC-irst (1998)
21. Serafini, L., Tamilin, A.: Drago: Distributed reasoning architecture for the semantic web. Technical report, ITC-irst (2004)
22. Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., Stuckenschmidt, H.: Contextualizing ontologies. Journal of Web Semantics **1** (2004)
23. Brewka, G., Gottlob, G.: Well-founded semantics for default logic. Fundamenta Informaticae **31** (1997) 221–236
24. Giunchiglia, F., Serafini, L.: Multilanguage hierarchical logics. AI (1994) 29–70
25. Serafini, L., Tamilin, A.: Local tableaux for reasoning in distributed description logics. (In: Proc. of the 2004 Intl. Workshop on Description Logics (DL2004))
26. Heymans, S., Van Nieuwenborgh, D., Vermeir, D.: Guarded Open Answer Set Programming with Generalized Literals. In Dix, J., Hegner, S., eds.: FoIKS 2006. (2006) 179–200
27. Lifschitz, V., Turner, H.: Splitting a logic program. In Hentenryck, P.V., ed.: Proc. of Eleventh Intl Conf. on Logic Programming, San Jose, Calif. (1994) 23–37
28. Rosati, R.: On the decidability and complexity of integrating ontologies and rules. **3** (2005) 41–60

Logic as a power tool to model negotiation mechanisms in the Semantic Web Era

Azzurra Ragone
a.ragone@poliba.it

SisInfLab, Politecnico di Bari, Bari, Italy

1 The research problem

The key theme of our research is the modeling of multi-attribute negotiation scenarios with the aid of logic languages, going from very simple (and not so expressive) languages as propositional logic up to more expressive logics, such as Description Logics (DLs).

The approach to multi-attribute negotiation we are investigating on exploits logic languages at least in two ways: (1) to model, through an ontology, relations between attributes to be negotiated and (2) to characterize buyer and seller preferences. Some of the advantages of such an approach are intuitive: the possibility to model disjontness, implication, utilities on bundle of issues are all useful in settings where negotiation is not limited to undifferentiated goods but is based on complex descriptions that require adequate negotiation mechanisms to produce —in an automated way— fair deals.

The research problem is both challenging and very timely for the Semantic Web initiative, because of the undeniable importance of e-commerce and negotiation over the Web and because Description Logics are at the core of Semantic Web languages. Indeed, many recent research efforts have been focused on automated negotiation in various contexts, including e-marketplaces, resource allocation settings, online auctions, supply chain management and, generally speaking, e-business processes. We think that, as it will be outlined in the next sections, DLs can be the pivotal tool in modeling negotiation mechanisms in the tumultuous Semantic Web arena.

2 The state of art

Automated bilateral negotiation between agents has been widely investigated, both in artificial intelligence and in microeconomics research communities. AI-oriented research has usually focused on automated negotiation between agents and on designing high-level protocols for agent interaction. Agents can play different roles: act on behalf of buyer or seller, but also play the role of a mediator or facilitator. In the following we give a brief overview of logic-based approaches to automated negotiation, comparing our approach with existing ones and highlighting differences. In [1] the use of propositional logic in multi-issue negotiation was investigated, while in [2] weighted propositional formulas in preference modeling were considered. However, in such works, no semantic relation between issues is taken into account. In our approach we adopt a logical theory, *i.e.*, an ontology, which allows one *e.g.*, to catch inconsistencies between demand and supply, model implication, find out a feasible agreement in a bundle, which are fundamental issues to model an e-marketplace. We borrow from [11] the definition of agreement as a model for a set of formulas from both agents. However, in [11] only multiple-rounds protocols are studied, and the approach leaves the burden to reach an agreement to the agents themselves, although they can follow a protocol. The approach does not take preferences into account, so that it is not possible to guarantee that the reached agreement is Pareto-efficient. Our approach, instead, aims at giving an *automated* support to negotiating agents to reach Pareto agreements. With reference to the work presented in [12], adopting a propositional logic setting, *common knowledge* is considered as just more entrenched preferences, that could be even dropped in some deals. We adopt a *knowledge base*, or ontology $\mathcal{T}$, of formulas which are common knowledge for both agents, whose constraints must always be enforced in the negotiation outcomes. Moreover we use *additive utilitites* over formulas: this allows an agent to make compensations between its requests and its concessions, while in [12] the concession of a more entrenched formula can never be compensated by less entrenched ones, no matter how many they are. Finally we devised a *protocol* which the agents should adhere to while negotiating; in contrast in [12] a game-theoretic approach is taken, presenting no protocol at all, since communication between agents is not considered. To the best of our knowledge our approach is the first one using DLs to design a logic-based negotiation mechanism, ensuring a greater expressiveness w.r.t. propositional logic. Moreover, w.r.t. to non-logic-based approaches, the use of an ontology $\mathcal{T}$ allows exploiting inference services that are used in the actual negotiation mechanisms.

3 Expected contributions

Bilateral negotiation is a challenging problem, which finds applications in a number of different scenarios, each one with its own peculiarities and issues. Among others, the approach can suitably model negotiation in e-marketplaces. Clearly, here we do not deal with simple marketplaces of commodities and undifferentiated goods, where only price, time or quantity have to be taken into account. We refer to e-marketplaces dealing with complex products (*e.g.,* computers, automobiles, houses, and so on) where both offers/requests are referring to goods/services that cannot be simply described in a machine understandable way without the help of some Knowledge Representation (KR) language. When a potential buyer browses *e.g.,* an automobile e-marketplace, she looks for a car fulfilling her needs and/or wishes, so not only the price is important, but also warranty or delivery time, as well as look, model, comfort and so on. In such domains it is harder to model not only the negotiation process, but also the request/offer descriptions, as well as finding the best suitable agreement. Furthermore preferences can refer to (1) bundle of issues, *e.g., Sports car with navigator pack* where both the meaning of *sport car* and *navigator pack* are in the ontology; or preferences can be (2)*conditional* ones – when issues are inter-dependent *i.e.,* the selection of one issue depends on the selection made for other issues – *e.g., I would like a car with leather seats if its color is black*. In such a cases some kind of logical theory (ontology), able to let users express their needs/offers, could surely help. Also, when descriptions refer to complex needs, we should take into account preferences, distinguishing them from hard mandatory constraints (**strict requirements**), e.g., *I would like a black station wagon,* *preferably with GPS system*[1]. The possibility to handle some of the above mentioned issues in some electronic facility may help not only in the discovery/matchmaking stage of a transaction process, thus selecting most promising counterparts to initiate a negotiation, but also in the actual negotiation stage. Obviously, the one described above is not the only feasible scenario to apply the approach proposed, since the negotiation framework we propose is very general and can be applied to many other negotiation scenarios where resource descriptions have to be modeled through KR languages, as *e.g.*, in (web)service scenarios.

4 Research methodology

In the early stage of our research we started modeling preferences and goods/services descriptions using propositional logic. In [6] we presented the theoretical framework, which makes use of a facilitator to compute, through a one-shot negotiation protocol, some particular Pareto-efficient outcomes – the ones which maximize the social welfare and the product of utilities.

Differently from well-known approaches that describe issues as uncorrelated; we represented buyer's request, seller's supply and their respective preferences as formulas endowed with a formal semantics. By modeling preferences as formulas it is hence possible to assign a utility value also to a bundle of issues, which is obviously more realistic than the trivial sum of utilities assigned to single elements in the bundle itself.

Afterward the approach has been extended and generalized and also complexity issues were discussed [5]: we proved the computational adequacy of our method by studying the complexity of the problem of finding Pareto-efficient solutions in a propositional logic setting, in particular we proved that both problems – the one maximizing the product and the one maximizing the sum of utilities – are NPO-complete problems. A further improvement has been presented in [9], where we extended the framework, so that it is possible to handle, in a homogeneous setting, both numerical features and non-numerical ones. The framework makes possible to formally represent typical situations in real e-marketplaces such as "*if I spend more than 20000 € for a sedan then I want a navigator pack included*" where both numerical (price) and non-numerical (sedan, navigator pack) issues coexist. To this aim we introduce $\mathcal{P}(\mathcal{N})$, a propositional logic extended with *concrete domains*, which allows to: model relations among issues (both numerical and not numerical ones) via logical entailment: *e.g.*, the seller can state that if you want a car with a GPS system you have to wait at least one month: $(\texttt{GPS\_system} \Rightarrow \texttt{deliverytime} \geq 31)$; as well as preferences can involve only numerical ones: *e.g.*, the buyer can state that she would be willing to spend more than 25000 € for a sedan only if there is more than a two years warranty $[(\texttt{price} > 25000) \Rightarrow (\texttt{year\_warranty} > 2)]$.

In the approach we proposed, buyer and seller reveal their preferences to a mediator, which compute Pareto-efficient agreements solving a multi objective optimization problem (MOP). Actually, we solve a multi objective optimization problem as we try to make buyer and seller equally satisfied, maximizing different utility functions, both of the buyer and the seller.

[1] Strict requirements, in contrast with preferences, are constraints the buyer and the seller want to be necessarily satisfied to accept the final agreement, while preferences are issues they may accept to negotiate on.

In addition to the set of functions to maximize, in a MOP there are also a set of constraints that have to be satisfied. In our setting, we have three different sets of constraints, coming from (1) the ontology, (2)the strict requirements (see Section 3) and (3) the disagreement thresholds [2]. The returned solution to the MOP is the agreement proposed to the buyer and the seller. Notice that a solution to a MOP is always Pareto optimal.

The negotiation mechanisms described so far model a bargaining scenario where agents—acting on behalf of buyer and seller—reveal their preferences to a mediator, which has the burden of collecting information and proposes a fair agreements to both participants. The intervention of a mediator is due to the fact that usually bargainers may not want to disclose their preferences or utility function to the other party, but they can be ready to reveal these information to a trusted automated mediator helping negotiating parties to achieve efficient and equitable outcomes. Therefore we proposed a one-shot protocol with the intervention of a mediator suggesting to each participant the solution which is Pareto-efficient. For what concerns strategy, the players can choose to accept or refuse the solution proposed by the mediator; they refuse if they think possible to reach a better agreement looking for another partner, or another shot, or for a different e-marketplace.

However in some cases it is not possible to rely on a mediator, so a decentralized approach has to be adopted —instead of a centralized one— where agents negotiate without the help of a mediator. Obviously, in such cases it can be difficult to design negotiation mechanism leading to Pareto-efficient agreements [4]. We are currently investigating some alternative negotiation mechanisms without the presence of a mediator.

The need for more expressive languages to adequately model negotiation frameworks led us to [7, 8] move from propositional logic to DLs.

We are, to the best of our knowledge, the first ones proposing a DLs-based approach to model multi-attribute bilateral negotiation. Being, in general, much more expressive than *e.g.*, propositional logic, DLs allow to model complex preferences on bundles of interrelated issues and to exploit inference services —such as satisfiability and subsumption— available in optimized reasoners. Satisfiability is useful to catch inconsistency between agent's preferences w.r.t. the ontology $\mathcal{T}$, *i.e.*, inconsistent goals cannot be in the same agreement, (*e.g.*, agents cannot agree on A and B at the same time if in $\mathcal{T}$ A is defined as disjoint from B). Through subsumption it is possible to discover if an agent's goal is satisfied by a goal of the opponent's one even if it does not immediately appears at the syntactic level.

In [7] we propose a logic-based *alternating-offers* protocol, inspired by Rubinstein's one [10]. Our protocol merges both Description Logics formalism and reasoning services, and utility theory, to find the most suitable agreements. We have also implemented a prototype to carry out some preliminary experiments with a buyer negotiating with multiple sellers at the same time, as would be in a real e-marketplace.

In [8] we propose a novel negotiation mechanism designed to model scenario with *fully* **incomplete information**. Actually, while in [7] we consider a scenario with *partial* incomplete information—the agents know the goals (preferences) of the other agents ignoring the utility value assigned to them—in [8] we consider agents keeping as private

[2] Thresholds are the minimum utility that each agent requires to pursue a deal. Minimum utilities may incorporate an agents attitude toward concluding the transaction, but also overhead costs involved in the transaction itself, e.g., fixed taxes.

information both their goals and their worths [3]. The protocol we propose in [8] is able to deal with such incomplete information without forcing agents to reveal either their goals or utility functions, so it suits all scenarios where agents are not willing to reveal private information or when it is hard to design a truthful revelation mechanism [4]. We prove that the proposed protocol converges if the DL adopted to model buyer's and seller's goals is constrained to satisfy the so-called finite implicants property. Such a protocol allows agent to use different strategies: we introduce and discuss two possible strategies, highlighting their properties. We are planning the implementation of a prototype and test-beds to numerically evaluate best strategies to adopt w.r.t. the negotiation mechanism.

5 Conclusion and Future Work

We have presented new approaches to automate logic-based multi-attribute negotiation. The research started from the investigation of propositional logic as a language to model agent proposals. Propositional logic, equipped with an ontology, allows modeling bundle of preferences and implication between them. Differently from well-known approaches that describe issues as uncorrelated; we represent buyer's request, seller's supply and their respective preferences as formulas endowed with a formal semantics. By modeling preferences as logical formulas it is hence possible to assign a utility value also to a bundle of issues, which is obviously more realistic than the trivial sum of utilities assigned to single elements in the bundle itself. In the second year of my PhD we moved to DLs to express agent's preferences. DLs in fact allow a greater expressiveness, remaining decidable. We have studied different negotiation mechanisms with the presence of a mediator [6, 5, 9] and without a mediator, with partial incomplete information [7] or fully incomplete information [8]. For each mechanism we illustrated the protocol followed by the agents and one or more strategies agents can adopt depending on the designed protocol. We have also implemented a prototype to carry out some initial experiments. In future work we are planning to validate our approach with agent-based simulations, and for the DLs-based frameworks [7, 8] we are also setting up an analysis of the game theoretic properties, as related properties of the negotiation protocols (*e.g.*, Pareto-efficiency), equilibrium strategies or properties of the agents (*e.g.*, individual rationality).

References

1. S. Bouveret, M. Lemaitre, H. Fargier, and J. Lang. Allocation of indivisible goods: a general model and some complexity results. In *Proc. of AAMAS '05*, pages 1309–1310, 2005.
2. Y. Chevaleyre, U. Endriss, and J. Lang. Expressive power of weighted propositional formulas for cardinal preference modeling. In *Proc. of KR 2006*, pages 145–152, 2006.
3. J.S. Rosenschein and G. Zlotkin. *Rules of Encounter*. MIT Press,, 1994.
4. Sarit Kraus. *Strategic Negotiation in Multiagent Environments*. The MIT Press, 2001.
5. A. Ragone, T. Di Noia, E. Di Sciascio, and F.M. Donini. A logic-based framework to compute pareto agreements in one-shot bilateral negotiation. In *Proc. of ECAI'06*, pages 230–234, 2006.
6. A. Ragone, T. Di Noia, E. Di Sciascio, and F.M. Donini. Propositional- logic approach to one-shot multi issue bilateral negotiation. *ACM SIGecom Exchanges*, 5(5):11–21, 2006.

7. Azzurra Ragone, Francesco Colasuonno, Tommaso Di Noia, Eugenio Di Sciascio, and Francesco M. Donini. Logic-based alternating-offers protocol for automated multi-issue bilateral negotiation in P2P e-marketplaces. Submitted for publication, 2007.
8. Azzurra Ragone, Tommaso Di Noia, Eugenio Di Sciascio, and Francesco M. Donini. Description Logics for Multi-issue Bilateral Negotiation with Incomplete Information. Submitted for publication, 2007.
9. Azzurra Ragone, Tommaso Di Noia, Eugenio Di Sciascio, and Francesco M. Donini. When price is not enough: Combining logical and numerical issues in bilateral negotiation. In *proc. of International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007)*. ACM Press, May 2007. to appear.
10. A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50:97–109, 1982.
11. M. Wooldridge and S. Parsons. Languages for negotiation. In *Proc of ECAI '00*, pages 393–400, 2000.
12. Dongmo Zhang and Yan Zhang. A computational model of logic-based negotiation. In *Proc. of the AAAI 06*, pages 728–733, 2006.

6 Acknowledgments

I would like to acknowledge Tommaso Di Noia, Eugenio Di Sciascio and Francesco M. Donini for their guidance, support and fruitful discussions.

Improving the Usability of Large Ontologies by Modularization

Anne Schlicht

Universität Mannheim, Germany
anne@informatik.uni-mannheim.de

1 Problem Definition

Recently, the benefits of modular representations of ontologies have been recognized by the semantic web community. With the growing utilization of ontologies in almost all branches of science and industry not only the number of available ontologies has increased considerably but also many widely used ontologies have reached a size that cannot be handled by the available reasoners (like the ontology for Diagnoses for Intensive Care Evaluation (DICE)), some cannot even be loaded into an standard editor e.g. the Foundational Model of Anatomy ontology (FMA). When system memory and computation time cannot be extended anymore the only feasibility to use existing tools is partitioning of the large ontology into smaller parts. Realization of this divide-and-conquer approach requires an infrastructure that supports reasoning over a distributed ontology and partitioning of a large ontology into a distributed ontology.

2 Related Work

The state of the art in ontology engineering is the usage of monolithic ontologies. Large ontologies are engineered by teams using editors like Protégé [1] with no support for modularization apart from the "owl:imports" construct that copies all axioms of one ontology into another. Maintenance and usage of ontologies is completely centralized in opposition to the idea of knowledge sharing that initiated the utilization of ontologies. There are tools for ontology development and a couple of reasoners (RACER [2], PELLET [3]) available but approaches to modular ontologies are still in the fledgling stages.

A modular ontology formalism is defined by every approach of mapping or importing ontologies because the modules of an ontology can be connected by links or import declarations.
The only tool that implements a distributed setting for DL i.e. performs reasoning tasks using different instances of a reasoner for different modules is the tableaux-reasoner DRAGO [4]. It uses the DL-reasoner pellet [3] and the modular ontology formalism Distributed Description Logic (DDL) [5]. This reasoner scales reasonably well but the underlying formalism is designed for the integration of existing ontologies. Consequently global completeness is not a desired feature of DDL as it is for representing an existing monolithic ontology in a modular form. Nevertheless, this approach provides a good basis for distributed reasoning in a decompositional setting.

$\mathcal{E}$-connections [6] were designed for composition, they can be used with the centralized DL-reasoner PELLET [3]. Using $\mathcal{E}$-connections for decomposition is implemented based on conservative extensions [7, 8] in the ontology editor SWOOP [9], but frequently fails partitioning i.e. the result is very often similar to the original ontology. Using $\mathcal{E}$-connections with another partitioning algorithm would suffer from the required semantical disjointness of local domains and impossibility to model subsumption links.
Package based description logic [10] uses modularization to enable collaborative development of ontologies. This approach is application focussed and provides an editor. It is structured as hierarchical owl import, extended by visibility management for concepts. Drawbacks of this approach are its limitation of module relations to owl imports and restriction to the $\mathcal{ALC}$ subset of DL.
Furthermore, some development tools for large ontologies implement ideas of modularization (e.g. "microtheories" of the CYC ontology [11]) but only with hierarchical i.e. transitive reuse of complete modules and customized for one specific ontology.
Every approach to modular ontologies imposes restrictions on the connections between modules like using a single link property or not being able to represent concept subsumption. Depending on the application these restrictions are often to strong for information preserving modularization of existing ontologies.
A promising reasoning algorithm was proposed independent of research on description logics [12]. This method improves performance of common first order resolution by partitioning the clauses. Obviously, the approach can be adapted for ontologies, the performance on description logic will be evaluated. A modification of partition based first order reasoning for propositional logic in a peer-to-peer setting was implemented in the SOMEWHERE system. It scales well and provides anytime reasoning but the poor expressivity limits applicability for DL.

3 Expected Contribution

The contribution of the thesis is an infrastructure for modular ontologies that improves usability of large[1] OWL-DL ontologies, solving the problems mentioned above. In particular the work will provide

- Representation
- Distributed Reasoning
- Partitioning Algorithms

Representation of modular ontologies includes definition of the connections between modules and communication protocol. The ontology modules and the links between them will be formulated in OWL-DL.
Reusing a large ontology usually requires first extracting the symbols and axioms that are relevant for the reusing application. On a modularized ontology this expensive and sophisticated extraction process is substituted by the much simpler task of selecting relevant modules.

[1] An ontology is considered "large" if the number of concepts and properties or the complexity of the structure impedes utilization with state of the art editors or reasoners.

4 Methodology and Status

Problem Definition. The problem that has to be solved is enabling usage and maintenance of large ontologies which is hardly possible at the moment.
Identification of Criteria for the Solution. Prior to analysing and evaluating the existing approaches we defined and classified the criteria for a good solution of the problem i.e. the criteria for a good modular ontology infrastructure [13].
Analysis of other Approaches. Based on the criteria and their importance we are now analysing the existing approaches guided by a translation of the modular ontology formalisms to DL. Now the main weaknesses and difficulties of approaches to modularity are detected. The results are related to properties of ontologies because performance of approaches generally depends e.g. on the ontology language or depth of the hierarchy.
Goal Definition. Guided by the analysis of existing approaches and available tools we defined the part of the problem that will be addressed in the thesis. This goal is an infrastructure for modular ontologies that supports distributed reasoning and partitioning.
Evaluation Framework. In order to focus research it is helpful to set up the evaluation at a early stage so it can guide the work. Evaluating the results of this work on modularization consists of verifying the correctness of the implemented algorithms and comparing the performance to state of the art centralized and distributed reasoners.
Evaluate other Approaches. After the evaluation is set up existing approaches can be evaluated to further reveal virtues and drawbacks of design alternatives.
Design. The preceding analysis and clear definition of the goal shows on which formalism the work will be based and which tools are to be reused and extended. The different parts of the work are listed and related to conclude a plan, that is then carried out.
Evaluation. The last step is evaluation of correctness and performance of the implementation using ontologies of different size, language, expressivity and density.
Developing complex algorithms is usually an iterative process. Evaluation reveals weaknesses and bugs leading one step back to redesigning the algorithm accordingly. Sometimes it may even be necessary to readjust the goals if they turn out to be to hard or are fully achieved by other implementations.

Problem definition and identification of criteria where carried out within the past nine month since I started my PhD. Currently I am analysing other approaches and defining the goal of the thesis.

5 Approach

For the development of an infrastructure for modular ontologies the crucial decision is on the linking language to be used. To guarantee the right choice we analyse the existing approaches with respect to their expressivity, tractability and extendability. This analysis is based on a translation of the mapping languages DDL and $\mathcal{E}$-connections as well as the import approaches P-DL and Semantic Import to common description logic. The mapping approaches are already translated by their developers [14, 5], we are working on the translation of import approaches.
The most promising approach though still incomplete is Semantic Import [15], this

work will be continued in cooperation with the inventors to develop a formalism for modular ontologies. Representation in DL will presumably reveal that import and mapping are semantical equivalent i.e. they can be defined in terms of each other. On the other hand the differences in expressivity of the different approaches and the corresponding computational properties are much more clear when formulated in DL. Every approach imposes certain syntactic restrictions on an ontology that is to be converted to that type of modular ontology. Mainly these are restrictions on the ontology language or the mapping/import language. For example local domains must be definable such that there are no properties connecting elements of different local domains[2], or (in case of $\mathcal{E}$-connections) these properties cannot be transitive. Modular ontologies are not formulated in unrestricted DL because it is very hard (maybe impossible) to find a tractable reasoning algorithm for unrestricted DL in a distributed setting. Especially the combination of intersecting local domains with the above type of property raises difficulties. Thus developing a modular ontology formalism inevitably means trading completeness for tractability. There is not a single formalism that is the right trade-off for all situations but a set of different formalisms that reflect the relative importance of completeness vs. tractability. The existing approaches can be viewed as part of this set, but there are many other restrictions to evaluate on different reasoning tasks for a better trade-off and to fill the gap for less restricted modular DL.

The second part of the work is the development of a distributed reasoner that can handle large distributed ontologies. For unrestricted distributed DL we cannot expect to find an efficient reasoning algorithm, but in some applications for which time is not a sparse resource even an unefficient satisfyability test would be very helpful.

There are two starting points for developing a global complete distributed reasoner. One is the reasoner DRAGO [4], a distributed tableaux-reasoner based on the DL-reasoner PELLET [3]. This reasoner scales reasonably well and is currently the only approach to actually distributed reasoning. Centralized reasoner inevitably fail on large ontologies because they cannot even load them. The other starting point to distributed reasoning are results from distributed first order reasoning [12], the tractability of these methods on DL will be investigated.

Distributed reasoning trades decreased memory requirements for increased communication time and some optimization methods that are used in centralized reasoning cannot be applied to a distributed knowledge base. Hence critical for the performance of distributed reasoning is the partitioning of the knowledge base because it determines communication time and the amount of time a module spends waiting for information from other modules. Effects of the partitioning were not considered for Drago because it was designed for the composition of existing ontologies. Thus, optimizing the partitioning for reasoning with Drago may greatly reduce computation time. Furthermore the time problem will be addressed by implementing an anytime reasoning algorithm that considers increasing parts of the modular ontology. In the first step only the current module is checked for inconsistencies, the next step includes all direct predecessors.

To facilitate application of the developed distributed reasoner we aim at providing it as plug-in interface for the ontology editor Protégé [1].

[2] i.e. property assertions $p(x, y)$, $p(x, z)$ are not permitted if y and z are from different local domains.

The first step of the evaluation is verification of the implemented reasoning algorithms using ontologies that are small enough for existing DL-reasoners to provide reference for comparision of reasoning tasks. After verification the performance of developed reasoners will be compared to centralized reasoning and related to the applied restrictions. Experiments will be carried out using very large real-life ontologies like the DICE ontology and the FMA ontology. The time requirements will be compared to those of Drago and other distributed reasoners that are available in the meantime. Based on this evaluation it will be decided which of the implemented algorithms provide a reasonable trade-off between completeness and tractability for what type of ontology.

References

1. Noy, N., Sintek, M., Decker, S., Crubezy, M., Fergerson, R., Musen, M.: Creating semantic web contents with protege-2000. Intelligent Systems **16**(2) (2001) 60– 71
2. Wessel, M., Möller, R.: A high performance semantic web query answering engine. In Horrocks, I., Sattler, U., Wolter, F., eds.: Proc. International Workshop on Description Logics. (2005)
3. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. Jounal of Web Semantics (2006) to appear.
4. Serafini, L., Tamilin, A.: DRAGO: Distributed reasoning architecture for the semantic web. In: Proc. of the Second European Semantic Web Conference (ESWC'05). (2005)
5. Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. Journal of Data Semantics **1** (2003) 153–184
6. Grau, B.C., Parsia, B., Sirin, E.: Combining owl ontologies using e-connections. Journal Of Web Semantics **4**(1) (2005)
7. Lutz, C., Walther, D., Wolter, F.: Conservative extensions in expressive description logics. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence IJCAI-07, AAAI Press (2007)
8. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: A Logical Framework for Modularity of Ontologies. In: Twentieth International Joint Conference on Artificail Intelligence (IJCAI). (2007)
9. Grau, B., Parsia, B., Sirin, E., Kalyanpur, A.: Automatic Partitioning of OWL Ontologies Using E-Connections. In: Proc. of Description Logic Workshop (DL). (2005)
10. Bao, J., Caragea, D., Honavar, V.: Towards collaborative environments for ontology construction and sharing. In: International Symposium on Collaborative Technologies and Systems. (2006)
11. Curtis, J., Matthews, G., Baxter, D.: On the effective use of cyc in a question answering system. In: IJCAI Workshop on Knowledge and Reasoning for Answering Questions. (2005)
12. Amir, E., McIlraith, S.: Partition-based logical reasoning for first-order and propositional theories. Artificial Intelligence **162**(1-2) (2005) 49–88
13. Schlicht, A., Stuckenschmidt, H.: Towards Structural Criteria for Ontology Modularization. In: Proc. of the ISWC 2006 Workshop on Modular Ontologies. (2006)
14. Grau, B.C., Kutz, O.: Modular ontology languages revisited. In: IJCAI Workshop on Semantic Web for Collaborative Knowledge Acquisition. (2007) to appear.
15. Pan, J.Z., Serafini, L., Zhao, Y.: Semantic import: An approach for partial ontology reuse. In: Proc. of the ISWC 2006 Workshop on Modular Ontologies. (2006)

A Non-traditional Inference Paradigm for Learned Ontologies[*]

Vít Nováček

Digital Enterprise Research Institute
National University of Ireland, Galway
E-mail: vit.novacek@deri.org

1 Main Thesis Focus

The purpose of this document is to give an overview of author's prospective doctoral thesis in terms of goals, plans, adopted methodology and current achievements. The thesis' general focus is the Semantic Web, AI, automatic ontology acquisition and reasoning.

When considering ontologies as general knowledge repositories, ideally reflecting substantial amount of information present on the web, it is obvious that developing them purely manually is infeasible task not only due to the extensive size of data, but also due to the highly dynamic nature of the environment. Therefore the need for automated methods of ontology creation and maintenance is well acknowledged in the community. However, there has been no explicit support for automatically learned ontologies in the main branches of research concerning inference in the Semantic Web.

We believe that efforts leading to bridging these two rather disparate lines of research are more than worthwhile and will prove beneficial for both automated ontology development and reasoning, considering the noisy, context-dependent and inconsistent character of mainly unstructured web data we *have to* deal with when making the Semantic Web real. The nature of this knowledge is hard to be captured by traditional (logical) reasoning paradigms that usually require quite extensively (and expensively) specified descriptions in order to allow any usable reasoning. We plan to develop an alternative formal semantics of the Semantic Web data and implement respective reasoning tool prototype that would be able to deal with this situation better in the context of ontology learning. This is reflected in the tentative thesis' title *A Non-traditional Inference Paradigm for Learned Ontologies.*

2 Motivations, Addressed Tasks and Proposed Solutions

Within implementation of the thesis topic prototype, we adhere to these required features:

- the ability to refine the learned knowledge on the fly by incorporation of a specific reasoning paradigm and respective tools;
- a query-processing mechanism able to infer valuable and useful knowledge from learned ontologies by tools basing on the same reasoning paradigm;

[*] This work has been supported by Academy of Sciences of the Czech Republic, 'Information Society' national research program, the grant AV 1ET100300419 and by the EU IST 6th framework's Network of Excellence 'Knowledge Web' (FP6-507482).

- a query-transformation layer that would allow to interface the system with the Semantic Web standard tools and languages (for evaluation and inter-operation purposes);
- a knowledge-transformation layer that would allow to export the knowledge in the Semantic Web standards (again, for evaluation and inter-operation purposes).

In the following overview of the respective tasks and solution sketches, we base on our ANUIC (*Adaptive Net of Universally Interrelated Concepts*) framework for representation of learned fuzzy ontologies [14, 15].

2.1 Task SW-1 (*reasoning support for ontology acquisition*)

To the best of our knowledge, there has been little effort dedicated to the development of methods that could refine a learned ontology dynamically on the fly by means of specifically tailored reasoning procedures. If a basic foundational and precise ontology for the given domain has been developed, it can be used as a top-level "seed" model for our ANUIC framework. The assertions (with weights initially set to 1.0) in this general seed will help refining the more specific dynamic insertions within ontology learning process (e.g. by decreasing weights of learned assertions that are inconsistent according to the seed ontology). The documents processed by ontology learning can contribute to the refinement of the weights by themselves – if there are certain more trusted or domain-relevant documents, the weights of the assertions learned from them should be favoured.

This will be accompanied by a mechanism of propagation of the weight changes in the vicinity of the influenced nodes in the semantic network induced by the ontology. Note that there will be no restriction on the propagation – even the seed ontology can be eventually changed if the empirical character of the field is different. The application of inherent rules (the idea introduced in the next section) will play as significant role as the seed model in the direct inference support of the acquisition process.

Evaluation of this task is quite straightforward – we can compare the ontologies learned with the inference support with ontologies learned by the same methods without the inference. Appropriate evaluation measures can be adapted according to [9, 3]. One possible option is to identify the differences and present them to potential users of the ontology and/or to an evaluation committee, elicitating the reasonability and usability of extensions/retractions caused by the reasoning process when compared to the "purely learned" ontology.

2.2 Task SW-2 (*reasoning with learned ontologies*)

The ontology reasoning research in the Semantic Web has been focused mainly on the development of rigorous knowledge representation models and related formalised procedures of logical inference. However, the models in question (namely OWL [1] ontologies) require an indispensable amount of expert human intervention to be built and maintained. This makes the knowledge management based on this kind of explicit representation very expensive, especially in dynamic and data-intensive domains (e.g. medicine), or even infeasible, if the experts are not always available (e.g. semantic desktop).

The scalable ontology learning methods can overcome the problem of large domains. Moreover, automatic bottom-up knowledge acquisition prevents the possible bias in hand-crafted ontologies. The price we have to pay is that we must be able to deal with the less complex, noisy, possibly imprecise and very probably inconsistent knowledge then. Nonetheless, there could be implicit knowledge worth to infer even in the learned ontologies if there is a substantial amount of data in them. A possible way to an alternative approach to reasoning with learned ontologies rests with the development of a new kind of "loose", yet formal semantics. This semantics will support both refinement of ontology learning results (Section 2.1) and full-fledged reasoning with and querying of the learned ontologies themselves.

The semantics has been worked out in three levels that are jointly contributing to the process of formal interpretation of the learned content[1]:

1. **Declarative** semantics reflects direct meaning of learned knowledge *declared* in the ANUIC network of fuzzy modelling primitives. Interpretation of a node at this level is based on fuzzy intersection of sets induced by ranges of its properties (this interpretation is crucial for establishment of fuzzy analogical mappings, among other things). We further plan to design a natural extension of the ANUIC model by simple IF-THEN rules treated exactly in the same dynamic manner as the relations between ANUIC concepts.
2. **Procedural** semantics comprises the formal aspects of *procedures* of rule execution and analogy retrieval, mapping and transfer in the underlying model. We plan to incorporate the AI methods of heuristic reasoning [16, 10] into the engine based on the improved fuzzy ANUIC model. Very valuable concept in this respect is the notion of analogical reasoning [12] and its fuzzy extension [2]. The latter can be further developed in the scope of our work with different notions of fuzzy similarity [22, 11]. For the implemented inference engine, we have to provide a respective query-transformation layer in order to interface our system with other Semantic Web frameworks and standards.
3. **Interlocutive** semantics allows to further specify and/or refine meaning of stored knowledge in dynamic interaction with users (human or artificial agents – e.g. other ANUIC-based reasoners fed with different data in similar or otherwise relevant domains).

The evaluation of this task remains more or less open problem for now. However, besides measuring the computational efficiency of the inference, we could formalise a measure of "usefulness" of answers to certain types of queries and compare our system to the similar ones in an application-oriented assessment trial.

3 Current Achievements

At this time, an automated ontology acquisition platform *OLE* (*Ontology LEarning*) has been developed before and within the work on the thesis topic itself. *OLE* processes natural language English documents (in plain text, HTML, PDF or PostScript) and extracts an ontology from them. It makes use of NLP and machine learning techniques. An ANUIC (*Adaptive Net of Universally Interrelated Concepts*) model has been proposed and initially implemented for the fuzzy representation of learned ontologies in

[1] Only very brief description is given here, partially also due to space restrictions. The topic of the three-level formal semantics is currently under thorough development within a conference submission.

OLE. The progress of this work has been documented in several refereed papers[2] and presented by the author of this document at the respective events.

A technique of so called conceptual refinement improving the results of initial ontology extraction methods has been proposed and implemented for the task of taxonomy acquisition. Under a certain interpretation, it boosts the precision of taxonomy acquisition methods by more than 150%. The preliminary results of this work form the major recently published or accepted achievements [14, 15] and were presented by the author of this document at the ESWC 2006 conference (an ICEIS 2007 presentation to come in June, 2007). This initial proposal and implementation of the natural and intuitive mechanism coping with autonomous assignment of fuzzy relevance measures to the general learned relations (which has been considered as an open problem in this respect [19]) forms the most tangible and strongly related basic groundwork of the thesis, aimed at reasoning in the proposed ANUIC model. Current progress is continually documented at the project's webpage[3].

4 Related Work

There are methods refining the ontology after the learning process, using external reference and pruning [5]. However, there are generally no suitable external resources for many practical domains, therefore our tool is more universal in this respect. Some approaches try to connect ontology learning and reasoning by transforming the learned knowledge into a shape acceptable by the "traditional" inference mechanisms. The *Text2Onto* tool removes inconsistent knowledge from the learned ontologies [7] in order to allow usual precise OWL reasoning. The approach in [8] translates ontologies acquired by application of Formal Concept Analysis into FOL formulas, which is even more simplistic. These approaches leave vast amount of the sense of the learned knowledge unrecognised (e. g. possible different contexts induced by consistent subsets, structural properties of the knowledge, implicit relations between concepts, etc.).

In [17], a fuzzy relational model of ontology is introduced. However, it is only very simple and IR-oriented one, with no proper semantics generally applicable in other domains. [6] focuses on mining knowledge from databases and uses for example fuzzy rules to refine the resulting ontologies. But the authors' concrete approach to this topic is rather unclear and the formal semantics is lacking again. There is an indirectly related research in fuzzy OWL [20] and fuzzy DL reasoning [21]. However, these approaches still exploit the "traditional" logics based knowledge representation, which we find inappropriate for reasoning with learned ontologies. AI methods of heuristic [16, 10] or analogical [12, 18] reasoning present alternative paradigms that have, however, not been connected to a mechanism of automatic real-world knowledge acquisition. This is a practical disadvantage our approach aims to tackle (among other things).

5 Selected Application Domains

Following the **medicine** use cases specified in [13, 4], the implementation of our framework for ontology learning and reasoning could massively help in the processing of the

[2] See `http://www.muni.cz/people/4049/publications` for the full list of author's publications to date.

[3] See `http://nlp.fi.muni.cz/projects/ole` – the top Google^{TM} result of the "*ontology acquisition*" query on December 15, 2006; a web interface to the system libraries is present there as well.

dynamically changing medical knowledge. After initial definition of the seed model, ontologies learned by our tool from the natural language in medical records and even from the databases (after a preprocessing) can integrate the newly coming knowledge with the current facts on a single formal and technical basis. Moreover, the efficient and robust reasoning in our model can support the everyday decision process of medical experts in purely automatic way, utilising even data that have not been covered by formal medical manually developed ontologies.

The **semantic desktop** domain is related to new topics that have appeared recently within the major Semantic Web and AI research activities like *CALO* project[4] in USA and/or *NEPOMUK* project[5] in EU. The main aim of the projects is the development of an intelligent layer on the top of the current personal desktop systems. Possible application of our work in the scope of the semantic desktop research efforts is especially in the field of dynamic and automatic knowledge acquisition from the "raw" data. The model and reasoning paradigm we plan to develop could help in efficient semi-automatic discovery of implicit relations in the personal data and thus improve the process of their semantic re-organisation, meta-data annotation and querying.

6 Conclusion and Future Work

We have presented our current results and a vision of our doctoral thesis in the context of the Semantic Web and AI. Some of the missing links in the contemporary research have been identified. We have argued importance of the respective research questions and analysed the tasks that can fill in the gaps then. Possible solutions and evaluation methods have been roughly outlined. Examples of concrete application domains have been sketched, showing the practical relevance of the topic.

The work on the thesis was formally started in March, 2006. Supposed term of the thesis submission is the beginning of the year 2009. We plan to deliver the complete elaboration of the proposed ANUIC uncertain KR model and its semantics by the end of the year 2007, together with respective extension of the ontology learning framework. During the year 2008, we plan to devise and implement basic set of rule-based heuristic and analogical reasoning methods for the prototype and evaluate it, summing up the results in the thesis.

References

1. S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. *OWL Web Ontology Language Reference*, 2004. Available at (February 2006): http://www.w3.org/TR/owl-ref/.
2. B. Bouchon-Meunier and L. Valverde. A fuzzy approach to analogical reasoning. *Soft Computing*, 3:141–147, 1999.
3. C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data driven ontology evaluation. In *Proceedings of LREC 2004*, 2004.
4. Marco Eichelberg (edited by). Requirements analysis for the ride roadmap. Deliverable D2.1.1, RIDE, 2006.
5. A. Gangemi, R. Navigli, and P. Velardi. Corpus driven ontology learning: a method and its application to automated terminology translation. *IEEE Intelligent Systems*, pages 22–31, 2003.

[4] See http://caloproject.sri.com/.

[5] See http://nepomuk.semanticdesktop.org.

6. Paulo Gottgtroy, Nikola Kasabov, and Stephen MacDonell. Evolving ontologies for intelligent decision support. In Elie Sanchez, editor, *Fuzzy Logic and the Semantic Web*, Capturing Intelligence, chapter 21, pages 415–440. Elsevier, 2006.
7. Peter Haase and Johanna Völker. Ontology learning and reasoning - dealing with uncertainty and inconsistency. In Paulo C. G. da Costa, Kathryn B. Laskey, Kenneth J. Laskey, and Michael Pool, editors, *Proceedings of the Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*, pages 45–55, NOV 2005.
8. Hele-Mai Haav. An ontology learning and reasoning framework. In Yasushi Kiyoki, Jaak Henno, Hannu Jaakkola, and Hannu Kangassalo, editors, *Information Modelling and Knowledge Bases XVII*, volume 136 of *Frontiers in Artificial Intelligence and Applications*, pages 302–309. IOS Press, 2006.
9. J. Hartmann, P. Spyns, A. Giboin, D. Maynard, R. Cuel, M. C. Suarez-Figueroa, and Y. Sure. Methods for ontology evaluation (D1.2.3). Deliverable 123, Knowledge Web, 2005.
10. Jerry R. Hobbs and Andrew S. Gordon. Toward a large-scale formal theory of commonsense psychology for metacognition. In *Proceedings of AAAI Spring Symposium on Metacognition in Computation*, pages 49–54, Stanford, CA, 2005. ACM.
11. Zsolt Csaba Johanyák and Szilvester Kovács. Distance based similarity measures of fuzzy sets. In *Proceedings of SAMI 2005*, 2005.
12. Boicho Kokinov and Robert M. French. Computational models of analogy making. In L. Nadel, editor, *Encyclopedia of Conginitve Science*, volume 1, pages 113–118. Nature Publishing Group, London, 2003.
13. Lyndon Nixon and Malgorzata Mochol. Prototypical business use cases (D1.1.2). Deliverable 112, Knowledge Web, 2004.
14. V. Nováček and P. Smrž. Empirical merging of ontologies – a proposal of universal uncertainty representation framework. In *LNCS*, volume 4011, pages 65–79. Springer-Verlag Berlin Heidelberg, 2006.
15. Vít Nováček. Imprecise empirical ontology refinement. In *Proceedings of ICEIS 2007, vol. Artificial Intelligence and Decision Support Systems*. Kluwer Academic Publishing, 2007. In press.
16. Praveen K. Paritosh. The heuristic reasoning manifesto. In *Proceedings of the 20th International Workshop on Qualitative Reasoning*, 2006.
17. Rachel Pereira, Ivan Ricarte, and Fernando Gomide. Fuzzy relational ontological model in information search systems. In Elie Sanchez, editor, *Fuzzy Logic and the Semantic Web*, Capturing Intelligence, chapter 20, pages 395–412. Elsevier, 2006.
18. Christian D. Schunn and Kevin Dunbarr. Priming, analogy and awareness in complex reasoning. *Memory & Cognition*, 24(3):271–284, 1996.
19. Amit Sheth, Cartic Ramakrishnan, and Christopher Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *International Journal on Semantic Web & Information Systems*, 1(1):1–18, 2005.
20. G. Stoilos, G. Stamou, V. Tzouvaras, J.Z. Pan, and I. Horrocks. Fuzzy owl: Uncertainty and the semantic web. International Workshop of OWL: Experiences and Directions, Galway, 2005, 2005.
21. Umberto Straccia. A fuzzy description logic for the semantic web. In Elie Sanchez, editor, *Fuzzy Logic and the Semantic Web*, Capturing Intelligence, chapter 4, pages 73–90. Elsevier, 2006.
22. Wen-June Wang. New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85:305–309, 1997.

Imprecise SPARQL: Towards a Unified Framework for Similarity-Based Semantic Web Tasks

Christoph Kiefer

Department of Informatics, University of Zurich,
Binzmuehlestrasse 14, CH-8050 Zurich, Switzerland
kiefer@ifi.unizh.ch

Abstract. This proposal explores a unified framework to solve Semantic Web tasks that often require similarity measures, such as RDF retrieval, ontology alignment, and semantic service matchmaking. Our aim is to see how far it is possible to integrate user-defined similarity functions (UDSF) into SPARQL to achieve good results for these tasks. We present some research questions, summarize the experimental work conducted so far, and present our research plan that focuses on the various challenges of similarity querying within the Semantic Web.

1 Motivation

Semantic Web tasks such as ontology alignment, semantic service matchmaking, and similarity-based retrieval depend on some *notion of similarity* (at least if they are not solely based on logic). Therefore, researchers still try to find sound *user-defined similarity functions* (UDSF) to achieve good results for these tasks. Finding good similarity functions is, however, data- and context-dependent, and needs to be reconsidered every time new data is inspected. Nonetheless, good UDSFs are crucial for the success of the above-mentioned Semantic Web tasks.

Furthermore, in recent years, query languages for the Semantic Web such as RDQL and SPARQL have gained increasing popularity. The current W3C candidate recommendation of SPARQL, however, does not support UDSF to analyze the data during query processing. The goal of this project is to overcome this limitation and to develop a *unified framework* based on SPARQL to solve similarity-dependent Semantic Web tasks. The proposed iSPARQL framework should be easy to use and easily extendable to allow for user-defined, task-specific similarity functions. The "i" stands for *imprecise* indicating that two or more resources are compared by using similarity measures.

We strive for a robust implementation of similarity querying for the Semantic Web and its integration into SPARQL. The proposed iSPARQL approach should have a high degree of flexibility in terms of customization to the actual Semantic Web task.

2 Related Work

RDF Retrieval. Siberski *et al.* [19] propose SPARQL extensions allowing the user to query the Semantic Web with preferences. New keywords (`PREFERRING`, `CASCADE`) are added to the official SPARQL grammar in order to favor query answers which match user-defined preference criteria. Finally, the answers which are not dominated by any other answers (optimal according to the defined preference dimensions) are returned to the user.

Ontology Alignment. The task of ontology alignment (aka *ontology mapping/matching*) is a heavily researched field within the Semantic Web. Noy and Musen [15] present the PROMPT framework – a suite of tools including iPROMPT and ANCHORPROMPT – simplifying the comparing, aligning, and merging of ontologies of different origins. Furthermore, Doan *et al.* [5] propose the GLUE system that assists the user in finding mappings between ontologies using techniques from machine learning. A different methodology is proposed by Ehrig and Staab in [6]: based on QOM, ontologies can be aligned on different layers focusing on different (modeling) aspects of ontologies. Euzenat and Valtchev [7] propose an approach that is based on a specialized similarity measure to compare OWL-lite ontologies. Last, in a more recent paper, Tous and Delgado [20] map nodes of ontologies to matrices which capture the relationships of the mapped nodes among each other. Finally, a graph matching algorithm is applied to find mappings between the ontologies under comparison.

Matchmaking/Discovery. Klusch *et al.* [14] propose OWLS-MX to perform service matchmaking which adopts both, pure logic-based and Information Retrieval (IR) based techniques for the needs of hybrid semantic service matchmaking. Furthermore, Hau *et al.* [10] propose a similarity measure to compare Semantic Web services expressed in the OWL-S language. In addition, Jaeger *et al.* [11] present an approach for matching service inputs, service outputs, a service category, and user-defined service matching criteria. The four individual matching scores are aggregated to result in an overall matchmaking score.

Query Optimization. Query optimization strategies have been developed to reduce the complexity of Semantic Web queries to boost their runtime performance. Ruckhaus *et al.* [16] propose to estimate the cost and cardinality of individual query predicates based on selectivity estimations taken from [18].

Similarity Joins (Data Integration). To perform data integration, Cohen [4] presents WHIRL and the notion of *similarity joins* by which data is joined on *similarity* rather than on *equality*. In WHIRL, the TF-IDF weighting schema from IR [1] is applied together with the cosine similarity measure to determine the affinity of text. Similar approaches are proposed by Gravano *et al.* employing *string joins* [8] and *text joins* [9] in order to correlate information from different databases and web sources respectively.

3 General Problem Areas/Gaps

Numerous Semantic Web tasks rely on some *notion of similarity*, either to *compare ontologies* (for alignment and/or integration), or to *compare services* (for matchmaking and/or discovery), or to *compare resources* (for querying and similarity-based retrieval) among others. All of the approaches presented in Section 2 tackle one of these tasks *individually* (*i.e.*, in their own specific way). None of the approaches present a unified framework to solve them all. We made the following observations:

- To solve these tasks, Semantic Web researchers still try to find sound *user-defined similarity functions* (UDSF), which are crucial for the *success* of these tasks. However, good similarity functions are data- and context-dependent, and generally not easy to find.
- SPARQL in *combination* with UDSFs could be used to solve individual tasks. However, traditional SPARQL does not support querying ontologies with UDSF. It is not clear what the optimal solution would look like: an extension of the official SPARQL grammar or the exploitation of "magic properties" (aka *virtual triples* or *property functions*) as supported in ARQ.[1]
- The semantics and complexity of UDSF-extended SPARQL queries are unclear. Hence, they should be elaborated and formally studied.
- UDSF statements add an additional layer of *complexity* to SPARQL queries. Therefore, an approach for optimizing queries containing UDSFs should be provided. This is particularly important when executing *web-scale queries*. In other words: do UDSF-queries have the potential to scale to the web?

4 Research Plan

4.1 Choice of Datasets and Evaluation Strategy

So far we have experimented with the two matchmaking/retrieval test collections OWLS-TC[2] and the OWL MIT Process Handbook[3]. For our preliminary optimization experiments we used SwetoDblp[4], which focuses on bibliography information of computer science publications. Furthermore, we worked with EvoOnt[5] – a set of ontologies to model the domain of object-oriented software source code. We will use these datasets for the evaluations of our proposed unified framework.

4.2 Current State of Our Research

RDF Retrieval. iRDQL [2] is our extension of traditional RDQL with similarity joins to determine the similarity of Semantic Web resources.[6] A limitation of iRDQL is that it allows to utilize only one similarity measure per query and

[1] `http://jena.sourceforge.net/ARQ/`

[2] `http://projects.semwebcentral.org/projects/owls-tc/`

[3] `http://www.ifi.unizh.ch/ddis/mitph.html`

[4] `http://lsdis.cs.uga.edu/projects/semdis/swetodblp/`

[5] `http://www.ifi.unizh.ch/ddis/evoont.html`

[6] All similarity measures are implemented in SimPack, our generic library of similarity measures for the use in ontologies (`http://www.ifi.unizh.ch/ddis/simpack.html`).

it does not perform any query optimization. A demonstration of our current prototype implementation iSPARQL is available at `http://www.ifi.unizh.ch/ddis/isparql.html`. We will use this prototype as a starting point (and benchmark) for the new framework to be accomplished within this PhD thesis.
Matchmaking/Discovery. In [12], the applicability of our iSPARQL prototype is evaluated for the task of Semantic Web service discovery within the OWL MIT Process Handbook.
Query Optimization. Our first steps toward Semantic Web query optimization are presented in [3]. The proposed OptARQ approach investigates SPARQL query optimization by means of a rule-based query optimization engine. Optimization techniques for UDSF-queries, however, are not covered by OptARQ.
Analyzing Software Repositories. To highlight the benefits and applicability of the proposed unified framework to different, initially *non-Semantic Web tasks*, we realized the Coogle [17] and EvoOnt [13] projects for the tasks of software evolution analysis and visualization as well as design flaws detection.

4.3 Our Approach – Next Steps

The aim of this work is the design, specification, implementation, and evaluation of a unified framework for similarity-based Semantic Web tasks. There are several goals to achieve: the first goal consists of a detailed revision of our preliminary work. This will answer the question if the virtual triple approach taken so far is sufficient to solve the remaining challenges of such a unified framework. The second goal is the formal elaboration of the iSPARQL grammar, its semantics and complexity. As a third goal, we investigate query optimization techniques to boost the performance of UDSF-queries. Finally, the whole iSPARQL model and implementation will be evaluated for applicability to different application tasks (see Section 2).

To achieve the goals, the following steps are planned: a revision of the current prototype with special attention to its usability, flexibility, customizability, and scalability; the specification of the iSPARQL model, particularly the complexity and semantics of UDSF-queries; the implementation of the unified framework; the investigation of UDSF-query optimization techniques; and an evaluation of the applicability to different similarity-based Semantic Web tasks in terms of testing, usability, customization, and performance measurement.

5 Conclusions

In this proposal we outlined the need of a unified framework to solve similarity-based Semantic Web tasks, such as ontology alignment, service matchmaking, and RDF retrieval. Our approach extends traditional SPARQL with user-defined similarity functions (UDSF). The semantics and complexity of SPARQL-based similarity queries will be formally elaborated and query optimization techniques proposed. This systematical assessment will answer the questions of what is the *range of tasks* that can be solved with the iSPARQL system, what is the *performance* to solve these tasks, and what is its potential to *scale to the web*. It is important to realize that these tasks provide a kind of "stress test" for the usefulness of our unified framework.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
2. A. Bernstein and C. Kiefer. Imprecise RDQL: Towards Generic Retrieval in Ontologies Using Similarity Joins. In *SAC 2006*, pages 1684–1689.
3. A. Bernstein, C. Kiefer, and M. Stocker. OptARQ: A SPARQL Optimization Approach based on Triple Pattern Selectivity Estimation. Technical Report ifi-2007.03, Department of Informatics, University of Zurich, 2007.
4. W. W. Cohen. Data Integration Using Similarity Joins and a Word-Based Information Representation Language. *TOIS*, 18(3):288–321, 2000.
5. A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the Semantic Web. *VLDB Journal*, 12(4):303–319, 2003.
6. M. Ehrig and S. Staab. QOM - Quick Ontology Mapping. In *GI Jahrestagung*, pages 356–361, 2004.
7. J. Euzénat, D. Loup, M. Touzani, and P. Valtchev. Ontology Alignment with OLA. In *ISWC 2004*, pages 333–337.
8. L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srivastava. Approximate string joins in a database (almost) for free. In *VLDB Journal*, pages 491–500, 2001.
9. L. Gravano, P. G. Ipeirotis, N. Koudas, and D. Srivastava. Text Joins in an RDBMS for Web Data Integration. In *WWW 2003*, pages 90–101.
10. J. Hau, W. Lee, and J. Darlington. A Semantic Similarity Measure for Semantic Web Services. In *WWW 2005*.
11. M. C. Jaeger, G. Rojec-Goldmann, G. Mühl, C. Liebetruth, and K. Geihs. Ranked Matching for Service Descriptions using OWL-S. In *KiVS 2005*, Informatik Aktuell, pages 91–102.
12. C. Kiefer, A. Bernstein, H. J. Lee, M. Klein, and M. Stocker. Semantic Process Retrieval with iSPARQL. In *ESWC 2007*.
13. C. Kiefer, A. Bernstein, and J. Tappolet. Mining Software Repositories with iSPARQL and a Software Evolution Ontology. In *MSR 2007*.
14. M. Klusch, B. Fries, and K. Sycara. Automated Semantic Web Service Discovery with OWLS-MX. In *AAMAS 2006*, pages 915–922.
15. N. F. Noy and M. A. Musen. The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping. *IJHCS*, 59(6):983–1024, 2003.
16. E. Ruckhaus, E. Ruiz, and M.-E. Vidal. Query Optimization in the Semantic Web. In *ALPSWS 2006*.
17. T. Sager, A. Bernstein, M. Pinzger, and C. Kiefer. Detecting Similar Java Classes Using Tree Algorithms. In *MSR 2006*.
18. P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access path selection in a relational database management system. In *SIGMOD 1979*, pages 23–34.
19. W. Siberski, J. Z. Pan, and U. Thaden. Querying the Semantic Web with Preferences. In *ISWC 2006*.
20. R. Tous and J. Delgado. A Vector Space Model for Semantic Similarity Calculation and OWL Ontology Alignment. In *DEXA 2006*.

Semiautomatic Creation of Semantic Networks

Lars Bröcker

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
Lars.Broecker@iais.fraunhofer.de

1 Introduction

The vision of the Semantic Web ist one of extending the World Wide Web of today to one "[..] in which information is given well-defined meaning, better enabling computers and people to work in cooperation." (Tim Berners-Lee in an article for the Scientific American in 2001). This promises an exciting future for the WWW.

The advantages for users and machines alike are eminent, many of the building stones like RDF or OWL are in place already. But why has the Semantic Web not been adopted by more content creators, more web sites? The main technological reason for this lies in the complexity associated with the creation of ontologies. Ontologies are, following a definition of T. R. Gruber, a formal, explicit specification of a shared conceptualization of a given domain[1]. As such, they are an essential part of every semantic web application, since they define the language used to express the view on the world. But their creation is a time-consuming and expensive endeavor that is beyond many organizations or communities. Most therefore stay away from the Semantic Web altogether. This severely handicaps the efforts of bringing about the vision of the Semantic Web, by preventing the attainment of a critical mass of content available using it.

1.1 Research Problem

What is needed is a means to generate a meaningful description of the semantics of content collections in such a way that it necessitates as little manual interaction as possible. The results may not be as distinguished as a manually created ontology, but they at least provide a way to utilize the benefits of the semantic web. Two main problems need to be tackled: first, the extraction of the semantic network inherent in the collection, and second, the design of a surrounding system being both versatile and easy to expand to accommodate new features, data stores, or services.

The first problem is one of automating ontology engineering. The goal here is to extract the main entities and their relations from the corpus in order to gain an understanding of the topics the corpus contains. This boils down to three tasks: entity recognition, relation discovery, and creation of the semantic net from the results of the first two tasks. While there are good tools available for entity recognition, relation discovery as of now has to do without. Scientific approaches in this area typically consider binary relationships, higher order relations get almost no coverage. The task of network-creation is a translation step from the entities and their relations into a language of the Semantic Web framework.

The second problem addresses use-case necessities. Many interesting collections are not static, but are subject to many changes (e.g. wiki-webs). In order to accommodate this, a semantic representation needs to be able to continuously monitor the corpus and adapt itself accordingly. Other requirements may result in the necessity for integration of additional services into the system.

1.2 Contribution

This thesis concentrates on the task of relation discovery in order to generate meaningful connections for the network, since there already are numerous good tools for Named Entity Recognition (e.g. GATE[2]) available. Accordingly, the first contribution is an algorithm that gathers n-ary relations ($n \geq 2$) in a text corpus between entities from a set of previously agreed upon concept classes. The second contribution is an architecture containing the algorithm, as well as facilities for the monitoring of dynamic collections, paired with adaptation of the network where necessary.

1.3 Use Cases

The envisioned system provides a semantic representation of the content of a document repository without changing its data, i.e. it provides a semantic wrapper around the collection. The wrapper supplies a semantic view on the topics of the collection that can be used for further processing, data exchange, or provision of sophisticated search interfaces.

The first application of the approach is part of an ongoing research project financed by the German Ministry of Research and Education (BMBF) called WIKINGER[3], where it is used to bootstrap and subsequently monitor a wiki-web for the domain of Contemporary History.

In a similar manner, media providers like broadcasters, newspapers, or news agencies could use this approach to better organize and tap the contents of their digital archives.

2 Approach

For the sake of brevity, only the approach concerning the creation of the semantic network will be described in detail. It is a process divided into five separate steps. In the first step, a set of core concept classes is defined, followed by the annotation of examples of these classes. They are used to train a Named Entity Recognition tool. Next, the corpus is segmented into sentences. Those containing less than two entities are discarded. The remaining sentences serve as input for an algorithm computing association rules on the entity classes. The association rules express the degree of association between classes using two measures: the confidence that there is an association, and its coverage of the collection. This allows different ranking approaches depending on the strategy to be followed.

Given an ordering of the rules, the next step iteratively analyses the set of sentences belonging to a given rule. Since one rule describes an unknown amount of different relations between its constituents, the task is to find a clustering of the set such that each cluster describes one single relation. Since the amount of relations is not known beforehand, hierarchical clustering has to be employed.

The next step provides labels for the relation clusters. They are presented to the domain experts for review, who can change or remove labels, entities, or relation clusters.

The final step collects all entities and relations and creates the semantic web from them. While the entity translation is straightforward, special care has to be taken in expressing the relations between them, since not all relations will be binary. Preservation of the n-ary relations requires the introduction of proxy entities into the net, in order to conform to the triple schema of RDF.

2.1 Results so far

System architecture using a service-oriented architecture.

Internal data representations allows inclusion of external data sources given a suitable transformer.

Versioned repositories for the internal data, allow change montoring, detection, and adaptation.

2.2 Results still to be achieved

Clustering and labeling different distance measures and vector representations are evaluated.

Translation into RDFS algorithm needs to be designed and implemented

Change Management the service responsible needs to be implemented.

2.3 Evaluation

Evaluation of the approach will be performed in the project WIKINGER. Domain experts will be on site to handcraft relation clusters. These will serve as ground truth for the automatically proposed relation clusters. Quality in a dynamic environment will be evaluated via periodical surveys when the system goes live in August of this year. In parallel, a similar setup for the domain of newspaper archives will be tested with the help of archive personnel from a newspaper company.

3 State of the Art

The approach presented of the thesis touches two areas of research: ontology learning and relation finding. This section highlights the approaches most relevant for this work.

3.1 Ontology Learning

Alexander Maedche from the AIFB in Karlsruhe describes a system called *Text-To-Onto*[4] that is used to aid ontology engineers in their work. Its objective is to find new concepts for the target ontology from domain taxonomies providing is-a relations, and hyponym relations gathered from texts using text mining methods. The candidate concepts are added manually to the ontology. An additional module deals with the discovery of non-taxonomic relations. It deducts possible relations from association rules. The module stops at this step and only considers concept-pairs.

Philipp Cimiano and Johanna Völker, also from the AIFB, present with *Text-2-Onto* an advanced system for the task of ontology learning from text. It holds the ontology in a so-called probabilistic ontology model (POM) that contains modelling primitives along with a confidence measure stating the probability of them being part of the ontology. A GUI allows manual changes to the ontology after the learned phase. The system reacts on changes in the corpus by only recalculating the parts of the ontology that are affected by the changes. Named entity recognition using GATE is performed on the collection, but only hyponym relations (kind-of) are extracted automatically from the texts.

3.2 Relation Learning

Takaaki Hasegawa et al. describe an algorithm for the discovery of relations in natural language texts[6], using named entity recognition with a small set of concept classes. This is followed by a per-sentence analysis of the corpus. All sentences containing two instances having a maximum distance of five words are considered for further processing. Finally, a cluster-analysis is performed on every class of pairs, resulting in clusters containing the different types of relation between pairs. Evaluation is done using a years worth of newspaper articles, and matching automatic performance against hand-picked relations. The best results (34 of 38 existing relations found) attain an F-measure of 60%.

Aron Culotta and Jeffrey Sorensen present an approach to relation extraction from texts using kernel methods [7]. The task is to extract previously learned binary relations from the corpus. This is achieved by first performing shallow parsing of a sentence and then using a kernel method on the smallest dependency tree containing both entities. This reduces the amount of words considered in the calculation of the kernel, thus reducing the amount of noise in the result. They reach 70% precision with 26% recall. Bunescu et al.[8] propose a variation of this approach: their kernels consider only the words on the shortest path between the two entities. Their evaluation is performed on the same data where they reach 71% precision with 39% recall.

3.3 Discussion

Text-To-Onto was developed as a tool for knowledge engineers, who are supposed to do the real modelling, and it shows. All additions to the ontology are performed manually, and while it contains a module for relation learning using association rules, it refrains from discovering the actual relations. Text-2-Onto uses an interesting storage model for the ontology, but is restricted to hyponym relations, thereby falling behind its predecessor with regard to relation discovery. The system described in this paper goes a step beyond these systems in two ways: it does not depend on the availability of ontology engineers, and it aims to discover all relevant relations contained in the text.

Hasegawa et al. use a clustering approach to find hitherto unknown relations but restrict themselves to pairs of entities, thus tearing apart relations of higher order that might have been present in the data. Their algorithm does not include a means to rank the pairs of prior to the clustering. The approaches by Culotta and Bunescu offer interesting possibilities for subsequent classification of relations, but cannot be used to discover them in the first place.

4 Conclusion

This paper summarizes the main topics of my PhD thesis. The approach promises to be a feasible way to bring the benefits of the Semantic Web to a larger audience, especially in those domains where creation of a specialized ontology is not feasible in the foreseeable future. The architecture has been designed such that it lends itself well for expansion in different ways. Inclusion of video or audio transcripts is an interesting option, since more and more such content finds its way onto the web. The inclusion of an easy interface allowing for the definition of new relations is another interesting expansion of the system, perhaps by graphical means using SVG or by an extended wiki-syntax as found in semantic wiki systems.

References

1. Gruber, T.R.: A translation approach to portable ontology specifications. In *Knowledge Acquisition*(5), 1993, pp. 199–220
2. Cunningham, H.:GATE, a General Architecture for Text Engineering. In *Computers and the Humanities*, vol 36, 2002, pp223 – 254
3. Bröcker, L.: WIKINGER – Semantically enhanced Knowledge Repositories for Scientific Communities. In: *ERCIM-News*, vol. 66, 2006, pp. 50–51.
4. Maedche, A.: The Text-To-Onto Environment. Chapter 7 in: *Maedche, A.: Ontology Learning for the Semantic Web*. Kluwer Academic Publishers, 2002.
5. Cimiano, P., Völker, J.: Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In *Proceedings of NLDB*, 2005.
6. Hasegawa, T., Sekine, S., Grishman, R.: Discovering relations among named entities from large corpora. In: *Proceedings of the 42nd Conf. of the ACL*, 2004. pp. 15–42
7. Culotta, A., Sorensen, J.: Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42nd Conf. of the ACL*, 2004. pp. 423–429.
8. Bunescu, R.C., Mooney, R.J.: A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of EMNLP 2005*, pp 724–731

Towards Cross-Media Document Annotation

Ajay Chakravarthy

Department of Computer Science, Regent Court, Portobello Street,
Sheffield S1 4DP
A.Chakravarthy@dcs.shef.ac.uk

Collecting and aggregating multimedia knowledge is of fundamental importance for every organisation in order to gain competitiveness and to reduce costs. It is possible that knowledge contained in just one medium E.G. text documents, does not carry the full evidence looked for. Therefore connecting information stored in more than one medium is often required. It is clear that current knowledge management technologies and practises cannot cope with such situations, as they mainly provide simple mechanisms (E.G. keyword searching). Currently knowledge workers manually pierce together the information from different sources. In this report we focus and envisage research methodologies that will enable the semantic enrichment of multimedia documents, both on multiple media and across media through annotation.

Annotation of a document is a complex and labour intensive task. So far, research has focused [1][2][4] on supporting the annotation of single media. Much less attention has been paid to the issue of annotating material across media. For this reason there is a growing interest in developing methodologies able to capture the content and the context of multimedia documents, in order to enable effective searching (and document-based knowledge management in general). Previous research in personal image management [6] and text annotation [3] demonstrated how annotating images or documents could be a way to organise information and transform it into knowledge that can be used easily later. Metadata enables the creation of a knowledge base which can then be queried as a way both to retrieve documents (via content and context) and to query the structured data (E.G. creating charts illustrating trends). We address many of these problems with AKTive Media[1] which is a system implemented during the PhD. AKTive Media is a user centric ontology based cross-media annotation system. The goal is to automate the process of annotation by means of knowledge sharing and reuse, thereby reducing user effort during the annotation process. The system actively queries web services and central annotational triple stores as a background service to look for context specific knowledge. The aim is to provide a seamless interface that guides the user through the process, reducing the complexity of the task. Language technologies and a web service architecture are adopted to provide a context specific annotation mechanism that uses suggestions inferred from both the ontology and from the previously stored annotations to help the user: the ontology is pre filtered to present only the top-level concepts (the most generic ones); The produced knowledge is then used as a way to establish connections with and to

[1] http://www.dcs.shef.ac.uk/~ajay/html/cresearch.html

navigate the information space: Example when the user annotates a part of an image of a car engine as “abrasion-damage” on a “crank-shaft” the system uses those annotations to retrieve other related images and documents. New relationships can then be established with the found knowledge, E.G. the damage can be related to other previous cases, and through free-text comments the relationship may be made explicit (E.G. this type of failure happens constantly on this blade in hot conditions, and this is proved by document x). AKTive Media has information extraction (IE) plug-ins built in (T-Rex)[5] which automate the annotation of textual documents. The main difference between our approach when compared to other state of the art annotation approaches [4][6] is that we use knowledge across media for annotation and also to further relate these annotation instances. This helps in greatly reducing user effort during manual annotation of documents, by providing intelligent suggestions derived from across media, the IE engine and from the central annotation server. The other major difference is that in AKTive Media, an effort is made to bridge the semantic gap between low level image features and semantically annotated metadata provided by users during annotation. We achieve this by providing means to index image collections and enable the user to query over the index using the visual content of the source image that is being annotated, the user can then use free hand mark-up over regions of the images to perform semi-automatic image segmentation and map high level ontology concepts to these segmented regions.

This research methodology has been deployed in various research projects including AKT, Memories for Life, X-Media and we have scheduled a detailed user evaluation in Rolls Royce UK at the end of year, for the annotation of strip reports.

Referencess

1. Ciravegna F., Dingli A., Petrelli D. and Wilks Y.: User-System Cooperation in Document Annotation based on Information Extraction. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), 1-4 October 2002 - Siguenza (Spain)
2. Dzbor M., Domingue J., Motta E. Towards a semantic web browser. Knowledge Media Institute, The Open University, Milton Keynes. UK. 2002
3. Handschuh S., Staab S., and Ciravegna F.. S-CREAM- Semi-automatic CREAtion of Metadata. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), 1-4 October 2002 - Siguenza (Spain), Lecture Notes in Artificial Intelligence 2473, Springer Verlag
4. Hendler J., Bijan P., Grovel M., Schain A., Golbek J., Halaschek-Wiener C., PhotoStuff – An Image Annotation Tool for the Sematnci Web, 2003. University of Maryland, MIND Lab, 8400 Baltimore Ave., College Park, MD 20742, USA 2NASA Headquarters, Washington, DC 20546, USA.
5. Iria, J. T-Rex: A Flexible Relation Extraction Framework. In Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK'05), Manchester, January 2005.
6. Kuchinsky A., Pering C., Creech M.L., Freeze D., Serra B., Gwizdka J., FotoFile: A Consumer Multimedia Organisation and Retrieval System, Proceedings of ACM CHI99 Conference on Human Factors in Computing Systems, 496-503, 1999.

Semantic Business Process Modeling [⋆]

Zhixian Yan[⋆⋆]

Digital Enterprise Research Institute (DERI) Innsbruck,
Innsbruck University, Austria
{zhixian.yan}@deri.com

Web services and BPM have become a combination research aiming at enterprize computing providing a more intelligent and interactive services as process-aware systems. In particular, with the emergence of BPEL4WS as a de-facto industrial standard for process grounding technology, Web services become the building blocks for the final process execution. However, really combining those two topics is still very hard as different perspectives (business and technical). With the development of semantics, especially semantic Web services, researchers propose SBPM to bridge the gap between business and technical levels. To achieve this, we need a comprehensive process modeling approach. Traditional process modeling has a long research history. From industrial perspective, the focuses of process modeling are pervasively on providing graphic-based modeling tools (Workflow or BPM suites) with various process notations, such as UML, BPMN and EPCs. Besides graphical modeling, language-based process description is another main emphasis, such as BPML, BPEL, XPDL etc. From academic perspective, there are also many formal concurrency theories supporting process automation and validation, such as Petri Net, Abstract State Machine, Process Algebra like Pi-Calculus, and some logic based AI models like Temporal Logic and Transaction Logic. However, neither industrial tools nor theoretical methods can completely support smooth combination between Web service and BPM. Therefore, the main motivation of this PhD research is to provide a semantic modeling framework for business processes, named Business Process Management Ontology (BPMO), which acts as the cornerstone of SBPM and the key transition role between business level and technical level.

Problem Statement Based on the vision of SBPM and the fundamental cornerstone about semantic process modeling, we provide following key issues as the problem statement ought to be involved and given appropriate solutions in this PhD research: (1)*Process Modeling Requirements* Modeling requirements (or called process description requirements) is the basis for the whole BPMO proposal, which needs to be determined first. Basically, we should answer "*what kind of concepts are involved?*" and "*what is the crucial functional requirements and nun-functional requirements need to be described for business processes?*". (2)*Process Modeling Architecture (Elements and Language)* Based on the previous determined requirements, we need a fully-fledged modeling architecture with comprehensive elements to cover all the requirements. A certain descrip-

[⋆] This work is funded by FFG SemBiz project, and special thanks to Emilia Cimpian, Michal Zaremba, Manuel Mazzara, Ying Ding

[⋆⋆] First year of PhD research

tion language is needed to describe and store all the elements involved in process modeling. (3)*Formal Process Modeling Approach* The distinguished advantage of semantics is machine-processable and further to support (semi-) automation. To completely achieve automation potential of semantics, traditional process formal works like ASM and Petri Net may be useful and can be refined with more semantics support. (4)*Legacy Process Integration* Integration with legacy system is crucial in real-world applications. Therefore, semantic process modeling ought to provide interface to integrate traditional processes modeled by non-semantic notations like BPMN, UML. (5)*Graphic Process Modeling Suite* The BPMO framework needs the grounding model suite, providing friendly graphic interfaces for both technical experts and businessmen. The distinguished modeling suite can really embody the semantics transition role between business level and technical level.

Proposed Approaches As the BPMN vision involves both business and technical levels, BPMO is a broad and cross-discipline topic. Basically, the semantic technology, esp. the semantic related description methodology is the main applied approach for this PhD thesis. However, there are four main general approaches can be referred to: (1)*Requirement Engineering.* To determine the process description requirements as the primary step for the BPMO framework research, some arbitrary requirements engineering techniques can be applied, such as determining system boundaries, stakeholders, goals etc. by analyzing real-world business use cases. (2) *Semantic Web Service.* The objective of applying semantics in Web services is to enable automatic service discovery, composition, invocation, interoperation etc. Business process has similar context and requirements. Among so many semantic web services activities, we mainly refer to the WSMF framework, especially its conceptual model WSMO. (3)*Formal Process Model.* We realize the importance of formal model to help process validation and automatic discovery/composition. We have briefly surveyed many existing formal process models, such as Petri Net (modeling workflow patterns), ASM, Pi-Calculus, and Cuncurrency Transaction Logic. It's not so easy to make an absolute choice among those formal methods. But so far, ASM and PetriNet are on the top list for its sound semantics and graphic process modeling support. 4) *Process Grounding Technology.* Although this PhD research is mainly focuses on the modeling context, some grounding technologies will also be considered especially the emerging de-facto standard BPEL.

Excepted Contribution This PhD work aims at investigating issues and making following contributions: (1) Specifying semantic description requirements for business processes, involving the whole BPM lifecycle. (2) Providing a fully-fledged semantic business process modeling framework BPMO, which provides the cornerstone for the SBPM vision and makes it feasible. (3) Based on semantic foundation, together with some formal process models, BPMO can enable (semi-)automatic process discovery/composition/invocation. (4) Besides the above scientific contributions, technically, this work can provide the integration with existing process systems, based on traditional notations such as BPMN, EPCs, and also grounding technology like BPEL.

Towards Open Ontology Engineering

Katharina Siorpaes
Digital Enterprise Research Institute (DERI), University of Innsbruck, Austria

1 Motivation

Even though ontologies are widely regarded as the backbone of the Semantic Web and the research in the area is very active, only few well-maintained and useful ontologies can be found when searching the Web. The reasons for this phenomenon are discussed in [1] who identifies four bottlenecks: first, many relevant domains of discourse are subject to a high degree of conceptual dynamics. Second, using and building ontologies is not reasonable if the cost of building an ontology is higher than its benefit. Third, a prerequisite for using an ontology and thus committing to its view of the world is to exactly understand its exact ontological commitment and the meaning of concepts and relations. This is hampered by the fact that most ontologies are built by a group of engineers and the user community does not have control over the evolution of the ontology due to the lack of efficient tool support for a broad audience with only limited ontology engineering skills. Fourth, existing standards specifications and all kinds of controlled vocabularies, which ontologies could re-use, are subject to intellectual property rights.

A community-oriented approach has several advantages towards an isolated, engineering-oriented approach: A community can keep up with the pace of conceptual dynamics in a domain more easily and it is cheaper for a community to collaboratively work on a specification of an ontology than for a group of ontology engineers as the workload is spread amongst the members. Finally, a community-agreed specification of a conceptualization will more likely be used and further developed. The idea of wikis is to allow a wide range of users to contribute to the content of the Web without requiring more than basic Web editing skills. The enormous success of the online encyclopedia Wikipedia[1] has proven the efficiency of wiki infrastructure. In my thesis, I take the following approach to collaborative and open ontology building tackling the problem of ontology maintenance in dynamic domains: I propose (1) the design of a lightweight user interface aligned with the wiki philosophy, (2) the re-use of data produced by social software, such as folksonomies, as well as other Web resources in domain ontologies, and (3) functionality that supports the community in achieving consensus.

2 Related Work

The related work can be divided into the following areas: **Collaborative ontology engineering:** [3] describes Tadzebao and WebOnto. [4] describe the DILIGENT knowledge process where ontology evolution and collaborative concept mapping are applied to deal with conceptual dynamics of domains. The ontology editor Protégé[2] is also available in a Web version [5]. **Semantic Wikis:** [6] describe Makna, a Wiki engine

[1] http://wikipedia.org/
[2] http://protege.stanford.edu/

that was extended with generic ontology-driven components that allow collaborative authoring, querying, and browsing Semantic Web information. IkeWiki [7] allows annotating links, typing of pages, and context dependent content adaptation. Platypus Wiki [8] aims at augmenting a wiki with semantics. The main difference to my thesis is that existing approaches aim at augmenting existing wiki content with semantics instead of using a wiki-like infrastructure as an environment for collaboratively building ontologies.

3 Methodology and Contribution

In my thesis, on which I have been working for five months now, I commit to the following research methodology: (1) Analysis of a trade-off between expressivity and tangibility of an ontology meta-model suitable for a broad audience. (2) Combination of external resources: (a) the statistical analysis of folksonomies and associated usage data, (b) Web resources, such as Google or Wikipedia, (c) terminological resources, and (d) ontology mapping and matching techniques. (3) Development of functionality that supports the community in achieving consensus. (4) Application of various techniques for visualization of ontologies and user interfaces to foster comprehensibility. (5) Evaluation of the prototype by (a) comparing the performance of community-driven, wiki-based ontology building to the performance of the traditional, engineering-oriented approach and (b) undertaking a usability study.

4 Expected Impact

The approach towards ontology building described in this paper is supposed to enable more users to contribute to the creation and maintenance of ontologies by (1) providing an easy-to-use, wiki-based user interface, (2) re-using various external resources in domain ontologies, and (3) supporting the community in achieving consensus, in order to yield more relevant, up-to-date ontologies.

Acknowledgments. This work has been funded by the Austrian BMVIT/FFG under the FIT-IT project myOntology (grant no. 812515/9284).

References

1. Hepp, M., Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies. IEEE Internet Computing, 2007. **11**(7): p. 96-102.
2. Adida, B. and M. Birbeck. RDFa Primer 1.0. Embedding RDF in XHTML. W3C Working Draft 16 May 2006.
3. Domingue, J. Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web. In 11th Knowledge Acquisition for Knowledge-Based Systems Workshop. 1998. Banff, Canada.
4. Vrandecic, D., et al., The DILIGENT knowledge process. Journal of Knowledge Management, 2005. **9**(5): p. 85-96.
5. Knublauch, H., et al., The Protege OWL Plugin: An Open Development Environment for Semantic Web Applications, in International Semantic Web Conference (ISWC). 2004, Springer: Hiroshima, Japan.
6. Dello, C., E. Paslaru Bontas Simperl, and R. Tolksdorf. Creating and using semantic content with Makna. in Wiki meets Semantics workshop at the ESWC2006. 2006. Budva, Montenegro.
7. Schaffert, S. IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. in 1st international workshop on Semantic Technologies in Collaborative Applications STICA06. 2006. Manchester, UK.
8. Campanini, S.E., P. Castagna, and R. Tazzoli. Platypus Wiki: a Semantic Wiki Wiki Web. in 1st Italian Semantic Web Workshop Semantic Web Applications and Perspectives (SWAP). 2004. Ancona, Italy.

Distributed SPARQL Query Processing enabling Virtual Data Integration for E-Science Grids

Andreas Langegger

Institute of Applied Knowledge Processing
Johannes Kepler University Linz, Austria
http://www.faw.at
al@jku.at

1 Introduction

For scientific collaboration, sharing data between different parties is fundamental. Grids, originally developed for high-performance and parallel computing, enable the sharing of distributed resources across institutional boundaries by providing a security infrastructure and standardized Grid-services. Because data is usually stored in different information systems and schemes, at the moment they have to be prepared and manually aligned to a common schema. Knowledge about data structures and semantics is a precondition to be able to integrate data sources. To enable virtual integration, several concepts have been proposed in the field of distributed and federated database systems. For the integration of heterogeneous information systems, the mediator-wrapper architecture can be used. In order to fulfill the requirements of a Grid-based data integration middleware for distributed, heterogeneous data sources, several concepts introduced in the Semantic Web community have been considered. The Resource Description Framework (RDF) is well suited for global schema management. It is simple, supports modularization of commonly used semantics by the ontology layer, and allows for reasoning. A standardized query language (SPARQL) is currently being developed.

2 Related Work and Proposed Approach

The use of ontologies for data integration is not new [1]. However, there is currently no approach which enables the integration of distributed, heterogeneous data sources by a SPARQL query processor. Related work can be divided into several domains: RDF triple stores [2, 7], RDF-based query algebra and processing [4], schema mapping, data integration [6], as well as distributed query processing [5, 8]. For some specific wrappers existing mapping frameworks can be applied, like for example D2R-Map [3] for relational database systems.

At the moment RDF and the other Semantic Web layers are not well suited for virtual data access. Although SPARQL provides a communication protocol, queries are executed on local sites only. In this PhD thesis, a new approach will be proposed to support query planning and execution in a distributed environment.

Global queries are processed by a mediator, which computes the optimal query plan by iterative dynamic programming. Wrappers for different information systems provide specific access and data manipulation functions. Depending on wrapper capabilities, multiple-set operations (e.g. join) can be executed locally or at the mediator. An iterator-based approach and concepts like row blocking, semi-joins, etc. are desirable to improve query processing performance.

The middleware is being developed within the Austrian Grid Project. There is also tight cooperation with several application workpackages. One of the prototype applications will be a *Virtual Observatory* for solar phenomenons developed together with the Kanzelhöhe Solar Observatory.

3 Outlook

Currently, there is no approach for virtual data integration based on systematic SPARQL query processing. Queries are either executed locally or targeted against single sites. Within this PhD thesis, a query processor will be developed based on the mediator-wrapper architecture, enabling virtual integration of heterogeneous, distributed data sources. The impact and sustainability is expected to be high in future.

Acknowledgements This work is supported by the *Austrian Grid Project*, funded by the Austrian *Federal Ministry for Education, Science and Culture* under contract GZ 4003/2-VI/4c/2004.

References

1. Vladimir Alexiev, Michael Breu, Jos de Bruijn, Dieter Fensel, Ruben Lara, and Holger Lausen, editors. *Information Integration with Ontologies: Ontology Based Information Integration in an Industrial Setting.* Wiley & Sons, 2005.
2. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: Implementing the Semantic Web Recommendations. In *Proceedings of the International World Wide Web Conference*, page 74. Hewlett Packard Labs, 2004.
3. Chris Bizer and Richard Cyganiak. D2R Server – Publishing Relational Databases on the Semantic Web. In *5th International Semantic Web Conference*, 2006.
4. Richard Cyganiak. A relational algebra for SPARQL. Technical Report HPL-2005-170, HP Labs, Bristol, UK, 2005.
5. Donald Kossmann. The state of the art in distributed query processing. *ACM Comput. Surv.*, 32(4):422–469, 2000.
6. Gio Wiederhold. Mediators in the Architecture of Future Information Systems. In A.R.Hurson, M.W.Bright, and S.H.Pakzad, editors, *Multi-database Systems: An Advanced Solution for Global Information Sharing.* IEEE Press, 1993.
7. S. Harris and N. Gibbins. 3store: Efficient Bulk RDF Storage. In *Proceedings of the First International Workshop on Practical and Scalable Semantic Systems*, Oct 2003.
8. M. Nedim Alpdemir, Arijit Mukherjee, Anastasios Gounaris, Norman W. Paton, Paul Watson, Alvaro Fernandes, and Jim Smith. OGSA-DQP: A Service-Based Distributed Query Processor For The Grid. In *Proceedings of the Second e-Science All Hands Meeting*, 2003.

Research on collaborative information sharing systems

Davide Eynard

Politecnico di Milano
Dipartimento di Elettronica e Informazione
Via Ponzio 34/5, 20133 Milano, Italy
eynard@elet.polimi.it

Collaborative systems are systems designed to help people involved in a common task achieve their goals. They are widely used today, and they're gaining a great consensus both inside corporations and on the World Wide Web. There are many kinds of collaborative systems, such as Wikis (like Wikipedia), blogs, tag-based systems (like Flickr, del.icio.us and Bibsonomy) and even collaborative maps (as in Google Maps). One of the main reasons of this success is that, as applications are becoming more and more data-driven, spontaneous user participation adds value to a system because it helps in creating a new, unique and hard to recreate source of data [1].

The main objective of this research project is to study collaborative systems and the possibility to enhance them through semantics. The aim of a contamination between these systems and Semantic Web technologies is twofold: on one side, we think that the huge quantity of information created by the participation of many users can be better managed and searched thanks to added semantics; on the other side, Semantic Web community can exploit spontaneous collaboration to increase the amount of knowledge described through formal representations, making it available to many other applications. Between the many different collaborative systems currently available we chose a couple of families which, in our opinion, presented the most interesting open problems. On one side, we approached Wikis and their semantic extensions [2–4]. On the other side we studied tag-based systems (also called *folksonomies*), with a particular attention to *social bookmarking* web sites, highlighting their advantages and their limitations[5–9].

One of the main problems which characterizes Wiki systems is that published information is unstructured, hard to search and manage. Current research on Semantic Wikis is trying to address this problem through formal descriptions of Wiki contents. Folksonomies have limits which are mostly due to their self-moderation: lack of precision, lack of recall, gaming (that is, anyone can pollute the system intentionally with wrong information), and lexical ambiguities [8], which do not allow to easily extract meaning from tags in ways other than statistics and clustering. To address Wiki limitations, we are working on a model which uses different ontology layers to describe not only the contents, but also the context (that is, the processes and the dynamics between users inside the wiki) and the system itself. This would make the system not only more interoperable with other applications, but also more easy to shape, so it would better suit the needs of particular *communities of practice*[10]. For what concerns folksonomies, we decided to extend them with semantics in two different ways. On one side, using ontologies to describe them in a formal way: through these *folksologies*[11, 12], it is possible to model tag-based systems allowing for interoperability on different levels (inside the single user space, within a system, or between different systems and users). On the other side, using ontologies to describe folksonomy contents rather than structure, mapping user tags inside it: this would allow users to both have a quick, bottom-up, easy to use tag space and a more formalized, top-down hierarchical view of their tags.

At the present time we have implemented a tool which maps tags from del.icio.us inside Wordnet ontology and provides a new way to browse them: this allowed us to address some of the main problems which are typical of folksonomies, such as lack of recall and lexical ambiguities. We have also developed a fuzzy model to describe tag-based systems, which allowed us to get more accurate results through advanced fuzzy queries and to formally describe properties of some particular classes of tags. Currently, we are working on a Semantic Wiki prototype which implements the model previously described, also allowing users to tag its contents and map their tags inside a domain ontology.

List of References

1. Tim O'Reilly. What is web 2.0. design patterns and business models for the next generation of software. September 2005.
2. Bo Leuf and Ward Cunningham. *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley, 2001.
3. Max Völkel and Sebastian Schaffert, editors. *Proceedings of the First Workshop on Semantic Wikis – From Wiki To Semantics*, Workshop on Semantic Wikis. ESWC2006, June 2006.
4. Michel Buffa, Gal Crova, Fabien Gandon, Claire Lecompte, and Jeremy Passeron. Sweetwiki : Semantic web enabled technologies in wiki. In Völkel and Schaffert [3].
5. Clay Shirky. Ontology is overrated: Categories, links, and tags, 2005. `http://www.shirky.com/writings/ontology_overrated.html`.
6. Ellyssa Kroski. The hive mind: Folksonomies and user-based tagging, Dec 2005. `http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/`.
7. Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, Apr 2005.
8. Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.
9. Hana Shepard, Harry Halpin, and Valentin Robu. The dynamics and semantics of collaborative tagging. In *Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006)*, 2006.
10. Etienne Wenger. *Communities of Practice. Learning, meaning, and identity.* Cambridge University Press, New York, Port Chester, Melbourne, Sydney, 1998.
11. Tom Gruber. Tagontology - a way to agree on the semantics of tagging data, 2005. `http://tomgruber.org/writing/tagontology-tagcamp-talk.pdf`.
12. Stefano Mazzocchi. Folksologies: de-idealizing ontologies, April 2005. `http://www.betaversion.org/~stefano/linotype/news/85/`.

Triple Space Computing for Semantic Web Services

Omair Shafiq
Digital Enterprise Research Institute (DERI),
University of Innsbruck (UIBK)
6020 Innsbruck, Austria.
omair.shafiq@deri.org

Abstract. This thesis will address how to enable Triple Space Computing as a communication paradigm for Semantic Web Services. Currently, Semantic Web Services are following a message based communication paradigm. Triple Space Computing is envisioned as communication and coordination paradigm for Semantic Web Services which is an extension of tuple space computing to support RDF and then use it for communication based on the principle of persistent publication and read of data. Web Service Modeling Ontology (WSMO) is our conceptual model for Semantic Web Services. Web Service Execution Environment (WSMX) is one of the reference implementations of the WSMO conceptual model. The paper presents an overview of technical insights about integration of WSMX with Triple Space Computing and proposes that how WSMX can use Triple Space computing for its communication and coordination in terms of dynamic components management, external communication management, resource management and coordination of different interconnected WSMXs.

1 Introduction

Triple Space Computing (TSC) [1] has been proposed that defines the technologies and settings needed to develop a new paradigm for Web service communication that complies with the basic principles of the Web, i.e. stateless communication of resources, persistent publication of resources, unique identification of resources and non-destructive read access to resources [5]. The Triple Space Computing further adds compatibility with Web design principles, thus overcoming the deficiencies of message-based communication. This thesis addresses

In order to overcome drawbacks in existing communication paradigm of Semantic Web Services, integration of Triple Space Computing becomes a necessity. It will help Semantic Web Services to conform to the principles of Web by allowing communication based on persistent publication and read of semantic data in form of RDF triples over Triple Space. It will allow the reuse of information while communicating as information is published persistently. Asynchronous communication will allow SWS to work in distributed environments like Web. It will help in logging the results of time consuming processes so that it can be reused where required.

2 A roadmap for Triple Space Computing in Semantic Web Services

This thesis will integrate the Triple Space Computing with WSMX [2] by analyzing that how and where exactly the two technologies fit together. The integration has been proposed [4] as three major entry points which are (1) enabling components management in WSMX using Triple Space Computing, (2) External communication grounding in WSMX using Triple Space Computing and (3) Resource Management in WSMX using Triple Space Computing. This integration will be used further to enable the communication of different inter-connected WSMX and then to build an application scenario to show the viability. Each of the integration entry points have been described in subsections below.

WSMX has a management component [3] that manages the over all execution by enabling coordination of different components based on some execution scenario [5] specified by user in Goal. In this way there is a clear separation between business and management logic in WSMX. The individual components have clearly defined interfaces and have component implementation well separated with communication issues. Each component in WSMX have wrapper to handle the communication issues. The WSMX manager and individual components wrappers are needed to be interfaced with Triple Space in order to enable the WSMX manager to manage the components over Triple Space. The communication between manager and wrappers of the components will be carried out by publishing and subscribing the data as a set of RDF triples over triple space. The wrappers of components that handle communication will be interfaced with Triple Space middleware.

WSMX acts as a semantic middleware between users and real world web services [3]. Currently, due to existence of message oriented communication paradigm, users communicate with WSMX and WSMX communicate with Web Services synchronously. The external communication manager of WSMX is needed to provide a support to communicate over Triple Space. The interfaces for sending and receiving external messages by WSMX are needed provide a grounding support to alternatively communicate over Triple Space. This needs to be resolved by addressing several issues, i.e. invoker component in WSMX is needed to support Web Services Description Language (WSDL) and Simple Object Access Protocol (SOAP) communication binding over Triple Space. The Entry point interfaces will be interfaced with Triple Space middleware in order to provide the glue between existing Web Services standards and Triple Space Computing.

WSMX contains different repositories to store ontologies, goals, mediators and web services descriptions as WSML based files [3]. The internal repositories of WSMX are needed to be made optional and enable to store the WSML based data as set of RDF named graphs in Triple Space Storage. This is mainly concerned with transforming the existing representation of data in form of WSML into RDF representation. The repository interfaces are needed to be interfaced with Triple Space middleware.

After enabling WSMX with Triple Space Computing, the next step will be to enable the communication and coordination of different WSMXs over Triple Space, i.e. forming a cluster of different interconnected WSMX nodes to support distributed service discovery, selection, composition, mediation, invocation etc. The management component in WSMX is will be enhanced to coordinate with WSMX managers in other WSMXs over Triple Space to form a cluster.

References

1. D. Fensel, Triple-space computing: Semantic Web Services based on persistent publication of informatio: In Proceedings of the IFIP International Conference on Intelligence in Communication Systems, INTELLCOMM 2004, Bangkok, Thailand, November 23-26, 2004.
2. C. Bussler et al, Web Service Execution Environment (WSMX), W3C Member Submis-sion, June 2005. Available at http://www.w3.org/Submission/WSMX
3. M. Zaremba, M. Moran, T. Haselwanter, WSMX Architecture, D13.4v0.2 WSMX Working Draft.
4. O. Shafiq, R. Krummenacher, F. Martin-Recuerda, Y. Ding, D. Fensel, "Triple Space Computing middleware for Semantic Web Services", The MWS Workshop at 10th IEEE International Enterprise Computing Conference (EDOC 2006), 16-20 October 2006, Hong Kong.
5. T. Haselwanter, Maciej Zaremba and Michal Zaremba. Enabling Components Management and Dynamic Execution Semantic in WSMX. WSMO Implementation Workshop 2005 (WIW 2005), 6-7 June, Innsbruck, Austria.

Reasoning with Large Data Sets

Darko Anicic

Digital Enterprise Research Institute (DERI), University of Innsbruck, Austria
darko.anicic@deri.org

Abstract. Efficient reasoning is a critical factor for successful Semantic Web applications. In this context, applications may require vast volumes of data to be processed in a short time. We develop novel reasoning techniques which will extend current reasoning methods as well as existing database technologies in order to enable large scale reasoning. We propose advances and key design principles primarily in: making an efficient query execution plan as well as in memory, storage and recovery management. Our study is being implemented in Integrated Rule Inference System (IRIS) - a reasoner for Web Service Modeling Language.

1 Problem Statement

The Web Service Modeling Language WSML[1] is a language framework for describing various aspects related to Semantic Web (SW) services. We are developing IRIS[2] to serve as a WSML reasoner which handles large workload efficiently.

Current inference systems exploit reasoner methods developed rather for small knowledge bases [2]. These systems[3], although utilize mature and efficient relational database management systems (RDBMSs) and exploit a number of their evaluation strategies (e.g., query planning, caching, buffering etc.), cannot meet requirements for reasoning in complex SW applications. Reason for this is found in the fact that database techniques are rather developed for explicitly represented data, and need to be extended for dealing with implicit knowledge.

In this work we investigate a framework which generalizes relational databases by adding deductive capabilities to them. RDBMSs suffer some limitations w.r.t the expressivity of their language. Full support for recursive views is one of them [3]. Further on, negation as failure is recognized as a very important nonmonotonic property for the Semantic Web. RDBMSs, although deal with negation as failure, can not select a minimal fixpoint that reflects the intended meaning in situations where the minimal fixpoint may not be unique. Our framework, although exceeding capabilities of RDBMSs, does not compromise their performance.

Current reasoners cannot cope with large data sets (i.e., relations larger than system main memory). Hence a reasoner needs to deal effectively with portions of

[1] WSML: http://www.wsmo.org/TR/d16/d16.1/v0.2/.

[2] IRIS: http://sourceforge.net/projects/iris-reasoner/.

[3] Reasoners which utilize persistant storage: KAON2, Aditi, InstanceStore, DLDB.

relations (possible distributed over many machines), and sophisticated strategies for partition-level relation management are required. Consequently, a relevant topic for our present and future work is: *The development of effective optimization algorithms as well as distribution and memory management strategies for reasoning with large data sets.*

2 Efficient Large Scale Reasoning: an Approach

We will now give a short overview of our approach to achieving effective reasoning with large data sets.

Unlike other inference systems[4], which utilize SQL to access existential relations, we tightly integrate IRIS with its storage layer (i.e., rules are translated into relational algebra expressions and SQL is avoided as an unnecessary overhead). We extend embedded RDBMS query optimizer (which is rather designed to be used for extensional data) for derived relations. The estimation of the size and evaluation cost of the intensional predicates will be based on the adaptive sampling method [4, 1], while the extensional data will be estimated using a graph-based synopses of data sets similarly as in [5]. Further on, for reasoning with large relations, run time memory overflow may occur. Therefore in IRIS we are developing novel techniques for a selective pushing of currently processed tuples to disk. This technique will be further extended for data distributed over many disks (e.g., a cluster of machines). Such techniques aim to enable IRIS to effectively handle large workload which cannot fit in main memory of the system.

Our framework comprises a recovery manager and thus features fault-tolerant architecture. Using logging and replications we ensure that, when a crash occurs, the system may continue with an ongoing operation without loss of previously computed results.

3 Acknowledgment

I am grateful to Michael Kifer and my supervisors: Stijn Heymans and Dieter Fensel for their help in the work conceptualization and insightful discussions.

References

1. M. E. Vidal E. Ruckhaus and E. Ruiz. Query evaluation and optimization in the semantic web. In *ALPSWS2006 Workshop, Washington, USA.*
2. Dieter Fensel and Frank van Harmelen. Unifying reasoning and search to web scale. *IEEE INTERNET COMPUTING*, page 3, 2 2007.
3. Michael Kifer, Arthur Bernstein, and Philip M. Lewis. *Database Systems: An Application Oriented Approach.* Addison-Wesley, Boston, MA, USA, 2005.
4. R. J. Lipton and J. F. Naughton. Query size estimation by adaptive sampling. In *PODS '90*, NY, USA.
5. J. Spiegel and N. Polyzotis. Graph-based synopses for relational selectivity estimation. In *SIGMOD '06*, NY, USA.

[4] KAON2, QUONTO, InstanceStore and DLDB exploit SQL for querying.

Towards a Semantic Wiki for Science

Christoph Lange

Computer Science, Jacobs University Bremen, ch.lange@iu-bremen.de

Collaborative work environments (CWEs) for scientific knowledge have many applications in research and education. In recent years, successful platforms open for anyone appeared on the web, e. g. *Wikipedia* and *PlanetMath*, a wiki particularly tailored to mathematics, or *Connexions*, a CMS for general courseware[1]. Thanks to flexible content creation and linking, similar systems also support corporate knowledge management, but they lack services desirable for effective *scientific knowledge management*. For example, full text search is not suitable for mathematical or chemical formulae[2], and tagging pages does not help to find unproven theorems about triangles. Current semantic wikis [5] solve the latter problem by typing pages and links with terms from ontologies, but they do not support formula search, which would require *structural semantic markup* (SSM), a common approach in mathematical knowledge management.

Further semantic services that have been realised on the Semantic Web, but not yet in open CWEs, include dependency maintenance across changes and learning assistance by suggesting direct and indirect prerequisites to the scholar. How can the knowledge that *is available* in CWEs (e. g. the RDF graph behind a semantic wiki) be used for more than just displaying navigation links, some editing assistance, and semantic search? I will investigate whether a CWE can be turned into an integration platform for semantic services by first creating a uniform ontology abstraction layer *at its core*[3] and prototype such an application that supports SSM formats for various scientific domains based on the semantic *IkeWiki* [3], as wikis particularly support the stepwise formalisation workflow required for scientific SSM (cf. [3,1])[4].

SSM, already having many applications in mathematics (e. g. in the context of the OMDoc XML format [1]), is currently being extended towards other sciences. Research conducted in our group showed that a three-layered model of knowledge can be assumed in mathematics and physics, and probably in most other sciences: *Objects* (symbols, numbers, equations, molecules, etc.), *statements* (axioms, hypotheses, measurement results, examples, with relations like "proves", "defines", or "explains") and *theories* (collections of interrelated statements, defining the context for symbols) [1]. For Semantic Web software, these classes and relations need to be formalised in an *ontology*; I will base my system on the ontologies behind scientific markup languages, and, following the

[1] See http://www.{wikipedia,planetmath,cnx}.org.

[2] $c = \sqrt{a^2 + b^2}$ can mean the same as $x^2 + y^2 = z^2$.

[3] Ontology support is mostly *optional* in current systems.

[4] The related se(ma)^{2}wi [6] system is an experiment with a *Semantic MediaWiki* fed with mathematical knowledge formatted in OMDoc. The semantic structure of the formulae and the links between pages is lost during this conversion, though.

assumption that sciences have common traits like the notion of a "theory" or a "dependency" relation among theories, a generic *upper* ontology of these. To date, merely part of the ontologies behind SSM formats are given as human-readable specifications; I will formalise and generify them in OWL. In a scientific CWE, one page would usually contain one statement, one small theory, or a course module aggregating a few of them. A generic mapping mechanism between XML schemata and ontologies will be applied to *extract* knowledge that is relevant for semantic services from those XML pages to an RDF representation.

As SSM is inherently hard to edit manually, the interaction with the semantic services will be designed in a user-centered way, where the benefits of services like enhanced search and navigation are shared with the users in order to motivate them to contribute. One such service is an ontology-based auto-completion of link targets in the editor. Not all page names starting with the letters typed so far are suggested, but only those pages whose type matches the range of the relation the current link represents. Further planned services include a learning assistant that suggests to explore transitive dependencies, a dependency maintenance assistant, as well as connecting the system to external services already available, e. g. *MathWebSearch*[5]. A preliminary classification suggests that most of the cross-domain services can indeed be modeled on top of the abstraction layer provided by the above-mentioned upper ontology; a formal analysis of the demands of the services on knowledge representation will follow. A challenge is, however, making the different levels of reasoning required by the services (plain triple query for auto-completion vs. computing compositions of relations for dependency management) work smoothly in an inherently inconsistent collaborative setting.

An existing prototype of a wiki for OMDoc [2], featuring basic functionality like page editing, rendering as XHTML+MathML and typed navigation links from a user's perspective, and a basic OMDoc/XML to RDF mapping from a knowledge representation perspective, will be completely redesigned by introducing a generic ontology-based abstraction layer and integrating semantic services on top. It will be evaluated in a cross-domain case study with scientists and in an educational case study with students, leading to feedback for the ontology design. If the abstraction layer approach does facilitate the design and integration of semantic services that increase benefit and reduce users' investment, improving other CWEs, even in non-scientific domains, in a similar way will become possible.

1. M. Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. Number 4180 in LNAI. Springer, 2006.
2. C. Lange. SWiM – a semantic wiki for mathematical knowledge management. Technical report, Jacobs University Bremen, 2007.
3. S. Schaffert. Semantic social software – semantically enabled social software or socially enabled semantic web? In Sure and Schaffert [4].
4. Y. Sure and S. Schaffert, editors. *Semantics: From Visions to Applications*, 2006.
5. M. Völkel, S. Schaffert, and S. Decker, editors. *1st Workshop on Semantic Wikis*, volume 206 of *CEUR Workshop Proceedings*, Budva, Montenegro, June 2006.
6. C. Zinn. Bootstrapping a semantic wiki application for learning mathematics. In Sure and Schaffert [4].

[5] http://search.mathweb.org

Ontology-Driven Management of Space Middleware

Reto Krummenacher

Digital Enterprise Research Institute, University of Innsbruck, Austria
reto.krummenacher@deri.org

Abstract. Recent work in the field of middleware technology proposes semantics-aware tuplespaces as a tool for coping with the *scalability*, *heterogeneity* and *dynamism* issues arising in distributed environments such as the (Semantic) Web. The fact that (Semantic) Web services communicate by synchronous message exchanges initiated *triplespace computing*. The aim was to bring the Web's "persistently publish and read" paradigm to service computing. Based on experiences with ontologies in traditional middleware we argue that ontology-driven management will be a major asset of semantic tuplespaces compared to traditional ones. In this research we look at ontology-based metadata to enhance semantic space infrastructures to become reflective middleware.[1]

1 Introduction

Middleware is software that connects distributed components and applications. (Semantic) Web services are seen to be a promising and currently widely researched middleware approach in particular for large scale systems. Unfortunately Semantic Web services have inherited the Web service communication model, which is based on synchronous message exchange, thus incompatible with the architectural model of the Web. Analogously to the conventional Web, truly Web-compliant service communication should be based on persistent publication in order to allow the communicated data to outlive the services publishing or consuming it. Recent middleware technology proposes semantics-aware tuplespaces as a tool for coping with the *scalability*, *heterogeneity* and *dynamism* issues of large scale open systems. Our proposition is *triplespace computing*: RDF *triple*s create a natural link from the *space*-based *computing* paradigm into the (Semantic) Web.

2 Research Problem

Various semantics-aware space projects matured the semantic coordination and data models; see TSC [1], Semantic Web Spaces [4], and the joint successor TripCom. These projects do however not yet sufficiently address the non-functional properties of middleware, e.g. distribution, scalability, reliability or dynamism.

A critical concept to deal with management issues in the absence of centralized control is metadata. Ontology-based metadata seem to be the natural choice for triplespace computing. In fact ontology-driven middleware management is seen to be one of the

[1] Supported by TSC (FIT-IT, tsc.deri.at) and TripCom (IST-4-027324-STP, www.tripcom.org)

major assets of semantics-aware tuplespaces over traditional approaches. Still, the aforementioned projects largely neglect this fact or have failed to show the procedures. This leads to our main research question:

Main Question: *what does an ontology-based metadata vocabulary have to incorporate and how can it be modelled in order to enhance triplespaces to become reflective for large scale open systems?*

The main question is divided into three sub-questions:

Q.1 *how to ontologize the space middleware and the data in order to provide reflective management of core non-functional properties, in particular distribution?*

Q.2 *what requirements result from personalization and usage-awareness and which additional metadata modules are needed?*

Question Q.3 is concerned with the application of the vocabularies to an existing implementation and the evaluation of its usability to the distribution property:

Q.3 *how can metadata be acquired and provided to participating nodes in order to improve the data distribution within the reflective middleware?*

Improving the distribution influences at ones the overall performance (network-driven distribution) and the search efficiency (usage-driven distribution).

3 Expected Contribution

First ideas for a management ontology were developed in TripCom [2]. In [3] we outline key factors for context modeling ontologies: traceability, comparability, logging and quality of data, extensibility, genericity, completeness and scalability. Similar criteria will form the basis for our management ontologies. To increase the interoperability and adoptability we investigate the mapping of our vocabularies to foundational ontologies. Such mappings foster wider understanding which is crucial for data coordination.

As outcome of this research we expect a metadata infrastructure that is tailored to the management processes present in reflective space middleware. We seek ontologies in the domain of distributed (semantic) information management on the one hand and adaptability, personalization on the other. As it is advisable to use small and simple ontologies that are easier adopted and reused in the large, we will come up with well-integrated, but distinct ontologies for the respective tasks – in particular distribution. Therewith we expect to significantly contribute to the success of triplespace computing by enhancing the management procedures with our ontologies that go beyond the ones of TripCom.

References

1. D. Fensel, R. Krummenacher, O. Shafiq, E. Kuehn, J. Riemer, Y. Ding, and B. Draxler. TSC - Triple Space Computing. *e&i Elektrotechnik und Informationstechnik*, 124(1/2), Feb. 2007.
2. R. Krummenacher, E. P. B. Simperl, V. Momtchev, L. Nixon, and O. Shafiq. Specification of Triple Space Ontology. TripCom Project Deliverable D2.2, March 2007.
3. R. Krummenacher and T. Strang. Ontology-Based Context Modeling. In *Workshop on Context-Aware Proactive Systems*, June 2007 (forthcoming).
4. L. J. B. Nixon, E. P. B. Simperl, O. Antonenko, and R. Tolksdorf. Towards Semantic Tuplespace Computing: The Semantic Web Spaces System. In *22nd ACM Symposium on Applied Computing*, March 2007.

Intelligent Search in a Collection of Video Lectures

Angela Fogarolli

University of Trento,
Dept. of Information and Communication Tech.,
Via Sommarive 10, 38050 Trento, Italy
afogarol@dit.unitn.it

1 Abstract

In recent years, the use of streamed digital video as a teaching and learning resource has become an increasingly attractive option for many educators as an innovation which expands the range of learning resources available to students by moving away from static text-and-graphic resources towards a video-rich learning environment. Streamed video is already widely used in some universities and it is mostly being used for transmitting unenhanced recordings of live lectures.

What we are proposing is a way of enriching this video streaming scenario in the eLearning context. We want to extract information from the video and the correlated materials and make them searchable. Thus, the aim of this thesis is to create a semantically searchable collection of video lectures.

In the literature, surprisingly little information can be found about speech and document retrieval in combination with lecture recording. There are interesting examples of e-lecture creation and delivery e.g. [5], audio retrieval of lecture recording [3] that explore automatic processing of speech, or systems such as the eLecture portal [1] which indexes the audio and also the text of the lecture slides. But to the best of our knowledge there is no system which combines and synchronizes the different modalities in a searchable collection.

What we propose is enabling the search and navigation through the different media types presented in a frontal lecture with the addition of the video recording. In video indexing domain Snoek and Worring [4] have proposed to define multimodality as "the capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels". The channels or modalities of a video document described in [4] are the visual, auditory and textual modality.

We believe that using more than one modality – as explained in [2] – could increase productivity also in the context of e-learning, where it is really frequent to scan for information. Our main focus is to enable search on two modalities; in particular we will index the auditory modality of video lecture content based on transcription obtained with automatic speech recognition tools and on textual modality using text indexing on the related materials. Furthermore, we do not just want to present an enhanced version of the current state of the art in

e-lecture retrieval but we also envision to add semantic capabilities to the search functionalities in order to provide a superior learning experience to the student (personalized search, personalized learning path, relevant contextualization, automatic video content profiling...).

The application we are proposing is a way to provide a tool for students to enable more flexibility in e-Lectures consumption. The student could seek inside a collection of learning lectures and related materials (desktop activities recording, PowerPoint presentations, interactive whiteboard tracks...). The search could also be personalized to meet the student demands. For each hit the system would display the lectures video-recording and the temporally synchronized learning materials. Another benefit we want to archive using Semantic Web techniques is to present a profile information of the content of the video lecture, this could lead to an improvement of the state of the art in the video-indexing field allowing automatic profile annotation of the content of the video.

The research work specifically addressed by this thesis will investigate the following challenges:

- Finding an innovative way for mastering the gap between information extraction and knowledge representation in our context. For each video and related learning resource an RDF representation would be extracted. The created graph would be navigated during the search task to find the requested information and suggest related topics using ontology linkage. We will use ontologies for high level lecture description and query understanding.
- Automatic content description of the presented learning material. A textual description of the content of a video result, lecture or course would be presented at the user. This could be realized presenting a profile information of the knowledge extracted from the video and the related material.
- Evaluation of the tool value for improving the student performance and for shortening the learning time we will conduct before and after the semantic enhancement.

References

[1] Christoph Hermann, Wolfgang Hürst, and Martina Welte. The electure portal: An advanced archive for lecture recordings. In *Informatics Education Europe Conference*, 2006.

[2] D. Jones, Shen A., and D W., Reynolds. Two experiments comparing reading with listening for human processing of conversational telephone. In *Interspeech 2005 Eurospeech Conference*, 2005.

[3] A. Park, T. Hazen, and J. Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *ICASSP*, March 2005.

[4] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. In *Multimedia Tools and Applications*, number 25, pages 5–35, 2005.

[5] Z. Zhu, C. McKittrick, and W. Li. Virtualized classroom automated production, media integration and user-customized presentation. In *Multimedia Data and Document Engineering*, July 2004. Semantic Web.

A Collaborative Semantic Space for Enterprise

Alexandre Passant

Université Paris IV Sorbonne, Laboratoire LaLICC, Paris, France
alexandre.passant@paris4.sorbonne.fr
EDF, Recherche et Développement, Clamart, France
alexandre.passant@edf.fr

Abstract. This abstract introduces a new kind of corporate knowledge management system, using a Semantic Web layer on the top of existing Web 2.0 tools in order to provide value-added services to end-users.

1 Motivations and research problem

EDF R&D[1] is a research center dedicated to energy domain. Due to its corporate culture and the fields it deals with, there is a real difficulty to make people share their knowledge within the company. In order to solve these problems and incite people to better exchange information, a corporate Web 2.0 platform - including blogs, RSS feeds and wikis - was recently introduced. Yet, these tools quickly showed some limitations regarding information integration, capitalization and retrieval. Indeed, if they provide efficient ways to publish information, they raised various issues as informations heterogeneity, re-usability of created data, ways of consuming information depending the user point of view...

This Ph.D. work focuses on how existing Web 2.0 tools can be part of the Semantic Web to (1) populate domain ontologies and immediately get benefits from these ontologies, their instances and relations among them to produce value-added tools and mash-up interfaces and (2) share a common model to describe information and index content in order to let users efficiently retrieve and exchange information; creating what we call a Collaborative Semantic Space. Among others, some questions to be answered in this Ph.D. work are: how can folksonomies be integrated with the Semantic Web and what such an approach can offer to tag-based search interfaces ? What about knowledge extraction from blogs and wikis and ontology population, in both editing and querying ? What kind of interfaces and services can prove the usefulness of the Semantic Web and domain ontologies in an industrial context ?

2 Proposed Approach and Contract with Existing Ones

In order to solve the issues mentioned before, our approach is similar to the RDF bus[1] architecture, since we have (1) a set of ontologies designed to represent both the documents and their content, (2) add-ons to existing and already

[1] Eléctricité de France Recherche et Développement, see http://rd.edf.fr.

used tools to provide RDF export of their data and (3) a triple-store to centralize triples and provide exports thanks to services plugged to its SPARQL endpoint. This, we did not created a new Semantic Web integration framework as CoMMA[4] or SCORE[6] but focused on adding a Semantic Web layer on the top of existing services. These add-ons (1) automatically translate data to a common format using the SIOC ontology and (2) provide semi-automatic ways to populate or link to domain ontologies, keeping user interfaces as simple as possible.

Regarding semantic blogging[3], we proposed a way to create a bridge between folksonomies and ontologies in order to solve problems they raised and offer a better search experience, as topics suggestion[5]. About wikis, we are currently working on a templated semantic wiki engine to let anyone create ontology instances and relations between them, without learning a specific syntax, what we think is a key feature for the adoption of the Semantic Web by end-users.

Regarding ontologies, we distinguish ontologies that represent the internal architecture of the system and the ones that represent content. Rather that defining a specific internal ontology as in CoMMA, we decided to use SIOC - an ontology for online communities, in which we have been involved - as a core of our system. In order to describe business data, we decided to use various ontologies as FOAF, DOAP or the geonames.org one, mapped with DOLCE[2] to have a stronger formalism behind. Thus, our system can import external resources without data integration issues, creating a link between open RDF data and enterprise information systems.

Finally, regarding data storage and exports, we decided to use a system providing a SPARQL endpoint so that new services could be easily plugged over HTTP, providing different ways to query, visualize or combine data for users.

3 Conclusion and Future Works

Right now, we have provided the basis for this Collaborative Semantic Space, that let us see how existing services can be integrated thanks to Semantic Web technologies, and what it can offer to end-users.

Among our future works, we will use the ontology to automatically index RSS feeds that users are subscribed to, and see how it can help to create virtual feeds depending on users interests, that can help to solve the problem of evolving annotations on the Semantic Web. Another part of the work will be to see how ontologies can help to find social networks within this Collaborative Semantic Space. For example, we would like to be able to find all engineers interested in european companies working on tidal energies. Finally, since we can add services to our system thanks to the use of its SPARQL endpoint, another goal will be to provide new and unforeseen services and query interfaces for RDF data.

References

1. T. Berners-Lee. Putting the Web back in Semantic Web. http://www.w3.org/2005/Talks/1110-iswc-tbl/, May 2005. Keynote presenta-

tion at ISWC 2005.

2. S. Borgo, A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. WonderWeb deliverable d18 – ontology library. Technical report, ISTC-CNR National Research Council, Institute of Cognitive Sciences and Technology, Padova, Italy, 2003.
3. S. Cayzer. What next for Semantic Blogging? Technical Report HPL-2006-149, Hewlett-Packard Laboratories, Bristol, UK, Oct. 2006.
4. F. Gandon. Agents handling annotation distribution in a corporate semantic web. *Web Intelligence and Agent System*, 1(1):23–45, 2003.
5. A. Passant. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In *International Conference on Weblogs and Social Media*, March 2007.
6. A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke. Managing semantic content for the web. *IEEE Internet Computing*, 6(4):80–87, 2002.

A Directed Hypergraph Model for RDF

Amadís Antonio Martínez Morales[1] and María Esther Vidal Serodio[2]

[1] Universidad de Carabobo, Venezuela, E-mail: aamartin@uc.edu.ve
[2] Universidad Simón Bolívar, Venezuela, E-mail: mvidal@ldc.usb.ve

RDF is a proposal of the W3C to express metadata about resources in the Web. The RDF data model allows several representations, each one with its own limitations at expressive power and support for the tasks of query answering and semantic reasoning. In this paper, we present a directed hypergraph model for RDF to represent RDF documents efficiently, overcoming those limitations.

1 Related Work

An RDF document can be viewed as a graph: nodes are resources and arcs are properties. Formally [4], suppose there are three infinite sets: $\mathcal{U}$ (URIs), $\mathcal{B} = \{b_j : j \in \mathbb{N}\}$ (blank nodes), and $\mathcal{L}$ (literals). $(s, p, o) \in (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{B} \cup \mathcal{L})$ is an RDF triple, s is called the subject, p the predicate, and o the object. An RDF graph T is a set of RDF triples. The universe of T, $univ(T)$, is the set of elements of $\mathcal{U} \cup \mathcal{B} \cup \mathcal{L}$ that occur in the triples of T. $sub(T)$ (resp. $pred(T)$, $obj(T)$) is the set of all elements in $univ(T)$ that appear as subjects (resp. predicates, objects) in T. RDF graphs allow several representations: labeled directed graphs (LDG) [4,7], undirected hypergraphs (UH) [5], and bipartite graphs (BG) [5,6].

In the LDG model, given an RDF graph T, nodes in V are elements of $sub(T) \cup obj(T)$, and arcs in E are elements of $pred(T)$ [4,7]. Each $(s, p, o) \in T$ is represented by a labeled arc, $s \xrightarrow{p} o$. The number of nodes and arcs for LDG representation is $|V| \leq 2|T|$ and $|E| = |T|$ [5]. This approach may violate some of the graph theory constraints. Thus, while LDG model is the most widely used representation, it can not be considered a formal model for RDF [1].

In the UH model, given an RDF graph T, each $t = (s, p, o) \in T$ is a hyperedge in E and each element of t (subject s, predicate p, and object o) is a node in V. The number of nodes and hyperedges for UH representation is $|V| = |univ(T)|$ and $|E| = |T|$ [5]. However, UH represent RDF documents as a generalization of undirected graphs, losing the notion of direction in RDF graphs, which impacts the task of semantic reasoning. Besides, it can be hard to graphically represent large RDF graphs, like the museum example [2].

In the BG model, given an RDF graph T, there are two types of nodes in V: statement nodes St (one for each $(s, p, o) \in T$) and value nodes Val (one for each $x \in univ(T)$). Arcs in E relate statement and value nodes as follows: Each $t \in St$ has three outcoming arcs that point to the corresponding node for the subject, predicate, or object of the triple represented by t. The number of nodes and arcs for BG representation is $|St| = |T|$, $|Val| = |univ(T)|$, and $|E| = 3|T|$ [5,6]. While BG satisfy the requirement of a formal model for RDF, issues such as reification, entailment and reasoning have not been addressed yet [1].

2 Proposed Solution

Directed hypergraphs (DH) have been used as a modeling tool to represent concepts and structures in many application areas: formal languages, relational databases, production and manufacturing systems, public transportation systems, between others [3]. RDF DH are formally defined as follows:

Definition 1. *Let T be an RDF graph. The RDF directed hypergraph representing T is a tuple $\mathcal{H}(T) = (W, E, \rho)$ such that: $W = \{w : w \in univ(T)\}$ is the set of nodes, $E = \{e_i : 1 \leq i \leq |T|\}$ is the set of hyperarcs, and $\rho : W \times E \rightarrow \{s, p, o\}$ is the role function of nodes w.r.t. hyperarcs. Let $t \in T$ be an RDF triple, $e \in E$ an hyperarc, and $w \in W$ a node such that $w \in head(e) \cup tail(e)$. Then the following must hold: (i) $(\rho(w, e) = s) \Leftrightarrow (w \in tail(e)) \wedge (w \in sub(\{t\}))$, (ii) $(\rho(w, e) = p) \Leftrightarrow (w \in tail(e)) \wedge (w \in pred(\{t\}))$, and (iii) $(\rho(w, e) = o) \Leftrightarrow (w \in head(e)) \wedge (w \in obj(\{t\}))$*

The information is only stored in the nodes, while hyperarcs preserve the role of nodes and the notion of direction in RDF graphs. Thus, the space complexity of our approach must be smaller than the complexity of representations in section 1. The number of nodes and hyperarcs for DH representation is $|W| = |univ(T)|$ and $|E| = |T|$. Besides, concepts and algorithms of hypergraph theory could be used to manipulate RDF graphs under this representation.

3 Preliminary Experimental Results

Labeled directed graph (LDG) and directed hypergraph (DH) representations were studied over twenty synthetic simple RDF documents, randomly generated using an uniform distribution. Around 33% of the resources simultaneously played the role of subject, predicate, or object. Document sizes were increased, ranging from 1000 to 100000 triples. We used two metrics in this study: (a) the space in memory required to store information, measured as the number of nodes and arcs and (b) the number of comparisons required to answer an elemental query. In both cases, the LDG approach showed a trend of linear dependence on the size of the document, while DH exhibited a more independent behavior.

4 Conclusions and Future Plan

We proposed a directed hypergraph model for RDF. Initial results make us believe that our approach scales better than existing representations. In the future, we propose to: (1) extend this representation for RDFS graphs, (2) develop query evaluation algorithms for conjunctive and SPARQL queries, (3) study the impact of this model on the tasks of query answering and semantic reasoning, and (4) conduct empirical studies to analyze the goodness of our approach.

References

1. F. Dau, "RDF as Graph-Based, Diagrammatic Logic", *Proceedings of ISMIS'06*, pages 332–337, 2006.
2. FORTH Institute of Computer Science (2003), "RQL v2.1 User Manual". [Online] `http://139.91.183.30:9090/RDF/RQL/Manual.html`
3. G. Gallo, G. Longo, S. Pallottino, and S. Nguyen, "Directed Hypergraphs and Applications", *Discrete Applied Mathematics*, Vol. 42, No. 2, pages 177–201, 1993.
4. C. Gutierrez, C. A. Hurtado, and A. O. Mendelzon, "Foundations of Semantic Web Databases", *Proceedings ACM Symposium on PODS*, pages 95–106, 2004.
5. J. Hayes, *A Graph Model for RDF*, Diploma Thesis, Technische Universitt Darmstadt / Universidad de Chile, 2004.
6. J. Hayes and C. Gutierrez, "Bipartite Graphs as Intermediate Model for RDF", *Proceedings of the ISWC'04*, pages 47–61, 2004.
7. G. Klyne and J. J. Carroll (2004), "Resource Description Framework (RDF): Concepts and Abstract Syntax", W3C Recommendation. [Online] `http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`

Applying Semantic Technologies to the Design of Open Service-oriented Architectures for Geospatial Applications

Thomas Usländer

Fraunhofer IITB, Fraunhoferstr. 1, D-76131 Karlsruhe
thomas.uslaender@iitb.fraunhofer.de

1 Research Problem

Up to now, there is no established methodology for the design of a geospatial service-oriented architecture (SOA), e.g., for environmental risk management applications. However, there are key design guidelines and constraints imposed by corresponding standards of ISO and the Open Geospatial Consortium (OGC). Standards exist on both the abstract (i.e. platform-neutral) and the concrete (i.e. platform-specific) level, e.g. Web services, but still focus on syntactic interoperability.

An example motivates the application of semantics: As part of a forest fire risk assessment process in Spain the need to access to "vulnerable infrastructure in Catalonia" has been identified. The abstract service platform offers the capability of a generic feature (object) access service that supports queries with geospatial filters. Currently, it is up to the SOA designer to establish a conceptual connection between "infrastructure in Catalonia" and "features". An ontological approach that knows the subsumption chain ("road" is-a "infrastructure element" is-a "feature") and knows that "Catalonia" is a geographical concept would help in the "early service discovery" and would open up new perspectives for (semi-)automated service engineering.

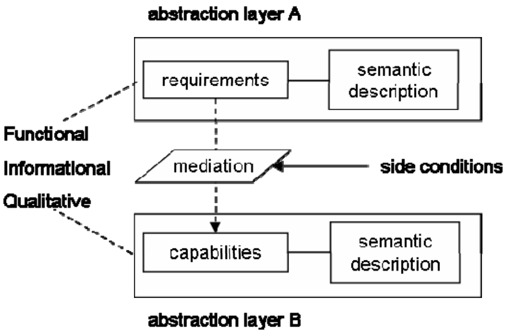


Fig. 1: Mapping of requirements to capabilities

A generic solution to such a design problem leads to the scientific kernel problem of semantically matching requirements of one abstraction layer A to capabilities of another abstraction layer B (see **Fig**. 1), taking side conditions explicitly into account. This kernel problem iteratively occurs when user requirements are broken down to capabilities of the next level. The task of mediation as a generic mechanism to bridge the gap between heterogeneous descriptions and/or expectations [2] plays a key role.

The thesis proposes a semantic SOA modelling framework (MFgeo) as a solution.

2 Methodology

Five different ontology types are proposed in [1] that contribute to forming a geospatial system. With MFgeo the thesis proposes a complementary, geospatial SOA design ontology as missing link for the design phase targeted at analysts and architects of geospatial applications. Emerging semantic web services frameworks such as WSMO, OWL-S or WSDL-S form the baseline of the methodology and will be considered in the context of existing geospatial ISO/OGC standards. MFgeo will support

1. annotation of informational, functional and qualitative requirements and discovery of capabilities triggered by domain, service and quality of service ontologies,
2. an iterative design process with a flexible mediation technique of requirements and capabilities taking side constraints, e.g. compliance to OGC standards and re-use of existing information and service models, explicitly into account,
3. means to document the design process enabling traceability of the user requirements and validation using reasoning tools, and
4. the specification of policies to monitor and control the operation of deployed service networks.

3 Current Results and Planning

Result so far is the architecture specification of the European Integrated Project ORCHESTRA [3] accepted as OGC discussion paper that has extended the OGC Reference Model by 1) a common meta-model approach for the service and information viewpoint, 2) the modelling of the mapping from the abstract to the concrete service platform, 3) a meta-information schema enabling semantic descriptions of geospatial resources, and 4) the consideration of policies in the engineering step of service networks. The current work focuses on semantic extensions of [3] followed by the design of MFgeo in 2008. The approach will be assessed by using MFgeo for an alternate ontology-driven design of an existing ORCHESTRA pilot application.

References

1. Rodriguez M.A., Cruz, I.F., Egenhofer, M.J. and Levashkin, S. (Eds.). GeoSpatial Semantics. First International Conference GeoS, Mexico City, 2005, LNCS 3799, 2005.
2. OASIS Semantic Execution Environment TC. Reference Model for Semantic Service Oriented Architecture. Working Draft 0.1, 2006, http://www.oasis-open.org.
3. Usländer, T. (Ed.). Reference Model for the ORCHESTRA Architecture (RM-OA) V2 (Rev 2.0). OGC 07-024, http://www.eu-orchestra.org/publications.shtml#OAspecs, 2007.

Pattern-based Ontology Construction

Eva Blomqvist

Jönköping University, Jönköping, Sweden
blev@jth.hj.se

Abstract. Large and complex enterprise systems face the same kind of information processing problems that exist on the web in general, and constructing an ontology is a crucial part of many solutions. Construction of enterprise ontologies needs to be at least semi-automatic in order to reduce the effort required, and another important issue is to introduce further knowledge reuse in the process. In order to realise these ideas the proposed research focuses on semi-automatic ontology construction, based on the methodology of case-based reasoning.

1 Introduction

When developing semantic applications for enterprises, constructing the enterprise application ontologies is a crucial part. Manual ontology engineering is a tedious and complex task. Another issue is knowledge reuse, common practises of the business world should be exploited as well as drawing on best practises in ontology engineering. By combining patterns with a case-based reasoning view, we aim at developing a novel semi-automatic ontology construction approach.

2 Background and Related Work

Our research focuses on application ontologies within enterprises, mainly for structuring and retrieval of information. We view an ontology design pattern as an ontology template, which is self-contained, comprised of a set of consistent ontology primitives, and intended to construct a part of some ontology. Related work on ontology patterns focus mainly on templates for manual use (like in [1]).

Recent developments in ontology engineering involve ontology learning (OL) as in [2], [3] and [4]. A major problem is that much of the information in a company is not explicitly stated, this is one issue where patterns can be of assistance. Case-based reasoning (CBR) is trying to mimic human behaviour, using previous experience to solve new problems. A case is a problem situation, previously solved cases are stored in the case base for reuse. The CBR process is viewed as a cycle of four phases: retrieval, reuse, revision and retaining cases.

3 Research Hypotheses

In our research some specific research questions have been derived, that can then be reformulated as the following hypotheses:

- CBR gives a framework for further automation of the ontology construction process, compared to related semi-automatic approaches that exist today.
- Using the CBR methodology (with patterns) can improve the quality of the generated ontologies compared to existing semi-automatic approaches.
- Automation reduces the total construction effort.
- Domain knowledge and engineering experience can be reused through patterns.

To verify the hypotheses the proposed method must be evaluated and compared to manual approaches as well as the related OL approaches stated earlier. The result produced by the method must be evaluated and shown to be of better quality compared to the result of related semi-automatic approaches.

4 Proposed Approach

The basis of a CBR approach is the case base and its content. In our approach the case base corresponds to a pattern catalogue (pattern base). The design patterns are represented as small ontologies and the architecture patterns are sets of constraints on the combination of design patterns, and may also include connections to specific design patterns.

The retrieval phase constitutes the process of analysing the input text corpus and deriving its representation, then matching this to the pattern base and selecting appropriate patterns. The reuse phase concerns the reuse and adaptation of the patterns, combining them into a first ontology. The revision phase includes extending the ontology, based on evaluation results. Retaining patterns includes the discovery of new patterns as well as improving existing patterns.

In our approach there is uncertainty inherent in all the described steps. For example each ontology primitive of the input representation have a certain degree of confidence associated, and the patterns are in themselves associated with a certain level of confidence. The levels of confidence are transferred onto the constructed initial ontology and can be used when evaluating it.

The main contributions of this approach is envisioned as both further automation of the ontology construction process, but in addition an increased quality of the produced ontology, as compared to other existing OL approaches. This increased quality will mainly be due to the use of patterns and the presence of an evaluation and revision phase in the method.

References

1. Gangemi, A.: Ontology Design Patterns for Semantic Web Content. In: Proceedings of ISWC 2005. Volume 3729 of LNCS., Springer (2005) 262–276
2. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer Science (2006)
3. Fortuna, B., Grobelnik, M., Mladenic, D.: Semi-automatic Data-driven Ontology Construction System. In: Proc. of IS-2006, Ljubljana, Slovenia (2006)
4. Iria, J., Brewster, C., Ciravegna, F., Wilks, Y.: An Incremental Tri-partite Approach to Ontology Learning. In: Proc. of LREC2006, Genoa (2006)

Semantic (Group Formation)

PhD Research Proposal[*]

Asma Ounnas[†]

School of Electronics and Computer Science
University of Southampton, UK
ao05r@ecs.soton.ac.uk

1 Motivation

For decades, group formation has been a subject of study in many domains. In learning, teachers form groups of students for different types of collaborative activities. For the formation to be efficient, teachers need take into account any constraints that can influence the performance of the group as a whole and that of the individuals within the group, such as students' previous experience, gender, nationality, and interests. The formation of groups in this context involves the creation of balanced groups in terms of expected performance in addition to maximizing each individual's goal from the collaboration. As the number of formation's constraints grows, forming groups that satisfy these constraints increases in complexity. We know that the Semantic Web (SW) aims at providing a promising foundation for enriching resources with well defined meanings and making them understandable for programs and applications. The potential of the SW in this context has allowed the semantic formation of social networks to be successful [1]. From this point, we trust that the problem of constraint group formation can as well be solved using SW technologies. The question is how to apply the SW vision to the problem, and take the most of its potential to apply it in real life applications such as e-learning. In particular, the problem can be formulated as how can we generate optimal groups by reasoning over possibly incomplete data about the students.

2 Research Overview and Essential Questions

Since forming groups of students with attention to constraint satisfaction is not a simple task for the teacher to do manually, especially for a large number of students, the proposed research is intended to investigate the automation of constrained group formation. In order to cover different types of collaborative activities, we consider the formation of different types of groups including: Teams, Communities of Practice (CoPs), and Social Networks (SNs). We believe that by reasoning on learners' profiles and the teacher's constraints, we can achieve a powerful foundation for automated group formation. With respect to SW concepts, our present and future work intends to give appropriate answers to the following questions: What do we model for the formations of different types of groups? How do we enable the teacher to get the group formation they want? How do we achieve that formation? And how effectively we achieved it? Due to their self-organized nature, for formation of CoPs and SNs to be effective, the instructor has to provide a degree of dynamic self organization within these groups. In this research we address the question of how do we enable the dynamic formation of instructor-initiated CoPs and SNs? If we do not have all the required information about the users, how do we process the formation with incomplete data? Can we find this data or similar data and substitute it to maintain the robustness of the grouping? Where can we get this data, and what type of data should it be? If we substitute the data, how significant is the measurement of the correlations between the required data and the alternative one?

3 Research Methodology

[*] Supervisors: Dr. Hugh C Davis and Dr. Dave E Millard, School of Electronics and Computer Science, University, Southampton, Emails: {hcd, dem}@ecs.soton.ac.uk

[†] Second Year of PhD research, approximate defense time is January 2009.

To answer the research questions and examine the soundness of the assumed hypotheses, we aim at building a Semantic Web based system that allows the instructor to automatically form different types of groups. The formation of the groups generated by the system will then be evaluated based on the quality of the generated groups, and the robustness of the formation in case of incomplete data.

1. Research Implementation: The system will have three main components:

1. The Ontology: called Semantic Learner Profile (SLP), the ontology is an extension of the FOAF vocabulary that aims at providing semantic data about the learner [2] for the formation of all types of groups. Each student has an extended foaf file that can be updated at any time. This allows them to publish data about themselves using a URI, which enables the data to be referred to from any dataset. An interface based on foaf-a-matic (http://www.ldodds.com/foaf/foaf-a-matic.html) will be provided to facilitate the creation of these profiles. Since FAOF allows the users to define their friends, social connections can be made for CoPs and SNs formation. As the students can modify their friends' list at any time, the relationships links between them allow a dynamic formation which provides the groups with a degree of self-organisation.

2. The Instructor Interface: The teachers will be allowed to choose the constraints they want to base the formation on. They will be provided by an option that enables them to set constraints on those values and the relationship between those values. The interface will also enable the instructor to rank the importance of these constraints to enable the system to manage compromises based on these priorities.

3. The group generator: The group generator will be supported by a set of rules that represent different formation algorithms that allows reasoning on the data provided by the learners and the teacher in order to generate effective groups. The system will be empowered by Jena inference engine and SAPRQL for querying over the data. To allow an effective grouping, students are to be encouraged to create meaningful descriptions of themselves with as much details as they can. In case they do not provide all required data for a formation, the instructor will be supported by an option that enables the system to use Semantic Web mining techniques to look for the missing data in the web and form correlations to the required data. Moreover, we need to address the data provenance, especially if it is extracted from blogs and web pages.

2. Research Evaluation: To evaluate the system and hence the research hypotheses, we intend to test the system on real life data by forming groups of students taking a software engineering course (SEG) in the University of Southampton. To ensure the system is tested for different groupings we also use randomly generated data, and a simulated population of students. For this, a person generator is created. The efficiency of the system will be measured based on the quality of the formation provided by the system which involves: to what degree did each generated group meet satisfied the constraints, how many groups satisfied the constraints, and what is the systems confidence in generating successful grouping. The same measures will be applied to evaluate the system's capability to form groups with incomplete data.

4 Current Status

So far, we implemented the SLP ontology, and the random person generator. Both the student interface and simulated data are currently under development. To support the creation of the simulated data and prepare for the evaluation of the semantic formation on the real life data next year, we are currently running an observational study based on two questionnaires one to get information about the student, and the other to evaluate the group formation. The questionnaires are given to the students taking the SEG course this year who have already been grouped manually by the teacher based on their previous grades and gender. This observational pre-study will enable us to compare the results of this manual formation with the automated semantic formation, which is intended to run as a controlled study on the same course next year. Moreover, the pre-study will help in getting information about the students' population for the creation of the simulated data. For our future work, the core components of the semantic formation system are to be implemented so that the hypothesis of the research can be evaluated. Future work will include more research on managing group formation with incomplete data.

5 References

1. Golbeck, J., Parsia, B. & Hendler, J, Trust Networks on the Semantic Web, Proc. of CIA. Helsinki, Finland, 2003.
2. Ounnas, A., Davis, H. C. and Millard, D. E. (2007) Semantic Modeling for Group Formation. In Proceedings of PING workshop at the UM2007, Corfu, Greece.

Combining HTN-DL Planning and CBR to compound Semantic Web Services

Antonio A. Sánchez-Ruiz, Pedro A. González-Calero, Belén Díaz-Agudo

Dep. Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid, Spain
email: {antsanch}@fdi.ucm.es {pedro,belend}@sip.ucm.es

1 Introduction

Semantic Web Services (SWS) are distributed and reusable software components that are described using standard formal languages like SWDL or OWL-S. SWS can be automatically discovered, invoked and combined. Complex applications can be built combining different Web Services and therefore, it is important to provide assisting tools to help in the composition process [6].

Planning techniques can be used to find the flow of services that accomplish a specific task. Several approaches have been tried in software component composition [4], but all of them have a common requirement: the domain must be completely formalized, and this is very difficult in real domains. Case Based Planning (CB Planning) [2] tries to solve this deficiency using cases that represent past experiences, i.e., plans that were used to solve previous problems. On the other hand, HTN-DL planning [5] is a very new approach that combines the power of hierarchical planning with the inference capabilities of Description Logics.

In my thesis I propose to combine CB Planning and HTN-DL to obtain a hierarchical planner that utilizes the best of both worlds.

2 Related work

The problem of web service composition has been studied extensively in recent years [6, 4]. Hierarchical planning (HTN Planning) [1] is a modern type of planning that tries to resolve problems by dividing them into simpler subproblems. HTN planning has been used successfully in complex domains, like SWS composition [7, 3].

HTN-DL [5] is a new HTN extension in which the domain, the problem and the current state are described using an ontology in OWL. HTN-DL works with the Open World Assumption and takes advantage of the inference capabilities of Description Logics (DL) in the planning process. Furthermore, it can work directly form a description in OWL-S of the available SWS.

3 My proposed approach: Case-Based HTN-DL Planning

The main drawbacks of HTN-DL are that it is much slower than classical planning and that needs an exhaustive domain description.

Case-Based Planning [2] adapts cases or past experiences to solve new problems. They key idea is that similar problems usually have similar solutions. The main features of CB Planners are: they can solve problems even without an exhaustive description of the domain because the cases can store implicit knowledge about the domain (maybe the validity of plans can not be checked, but the planner can guest its validity based on previous experiences); they can enhance the performance and accuracy with use, by just learning new experiences (cases); and they use the cases as heuristics in order to find solutions exploring a small part of the search space (these heuristics can improve as more quality cases are available).

In my thesis I propose to combine Case Based Planning and HTN-DL in order to obtain the best of both worlds (CB HTN-DL Planning) and apply these ideas to compound SWS. The main features of this new approach are: it works with the Open World Assumption using the DL inference capabilities; it works directly with the OWL-S descriptions of the SWS; it will be able to work without a complete description of the domain; it can use the cases as heuristic to guide the search and enhance the performance; and the planner will presumably improve the performance and accuracy with use because new cases will be learned.

The thesis will have 3 different parts: the formalization of the planning theory behind CB HTN-DL, the development of an example application in a real environment, and the evaluation of the results.

References

1. K. Erol, J. A. Hendler, and D. S. Nau. UMCP: A sound and complete procedure for hierarchical task-network planning. In *Artificial Intelligence Planning Systems*, pages 249–254, 1994.
2. K. J. Hammond. *Case-based planning: viewing planning as a memory task*. Academic Press Professional, Inc., San Diego, CA, USA, 1989.
3. U. Kuter, E. Sirin, B. Parsia, D. S. Nau, and J. A. Hendler. Information gathering during planning for web service composition. *J. Web Sem.*, 3(2-3):183–205, 2005.
4. J. Peer. Web service composition as ai planning - a survey. Technical report, University of St. Gallen, March 2005.
5. E. Sirin. *Combining description logic reasoning with ai planning for composition of web services*. PhD thesis, University of Maryland, 2006.
6. B. Srivastava and J. Koehler. Web service composition — current solutions and open problems. In *ICAPS 2003*, 2003.
7. D. Wu, B. Parsia, E. Sirin, J. A. Hendler, and D. S. Nau. Automating daml-s web services composition using shop2. In *International Semantic Web Conference*, pages 195–210, 2003.

Ontology Alignment Specification Language

François Scharffe
francois.scharffe@deri.org,

Digital Entreprise Research Institute
University of Innsbruck

Abstract. Ontology mediation is one of the key research topics for the acomplishment of the semantic web. Different tasks can be distinguished under this generic term: instance transformation, query rewriting, instance unification, ontology merging or mapping creation. All first four tasks require a mapping specification between the ontologies to be mediated. Mapping creation using tools and algorithms is outputting such a specification. We argue in this thesis proposal that a specific language to express mapping specifications is needed. This proposal presents arguments why such a language is needed, introducing particularly the concept of mapping patterns, based on a study of the frequent mismatches arising when trying to mediate between ontologies. Such a language is then proposed and its applicability is demonstrated for three scenarios: a graphical tool for ontology mapping, an output format for ontology matching algorithms and a merging algorithm. We also give first results on the language design, mainly represented by an alignment ontology.

1 Contributions of the proposed thesis

With this thesis we expect to achieve the following goals:

- Having a language able to model ontology mappings patterns.
- Having demonstrated this language is of practical use by using it for concrete mediation tasks.
- Providing usable tools around the language (API, patterns library)
- Having the mapping language being used as a standard way for representing and exchanging ontology mappings

The language referenced in this proposal is already used by two ontology mapping tools: a graphical mapping editor a text editor to edit mapping documents. The mapping language is compatible and extends the ontology alignment format used as part of the Ontology Alignment Evaluation Initiative[1]. The Alignment Format allows to express simple mappings while the mapping language is more expressive. A few algorithms are actually able to detect such complex mappings. We developped a merging algorithm[1] able to automatically merge a set of ontology in a network, given one to one mappings between ontologies. As today the language is reaching some maturity given the feedback of the graphical mapping tool implementation and the support of the ontology alignment format. Next steps are given in Section 3.

[1] http://oaei.ontologymatching.org

2 Results

We defined based on a list of requirements a mapping language specified in semantic web standard RDF, and using an OWL ontology in its last version. We also maintain a Lisp-style syntax, more convenient to be read. Due to limited place we will not present the syntax in this document but strongly encourage the reader to look at the language specification and ontology[2]. A Java API providing methods to parse and export mapping documents as well as giving an in-memory representation is under development. We also developped a library of common mapping patterns and currently adapt it to rdf graph patterns. This library is available in [2]. Parts of this work are published in [3, 4, 1, 5].

3 Conclusions and Future Work

We have studied a set of requirements a mapping language should have and compared state of the art formalisms to represent mapping with this requirements. From this we have designed an ontology mapping language answering the given requirements at best. We actually have a rather stable syntax for the language and propose a library of common mapping patterns and an API to deal with mapping language constructs. We also aligned the language with the Alignment Format and propose an algorithm to merge a set of ontology on provided mappings. We currently work on a SPARQL based mediation engine, able to transform instances from one ontology to another at the RDF level. We plan to finish this thesis before the end of the current year and therefore need to start the writing task soon. We also plan to push the mapping language towards standardization.

References

1. Scharffe, F.: Dynamerge: A merging algorithm for data integration on the web. In: Proceedings of the Conference on Database Systems for Advanced Applications, SWIIS workshop. (2007)
2. Scharffe, F., de Bruijn, J., Foxvog, D.: D4.2.2 ontology mediation patterns library v2. Technical report, SEKT (2005)
3. Scharffe, F., de Bruijn, J.: A language to specify mappings between ontologies. In: Proc. of the Internet Based Systems IEEE Conference (SITIS05). (2005)
4. Scharffe, F.: Instance transformation for semantic data mediation. In: Proc. of the Int. Semantic Web and Web Services Conference SWWS'06. (2006)
5. Scharffe, F.: Schema mappings for the web. In: Proc. of the Int. Semantic Web Conference ISWC'06. (2006)

[2] **http://www.omwg.org/TR/d7/**