Knowledge Gardening in the Web of Data

Trond Aalberg[1], Fabien Duchateau[2], Mohand-Said Hacid[2], Nicolas Lumineau[2], and Naimdjon Takhirov[1]

[1] Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
{takhirov,trondaal}@idi.ntnu.no
[2] Université Lyon 1, LIRIS, UMR5205, Lyon, France
{fduchate,mhacid,nluminea}@liris.cnrs.fr

1 Introduction and Context

Cultural institutions have the crucial goal of storing, managing and disseminating knowledge about our culture to a worldwide audience of people. Museums, libraries, archives and other vendors of cultural heritage information are currently in the process of exposing their data as Linked Open Data (LOD) to encourage reuse of this information in semantic aware services which utilize the synergy that can be achieved when combining and integrating complimentary information resources. Linked Open Data is based on the the standards and formats that have been developed for the Semantic Web, but even though we have the technological framework, there are many challenges that must be solved to achieve true automatic integration and interoperability between a potentially very large number of loosely connected information sources.
The main objective of this project **KOGAR** (**Kno**wledge **Gar**dening in the Web of Data) is to support the process of integration and alignment of heterogeneous, but related information sources within the area of cultural heritage. This project aims at proposing a novel communication protocol to support the process of semantic mappings in order to ease the interconnection between a large number of semantic data providers. The performance of this protocol will be evaluated and will be compared.

Keywords: Semantic Web, Linked Open Data, Peer-to-Peer, Cultural Heritage, Knowledge Gardening, Web of Data

2 Project Description

This part covers the scientific objectives and the organization of the project KOGAR.

2.1 Scientific Objectives of the Collaboration

To infer new knowledge and to target a broader audience in our context, we could believe that the Linked Open Data cloud[3] (LOD), which nowadays contains hundreds of interconnected knowledge bases exposing billions of semantic triples, is a key solution for the cultural institutions. Yet, the current version of the Linked Open Data cloud suffers from a centralized point of view. Indeed, DBpedia[4] is considered as the nucleus of the LOD, and many LOD

[3] LOD cloud, `http://linkeddata.org/`

[4] DBpedia, `http://www.dbpedia.org/`

knowledge bases only link their entities to the corresponding one from DBpedia. In addition, the data stored on LOD is seldom updated. Such centralized environment is not appropriate when a large number of distributed and autonomous data sources are continuously updating their data. Rather, all cultural institutions can be seen as a large peer-to-peer network : each institution is a peer which exposes and manages its own data and it can query the data sources of the other peers to provide advanced navigation and semantic enrichment for users. In this vision, the mapping, i.e., a correspondence between two entities, is a crucial concept that serves as a basis in our network. Thus, the objectives of our collaboration are threefold:

- Demonstrate by simulation on real data the feasibility of an interconnected semantic peer-to-peer network with a novel communication protocol based on the mappings
- Design efficient methods and techniques for gardening the network (i.e., facilitate the tagging and annotating, connect entities on-the-fly, maintain the consistency of the data, evaluate the degree of trust of each peer, etc.) and for improving the quality of query answers
- Deploy this network and use the associated methods in the cultural heritage domain. The Norwegian team owns data from the National Norwegian Library which have been converted into the semantic FRBR[5] model

The main problems that we tackle in the context of the KOGAR project can be summarized as follows:

- **Large scale aspect.** In practice, the approach has to be scalable due to the potential large number of peers (museums, libraries, LOD data sources, etc.)
- **Dynamicity.** As each peer can join or update its data anytime or quit the network, the different techniques to be proposed have to feature the incremental property. In other words, a new peer first discovers some mappings with a few peers, and these mappings are refined and extended later
- **Entity matching.** Two institutions can have data about the same entity. In this case, an equivalence link between these entities must be "drawn"
- **Semantic integration.** Detecting an equivalence link between two entities is not sufficient for some applications. The two entities should fusion, i.e., their attributes and associated values have to be merged when necessary. The quality of this semantic integration is crucial, for instance the respect of the completeness and minimality criteria
- **Knowledge gardening.** In a self-organizing network, the task of knowledge gardening refers to the discovery of inconsistencies, using reasoning for instance, to the storage of the alignment for reuse, to assess the degree of trust between peers, etc.

2.2 Work Planning

Table 1 provides an overview of the planning for the project. During the first three months, the teams will study the most recent works in the different domains (ontology alignment, distributed semantic systems, knowledge sharing for cultural heritage). A first exchange will allow us to share and set up ideas to facilitate the dispatching of the work between the teams. The second step of the project deals with the implementation part, and since Naimdjon

[5] Functional Requirements for Bibliographic Records

Takhirov and Fabien Duchateau are the main developers of the project, an exchange is planned during that period. Another exchange will enable us to finalize the writing of the papers. The last month is dedicated to the evaluation of our work to detect the opportunities for submitting new proposals.

Starting Date	Ending Date	Activity
01/01/2013	31/03/2013	State of the art and evaluation of similar works • *Exchange (Norwegian team traveling to France): dispatching upcoming work*
01/04/2013	30/11/2013	Implementation of the chosen solutions, validation, writing of papers • *Exchange (Fabien Duchateau traveling to Norway): implementation* • *Exchange (French team traveling to Norway): writing papers*
01/12/2013	31/12/2013	Evaluation of our work and perspectives in terms of research and funding • *Exchange (Norwegian team traveling to France): writing the final report and new proposals*

Table 1. Organization of the project

2.3 Expected Results

The expected results of this KOGAR project are:

- Demonstrate by simulation on real data the feasibility of an interconnected semantic peer-to-peer network with a novel communication protocol based on the mappings, without any central knowledge base and which enables the dynamicity of the data sources
- Compare the performance (completeness, mappings quality, etc.) of our proposal with existing solutions, and interact with the SEALS project[6] for the evaluation part
- Submit at least two papers in well-ranked conferences: the former will present our approach to incrementally align ontologies and refine the mappings over time. The latter would demonstrate the feasibility of our system, especially the P2P network based on the mappings. The targeted conferences are in three research areas, the Semantic Web (e.g., ESWC, ISWC), Digital Libraries (e.g., JCDL, TPDL) and Data Management (e.g., ICDE, CIKM)

We plan to organize and host a small workshop on Knowledge Gardening, locally in Lyon. The technical motivation for a such event is to contribute to disseminate our work by sharing results with experts in the field, as well as to inseminate such experts in the field with our own results and perspectives.

A successful project would encourage us to submit new types of funding proposals at the European level for instance. In that case, the experience gained from this project will give us more credibility when searching for partners such as European laboratories and cultural institutions such as museums and libraries.

3 Participants

This part of the proposal describes the participants involved in the project and the existing cooperation.

[6] `http://www.seals-project.eu/`

3.1 Context of the Cooperation and Existing Relationships

In 2010, ERCIM[7] granted Fabien Duchateau with a 18-months postdoctoral research fellowship. Half of this fellowship was conducted in Trond Aalberg's team at NTNU. During that period, Trond, Naimdjon and Fabien closely worked on the application of semantic technology for cultural heritage. More specifically, they have demonstrated the possibility to efficiently convert legacy cultural data into a semantic model called FRBR [8, 9]. They have also proposed an approach for linking entities to the Linked Open Data cloud both for reusability and for semantic enrichment [2, 7].

Since September 2011, Fabien Duchateau was hired as associate professor at the Lyon University, where he joined Mohand-Said Hacid and Nicolas Lumineau in the database research group. This group mainly aims at studying data management issues in distributed environments at a conceptual level [4] and at a query level [3]. They have demonstrated the interest of exploiting the collaboration between peers to align a large number of OWL Ontologies distributed through a P2P network [5, 6]. Moreover, they contributed to the Tarchna Project[8] (European project: culture 2000) that has contributed to ease the accessibility of archaeological data. In the meanwhile, the ongoing cooperation between Trond Aalberg's team and Fabien Duchateau led to the development of a knowledge extraction tool to populate a semantic knowledge base with binary relationships (paper under submission to CIKM).

3.2 Interest in the Collaboration and Synergy of the Teams

The Norwegian team has an expertise in digital libraries and cultural heritage. The main focus of the research has been within the various areas of digital libraries: architecture of digital libraries, collection development, modeling and linking of information objects, information seeking and retrieval, and how to use and integrate Digital Libraries into a working environment. The group has been an active participant in the EU funded DELOS cooperation[9] since the beginning. The objective of the DELOS Working Group is to promote research into the further development of digital library technologies. DELOS Network of Excellence aiming to provide an open context in which an international research agenda for future research activities in the digital libraries domain can be developed and continuously updated.

On the French side, the data management problem is a primary topic of interest. More specifically, the French participants have been involved in projects dealing with the schema matching and the ontology alignment issues [1, 5]. These works were mainly part of projects at different levels, for instance the National FORUM[10] project (2005-2008) or the EU Tarchna project, which has revealed the significant concerns and needs from the cultural heritage scientists to preserve and update their data.

Finally, both teams have members working in the Semantic Web domain, thus strengthening the synergy of the cooperation. Indeed, the main issue to be solved deals with the ontology alignment, the semantic integration and the consistency of the semantic data, which involve reasoning over this data.

[7] European Research Consortium for Informatics and Mathematics, `www.ercim.eu`

[8] `http://books.google.fr/books?id=QRJWzNl6ymOC&printsec=frontcover`

[9] `http://www.ercim.org/delos/`

[10] `http://www.lirmm.fr/FORUM/`

3.3 Planned Exchanges between Teams

The main exchange deals with the researchers. Similarly to the visiting postdoctoral visit of Fabien Duchateau in 2010, we believe that the proximity of the researchers enables a greater achievement step towards high-impact contributions. In other words, such an exchange has a significant impact for focusing on the issues described in the project while fostering discussions and novel ideas. There is no planned exchange in terms of material and documentation.

3.4 Details of the Work during Collaboration

The project will be supervised by Trond Aalberg and Fabien Duchateau. Naimdjon Takhirov will be the main developer on the project. He will closely work with Fabien Duchateau and Nicolas Lumineau using collaboration tools both for the implementation and the writing of the papers. Trond Aalberg and Mohand-Said Hacid will mainly participate in the writing of the papers and the sharing of their experiences and knowledge in their respective domains.

4 Perspectives

In terms of European perspectives, our project is related to the topics of the 7^{th} framework. Therefore, we could submit in one or two years a bigger project which would involve more partners. In that case, the European project *Tarchna* led by Mohand-Said Hacid could facilitate the identification of these potential partners.
In terms of industrial perspectives, the oil and energy companies - especially in Norway - have many standards for documenting their data. However, there is currently a need of alignment for sharing some of these information. Thus, they are looking for techniques to add a semantic layer to their data in order to expose them. The results of our project could provide these companies with some useful semantic data solutions for adaptation and communication.

References

1. Z. Bellahsene, S. Benbernou, H. Jaudoin, F. Pinet, O. Pivert, F. Toumani, S. Bernard, P. Colomb, R. Coletta, E. Coquery, F. de Marchi, F. Duchateau, M.-S. Hacid, A. Hadjali, and M. Roche. Forum: a flexible data integration system. *SIGMOD Record*, 39(2):11–18, 2010.
2. F. Duchateau, N. Takhirov, and T. Aalberg. Frbrpedia: a tool for frbrizing web products and linking frbr entities to dbpedia. In *Joint Conference on Digital Libraries*, 2011.
3. Y. Gripay, F. Laforest, F. Lesueur, N. Lumineau, J.-M. Petit, V.-M. Scuturici, S. Sebahi, and S. Sabina. Colistrack: Testbed for a pervasive environment management system. *15th International Conference on Extending Database Technology (EDBT) [demo]*, 2012.
4. N. Lumineau, F. Laforest, Y. Gripay, and J.-M. Petit. Extending conceptual data model for dynamic environment. *Conceptual Modeling - ER 2011, 31st International Conference, ER 2012, Florence, Italy, 2012.*, 2012.
5. N. Lumineau and L. Médini. Simtole : un simulateur p2p dédié à l'alignement d'ontologies à large échelle. In S. B. Yahia and J.-M. Petit, editors, *EGC*, volume RNTI-E-19 of *Revue des Nouvelles Technologies de l'Information*, pages 633–634. Cépaduès-Éditions, 2010.
6. L. Médini, N. Lumineau, N. Dieudonné, and L. Vallier. Ontology alignment in browser-based p2p nodes (demo). *23ième Journées Francophones sur d'Ingénieurie des Connaissances (IC'2012), Paris, France. 2012.*, 2012.
7. N. Takhirov, F. Duchateau, and T. Aalberg. Linking frbr entities to lod through semantic matching. In Springer, editor, *Theory and Practice of Digital Libraries*, 2011.
8. N. Takhirov, F. Duchateau, and T. Aalberg. Supporting frbrization of web product descriptions. In Springer, editor, *Theory and Practice of Digital Libraries*, 2011.
9. N. Takhirov, F. Duchateau, T. Aalberg, and M. Zumer. Frbr-ml: A frbr-based framework for semantic interoperability. In I. Press, editor, *Journal of Semantic Web*, 2012.