A Meta-data Driven Platform for Semi-automatic Configuration of Ontology Mediators

Manuel Fiorelli, Maria Teresa Pazienza, Armando Stellato

University of Rome, Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
{fiorelli, pazienza, stellato}@info.uniroma2.it

Abstract

Ontology mediators often demand extensive configuration, or even the adaptation of the input ontologies for remedying unsupported modeling patterns. In this paper we propose MAPLE (MAPping Architecture based on Linguistic Evidences), an architecture and software platform that semi-automatically solves this configuration problem, by reasoning on metadata about the linguistic expressivity of the input ontologies, the available mediators and other components relevant to the mediation task. In our methodology mediators should access the input ontologies through uniform interfaces abstracting many low-level details, while depending on generic third-party linguistic resources providing external information. Given a pair of ontologies to reconcile, MAPLE ranks the available mediators according to their ability to exploit most of the input ontologies content, while coping with the exhibited degree of linguistic heterogeneity. MAPLE provides the chosen mediator with concrete linguistic resources and suitable implementations of the required interfaces. The resulting mediators are more robust, as they are isolated from many low-level issues, and their applicability and performance may increase over time as new and better resources and other components are made available. To sustain this trend, we foresee the use of the Web as a large scale repository.

Keywords: ontology matching, metadata, language resource

1. Introduction

Mediation (Widerhold, 1994) acknowledges the "autonomy and diversity" of the networked information systems, thus enabling domain and application specialization, and allowing the evolution of the overall system by local experimentation of new schemas. As number and variety of available information sources increase, traditional data integration based on mediated schemas is superseded by the novel concepts of dataspace (Franklin, et al., 2005) and ongoing integration (Madhavan , et al., 2007). Integration becomes a sort of "background process" that reconciles data sources as a tighter connection between them is required for servicing user requests. Accordingly, while deploying knowledge representation techniques at the Web scale, the Linked Data (Berners-Lee, 2006) community does not aim at constructing a knowledge base, rather evolving the Web into a global dataspace (Heath & Bizer, 2011).

As it is not possible nor desirable to completely eliminate heterogeneity in large Web-scale systems, ontology mediation (Euzenat & Shvaiko, 2007) is an essential part of the Semantic Web (Shadbolt, et al., 2006; Berners-Lee, et al., 2001). The reconciliation of different ontologies is primarily driven by their linguistic grounding, which best reflects their intended meaning. More advanced methods combine structural, extensional and semantic features. Beyond information within the input ontologies, some approaches exploit external sources of information, such as the Web, Wikipedia, domain corpora, lexical resources and upper-ontologies.

Unfortunately, the emphasis on performance in shared test cases has somehow overshadowed the importance of the tasks that need to be accomplished beforehand, in order to produce quality alignments under the uncontrolled conditions found in real-world scenarios. In these circumstances even state-of-the-art techniques may fail, if their implementations are unable to completely exploit the information contained in the ontologies (Ritze & Eckert, 2012). Selecting the right mediator may be hard, as well, since results in evaluation campaigns hardly correlate with performance and ease of use in real applications.

In this paper we propose MAPLE (MAPping Architecture based on Linguistic Evidences), an architecture and software platform that promotes the development of mediators, which use external linguistic resources and access the ontologies through uniform interfaces. Automatic reasoning on various (linguistic) metadata supports the configuration of promising mediation processes, by assessing the usefulness of ontology mediators for each scenario, and by providing bindings for the required dependencies of the chosen mediator. Nonetheless, the user may actively participate in the configuration process, by validating and otherwise customizing the suggestions produced by MAPLE.

2. Motivation

Ontology mediation is the task of finding an alignment (i.e., a set of conceptual correspondences) between semantically related resources within two ontologies.

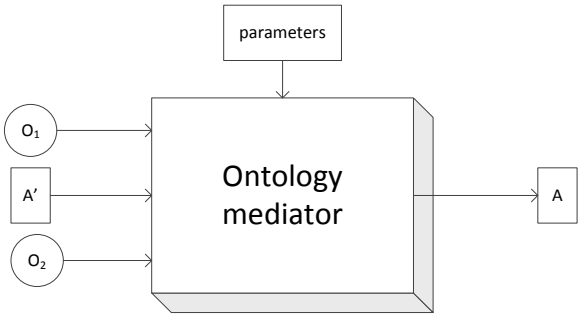

Figure 1: Black-box model of an ontology mediator

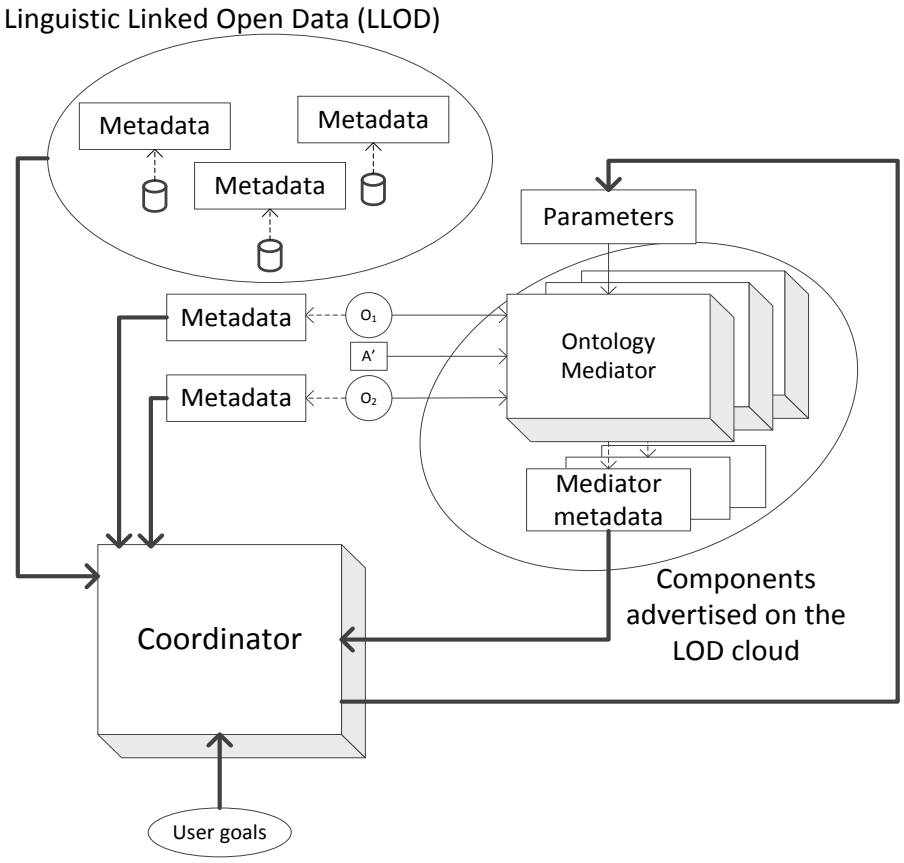

Figure 2: Wider perspective on the ontology mediation task

Figure 1 represents the black-box model of a generic ontology mediator conforming to the above definition. Notably, this model allows to start from a pre-computed alignment, or to configure a mediator by means of some parameters. Indeed, the interpretation of these parameters and the proper assignment of values to them are two challenging problems for the actual use of ontology mediators.

Going inside the black-box, mainstream research produced numerous algorithms, in order to raise the performance in shared tasks. The Ontology Alignment Evaluation Initiative[1] (OAEI) has organized annual evaluation campaigns, since 2004. The results of recent campaigns suggest that new paths have to be explored (Pavel & Euzenat, 2013) in order to continue making significant improvements in the near future. As no system clearly outperforms the others in all scenarios, the user has to choose the appropriate one for each situation. Unfortunately, it has been shown (van Ossenbruggen, et al., 2011) that shared test cases often do not help very much in predicating the behavior of the evaluated approaches, while it appears that in many real-world scenarios even simple mediators based on label matching may lead to effective (semi-automatic) solutions (Caracciolo, et al., 2012). Furthermore, even state-of-the-art techniques are useless if their implementations are unable to understand modeling patterns found in the input ontologies, in order to exploit their content.

In our opinion it is possible to overcome the above limitations, by following an alternative path. Instead of designing yet another mediation algorithm, we widened our perspective on the topic to embrace the environment and the tasks that prelude to a successful mediation strategy (see Figure 2).

From this viewpoint, we must solve a coordination task, i.e. finding the mediator that will achieve the user goals against the given ontologies, and properly configuring it to deal with the specific mediation scenario. The MAPLE coordinator performs this task, by exploiting metadata about the input ontologies, the mediators, the linguistic resources and other components. While the set of relevant features is probably endless, we focus mostly on linguistic aspects. In fact, our proposal builds on, extends and updates our previous work on linguistic coordination of communicating agents (Pazienza, et al., 2007). While the underlying principles remain the same, the current work differs substantially for its grounding in the Linked Data world and the shift to component-based architectures.

3. MAPLE Distributed Architecture

The MAPLE architecture (Figure 3) combines the principles of Linked Data for Web scale distribution, with the OSGi[2] specification for the modularization of Java systems.

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

[2] http://www.osgi.org/

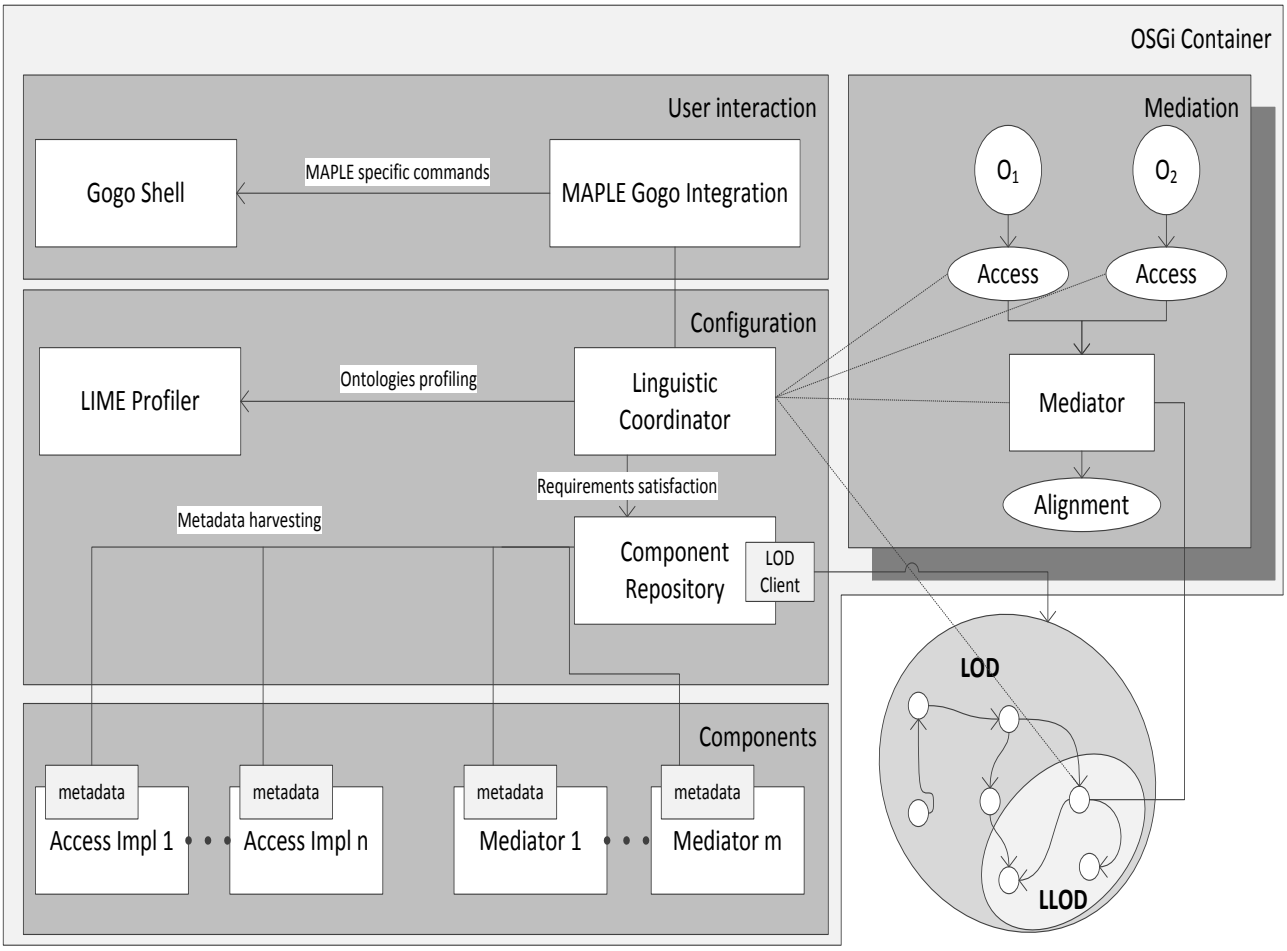

Figure 3: MAPLE distributed architecture

At the *user interaction* level MAPLE provides the user with an interactive shell, based on Apache Felix Gogo[3]. The shell offers an environment for interacting with MAPLE through a sophisticated yet concise command language. While the shell itself provides the core commands, MAPLE leverages its extendibility for providing additional commands related to its own capabilities.

These capabilities are mostly realized at the *configuration* level, which includes the Linguistic Coordinator. It is the component in charge of the coordination task, which embraces the following activities: become aware of the given ontology mediation scenario, choose the suitable ontology mediator, and configure it with required linguistic resources and concrete implementations of the interfaces provided by MAPLE.

The use of external linguistic resources (possibly retrieved from the Web) promotes the development of generic mediators, which are not constrained by the set of prepackaged resources. Furthermore, MAPLE provides the mediators with various interfaces, which support the access to the input ontologies at a more generic and abstract level, while hiding most of the details about their internal and often idiosyncratic organization. The use of these interfaces guarantees greater resilience, as MAPLE would supply the right implementation for each scenario.

The access interfaces are classified with respect to the kind of information they provide access to, i.e. conceptual and linguistic. The conceptual access may be either a flat list of the available concepts, or a hierarchical view of the underlying taxonomy. The linguistic access concerns, at very minimum, with the lexicalizations in a natural language, while a richer view might deal with differences at the terminological correspondence level, i.e. preferred vs alternative vs hidden labels.

These two orthogonal concerns are associated with distinct access interfaces, which can be lately combined together. In fact, a typical mediator would use a conceptual access interface for browsing the concepts, while depending on a linguistic access interface for obtaining their lexicalizations.

Obviously, the right implementation of these interfaces depends on the specific data models and modelling patterns found in the input ontologies. Moreover, there are circumstances in which some interfaces are not applicable, since the given ontologies do not contain the necessary information.

The Linguistic Coordinator performs its task, by using metadata about the input ontologies, the mediators, the third-party linguistic resources and the implementations of the aforementioned interfaces. These metadata are represented in RDF (Klyne, et al., 2004) through

[3] http://felix.apache.org/site/apache-felix-gogo.html

specialized vocabularies developed by us, and made available through SPARQL (Harris & Seaborne, 2013) endpoints. However, the input ontologies are usually profiled to automatically generate their descriptions. In fact, the LIME Profiler provides an alternative path, for handling the case, very common until the wide acceptance of our metadata vocabulary, in which the input ontologies are not provided with proper metadata by their publishers. In addition to remote SPARQL endpoints, the Component Repository manages a local endpoint, containing metadata harvested from locally installed components.
The Linguistic Coordinator delegates to the Component Repository the task of satisfying requirements (expressed in terms of the metadata vocabulary) through components either locally installed or publically available on the Web.

4. Metadata

A unified vocabulary, called LIME (LInguistic MEtadata) (Fiorelli, et al., 2013), concerns with the linguistic expressivity of ontologies and the characteristics of third-party linguistic resources. As the aim of LIME is to describe ontologies and other RDF datasets as a whole, we developed it as an extension of VoID[4], which provides an extensible framework for describing interlinked RDF datasets. Accordingly, the LIME metadata are attached to an instance of void:Dataset, which is a proxy that stands for the dataset being described.
The primary fact we are interested in is the set of supported natural languages, each one represented as a value of the property lime:language. This information roughly indicates the linguistic compatibility of the input ontologies, and the usefulness of a linguistic resource for aligning them. Furthermore, the metadata should describe the existence of links to linguistic resources, which can be regarded as a less ambiguous inter-lingua. We are also interested in various statistics about this linguistic content. For instance, the following RDF fragment expresses that in a given dataset (:dat) the 75% of owl:Classes has at least one lexicalization in Italian, and that on average they have 1.75 such lexicalizations:

```
:dat lime:languageCoverage [
  lime:lang "it";
  lime:resourceCoverage [
    lime:class owl:Class  ;
    lime:percentage 0.75 ;
    lime:avgNumOfEntries 1.75
  ]
] .
```

The property lime:linguisticModel holds each lexicon model for charactering an ontology in natural language, e.g. RDFS (Klyne, et al., 2004), SKOS (Miles & Bechhofer, 2009), SKOS-XL, OntoLex[5].
In contrast, we assume that the linguistic resources are represented in RDF complying with the OntoLex upper-model. In fact, the publication of linguistic resources through the Linked Open Data principles is leading to the formation of a subset of the LOD cloud, known as the Linguistic Linked Open Data cloud (LLOD) (Chiarcos, et al., 2012). Actually, we foresee to use the LLOD as a distributed repository of third-party linguistic resources.
The metadata about mediators include their compatibility with specific classes of mediation scenarios, the exploitation of different aspects of the linguistic and conceptual content of the input ontologies, and the use of linguistic resources.
While the architecture defines a finite set of interfaces for accessing the input ontologies, the set of implementations is in fact open-ended, as they can be supplied by extensions. These extensions must describe the capabilities of the provided implementations through dedicated metadata.
Both ontology mediators and implementations of MAPLE interfaces are packaged as OSGi bundles, possibly available in Web accessible OBR[6] repositories. Therefore, they require additional deployment metadata for the installation of the OSGi bundle, and the instantiation of concrete objects out of it.

5. Coordination Strategy

The Linguistic Coordinator is the component that assists the user in the configuration of a promising mediation process. Towards that goal, the coordinator implements a configuration workflow (Figure 4), which exploits the metadata introduced in the previous section, and encourages the participation of the end-user.
By first, the coordinator figures out the kind of mediation problem faced by the user, by considering the metadata about the input ontologies. If these metadata are not readily available, the coordinator depends on the LIME profiler, for automatically producing most of the metadata by means of statistical methods.
The metadata about the linguistic characterization of the input ontologies support the classification of the given ontology mediation scenario with respect to the degree of linguistic heterogeneity. When the input ontologies share the support for zero, one or more natural languages, the mediation scenario is classified as cross-lingual, monolingual and multilingual, respectively. Similar considerations apply to links to linguistic resources.
The metadata about the data models for the representation of linguistic and conceptual knowledge support the discovery of implementations, if any, of the interfaces defined by MAPLE. For instance, RDFS does not include the distinction between preferred and alternative labels, therefore the interface concerning with this distinction has no implementation for the RDFS lexicon model.
The coordinator formulates SPARQL queries, for discovering mediators that can handle the given mediation scenario, and exploit most of the available information (e.g., use the "richest" interfaces). These queries are executed by the Component Repository against the provided SPARQL endpoints. In addition to remote endpoints, there is a local one, which allows the retrieval of already installed mediators. Actually, the query is formulated in such a way that relaxed matches are allowed. For instance, while the input ontologies share the support for more than one natural language, monolingual mediators are acceptable, as well. However, the closer a mediator matches the coordinator needs, the higher it will be ranked in the list provided to the user for choosing the mediator to use.

[4] http://www.w3.org/TR/void/
[5] http://www.w3.org/community/ontolex/
[6] http://felix.apache.org/site/apache-felix-osgi-bundle-repository.html

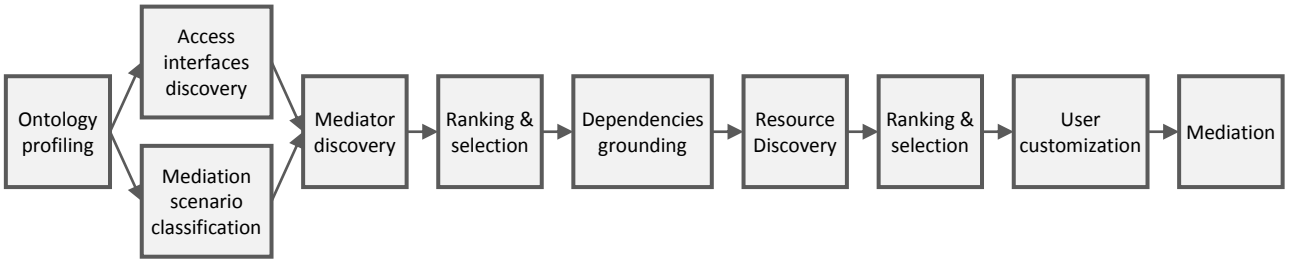

Figure 4: MAPLE Configuration Workflow

Once the user selected a mediator, the Component Repository is in charge of its deployment in the local execution container.

The process continues with the configuration of the mediator, by supplying it with resources that satisfy its requirements.

The first class of requirements includes interfaces defined by the architecture for abstracting the detailed organization of both the conceptual and the linguistic content. The platform is in charge of identifying and deploying the appropriate implementation of the interfaces required by the mediator.

Another class of requirements consist in the use of external linguistic resources. Mediators do not embed linguistic resources, nor do they specify exactly the external resources they depend on. In fact, mediators declare their dependency on generic classes of linguistic resources introduced by LIME, e.g. a bilingual dictionary. Moreover, these requirements are parameterized with respect to the characteristics of the input ontologies, e.g. a bilingual dictionary between the languages used in a cross-lingual mediation. Later, when a mediator is applied to a specific problem, its requirements are grounded by substituting the metadata of the input ontologies for the parameters, e.g. an Italian-English bilingual dictionary, if mediating an Italian ontology and an English one. These ground requirements are turned into SPARQL queries, for retrieving suitable linguistic resources. Discovered resources are ranked according to various criteria (e.g., their statistical footprint), and presented to the user for the ultimate choice.

At this stage, the mediator is configured, its dependencies are bound to concrete resources, and the user can further customize it before the actual mediation takes place.

6. Related Works

The Alignment API (David, et al., 2011) provides abstractions for alignments, matchers and evaluators. The reference implementation has two interesting dependencies: Ontowrap[7] and OntoSim[8]. The former relates to our access API as it provides various interfaces for different levels of interaction with an ontology, e.g. interacting with the hierarchy or not. OntoSim is a library for computing similarities between ontologies and ontology elements. While our abstractions overlap with those defined by the Alignment API, and in fact we directly use some of them, we provide a dynamic platform for deployment, configuration and execution of components, automatically bound to third-party resources.

Amalgame (van Ossenbruggen, et al., 2011) supports the interactive composition of mediators, filters and other components in complex workflows. For the moment, we focus on the configuration of individual mediators, rather than assembling complex workflows.

SEALS (Wrigley, et al., 2012) defines standards for describing, packaging, publishing and executing components related to semantic technologies. However, being focused on the systematic evaluation of these technologies, SEALS should not provide services that help them in performing their task.

R2R (Bizer & Schultz, 2010) is an RDF data translation architecture based on the composition of alignments made available in the LOD. While similar to our proposal in the emphasis on runtime discovery of resources, actually we complement R2R by seeding its composition algorithm with the necessary primitive alignments.

MOMA (Mochol & Jentzsch, 2008) is an architecture closely related to our proposal by the usage of metadata for the selection of ontology mediators. Our proposal differs substantially in the automatic provisioning of third-party linguistic resources and implementations of the interfaces hiding the details of conceptual and linguistic modeling. Input ontology metadata supporting MOMA and our approach are related to the concept of *schema feature* in works on self-configuring matching systems (Peukert, et al., 2012). However, we focus on metadata which express compatibility conditions amenable to symbolic manipulation, rather than generic clues for quantitative algorithms.

7. Conclusion and Future Works

MAPLE assists users in the definition of effective mediation processes, while developers are isolated from many low-level issues and encouraged to use linguistic resources in their algorithms.

By homogenizing the problem space, our methodology increases the robustness of ontology mediators, while the dependency on generic external linguistic resources (rather than embedded ones) guarantees greater flexibility.

At the time of writing the implementation of the proposed platform is not finished yet, as we are completing the infrastructure for accessing linguistic resources, and discovering both linguistic resources and other components from the Web. As a consequence of a Web scale discovery mechanism, we expect that applicability and performance of ontology mediators would increase over time as new and better linguistic resources and other components are made available. In fact, the impact of our

[7] http://alignapi.gforge.inria.fr/ontowrap.html
[8] http://ontosim.gforge.inria.fr/

proposal is largely influenced by the growth of the LLOD cloud, and by the adoption of LIME for providing metadata. Towards this goal, we proposed our vocabulary as the starting point for the development of a metadata module for the OntoLex specification.

8. Availability

At the time of writing we are working on the reference implementation of MAPLE, which is available at: http://art.uniroma2.it/maple/

9. Acknowledgements

This research has been partially supported by the EU project SemaGrow (Grant agreement no: 318497).

10. References

Berners-Lee, T. (2006, July 27). *Linked Data.* Retrieved October 6, 2013, from w3.org: http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J. A., & Lassila, O. (2001). The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American, 279*(5), 34-43.

Bizer, C., & Schultz, A. (2010). The R2R Framework: Publishing and Discovering Mappings on the Web. *Proceedings of the First International Workshop on Consuming Linked Data.* CEUR-WS.org.

Brickley, D., & Guha, R. V. (2004, February 10). *RDF Vocabulary Description Language 1.0: RDF Schema.* (B. McBride, Ed.) Retrieved March 22, 2011, from World Wide Web Consortium (W3C): http://www.w3.org/ TR/rdf-schema/

Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Keizer, J., & Jacques, Y. (2012, August Tuesday, 14). Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. *International Journal of Metadata, Semantics and Ontologies (IJMSO), 7*(1), 65-75. doi:10.1504/IJMSO.2012.048511

Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). *Linked Data in Linguistics.* Springer.

David, J., Euzenat, J., Scharffe, F., & Trojahn dos Santos, C. (2011). The Alignment API 4.0. *Semantic Web Journal, 2*(1), 3-10.

Euzenat, J., & Shvaiko, P. (2007). *Ontology Matching.* Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Fiorelli, M., Pazienza, M., & Stellato, A. (2013). LIME: Towards a Metadata Module for Ontolex. *2nd Workshop on Linked Data in Linguistics: Representing and Linking lexicons, terminologies and other language data.* Pisa, Italy.

Franklin, M., Halevy, A., & Maier, D. (2005, December). From databases to dataspaces: a new abstraction for information management. *SIGMOD Record, 34*(4), 27-33. doi:10.1145/1107499.1107502

Harris, S., & Seaborne, A. (2013, March 21). *SPARQL 1.1 Query Language.* Retrieved March 09, 2014, from World Wide Web Consortium - Web Standards: http://www.w3.org/TR/sparql11-query/

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology, 1*(1), 1-136.

Klyne, G., Carroll, J. J., & McBride, B. (2004, February 10). *Resource Description Framework(RDF):Concepts and Abstract Syntax.* Retrieved from Resource Description Framework(RDF) :Concepts and Abstract Syntax,W3C Recommendation: http://www.w3.org/TR/rdf-concepts/

Madhavan , J., Jeffery , S. R., Cohen , S., Dong , X., Ko, D., Yu, C., & Halevy, A. (2007). Web-scale Data Integration: You can only afford to Pay As You Go. *Proceedings of 7th Biennial Conference on Innovative Data Systems Research (CIDR2007).*

Miles, A., & Bechhofer, S. (2009, August 18). *SKOS Simple Knowledge Organization System Reference.* Retrieved March 09, 2014, from World Wide Web Consortium (W3C): http://www.w3.org/TR/skos-reference

Mochol, M., & Jentzsch, A. (2008). Towards a Rule-Based Matcher Selection. *Proceedings of the 16th International Conference on Knowledge Engineering: Practice and Patterns* (pp. 109-119). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-540-87696-0_12

Pavel, S., & Euzenat, J. (2013, January). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering, 25*(1), 158-176. doi:10.1109/TKDE.2011.253

Pazienza, M., Sguera, S., & Stellato, A. (2007, December 26). Let's talk about our "being": A linguistic-based ontology framework for coordinating agents. (R. Ferrario, & L. Prévot, Eds.) *Applied Ontology, special issue on Formal Ontologies for Communicating Agents, 2*(3-4), 305-332.

Peukert, E., Eberius, J., & Rahm, E. (2012). A Self-Configuring Schema Matching System. *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering* (pp. 306-317). Washington, DC, USA: IEEE Computer Society. doi:10.1109/ICDE.2012.21

Ritze, D., & Eckert, K. (2012). Thesaurus mapping: a challenge for ontology alignment? *Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012.* CEUR-WS.org.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006, May). The Semantic Web Revisited. *IEEE Intelligent Systems, 21*(3), 96-101. doi:10.1109/MIS.2006.62

van Ossenbruggen, J., Hildebrand, M., & de Boer, V. (2011). Interactive vocabulary alignment. *Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries* (pp. 296-307). Berlin, Germany: Springer-Verlag.

Widerhold, G. (1994). Interoperation, Mediation and Ontologies. *Proceedings International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogeneous Cooperative Knowledge Bases* (pp. 33-48). Tokyo, Japan: ICOT.

Wrigley, S. N., García-Castro, R., & Nixon, L. (2012). Semantic evaluation at large scale (SEALS). *Proceedings of the 21st international conference companion on World Wide Web* (pp. 299-302). New York, NY, USA: ACM.