Steffen Lohmann
Tassilo Pellegrini (Eds.)

I-SEMANTICS 2012
Posters & Demos

Proceedings of the I-SEMANTICS 2012
Posters & Demonstrations Track

i-semantics

8th International Conference on Semantic Systems
September 5-7, 2012, Graz, Austria

Title: Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track
Editors: Steffen Lohmann, Tassilo Pellegrini
ISSN: 1613-0073
CEUR Workshop Proceedings
(CEUR-WS.org)

Editors' addresses:

Steffen Lohmann

University of Stuttgart
Universitätsstraße 38
70569 Stuttgart
Germany
steffen.lohmann@vis.uni-stuttgart.de

Tassilo Pellegrini

St. Pölten University of Applied Sciences
Department of Economics
Matthias Corvinus-Straße 15
3100 St. Pölten
Austria
tassilo.pellegrini@fhstp.ac.at

Preface

I-SEMANTICS 2012 was the eighth edition of the *International Conference on Semantic Systems*. The conference provides a forum for the exchange of latest scientific results in topics such as Semantic Technologies, Semantic Web, and Linked Data. It has become a major event in the field, attracting more than 400 participants every year. As a conference that aims to bring together science and industry, I-SEMANTICS encourages scientific research and application-oriented contributions. As in previous years, the eighth edition of I-SEMANTICS took place in Graz, Austria, in early September.

The I-SEMANTICS Posters & Demonstrations Track complements the main conference track. It provides an opportunity to present late-breaking research results, smaller contributions, and innovative work in progress. It gives conference attendees the possibility to learn about on-going work and encourages discussions between researchers and practitioners in the field. For presenters, it provides an excellent opportunity to obtain feedback from peers.

For the first time, the Posters & Demonstrations Track had a separate call for contributions and reviewing process this year. A total of 23 papers were submitted to the track, each sent for review to three members of the program committee. Based on the reviews, we selected twelve posters and demos for presentation at the conference. The papers to these twelve contributions are included in the proceedings.

We thank all authors for their contributions and the program committee and additional reviewers for their valuable work in reviewing the submissions. We are also grateful to the staff of the Know-Center and Messe Congress Graz who supported us in the local organization of this track.

October 2012

Steffen Lohmann
(Posters & Demos Chair)

Tassilo Pellegrini
(Conference Chair)

Program Committee

Simone Braun, FZI Karlsruhe, Germany

Claudia D'Amato, University of Bari, Italy

Philipp Heim, TRUMPF Group, Germany

Aidan Hogan, DERI Galway, Ireland

Tim Hussein, University of Duisburg-Essen, Germany

Pablo Mendes, FU Berlin, Germany

Jindřich Mynarz, National Technical Library, Czech Republic

Teresa Onorati, Universidad Carlos III de Madrid, Spain

Alexandre Passant, DERI Galway, Ireland

Heiko Paulheim, TU Darmstadt, Germany

John Pereira, Salzburg Research, Austria

Thomas Riechert, University of Leipzig, Germany

Boris Villazón-Terrazas, Universidad Politécnica de Madrid, Spain

Additional Reviewers

Robert Isele, FU Berlin, Germany

Myriam Leggieri, DERI Galway, Ireland

Fabrizio Orlandi, DERI Galway, Ireland

Owen Sacco, DERI Galway, Ireland

Jodi Schneider, DERI Galway, Ireland

Andreas Schultz, FU Berlin, German

Contents

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 1-5, 2012.

A Novel Concept-based Search for the Web of Data

Melike Sah and Vincent Wade
Knowledge and Data Engineering Group, Trinity College Dublin, Dublin, Ireland
{Melike.Sah, Vincent.Wade}@scss.tcd.ie

Abstract. With the increasing volumes of data, access to the Linked Open Data (LOD) becomes a challenging task. Current LOD search engines provide flat result lists, which is not an efficient access method to the Web of Data (WoD). In this demo, we introduce a novel and scalable concept-based search mechanism on the WoD, which allows searching based on meaning of objects. In particular, the retrieved resources are dynamically categorized into UMBEL vocabulary concepts (topics) using a novel fuzzy retrieval model and resources with the same concepts are grouped together to form categories, which we call *concept lenses*. In addition, search results are presented with hierarchy of categories and concept lenses for easy access to the LOD. Such categorization enables concept-based browsing of the retrieved results aligned to users' intent or interests. Results categorization can also be used to support more effective personalized presentation of search results.

Keywords: Categorization, concept-based search, semantic indexing, fuzzy retrieval model, linked open data, UMBEL, scalability, demo.

1 Introduction

Linked Open Data (LOD) or the Web of Data (WoD) is becoming a de-facto for publishing structured and interlinked data according to a set of Linked Data principles and practices. The main promise of the LOD is providing rich Web-scale interlinked metadata, which can be consumed by Web applications in more innovative ways that was not possible before. However, as the number of datasets and data on the LOD is increasing, the challenge turn into finding and accessing the relevant datasets and data. Thus, LOD search engines are becoming more important to enable exploration and browsing of LOD data and search engines are crucial for the uptake of the WoD.

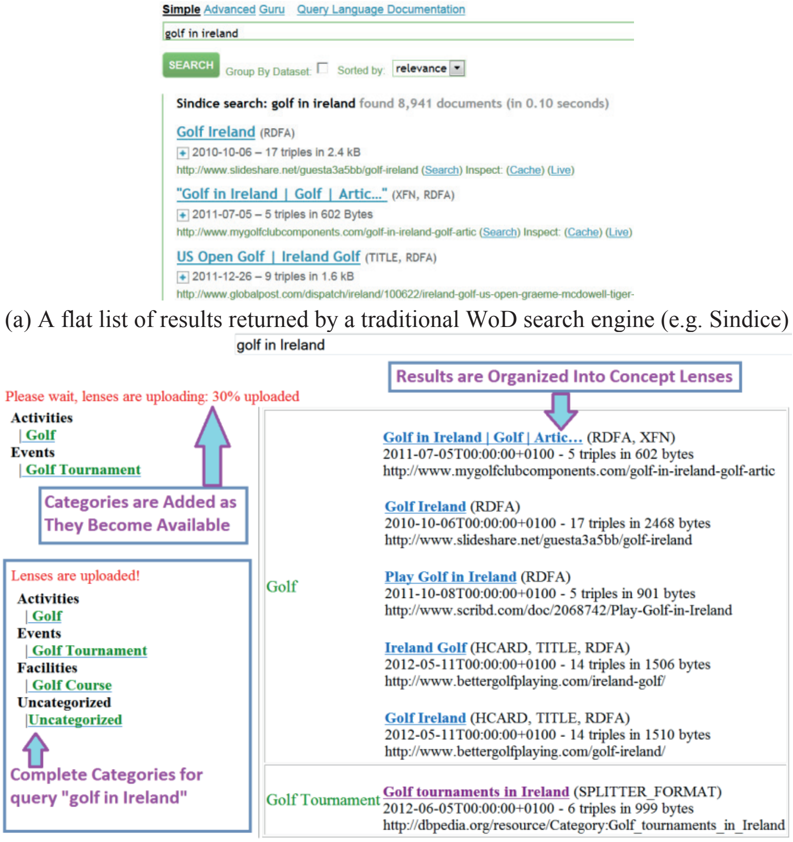
On the other hand, current WoD search engines and mechanisms, such as Sindice [1] and Watson [2], display the search results as ranked lists. In particular, they present the resource title or example triples about the resource in the search results. However, presentation of resource titles is not an efficient presentation method for the WoD since users cannot understand "what the resource is about" without opening and investigating the LOD resource itself. Sig.ma service or Sig.ma end-user application, attempts to solve this problem with a data mash-up based presentation paradigm by using querying, rules, machine learning and user interaction [3]. The user can query the WoD and Sig.ma presents rich aggregated mashup information about the retrieved resources. Sig.ma's focus is on data aggregation and it is not for search results presentation. Another search paradigm for the LOD is faceted search/browsing, which provide facets (categories) for interactive searching and browsing [4]. The main limitation of the faceted search mechanisms is that facet generation depends on specific data/schema properties of underlying metadata. Thus it can be challenging to generate useful facets to large and heterogeneous WoD [5]. It is evident that more efficient WoD search mechanisms are needed for the uptake of LOD by a wider community.

To overcome this issue, we introduce and demonstrate a novel concept-based search for the WoD using UMBEL concept hierarchy (http://umbel.org) and a novel fuzzy retrieval model [6][8]. In particular, the WoD is searched and the retrieved results are categorized into concepts based on their meaning. Then, LOD resources with the same concepts are grouped to form categories, which we call *concept lenses*. In this way, search results are presented using a hierarchy of categories and concept lenses, which can support more efficient access and browsing of results rather than flat result lists. Moreover, categories can be used for efficient personalization.

There are three unique contributions of our approach [6]: (1) For the first time, UMBEL is used for concept-based Information Retrieval (IR). (2) A second contribution is in novel semantic indexing and fuzzy retrieval model, which provides efficient categorization of search results in UMBEL concepts. (3) A minor contribution is the realization of a concept-based search realm to WoD exploration. Concept-based search has occurred in traditional Web. In this paper, we improve our previous work [6]: (1) With a more scalable system architecture, where the system performance can scale by using an indexing service at the server-side for dynamic categorizations. (2) In our previous work, a flat list of concepts was presented. In the current version, we improved the presentation by organizing concepts into a hierarchy, where users can locate relevant lenses using hierarchical organization. In this demo, we discuss the benefits of our approach compared to traditional WoD search engines using scenarios. In addition, we will discuss how the two main challenges, categorization accuracy and system performance, are resolved.

2 A Search Scenario on the Web of Data

To better illustrate the benefits of the proposed concept-based search, we describe a real life search scenario on the WoD (see our demo [8]). Assume Sue is knowledge engineer designing a website for "tourism in Ireland". She is designing an ontology to structure the site content and wants to populate the ontology with metadata and instances. Assume this ontology contains activities in Ireland, such as "golf". First, she searches for existing information on the WoD using "golf in Ireland" query. As shown in Figure 1(a), such a query may return many diverse results by a traditional WoD search engine (e.g. Sindice in this example). In this case, she needs to open and investigate large number of results for its suitability to her investigation. On contrary, when the same query is searched on our concept-based search, the results are automatically categorized and presented with hierarchy of categories and concept lenses as shown in Figure 1(b). In this case, Sue can discard irrelevant matches easily and can locate matching resources based on their concepts. In this example, Sue may notice that she can include classes and metadata about golf courses and golf tournaments in her ontology. In general, hierarchical search results clustering have the advantage of providing shortcuts to the items that have similar meaning. It also allows better topic understanding and favours systematic exploration of search results [7]. Our concept-based search is unique on the LOD to support such search results exploration on the WoD. Moreover, as seen in Figure 1, most of the results are Web pages that contain embedded metadata. Thanks to robust categorization, our approach is applicable to categorization of Web pages on the Web (in most cases we only use URL labels) as well as categorization of LOD resources on the Semantic Web.


Simple Advanced Guru Query Language Documentation
SEARCH
Group By Dataset: □ Sorted by: relevance

golf in ireland

Sindice search: golf in ireland found 8,941 documents (in 0.10 seconds)

Golf Ireland (RDFA)
2010-10-06 – 17 triples in 2.4 kB
http://www.slideshare.net/guesta3a5bb/golf-ireland (Search) Inspect: (Cache) (Live)

"Golf in Ireland | Golf | Artic..." (XFN, RDFA)
2011-07-05 – 5 triples in 602 Bytes
http://www.mygolfclubcomponents.com/golf-in-ireland-golf-artic (Search) Inspect: (Cache) (Live)

US Open Golf | Ireland Golf (TITLE, RDFA)
2011-12-26 – 9 triples in 1.6 kB
http://www.globalpost.com/dispatch/ireland/100622/ireland-golf-us-open-graeme-mcdowell-tiger-

(a) A flat list of results returned by a traditional WoD search engine (e.g. Sindice)

golf in Ireland

Results are Organized Into Concept Lenses

Please wait, lenses are uploading: 30% uploaded

Activities
| Golf
Events
| Golf Tournament

Categories are Added as They Become Available

Lenses are uploaded!

Activities
| Golf
Events
| Golf Tournament
Facilities
| Golf Course
Uncategorized
|Uncategorized

Complete Categories for query "golf in Ireland"

Golf

Golf in Ireland | Golf | Artic... (RDFA, XFN)
2011-07-05T00:00:00+0100 - 5 triples in 602 bytes
http://www.mygolfclubcomponents.com/golf-in-ireland-golf-artic

Golf Ireland (RDFA)
2010-10-06T00:00:00+0100 - 17 triples in 2468 bytes
http://www.slideshare.net/guesta3a5bb/golf-ireland

Play Golf in Ireland (RDFA)
2011-10-08T00:00:00+0100 - 5 triples in 901 bytes
http://www.scribd.com/doc/2068742/Play-Golf-in-Ireland

Ireland Golf (HCARD, TITLE, RDFA)
2012-05-11T00:00:00+0100 - 14 triples in 1506 bytes
http://www.bettergolfplaying.com/ireland-golf/

Golf Ireland (HCARD, TITLE, RDFA)
2012-05-11T00:00:00+0100 - 14 triples in 1510 bytes
http://www.bettergolfplaying.com/golf-ireland/

Golf Tournament

Golf tournaments in Ireland (SPLITTER_FORMAT)
2012-06-05T00:00:00+0100 - 6 triples in 999 bytes
http://dbpedia.org/resource/Category:Golf_tournaments_in_Ireland

(b) The same results are presented with categories and concept lenses by our approach

Fig. 1. Comparison of (a) traditional and (b) concept-based search for query "golf in Ireland"

3 Proposed Concept-based Search

System Architecture (Figure 2). Client-side is developed with Javascript and AJAX (parallel processing and incremental presentation for performance). Java Servlets are utilized at the server side where we use Jena for processing RDF and Lucene IR framework for indexing and implementation of categorization. Sindice Search and Sindice Cache APIs are used for searching the WoD and accessing RDF descriptions of LOD resources. In our approach, results that are retrieved by the Sindice Search are further processed to categorize into categories. For this purpose, features are extracted from LOD resources and matched to UMBEL concept descriptions using a fuzzy retrieval model [6]. Categorized LOD resources are cached to a local index for system performance and sent to client for presentations with categories and concept lenses. Our search mechanism can work with any query and on any dataset because of a proposed robust categorization method and broad concepts provided by UMBEL.

UMBEL Concept Vocabulary. UMBEL is sub-set of OpenCyc. It provides broad topics (~28,000 concepts) with useful relations and properties drawn from OpenCyc (i.e. broader/narrower classes, preferred/alternative/hidden labels). UMBEL concepts are also organized into 32 supertype classes (e.g. Event, Activities, Places, etc.), which make it easier to reason, search and browse. In traditional concept-based IR systems, the concept descriptions are indexed using a vector space model (i.e. term frequency, inverse document frequency – *tf*×*idf*). For more efficient representations, we applied a novel semantic indexing model; associated weight of the term to the concept depends on where the term appears in a structured concept description (i.e. in URI label, preferred/alternative labels, sub/super-concepts labels).

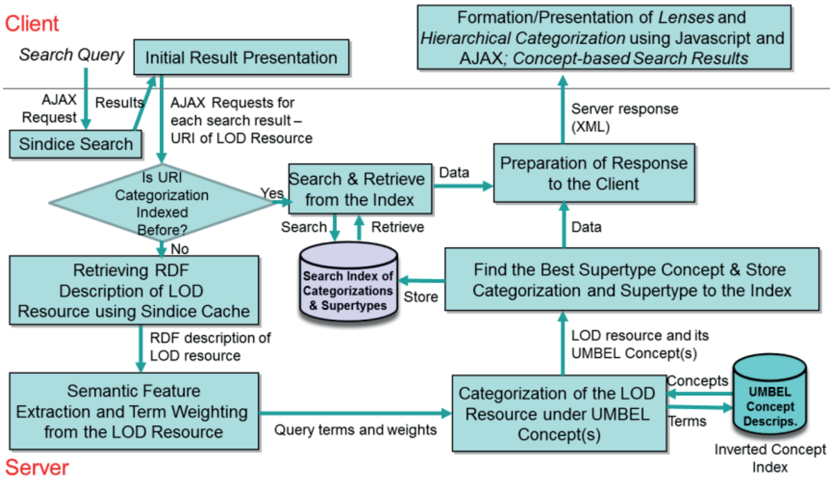

Client
Search Query
Initial Result Presentation
Formation/Presentation of *Lenses* and *Hierarchical Categorization* using Javascript and AJAX; *Concept-based Search Results*
AJAX Request
Results
AJAX Requests for each search result – URI of LOD Resource
Server response (XML)
Sindice Search
Is URI Categorization Indexed Before?
Yes
Search & Retrieve from the Index
Data
Preparation of Response to the Client
Search
Retrieve
Data
No
Retrieving RDF Description of LOD Resource using Sindice Cache
Search Index of Categorizations & Supertypes
Store
Find the Best Supertype Concept & Store Categorization and Supertype to the Index
RDF description of LOD resource
LOD resource and its UMBEL Concept(s)
Semantic Feature Extraction and Term Weighting from the LOD Resource
Query terms and weights
Categorization of the LOD Resource under UMBEL Concept(s)
Concepts
Terms
UMBEL Concept Descrips.
Inverted Concept Index
Server

Fig. 2. System Architecture

Feature Extraction from the Context of LOD Resources. In order to categorize LOD resources under UMBEL concepts, lexical information is mined from the common features of LOD resources, such as *URI*, *label*, *type*, *subject* and *property names*. Moreover, a semantic enrichment technique is applied to gather more lexical information from the LOD graph by traversing owl:sameAs links. From the extracted terms, stop words are removed and the terms are stemmed into their roots. Then, the terms are weighted according to their term frequency and where they appear in the LOD resources; i.e. terms that are appear in subject and type fields may provide more contextual information about the resource. Thus, they are weighted higher.

Categorization of LOD Resources. The extracted terms from the LOD resource is matched against UMBEL concept descriptions using a novel fuzzy-based retrieval model. Proposed fuzzy retrieval model generates a fuzzy relevancy score according to relevancy of a term to semantic elements (structure) of concept(s) ([6] for details): For example, UMBEL concepts are organized into a hierarchy of concepts. A concept may have relevant terms in concept, more specific terms in sub-concepts and more general terms in super-concepts. Instead of combining all the terms from the concept, sub-concepts and super-concepts, we weight term importance based on where they appear. Then, a fuzzy retrieval model combines term weights and a voting algorithm is applied to decide the final categorization of the LOD resource. Moreover, supertype class of the UMBEL concept needs to be found for hierarchical presentation of categories. In UMBEL, a concept might belong to more than one supertype class. We apply a voting algorithm, i.e. supertype class with the highest *tf*×*idf* rank of all LOD terms will be selected as the best representing supertype for that UMBEL concept.

Client-Side. At the client-side, a script (Javascript functions) processes the server responses and incrementally generates/updates hierarchical categories as well as concept lenses using AJAX. In this way, we prevent long delays in server responses.

Indexing for a Scalable Performance. In our approach, search results are processed in parallel for a scalable performance. In this paper, the system performance is enhanced further by adding a search index at the server-side. After the categorization, UMBEL and supertype concepts of the LOD URI are indexed. Since the index size affects search performance and the required disk space, we only index concept names without the base namespace. When a URI is requested, first the indices are searched; if URI has not been processed before, we apply dynamic categorization. Thus, we achieve significant decrease in network traffic and supply on-time categorizations.

Evaluations. Extensive evaluations are carried out to test the performance of our system on a benchmark of ~10,000 DBpedia mappings (see [6]). Evaluations showed that the proposed fuzzy retrieval model achieves very promising results ~90% precision, which is crucial for the correct formation of categories and the uptake of the proposed concept-based search. Moreover, the system performance can scale thanks to parallel processing and the use of search indices with minimum disk space.

4 References

1. Delbru, R., S., Campinas, G., Tummarello: Searching Web Data: an Entity Retrieval and High-Performance Indexing Model. Journal of Web Semantics, vol. 10, pp. 33-58, (2012)
2. D'Aquin, M., E., Motta, M., Sabou, S., Angeletou, L., Gridinoc, V., Lopez and D., Guidi: Toward a New Generation of Semantic Web Applications. IEEE Intelligent Systems (2008)
3. Tummarello, G., R., Cyganiak, M., Catasta, S., Danielczyk, R., Delbru and S., Decker: Sig.ma: live views on the Web of Data, Journal of Web Semantics, 8(4), pp. 355-364 (2010)
4. Heim, P., T., Ertl and J., Ziegler: Facet Graphs: Complex Semantic Querying Made Easy, Extended Semantic Web Conference (ESWC), LNCS, vol. 6088, pp. 288-302, (2010)
5. Teevan, J., S. T., Dumais and Z. Gutt.: Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web. Workshop on HCIR, (2008)
6. Sah, M., and V., Wade. A Novel Concept-based Search for the Web of Data using UMBEL and a Fuzzy Retrieval Model. Extended Semantic Web Conference (ESWC), (2012)
7. Carpineto, C., S., Osinski, G., Romano, and D. Weiss. A Survey of Web Clustering Engines. ACM Computing Surveys, 41(3), 2009.
8. A demo is available online at https://www.scss.tcd.ie/melike.sah/golf_demo.swf

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 6-11, 2012.

Matching Linked Open Data Entities to Local Thesaurus Concepts

Peter Wetz[1], Hermann Stern[1], Jürgen Jakobitsch[2], and Viktoria Pammer[1]

[1] Know-Center GmbH
Inffeldgasse 21a, 8010 Graz, Austria
{pwetz,hstern,vpammer}@know-center.at
http://www.know-center.at
[2] Semantic Web Company GmbH
Neubaugasse 1, 1070 Vienna, Austria
j.jakobitsch@semantic-web.at
http://www.semantic-web.at

Abstract. We describe a solution for matching Linked Open Data (LOD) entities to concepts within a local thesaurus. The solution is currently integrated into a demonstrator of the PoolParty thesaurus management software. The underlying motivation is to support thesaurus users in linking locally relevant concepts in a thesaurus to descriptions available openly on the Web. Our concept matching algorithm ranks a list of potentially matching LOD entities with respect to a local thesaurus concept, based on their similarity. This similarity is calculated through string matching algorithms based not only on concept and entity labels, but also on the "context" of concepts, i.e. the values of properties of the local concept and the LOD concept. We evaluate over 41 different similarity algorithms on two test-ontologies with 17 and 50 concepts, respectively. The results of the first evaluation are validated on the second test-dataset of 50 concepts in order to ensure the generalisability of our chosen similarity matches. Finally, the overlap-, TFIDF- and SoftTFIDF-similarity algorithms emerge as winners of this selection and evaluation procedure.

Keywords: linked open data, dbpedia, thesaurus, similarity, evaluation, concept matching

1 Introduction

A solution for matching Linked Open Data (LOD) entities to concepts within a local thesaurus is specified in this paper. The solution is currently integrated into a demonstrator of the PoolParty thesaurus management software. The underlying motivation is to support thesaurus users in linking locally relevant concepts in a thesaurus to descriptions available openly on the Web. This "linking" has a technical (realising an RDF triple) and a conceptual (realising that other, possibly complementary, descriptions of the same entity exist) component. The strategy of linking LOD entities to a local thesaurus uses the concepts of Linked Data to expand and enrich the information stored in the thesaurus ultimately leading to a more valuable knowledge base.

Since the maturing of Semantic Web technologies, and the massive emergence of LOD repositories in many domains[3], the LOD cloud presents a valuable source of knowledge. When managing a thesaurus, this source can be tapped into, either loosely in the sense of exploring additional, openly available information, or by creating an RDF triple that technically links the local concept to a LOD entity.

Naturally, others have explored the challenges and possibilites around concept matching (e.g., in the field of schema matching and ontology matching). Specifically for interlinking LOD entities, Raimond et al. [2] for instance describe two naïve approaches using literal lookups to interlink music datasets as well as an explorative graph matching algorithm based on literal similarity and graph mappings. Waitelonis and Sack [3] use matching algorithms to map labels of their yovisto video search engine to DBpedia entities. Mendes et al. [1] describe with DBpedia Spotlight a service that interlinks text documents with LOD from DBpedia. Similar to these works, we experiment with a mixture of string similarity and exploiting the graph nature of both the local thesaurus and LOD entities.

In the live demo, participants will be able to create a new thesaurus with PoolParty, or use an existing thesaurus, and enrich it with the presented matching algorithm with LOD entities from DBpedia. Participants will thus be able to gauge the usefulness of such a semi-automatic data linking themselves.

2 Problem Statement

The problem which we describe the solution for in this paper is the following: Given a specific concept in a local thesaurus, and a list of potentially matching LOD entities, which LOD entity is most similar to the local thesaurus concept?

We assume that typically both the local concept and the given list of LOD entities have a context, i.e. will have additional properties that describe them, such as a verbal description, a categorisation etc. We delegate the task of finding "potentially matching LOD entities" to a LOD lookup service, that queries the LOD cloud with a request that stems from the local concept's label.

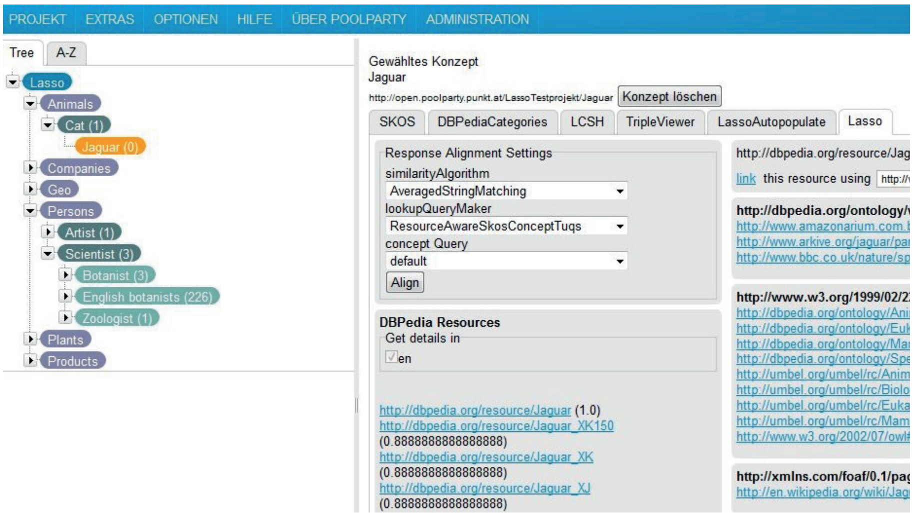
This approach can be called *interlinking of entities*, *alignment of entities*, *semantical enrichment of data*, *augmenting data with LOD* or *entity reconciliation.*

[3] media - `http://data.nytimes.com/`, geography - `http://www.geonames.org/`, encylcopedic knowledge - `http://dbpedia.org`

3 Solution

A *lookup service* is responsible for finding potentially matching LOD entities by matching concept labels to labels of LOD entities. This lookup service can be used with any LOD SPARQL[4] endpoint. We also investigated on how much context information should be taken into account when querying for potentially matching LOD concepts.

Contextual information is integrated by adding the literal string values of connected properties of the query's entity into the similarity comparison process. In the SKOS[5] syntax, which all thesauri of this system are based on, these properties are represented as *broader* (describing hierarchically more general entities), *narrower* (describing hierarchically more specific entities) and *related* (describing similar entities) links to other entities in the same thesaurus. This additional information describes the entity in more detail, furthermore helping to deal with ambiguous terms and getting more precise results. In our current implementation we take into account broader, narrower and all related concepts of the local thesaurus concept which lead to satisfying results. In theory it is also possible to only use a subset of these contextual properties.



Gewähltes Konzept
Jaguar
http://open.poolparty.punkt.at/LassoTestprojekt/Jaguar

Response Alignment Settings
similarityAlgorithm: AveragedStringMatching
lookupQueryMaker: ResourceAwareSkosConceptTuqs
concept Query: default

DBPedia Resources
Get details in
☑en

http://dbpedia.org/resource/Jaguar (1.0)
http://dbpedia.org/resource/Jaguar_XK150 (0.8888888888888888)
http://dbpedia.org/resource/Jaguar_XK (0.8888888888888888)
http://dbpedia.org/resource/Jaguar_XJ (0.8888888888888888)

http://dbpedia.org/resource/Jag
link this resource using http://

http://dbpedia.org/ontology/
http://www.amazonarium.com.
http://www.arkive.org/jaguar/pa
http://www.bbc.co.uk/nature/sp

http://www.w3.org/1999/02/2
http://dbpedia.org/ontology/Ani
http://dbpedia.org/ontology/Euk
http://dbpedia.org/ontology/Ma
http://dbpedia.org/ontology/Spe
http://umbel.org/umbel/rc/Anim
http://umbel.org/umbel/rc/Biolo
http://umbel.org/umbel/rc/Euka
http://umbel.org/umbel/rc/Mam
http://www.w3.org/2002/07/owl#

http://xmlns.com/foaf/0.1/pag
http://en.wikipedia.org/wiki/Jag

Fig. 1. DBpedia Lookup for concept "Jaguar".

As can be seen in Figure 1 the entity *Jaguar* has *Cat* as a broader concept. This relation will pour into the query as contextual information, which will yield the animal called Jaguar as a result. If the query would be triggered by

[4] `http://www.w3.org/TR/rdf-sparql-query/`

[5] `http://www.w3.org/2004/02/skos/`

choosing the *Jaguar* from the *Products* branch of the thesaurus, and therefore using other contextual information, the famous car would be on top of the results.

3.1 Implementation

The *concept matching algorithm* is responsible for comparing the local concept and the potentially matching LOD entities based on a similarity algorithm. Depending on the similarity algorithm, strings of labels and of different properties are compared to each other. Coefficients are calculated and the resulting similarity values determine the ranking of the LOD entities (the highest ranking is the LOD entity most similar to the local concept). The ranked list of LOD entities is visible for users in the PoolParty demonstrator as described in the next paragraph.

In Fig. 1 we show a *classical Semantic Web disambiguation example* in terms of the PoolParty user interface of our concept matching algorithm: the user wants to connect a concept with the preferred label *Jaguar* to the appropriate counterpart in DBpedia. The concept matching algorithm grabs the labels of concepts that are connected to Jaguar, which are *cat* and *animal*. These labels are compared to the labels of candidate resources from DBpedia using the active similarity algorithm. We see that the resource *Jaguar* referring to the cat is on the top of this list, followed by several resources referring to cars of that name (see bottom middle part of Figure 1). The DBpedia facts of the selected Jaguar resource are displayed on the right hand side. If the top-ranked LOD entity indeed describes the same real-world entity as the local thesaurus concept, then these concepts can be linked through the graphical user interface, which technically corresponds to creating an RDF triple relating both concepts.

3.2 Method and Result of Selecting a Similarity Algorithm

We selected the default similarity algorithm which the final presented ranking is based on by comparing the performance of 41 similarity algorithms on two test-datasets of 17 and 50 concepts, respectively. The first dataset includes general ambiguous terms to enable testing of the algorithm's efficiency regarding disambiguation. To get further insight into the datasets please register a demo account[6] to be able to browse directly using the PoolParty System.

In order to ensure generalisability, we compared the performance of the best algorithms on the first dataset with their performance on the Reegle[7] thesaurus - which consists of concepts dealing with clean energy - by extracting and using 50 concepts. In both cases the algorithms *overlap*, *TFIDF* and *SoftTFIDF* performed very well (see Table 1).

[6] `http://poolparty.punkt.at/de/try-it/`

[7] `http://www.reegle.info/`

4 Discussion and Outlook

In our selection and evaluation procedure of similarity algorithms, the overlap algorithm worked very well for both test ontologies. It simply checks how many of the terms in the query are also found in each result and then calculates a coefficient. On the second and third rank there are similar algorithms, which only differ in parameters dealing with tokenisation (TFIDF and SoftTFIDF). The TFIDF algorithms calculate a so-called corpus of all words including the query and all results. Based on this corpus the relevancy of each result compared to the query is computed. Overall, an accuracy of about 80% can be achieved resulting in a meaningful and efficient linkage of local thesaurus entities with entities from remote LOD repositories. Additionally, our results indicate that the winning similarity algorithms will perform well also on ontologies of other domains.

In an implementation where all complexity should be hidden from the user, one of these algorithms would be selected as the default (and probably only) similarity algorithm. Alternatively, a “voting” mechanism that always involves all three algorithms is conceivable.

To sum up, the integration and usage of SKOS principles helping us to gain contextual information for the queries, the high accuracy of top ranked algorithms and the confirmation that the overlap and TFIDF algorithms work best are a major contribution to findings which have already been made in related work.

#	Algorithm	Points	#	Algorithm	Points
1	overlap	0,823	37	qGramsDistance (qg2)	0,507
2	overlap (ws)	0,823	38	MatchingCoefficient (qg3)	0,477
3	overlap (qg3)	0,765	39	levenshtein	0,470
4	TFIDF	0,749	40	NeedlemanWunsch	0,320
5	SmithWaterman	0,725	41	stringTFIDF	0,318

Table 1. A list of the top and bottom five ranked algorithms after both evaluations. *ws* means whitespace-tokenisation; *qg2* and *qg3* mean qgram2- and qgram3-tokenisation, respectively. *Points* is a relative number to 1. 1 meaning all results would have been ranked correctly. Please find the complete table at `http://bit.ly/O7ufgk`.

Acknowledgements. The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. This work has been co-funded by the FFG project LASSO (Fit-IT).

References

1. Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.
2. Yves Raimond, Christopher Sutton, and Mark Sandler. Automatic Interlinking of Music Datasets on the Semantic Web. 2008.
3. Jörg Waitelonis and Harald Sack. Augmenting Video Search with Linked Open Data. In *Proc. of Int. Conf. on Semantic Systems 2009, i-Semantics 2009*, 2009.

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 12-16, 2012.

HETWIN: Helping Evaluate the Trustworthiness of Web Information for Web Users Framework using Semantic Web Technologies

Jarutas Pattanaphanchai, Kieron O'Hara, and Wendy Hall

Electronics and Computer Science, Faculty of Physical and Applied Science,
University of Southampton,
Southampton, SO17 1BJ, United Kingdom
{jp11g09, kmo, wh}@ecs.soton.ac.uk

Abstract. Assessing the trustworthiness of information found on the Web is challenging because of two factors. First, there is a little control over publishing quality. Second, Web users have little information available on which to base judgment of the trustworthiness of Web information while they are interacting with it. This work addresses this problem by collecting and presenting metadata about authority, currency, accuracy and relevance to evaluate the trustworthiness of Web information during information seeking processes. In this poster, we propose the *HETWIN* application framework and present a design as a prototype tool that employs this framework for academic publications.

Keywords: Trust, Credibility, Information Quality, Semantic Web

1 Introduction

It is known that ordinary Web users base decisions on whether to trust information on the Web on heuristic factors (pertaining to its presentation and layout). However, as these heuristic factors are mainly based on surface level characteristics of the Web page, and such characteristics are easily disguised, Web users can arrive in the wrong conclusions about the trustworthiness of information they consume [3]. However, a number of studies have suggested that additional information such as the identity of the author (e.g. name, position), and the expertise of the author could potentially increase the Web users' confidence and help them to make better assessments than by using their own heuristic criteria alone [2, 4–6]. In particular, Bizer *et.al.* [1] proposed the TriQL.P browser, a RDF browser, that presents recommended RDF datasets that should be trusted based on trust policies. However, in their work, the user needs to go to a certain Web page, from which the browser can extract Semantic Web content. On the contrary, it is more useful to provide Web users with a tool with which they can look for the information they need while it *automatically* gathers the supportive information to help them evaluate the trustworthiness of Web information.

In order to address this problem, we propose a framework to help users evaluate the trustworthiness of Web information, called *HETWIN*, which, with information of the factors from the studies above, are selected to use in the framework. In addition, we propose a prototype tool, which employs the HETWIN framework implemented as a chrome extension. The prototype collects metadata using Semantic Web technologies and presents it in a useful way in the context of the users' search for information. In the following section, we explain the HETWIN architecture and display an example result from our prototype. Then, we present planned future work.

2 Helping Evaluate the Trustworthiness of Web Information for Web Users Framework

Our framework uses Semantic Web technologies to collect RDF data which is published alongside Websites and queried from SPARQL endpoints, which it then integrates to build metadata graphs. Then, these metadata graphs are used to create supportive data, which is presented to users in order to help them evaluate the trustworthiness of Web information. In this work, we assume that the RDF data on the Web or in the data store is accurate. Evaluation is based on a case study of the *ePrints* of the University of Southampton[1], which is an online repository of academic publications, in which the accuracy of the RDF data published is verified by authorized staff[2]. Our application framework, as shown in Figure 1, consists of three main modules.

- **The input module** accepts the user's search keywords and the domain of interest, which affects the type of information returned by the search. In this work, we defined four domains of interest (business, informational, news and personal). The input module extracts any RDF linked to from the web page. Also, our model evaluates the trustworthiness of the information every time the user interacts with the system. Therefore, the system obtains the most recent information at the time at which the evaluation is performed.
- **The trustworthiness criteria and metacollection module** is composed of two main components. The trustworthiness criteria comprises of four basic criteria: *authority, currency, accuracy, and relevance*, which the assessment of trustworthiness in each domain of interest is based. Each criterion provides the basic predicate keys of RDF that should be used to collect metadata. For instance, in the informational domain, trustworthiness is evaluated based on the authority criterion, using predicate key, *"dct:creator"*, the currency criterion, using predicate key, *"dct:date"*, the accuracy criterion which is based on the predicate key, *"bibo:status"*, and the relevance criterion which is based on data returned from querying using the predicate key, *"dct:title"* and *"dct:abstract"*. Alternatively, in the news domain, which still evaluates the trustworthiness of the information based on the same criteria, different or additional predicate keys might be used. For example, the authority criterion might use the predicate key *"dct:publisher"* in addition to the *"dct:creator"*.

[1] http://www.eprints.soton.ac.uk

[2] http://www.southampton.ac.uk/library/research/eprints/policies/eprints.html

Our framework allows one to add additional predicate keys or new domains by adding them into its configuration file. Therefore, the framework can adjust for use in different domains and can extend to new domains.
The metacollection component gathers metadata based on the predicates which are defined in the trustworthiness criteria. The collected metadata will be aggregated in order to build metadata graph. The basic approach of aggregating metadata assumes that the metadata from the four basic predicates have the same level of important for assessing the trustworthiness of Web information. In the case that the system needs the additional data, the system will add the additional data into the metadata graph after the basic metadata has been added.

- **The output module** displays the metadata graph in a human readable format to help the users assess the trustworthiness of Web information. In addition, it orders the results based on the relevance of the information to the user's query which is computed based on the frequent of the appearance of search terms in the title and the abstract and the expertise of the authors or creators of the information.

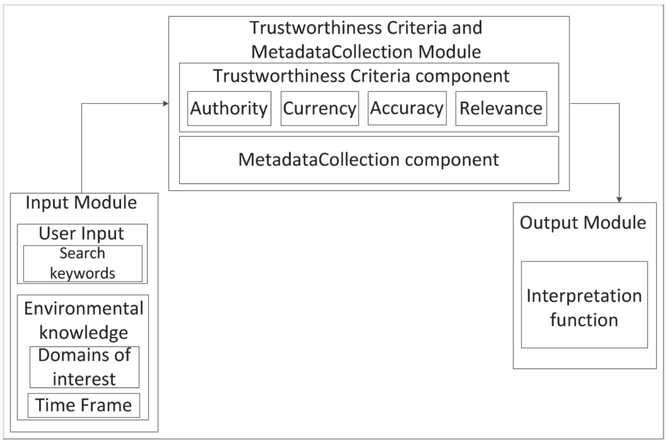
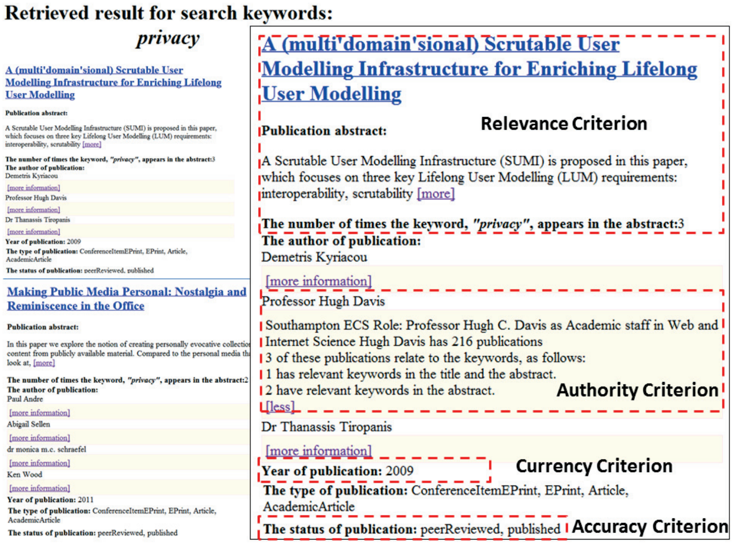


Fig. 1. HETWIN architecture

3 Results

We present example results of the output of our tool, using data from ePrints at University of Southampton. Specifically, we consider the publications of the School of Electronics and Computer Science, and we focus our evaluation on the informational domain. The result in Figure 2 displays the identifying details of the publication including its title, its abstract and the name of its authors.

Moreover, the results are ordered by the relevance of the information to the user's interests. Specifically, it shows the authors' full names and also their appellations. These can indicate the authority of the author, which represents their reputation in producing this content. In addition, it displays the detail of each author's publications, which itself is indicative of the author's expertise in the area. For example, if the author has several publications that relate to privacy, this implies that the author is not only interested in that area but also has expertise in it. In addition, the more publications that exist by that author that relate to the keywords, the more likely the author is to be an expert in that topic. The publication date indicates the currency of the publication. The system gathers the status of the publication for determining the accuracy of the information within it because, in the case of publications, there is a review process, which can help to evaluate the accuracy of the content. For example, if the publication has been peer-reviewed or published in an academic publication, it is likely to be accurate, and therefore trustworthy. In current work, the relevance criterion analyses the abstract and the title of the publication. If they contain the user's search keywords, the document is more likely to meet the user's needs. However, we consider another potential approach to evaluate relevance more efficiently than matching exact keywords; adopting an ontology concept for finding terms related to the user's keywords to match in key areas of the content such as the title or the first paragraph of content. This allows the framework match the relevant information to user's needs better.



Retrieved result for search keywords:
privacy

A (multi'domain'sional) Scrutable User Modelling Infrastructure for Enriching Lifelong User Modelling

Publication abstract:

A Scrutable User Modelling Infrastructure (SUMI) is proposed in this paper, which focuses on three key Lifelong User Modelling (LUM) requirements: interoperability, scrutability [more]

The number of times the keyword, *"privacy"*, appears in the abstract:3
The author of publication:
Demetris Kyriacou
[more information]
Professor Hugh Davis
[more information]
Dr Thanassis Tiropanis
[more information]
Year of publication: 2009
The type of publication: ConferenceItemEPrint, EPrint, Article, AcademicArticle
The status of publication: peerReviewed, published

Making Public Media Personal: Nostalgia and Reminiscence in the Office

Publication abstract:

In this paper we explore the notion of creating personally evocative collectio content from publicly available material. Compared to the personal media th look at, [more]

The number of times the keyword, *"privacy"*, appears in the abstract:2
The author of publication:
Paul Andre
[more information]
Abigail Sellen
[more information]
dr monica m.c. schraefel
[more information]
Ken Wood
[more information]
Year of publication: 2011
The type of publication: ConferenceItemEPrint, EPrint, Article, AcademicArticle
The status of publication: peerReviewed, published

A (multi'domain'sional) Scrutable User Modelling Infrastructure for Enriching Lifelong User Modelling

Publication abstract:

A Scrutable User Modelling Infrastructure (SUMI) is proposed in this paper, which focuses on three key Lifelong User Modelling (LUM) requirements: interoperability, scrutability [more]

The number of times the keyword, *"privacy"*, appears in the abstract:3
The author of publication:
Demetris Kyriacou
[more information]
Professor Hugh Davis

Southampton ECS Role: Professor Hugh C. Davis as Academic staff in Web and Internet Science Hugh Davis has 216 publications
3 of these publications relate to the keywords, as follows:
1 has relevant keywords in the title and the abstract.
2 have relevant keywords in the abstract.
[less]
Dr Thanassis Tiropanis
[more information]
Year of publication: 2009
The type of publication: ConferenceItemEPrint, EPrint, Article, AcademicArticle
The status of publication: peerReviewed, published

Fig. 2. Example of results using "privacy" as keyword and "informational" as a domain (the four criteria are shown in bold)

4 Conclusions

We proposed an application framework and prototype tool which helps users to evaluate the trustworthiness of Web information using Semantic Web technologies. The result from the prototype shows the supportive data in each criteria, which is explained to the user and can help them assess the trustworthiness of Web information. In future work, we will evaluate our framework by conducting a user survey. This survey will elicit information about how satisfied the users were with the system and how their approach to assessing trust has changed since using our system in comparison to an expert.

References

1. Bizer, C., Cyganiak, R., Gauss, T., Maresch, O.: The TriQL. P browser: Filtering information using context-, content-and rating-based trust policies. In: Proc. of the Semantic Web and Policy Workshop. vol. 7, pp. 12–20 (2005)
2. Fogg, B., Marshall, J., Laraki, O., Osipovich, A.: What makes Web sites credible?: a report on a large quantitative study. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems pp. 61–68 (2001)
3. Fogg, B., Soohoo, C., Danielson, D.R., Marable, L., Standford, J., Tauber, E.R.: How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. Proc. of the 2003 conference on Designing for user experiences pp. 1–15 (2003)
4. Rieh, S., Belkin, N.: Understanding judgment of information quality and cognitive authority in the WWW. In: the 61th annual meeting of the Am. Soc. for Inf. Sci. and Technol. vol. 35, pp. 279–89 (1998)
5. Tate, M.: Web Wisdom: How To Evaluate and Create Information Quality on the Web, Second Edition. CRC Press (2009)
6. Wathen, C.N., Burkell, J.: Believe it or not: Factors influencing credibility on the Web. Journal of the Am. Soc. for Inf. Sci. and Technol. 53(2), 134–144 (2002)

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 17-21, 2012.

Interlinking Media Archives with the Web of Data

Semantic inline annotation of online content

Dietmar Glachs, Sebastian Schaffert, Christoph Bauer

SalzburgResearch Forschungsgesellschaft m.b.H., Salzburg, Austria
{dietmar.glachs,sebastian.schaffert}@salzburgresearch.at
Österreichischer Rundfunk, Wien, Österreich
christoph.bauer@salzburgresearch.at

Abstract. Today's enterprises heavily rely upon accurate, consistent, and timely access to data. However, company data is typically scattered across multiple databases and file shares in a multitude of forms and versions. Moreover, an increasing amount of valuable background information is available outside the companies' influence and control. This situation is typical for many enterprise information integration scenarios, also in Austria's largest broadcasting media archive. Our demonstration argues for an information integration approach that uses semantic web principles to interlink archival media content of the Austrian Broadcasting Corporation (ORF) with the web of data and with internal knowledge resources to facilitate semantic search and to increase the user experience of browsing and discovering media content in the daily production workflow.

Keywords: Linked Enterprise Data, Linked Media, Semantic Media Archive

1 Introduction

The Linked Open Data (LOD) community project was initiated in 2007 by the W3C [1] and proposes the usage of standards like the Resource Description Framework (RDF) [2] for publishing datasets on the web in order to make them available for interlinking [3]. The number of datasets available, commonly referred to as the Linked Data Cloud[1] [4], is still growing and provides enterprises with the opportunity to interlink enterprise data with background information or to allow for disambiguation of concepts. Enterprises however still hesitate to use Linked Data in their value chain. Based on experiences with industrial partners, the main barriers in the adoption of Linked Data are (i) a rather new technology since accessing data from the Linked Data cloud is still cumbersome; (ii) the lack of complete solutions because Linked Data is still considered read-only and metadata-only whilst enterprise data is highly dynamic and increasingly includes multimedia content and (iii) the need of adapting established enterprise processes when using linked data [5].

[1] http://richard.cyganiak.de/2007/10/lod/

With this article we propose the integration of large datasets available on the web by following the Linked Data principles as outlined in [6] to enhance closed enterprise content with additional information from the Linked Open Data cloud [7]. This demo uses the Linked Media Framework (LMF[2]) [8], a platform for enterprise information integration. Based on Linked Data as well as Apache Stanbol[3] for content analysis, the LMF shows how to eliminate the entry barriers when using Linked Data in enterprises.

2 Semantic Media Archive

The Austrian Public Broadcaster's (ORF – Österreichischer Rundfunk) archive is the central repository for all video and audio material created by the ORF in the last 60 years and contains a vast amount of media content in different formats. The primary objective of the archive is to preserve audio/video content for potential future use and make it accessible to editors. When archiving new content, several archiving tools restricted to expert users are used; FESAD[4] as an example is used to manage video based content. However, for journalists, editors and program planners the archiving division uses a web based tool for federated search and investigation. For now, the work of describing the clips (e.g. annotating the content) is actually carried out solely by members of the archiving division. The users of the search tool currently cannot modify/annotate content in order to improve data quality or search confidence.

The main objective for the ORF is therefore to (i) provide additional information to the end users like editors and journalists, (ii) to allow simple annotation means which are not restricted to the archiving division and (iii) provide/integrate semantic search facilities for improved search results. As an integrated solution we integrated the LMF as Linked Media Server in the Archival Toolset of the ORF. In addition to the existing tools, the LMF provides extended semantic search facilities and also allows for interlinking of archival content with publicly available linked data sources. As shown in Fig. 1, the LMF extends the search tool mARCo by adding itself as an additional data source and by providing means for annotating mARCO search results. The annotations are then subject of future searches in mARCo.

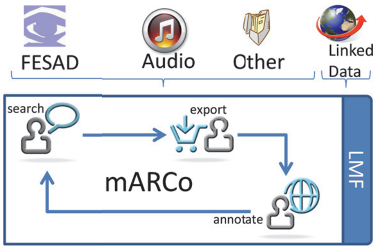
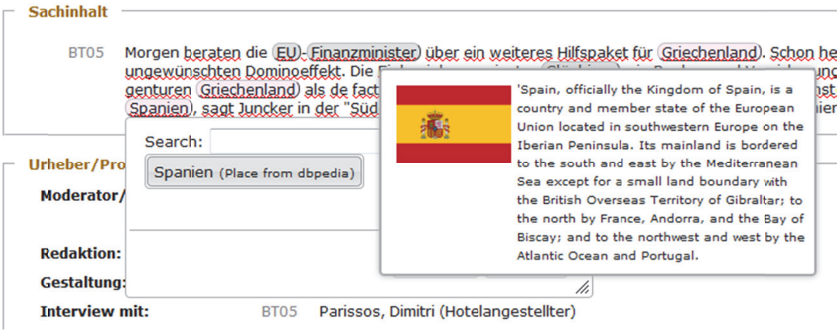


Fig. 1. Semantic Media Archive

[2] http://code.google.com/p/lmf/
[3] http://incubator.apache.org/stanbol/
[4] FESAD – Video Archival System used by ORF, ARD

2.1 Annotating Media Content

When browsing search results, editors or journalists are enabled to annotate the content. With the help of a special annotation plugin, the formerly "read-only" search result page becomes editable by injecting the annotation features into the web page. Parts of the page such as the content description are analyzed by Apache Stanbol[5]. As a result, eligible resource annotations are provided to the user as shown in Fig. 2.



Sachinhalt

BT05 Morgen beraten die EU-Finanzminister über ein weiteres Hilfspaket für Griechenland. Schon he
ungewünschten Dominoeffekt. Die
genturen Griechenland als de fact
Spanien, sagt Juncker in der "Süd

Search:

Spanien (Place from dbpedia)

'Spain, officially the Kingdom of Spain, is a country and member state of the European Union located in southwestern Europe on the Iberian Peninsula. Its mainland is bordered to the south and east by the Mediterranean Sea except for a small land boundary with the British Overseas Territory of Gibraltar; to the north by France, Andorra, and the Bay of Biscay; and to the northwest and west by the Atlantic Ocean and Portugal.

Urheber/Pro

Moderator/

Redaktion:

Gestaltung:

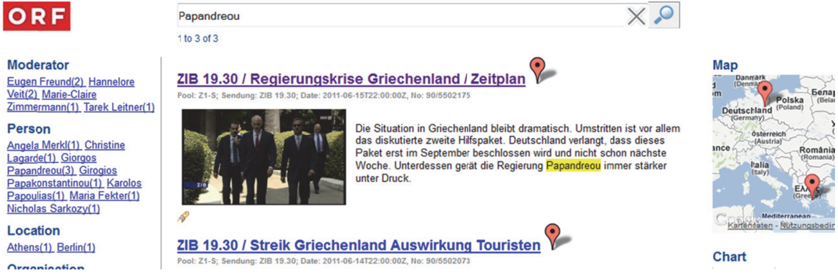
Interview mit: BT05 Parissos, Dimitri (Hotelangestellter)

Fig. 2. Annotation and interlinking interface

By selecting a suggestion, the journalist can review the proposal and finally annotate the content. The Linked Media Framework stores the annotation by means of SPARQL Update [10] and also collects the available properties of the referenced resource and thus makes the information immediately available for semantic search.

2.2 Semantic Media Search

The search experience can be improved by facilitating the semantic relations of the archived data. By using the semantic concepts of the data which are either production related (e. g. moderator, editor, program etc.) or content related (e.g. persons named or in video, content description, location of the clips content), it is possible to provide a faceted search as shown in Fig. 3, for example to narrow down the search, the user may select one or more facet properties shown in the search interface.



Papandreou

1 to 3 of 3

Moderator

Eugen Freund(2) Hannelore Veit(2) Marie-Claire Zimmermann(1) Tarek Leitner(1)

Person

Angela Merkl(1) Christine Lagarde(1) Giorgos Papandreou(3) Girogios Papakonstantinou(1) Karolos Papoulias(1) Maria Fekter(1) Nicholas Sarkozy(1)

Location

Athens(1) Berlin(1)

ZIB 19.30 / Regierungskrise Griechenland / Zeitplan

Pool: Z1-S; Sendung: ZIB 19.30; Date: 2011-06-15T22:00:00Z, No: 90/5502175

Die Situation in Griechenland bleibt dramatisch. Umstritten ist vor allem das diskutierte zweite Hilfspaket. Deutschland verlangt, dass dieses Paket erst im September beschlossen wird und nicht schon nächste Woche. Unterdessen gerät die Regierung Papandreou immer stärker unter Druck.

ZIB 19.30 / Streik Griechenland Auswirkung Touristen

Pool: Z1-S; Sendung: ZIB 19.30; Date: 2011-06-14T22:00:00Z, No: 90/5502073

Fig. 3. Search Demonstrator

[5] http://incubator.apache.org/stanbol

3 DEMO OUTLINE

The Linked Media Framework (LMF) serves as the backend whereas the both clients for search and annotation are lightweight JavaScript implementations using RESTful webservices for the communication with the backend service. The LMF is a service oriented framework which uses semi-structured data representation (RDF) and HTTP URLs as uniform resource identifier to store and identify resources, as recommended for Linked Data [6]. The demo we show at the conference will first show the Semantic Search Component as it is a fundamental part of the LMF and demonstrates the power and flexibility of using Semantic Web technologies for search and retrieval. We will then use a VIE bookmarklet[6] for the annotation of a typical ORF search result page which relies on concepts from DBPedia[7] and an internal SKOS[8] based thesaurus. Accepting proposed annotations with the LMF will immediately influence the search results and optionally add new concepts to an internal company thesaurus. In the production scenario, the LMF will also be tightly connected with the mARCo search facility and therefore will be part of the federated search component.

The LMF integrates/connects the linked data cloud as possible sources for background information and finally enables annotation by storing selected concepts in the (local) Linked Data server by means of SPARQL Update statements. In particular this annotation functionality will be subject of the demonstration given at I-Semantics to first show the where we will preload the LMF with a selection of news articles out of the Austrian Broadcasters Archive. The demonstration will also cover how the news articles are presented to journalists for annotation. Finally, the demonstration of the search interface is also available online at the NewMediaLabs demonstration site[9].

4 CONCLUSION

The potential of Linked Data in general and the Linked Media Framework as a platform for supporting semantic search has been proven in several projects. With this demonstration we aimed to outline its potential for the use in an Enterprise Information Integration scenario where Linked Data technology is used to support users in their daily work and to improve the amount and quality of content annotation. The latter directly leads to an improved search result with respect to precision which is a fundamental requirement in the news domain. Because of the smooth integration in existing processes, the functionality is offered as an optional add-on to the users. The improved search results as well as the provided background information are the inducement for the users to use the offered functionality. In contrast to the increasing number of semantic web case studies[10], the demonstrated scenario Linked Media Framework allows the publication of structured information as Linked Data and also

[6] http://szabyg.github.com/vie-annotation-bookmarklet/

[7] http://dbpedia.org

[8] http://www.w3.org/2004/02/skos/

[9] http://labs.newmedialab.at/ORF/orf/search/index.html

[10] http://www.w3.org/2001/sw/sweo/public/UseCases/

enables the full read-write management of the published data and in particular enables the full roundtrip of annotations for further usage during search and retrieval.

5 ACKNOWLEDGMENTS

The media content enhancement and the semantic search described in this paper were planned and developed in the Austrian research centre "Salzburg NewMediaLab - The Next Generation" (SNML-TNG). The centre is funded by the Austrian Federal Ministry of Economy, Family and Youth (BMWFJ), the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) and the Province of Salzburg. The demo content is taken from the ORF archive by courtesy of the Austrian Broacasting Corporation. The development of the LMF has been inspired by the needs & requests of our industrial partners. As a result, the Linked Media Framework currently serves several real-world scenarios.

6 References

1. Linking Open Data. 2010. W3C SWEO Community Project. Retrieved from http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
2. RDF: G. Klyne and J. J. Carroll. Resource description framework (RDF): Concepts and abstract syntax. Technical report, W3C, 2 2004
3. Bizer, C., Cyganiak, R., Heath, T. 2007. How to Publish Linked Data on the Web. Retrieved from http://www4.wiwiss.fuberlin.de/bizer/pub/LinkedDataTutorial
4. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems, 4(2)*, 1-22. Elsevier. Retrieved from http://www.citeulike.org/user/omunoz/article/5008761
5. Wood, D. 2010. Linking Enterprise Data. ISBN 978-1-4419-7664-2. DOI 10.007/978-1-4419-7665-9
6. Berners-Lee, T. 2006. Linked Data – Design Issues. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html
7. Bizer, C., Heath, T., Ayers, D., Raimond, Y. 2007. Interlinking Open Data on the Web (Poster). In 4th European Semantic Web Conference (ESWC2007), pages 802–815.
8. Kurz, T., Schaffert, S., Bürger, T. (2011). LMF – A Framework for Linked Media. In: Workshop for Multimedia on the Web (MMWeb2011).
9. Damjanovic, V., Kurz, T., Westenthaler, R., Behrendt, W., Gruber, A. and Schaffert, S. 2011. Semantic enhancement: The key to massive and heterogeneous data pools. In Proceeding of the 20th International IEEE ERK (Electrotechnical and Computer Science) Conference 2011, Portoroz, Slovenia.
10. Prud'hommeaux, E., & Seaborne, A. 2008. SPARQL Query Language for RDF. W3C working draft. Retrieved from http://www.w3.org/TR/rdf-sparql-query

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 22-25, 2012.

Visualizing and Editing Ontology Fragments with OWLGrEd

Renars Liepins[**], Karlis Cerans[*], Arturs Sprogis[**]

Institute of Mathematics and Computer Science, University of Latvia
{ Renars.Liepins, Karlis.Cerans, Arturs.Sprogis}@lumii.lv

Abstract. The OWLGrEd ontology editor allows graphical visualization and authoring of OWL 2.0 ontologies using a compact yet intuitive presentation that combines UML class diagram notation with textual Manchester syntax for class expressions. Here we show, how to integrate OWLGrEd with ontology module mechanism from OWL API to obtain on-demand ontology fragment visualization that is essential for many existing large ontologies that do not fit in a single reasonably perceivable UML class diagram.

Keywords: OWL ontologies, visualization, UML/OWL profile, OWLGrEd, ontology decomposition, ontology modules

1 Introduction

Intuitive ontology visualization is a key for their learning, exchange, as well as their use in conceptual modeling and semantic database schema design. A number of tools and approaches exist for rendering and/or editing OWL [1] ontologies in a graphical form, including UML Profile for OWL DL [2], ODM [3], TopBraid Composer [4], Protégé [5] plug-in OWLViz [6] and OWLGrEd [7,8]. The approaches of [2,3,7,8] use UML [9] class diagrams to visualize OWL ontologies. This is achieved by visualizing an independent hierarchy of ontology classes and then structuring the data and object property visualizations along the property domain and range classes. The OWL ontology constructions not having direct UML counterparts (e.g. class expressions, properties with more than one domain assertion, sub-property relations etc.) are usually handled by some auxiliary means in the notation and the editor. OWLGrEd uses textual OWL Manchester syntax [10] for class expressions where the graphical notation is not available or is not desired thus allowing compact and comprehensible presentation of up to medium-sized ontologies within a single diagram.

The main focus of this demo is on using the compact UML-style notation, offered by OWLGrEd, on large ontologies that do not fit within any reasonably-sized class diagram, or whose rendering appears to be too complicated due to a kind of “spider web” effect produced by many classes and relations. Its key idea consists in splitting the ontology into meaningful fragments of conceivable size and then visualizing each of the fragments in a separate diagram .

[**] Partially supported by Latvian 2010.-2013. National Research Program Nr.2 project Nr.5.

[*] Partially supported by European Union via European Regional Development Fund project 2011/0009/2DP/2.1.1.1.0/10/APIA/VIAA/112.

OWLGrEd already has the means to partition ontology into sub-diagrams (fragments) when authoring or reengineering an existing ontology. But there was no way to automatically partition an ontology that is imported into OWLGrEd for visualization. In this demo we will present an extension to OWLGrEd visualization capabilities, that allows automatic partitioning of an ontology into logical fragments. The addition is based on Automatic Decomposition [11] that was recently implemented in the OWL API[1]. The decomposition is based on signatures, i.e. for each fragment a user selects some entities that should be included in the fragment. Then the fragment is extended with all the logically relevant axioms for these entities. Finally all the fragments are rendered graphically in the OWLGrEd editor.

The demonstration shows (i) working with OWLGrEd tool to render and author OWL ontologies (ii) OWLGrEd extension to automatically partition ontology into logical overlapping fragments based on fragment signatures.

2 OWLGrEd Notation and Editor

OWLGrEd[1] provides a complete graphical notation for OWL 2 [1], based on UML class diagrams. We visualize OWL classes as UML classes, data properties as class attributes, object properties as associations, individuals as objects and cardinality restrictions on association domain class as UML cardinalities. It is easy to visualize also subclass and inverse properties notations. For the full OWL 2 construct coverage we enrich the UML class diagrams with the new extension notations, e.g. (cf. [7,8]):

- fields in classes for *equivalent class*, *superclass* and *disjoint class* expressions written in Manchester OWL syntax [10];
- fields in associations and attributes for *equivalent*, *disjoint* and *super* properties and fields for property characteristics, e.g., *functional*, *transitive*, etc.;
- anonymous classes containing *equivalent class expression* but no name (we show graphically only anonymous classes that need to have graphic representation in order to be able to describe other ontology concepts in the diagram);
- connectors (as lines) for visualizing binary *disjoint*, *equivalent*, etc. axioms;
- boxes with connectors for n-ary *disjoint*, *equivalent*, etc. axioms;
- connectors (lines) for visualizing object property restrictions *some*, *only*, *exactly*, as well as cardinality restrictions.

OWLGrEd provides option to specify class expressions in compact textual form rather than using separate graphical element for each logical item within class expression. If an expression is referenced in multiple places, it can optionally be shown as an anonymous class. An anonymous class is also used as a base for property domain/range specification, if this domain/range is not a named class.

Figure 1 illustrates some basic OWLGrEd constructs on a simple mini-University ontology, including different notation options for *EquivalentClasses* assertion, object property restriction and a comment. The notation is explained in more detail in [7].

[1] http://owlapi.sourceforge.net

[2] http://owlgred.lumii.lv/

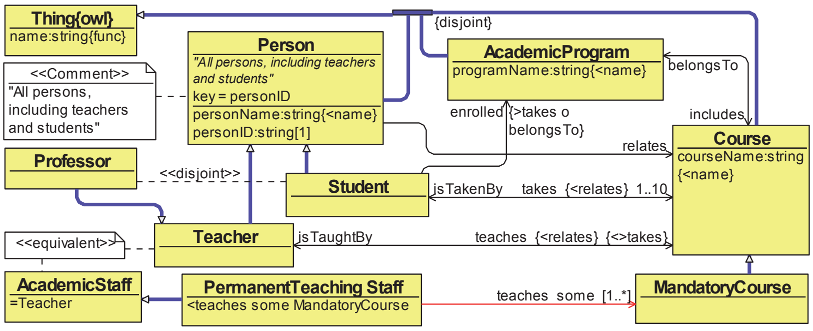


Fig. 1. Example: OWLGrEd notation for a mini-University ontology

The OWGrEd editor offers ontology interoperability (import/export) functionality with the Protégé 4.2 ontology editor [5]. The principal OWLGrEd usage ways are:

- ontology authoring (create and edit an ontology in OWLGrEd, then export it to Protégé to analyze and possibly submit it to other ontology processing tools)
- ontology visualization (an ontology that is imported from Protégé is displayed graphically to obtain a comprehensible visual view on it).

3 Visualizing Fragments of Ontology

The graphical form is ideal for understanding small ontologies, but for large ontologies it fast becomes overwhelming because of too many line crossings. For visualization of the ontology in the form of fragments an issue is to describe the fragments to be visualized since manual enumeration of all axioms to be included into a fragment would clearly be infeasible. Recently there has been work on signature-based automatic decomposition [11] of ontologies that allows specify ontology modules just in terms of their “core” terms/entities. The decomposition then finds all the axioms that are logically relevant for the given entities.

We have extended OWLGrEd editor with the Automatic Decomposition feature. A user can specify either a single ontology fragment, or a list of fragments covering the whole ontology that is to be visualized. The automatic decomposition then finds all the relevant axioms for each specified fragment thus allowing OWLGrEd showing the fragments visually in a graphical form.

As an example consider the schema.org ontology. It consists of about 300 classes, 110 object properties, 70 data properties and 310 subclass assertions. The ontology is clearly too large to be easily perceived as a single diagram. However, it would be feasible as well as meaningful to visualize fragments of the. For example, in the Figure 2 is shown a fragment that is centered on entities “Event”, “Product” and “Person”. Once the user has specified such an entity list, the tool automatically finds the relevant axioms for these entities and then shows this fragment graphically. It is possible to specify any number of such fragment signatures at a time and the tool will create visualization for each of them.

The experiments we have performed allows us to judge that the offered approach of combining of the traditional OWLGrEd ontology visualization means with ontology decomposition techniques would be a useful tool for the semantic technology community in ontology schema structure representation.

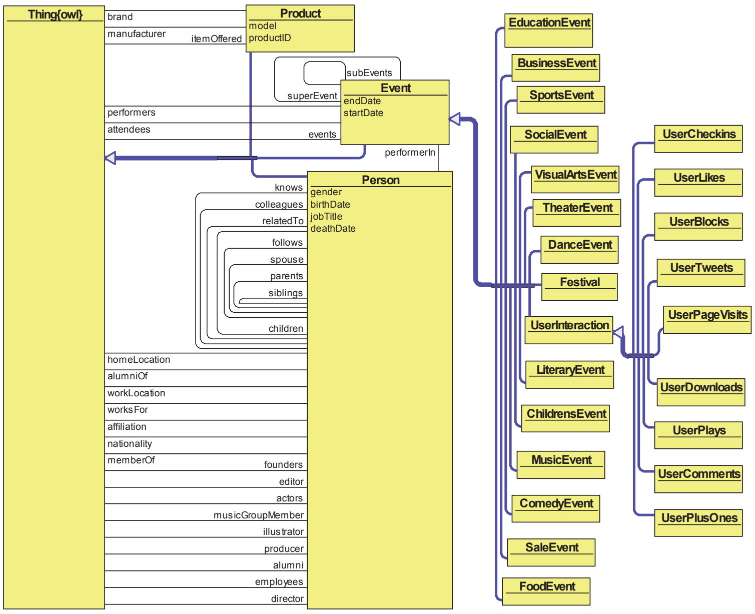


Fig. 2. Automatically extracted fragment of schema.org ontology based on a signature "Event, Product, Person".

References

1. Motik, B; Patel-Schneider P.F; Parsia B.: OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax, 2009
2. Brockmans, S., Volz, R., Eberhart, A., Löffler, P. Visual Modeling of OWL DL Ontologies Using UML, Proc. of ISWC 2004, LNCS 3298, pp. 198-213, 2004.
3. ODM UML profile for OWL, http://www.omg.org/spec/ODM/1.0/PDF/
4. TopBraid Composer, http://www.topquadrant.com/products/TB_Composer.html.
5. Protégé 4, http://protege.stanford.edu/
6. OWL Viz, http://www.co-ode.org/downloads/owlviz/
7. Barzdins, J.; Barzdins, G.; Cerans, K.; Liepins, R.; Sprogis, A.: OWLGrEd: a UML Style Graphical Notation and Editor for OWL 2. In Proc. of OWLED 2010, 2010.
8. Barzdins, J.; Cerans, K.; Liepins, R.; Sprogis, A.: UML Style Graphical Notation and Editor for OWL 2. In Proc. of BIR'2010, LNBIP, Springer 2010, vol. 64, p. 102-113, 2010.
9. Unified Modeling Language: Infrastructure, version 2.1. OMG Specification ptc/06-04-03, http://www.omg.org/docs/ptc/06-04-03.pdf
10. OWL 2 Manchester Syntax, http://www.w3.org/TR/owl2-manchester-syntax/
11. Klinov, P.; Vescovo, C.; Schneider, T.: Incrementally Updateable and Persistent Decomposition of OWL Ontologies. In Proc of OWLED 2012.

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 26-30, 2012.

Linked Open Data Infrastructure for Public Sector Information: Example from Serbia

Valentina Janev[1], Uroš Milošević[1], Mirko Spasić[1], Jelena Milojković[2], Sanja Vraneš[1]

[1]Mihailo Pupin Institute, University of Belgrade, Belgrade, Serbia
{valentina.janev, uros.milosevic, mirko.spasic,
sanja.vranes@pupin.rs}
[2]Statistical Office of the Republic of Serbia, Belgrade, Serbia
{jelena.milojkovic@stat.gov.rs}

Abstract. To improve transparency and public service delivery, national, regional and local governmental bodies need to consider new strategies to openning up their data. We approach the problem of creating a more scalable and interoperable Open Government Data ecosystem by considering the latest advances in Linked Open Data. More precisely, we showcase how an integrated and coherent collection of aligned state of the art software tools, the LOD2 Stack, can be used to deliver trusted, open and rich collections of interlinked datasets to the public. The usage of the Tool Stack is demonstrated on the case of one of the largest data providers in the Republic of Serbia – its Statistical Office.

Keywords. linked open data, open government data, infrastructure, tools, public sector, Serbia

1 Introduction

In order to improve efficiency in the provision of public services, increase transparency and interaction with citizens and society as a whole, but also create new businesses and job opportunities, both local and national governments need to find better strategies for delivering large amounts of trusted data to the public. The fact that the European Commission is investing considerable amounts of finances to overcome this problem is a strong indicator of its significance. As a direct example, consider the ISA (Interoperability Solutions for European Public Administrations) program for the period from 2010-2015 that has been assigned a budget of 164,1 million euros[1]. The program enables "the delivery of electronic public services and ensures the availability, interoperability, re-use and sharing of common solutions"[2]. To make government data truly open (for use and re-use), and increase transparency, it needs to be published in a non-proprietary, machine-readable format (e.g. RDF, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210).

[1] European Commission ISA Webpage, http://ec.europa.eu/isa/
[2] European Commission ISA Webpage, http://ec.europa.eu/isa/faq/faq_en.htm

In this paper, we will show why Linked Data is considered a promising approach to the above problem, and how the LOD2 Stack, a powerful set of software tools and components, can be used to lower the cost of addressing the challenges of publishing and integrating Open Government Data (OGD). The evaluation of the tools used in the *National Statistical Office* use case workflow (see Fig. 1) will be given in section 2. Section 3 discusses the achieved results in the process of integration of Serbian public data in the LOD cloud, with a special attention to the case of one of the largest data providers in the Republic of Serbia – its Statistical Office (SORS).

1.1 LOD2: The Project and the OGD Use Case

In the last few years the Linked Data paradigm has evolved as a powerful enabler for the transition of the current document-oriented Web into a Web of interlinked Data and, ultimately, into the Semantic Web. Aimed at speeding up this process, the LOD2 project ("Creating knowledge out of interlinked data", http://lod2.eu) partners have delivered the LOD2 Stack, "an integrated collection of aligned state of the art software components that enable corporations, organizations and individuals to employ Linked Data technologies with minimal initial investments" [1].

One of the LOD2 objectives is to showcase the wide applicability of the LOD2 Stack for building public services for ordinary citizens of the European Union. As partners of the LOD2 project, the *Mihailo Pupin Institute*'s team established the Serbian CKAN,[3] the first catalogue of this kind in the West Balkan countries, with a goal of becoming an essential tool for enforcing business ventures based on open data in this region. The RDF datasets cataloged with the Serbian CKAN (rs.ckan.net) are periodically harvested and synchronized at an international level with the PublicData.eu portal[4] and integrated into the LOD cloud.

2 Evaluation of LOD Tools and Technologies

The LOD2 Stack was evaluated for allowing governments and governmental agencies to publish their data based on open standards. Requirements identified for the National Statistical Office scenario [2] were grouped into the following types: *Data extraction and transformation*, *Domain-specific modeling*, *Data enrichment and interlinking*, *Data storage*, *Exploration and analysis*, and *Data and Service administration*. Table 1 shows how the LOD2 Stack responds to these requirements.

Vocabularies suitable for modeling statistical data in RDF format are the Data Cube vocabulary [3] which is fully compatible with the cube model that underlines SDMX[5], and VoID (Vocabulary of Interlinked Datasets, http://www.w3.org/TR/void/), an RDF based schema used to describe linked datasets.

[3] CKAN is a data catalogue system used by various institutions and communities to manage open data.

[4] PublicData.eu has been developed as a part of the LOD2 project.

[5] SDMX (Statistical Data and Metadata eXchange), http://code.google.com/p/publishing-statistical-data/wiki/Documentation.

Table 1. Overview of LOD2 Stack capabilities

Data Extraction and Transformation
In a case where direct central database access is enabled, the *D2R server* and D2RQ mapping language can be used to represent the content in RDF format (e.g. using the SPARQL endpoint). Otherwise, for data provided in Excel or XML format, *OntoWiki*‘s stat2RDF extension or the LOD2 *XSLT* processor can be used.
Domain-specific Modeling
The *PoolParty Thesaurus Manager* (PPT, http://lod2.poolparty.biz) tool for enterprise metadata management and linked data publishing is based on standard SKOS vocabulary and can be combined with text mining and linked data technologies. Additionally, knowledge models developed with PoolParty can be edited and enhanced with *OntoWiki* (http://ontowiki.net/) authoring tool.
Data Enrichment and Interlinking
These features are very important as a pre-processing step in integration and analysis of statistical data from multiple sources. The LOD2 tools such as *SILK* (http://www4.wiwiss.fu-berlin.de/bizer/silk) and *Limes* (http://aksw.org/Projects/LIMES) facilitate mapping between knowledge bases, while *GRefine* can be used to enrich the data with descriptions from DBpedia or reconcile with other information in the LOD cloud.
Data Storage
The LOD Cloud Cluster knowledge store for the LOD2 Project (http://lod.openlinksw.com) hosting 50 billion plus triples, consists of a *Virtuoso* clustered instance hosted on 8 server nodes at the *Sindice* Data Centre at DERI (NUIG)[4].
Exploration and Analysis
The LOD2 Stack offers tools such as SparQLed, *Sindice's assisted SPARQL editor* (http://sindicetech.com/sindice-suite/sparqled/) and the RDF Data Cube visualization component *CubeViz* (https://github.com/AKSW/cubeviz.ontowiki), that are of special importance for statistical data analysis and visualization.

3 Linked Open Data Example from Serbia

In an attempt to adopt the LOD2 Stack for the Statistical Office of the Republic of Serbia, over 100 datasets were extracted from the central statistics database (http://webrzs.stat.gov.rs/WebSite/public/ReportView.aspx), transformed into RDF, stored as RDF dump files on a local server (http://elpo.stat.gov.rs/lod2/) and registered with the Serbian CKAN. The data includes statistics from the *Prices*, *National accounts*, *Usage of Information and Communication Technologies,* and *Science*, *Technology and Innovation* domains (see [2] for more details). Performed activities can be summarized as follows.

Metadata Management. The statistics published by National Statistical Offices or Eurostat are organized by theme, presented in aggregate form by using a wide range of standard metadata (code lists). In the SORS Use case, a knowledge model was built where standard code lists were modeled using the SKOS vocabulary [2]. The model (http://lod2.poolparty.biz/) currently incorporates 12 concept schemas including the *NACE (revision 1 and revision 2)*, *COICOP*, and *SITC (revision 4)*, as well as other schemas used in SORS statistical publications, such as *geographical*, *time* and *statistical areas* code lists. In order to formalize the conceptualization of the *National accounts* domain, for instance, the ESA 95 (European system of accounts ESA, http://circa.europa.eu/irc/dsis/nfaccount/info/data/ESA95/en/titelen.htm) was used. In governmental organizations, the metadata management activity is carried out by users with administration permissions (depicted in Fig.1). Using Silk and LODGefine (http://code.zemanta.com/sparkica/) some of the code lists were interlinked with DBpedia and Eurostat code lists.

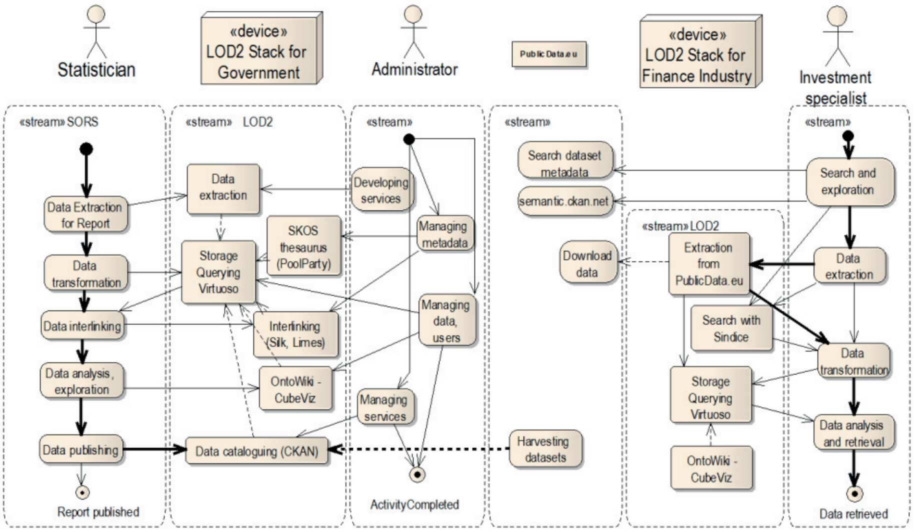
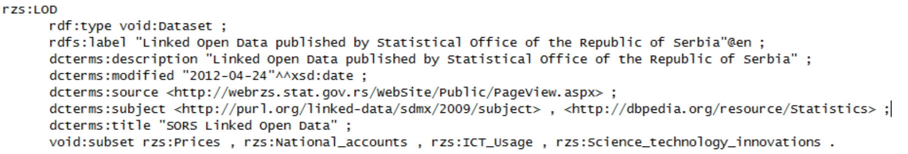


Fig. 1. Using LOD2 tools for publishing and consuming statistical data

The Serbian CKAN. The Serbian CKAN portal is deployed on a server with the following characteristics: Intel® Xeon® CPU 5140, dual core @ 2.33GHz 8GB RAM, Ubuntu 11.04, with kernel version: 2.6.38-12. The CKAN software was fully translated to Serbian, enabling support for two character sets (Latin and Cyrillic). Furthermore, a large number of dataset relationships have been defined, making the CKAN browsing and navigation experiences more comfortable. The Serbian CKAN is currently maintained by the *Mihailo Pupin Institute*'s team.

The SORS LOD Cloud. The SORS statistical data in XML form was passed as input to the XSLT processor and transformed into RDF using the aforementioned vocabularies (RDF Data Cube, SDMX-RDF, SKOS, Dublin Core Terms, VoID) and developed concept schemes. The VoID definition of the SORS LOD dataset is given in Fig.2. The SORS dataset (87.968 triples, see http://stats.lod2.eu/serbia) was also uploaded to the LOD Cloud Cluster knowledge store under the graph name http://elpo.stat.gov.rs/lod2/.



```
rzs:LOD
      rdf:type void:Dataset ;
      rdfs:label "Linked Open Data published by Statistical Office of the Republic of Serbia"@en ;
      dcterms:description "Linked Open Data published by Statistical Office of the Republic of Serbia" ;
      dcterms:modified "2012-04-24"^^xsd:date ;
      dcterms:source <http://webrzs.stat.gov.rs/WebSite/Public/PageView.aspx> ;
      dcterms:subject <http://purl.org/linked-data/sdmx/2009/subject> , <http://dbpedia.org/resource/Statistics> ;
      dcterms:title "SORS Linked Open Data" ;
      void:subset rzs:Prices , rzs:National_accounts , rzs:ICT_Usage , rzs:Science_technology_innovations .
```

Fig. 2. VoID description of the SORS LOD

4 Conclusion and Outlook

This paper contributes to the understanding of the LOD2 tools and technologies and discusses their use for publishing and consuming public sector information through the SORS Use case. The main lessons learnt from this study are:

- The Data Cube RDF vocabulary is mature enough to be used for publishing statistical data as it improves interoperability and allows comparison of data from different statistical sources.
- The LOD2 Stack provides a wide range of data transformation, enrichment and exploitation tools. However, advanced tools for analysis and visualization of statistical data are still under development.
- For publishers who currently only offer static files, Linked Data offers a flexible, non-proprietary, machine-readable means of publication that supports an out-of-the-box web API for programmatic access.
- The Serbian CKAN increases the visibility and accessibility of Serbian public sector data

We conclude that adoption of LOD2 tools and technologies leads to establishment of an interoperable Open Government Data ecosystem. Future work will include an analysis of the LOD2 Stack components for building custom applications for different LOD stakeholders.

Acknowledgements. The research presented in this paper is partly financed by the European Union (FP7 LOD2 project, Pr. No: 257943), and partly by the Ministry of Science and Technological Development of Republic of Serbia (SOFIA project, Pr. No: TR-32010). The Linked Open Data example was realized through close cooperation with the Statistical Office of the Republic of Serbia.

References

1. Auer, S., Martin, M., Frischmuth, P., Deblieck, B.: Facilitation the publication of Open Governmental Data with the LOD2 Stack. Share-PSI workshop, Brussels. Retrieved from http://share-psi.eu/papers/LOD2.pdf (2011)
2. Vraneš, S., Janev, V., Spasić, M., Milošević, U.: Establishment of the Serbian CKAN. LOD2 Deliverable 9.5.1, Institute Mihajlo Pupin (2012).
3. Cyganiak R., Reynolds D., Tennison J.: The RDF Data Cube vocabulary (July 14. 2010).
4. Williams, H., Boncz, P., Tummarello, G., Auer, S.: 50 Billion plus Triple LOD Cloud Hosted on the LOD2 Knowledge Store Cluster. LOD2 Deliverable 2.1.3 (2012).

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 31-37, 2012.

LODGrefine – LOD-enabled Google Refine in Action

Mateja Verlic

Zemanta d. o. o.
mateja.verlic@zemanta.com,
http://www.zemanta.com

Abstract. As a part of LOD2 project we developed several extensions for Google Refine - a simple, yet very powerful open-source tool for working with messy data, to make LOD a first-class citizen in this tool, which is currently tightly coupled with Freebase and has no support for DBpedia. LODGrefine is a version of Google Refine with integrated extensions developed by Zemanta and DERI, adding support for reconciliation with DBpedia, export to RDF, augmentation of data with columns from DBpedia and extraction of entities from full text. Use of LODGrefine will be demonstrated in three use cases.

Keywords: Semantic Web, data cleaning tools, Google Refine, LOD

1 Introduction

Data cleansing and linking are very important steps in the life cycle of linked data [1]. They are even more important in the process of creating new linked data. Data comes from different sources and it is published in many formats, either as XML, CSV, HTML, as a dump from relational databases, or in custom formats like JSON, obtained from different web services. By linking these different bits of data from various sources we can extract information otherwise hidden and in some cases even gain new knowledge.

Unfortunately, these steps are not always trivial for an average user, e.g. a statistician working with statistical government data; on the contrary, they pose a problem even for more skilled researchers working in the field of semantic web. If data is available online this doesn't necessary mean it is ready to be used in semantic applications. In most cases such assumptions are wrong; it is very likely that data has to be cleaned, because *everyone* can publish data on the Web. Taking care of quality of Web data is still one of the main challenges for Semantic Web applications [2]. Data cleansing, especially if done manually, is a very tedious and time consuming task, mostly due to the lack of good tools. Commercial products such as PoolParty [7] provide a wide range of functionalities (thesaurus management, text mining, data integration), but they may not be the best solution when dealing with smaller datasets (in comparison to huge datasets in big companies) and by less proficient users trying to convert data stored in Excel files or flat files.

A good and publicly available cleansing/linking tool should at least be able to: assist user in detecting inconsistent data, quickly perform transformations on a relatively large amount of data, export cleansed data into different formats, be relatively simple to use, and be available for different operating systems. Fortunately, there is one open-source (BSD licensed) solution available, which meets all the criteria mentioned above and even more. It was created especially for dealing with messy data, it is modular based and extendable, it works on all three major operating systems and it already provides functionalities to reconcile data against Freebase. This tool is Google Refine (GR) [4].

GR provides means to reconcile and extend data with data from Freebase, but not from DBpedia. By providing a LOD-friendy version of this tool (LODGrefine) supporting DBpedia we've made an important step towards making LOD a first-class citizen in this powerful, yet easy to use tool. LODGrefine has preserved all of the GR's cleansing and reconciliation functionalities and added new ones to make it even more useful for Semantic Web community.

2 From Google Refine to LODGrefine

GR is currently one of most powerful and user-friendly open-source tools for cleansing and linking data with Freebase. Support for faceted browsing and good filtering possibilities are its main assets, it works fast even when dealing with large amounts of data and it has a built-in support for Google Refine Expression Language (GREL), a special scripting language, which is easy to learn and use to transform data. The most important features of GR are the reconciling and extending data.

It is a server-client web application intended to run locally by one user. Instead of using a database to store imported data, it uses memory data-store, which is built up-front and optimized for GR operations. Its data cleansing and reconciliation abilities are tightly integrated with Freebase (Fig. 1) and making it support a different triplestore offering a SPARQL endpoint, e.g DBpedia, was not possible without implementation.

2.1 LOD extensions

Due to the modular nature of GR architecture it was not required to change the code of GR itself to make it LOD-enabled. We implemented extensions with additional functionalities. Maali and Cyganiak, researchers at Digital Enterprise Research Institute already developed RDF Refine extension [6] for GR, which can be used to reconcile data against any SPARQL endpoint or RDF dump and to export data as RDF based on a user-defined RDF schema.

Extensions (dbpedia-extension) developed by Zemanta complements functionalities of RDF Refine with ability to extend reconciled data with new

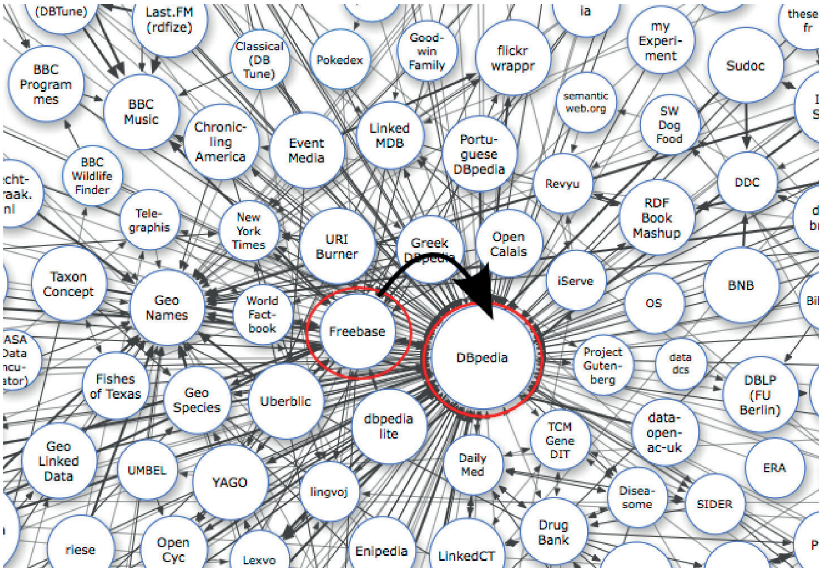


Fig. 1. With LODGrefine we closed the gap between Freebase and DBpedia in the LOD cloud [3].

columns based on data from DBpedia. It also supports extraction of entities from unstructured text using Zemanta API [9]. For example, if we extend reconciled data with description or biography property from DBpedia, we can extract different types of entities from this text and add in new columns to use it in RDF schema, which maps data in columns to nodes in linked graph.

Both extensions are free to use, their code is shared under the BSD License on Github (RDF Refine[1] [5], dbpedia-extension[2] [8]) and binary versions can be obtained from their home pages.

2.2 LODGrefine

To simplify the process of obtaining and installing LOD extensions we decided to integrate them into the latest version of GR (2.5-r2407) and name the LOD-enabled version LODGrefine. Although GR itself does not need any special installation, it is enough to unpack it and run it, the location of extensions depends on the operating system and it is more convenient, especially for first time users, if extensions are already integrated in the tool. Furthermore, we created a Debian package for LODGrefine, which will be integrated into LOD2 Stack, a stack of tools for managing the life-cycle of Lined Data.

[1] https://github.com/fadmaa/grefine-rdf-extension

[2] https://github.com/sparkica/dbpedia-extension

LODGrefine is available under Apache License 2.0 and can be freely downloaded either in binary format or as source code [8].

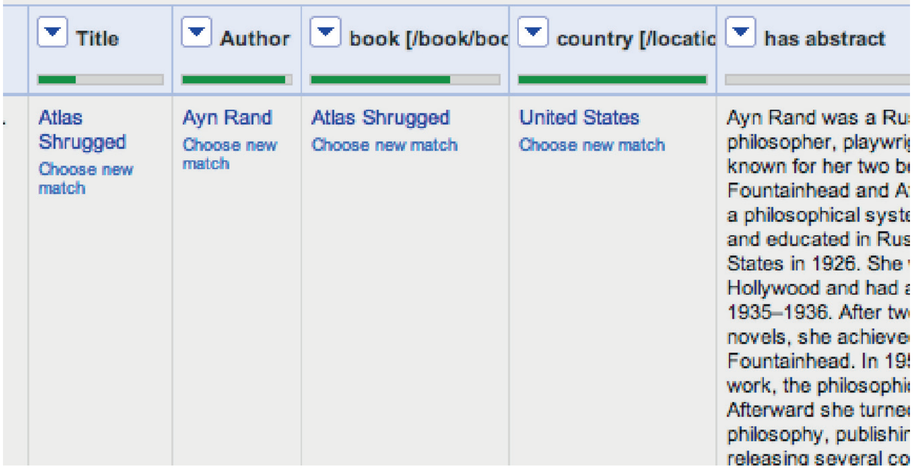
3 LODGrefine in action - use cases

For demonstration we prepared three use cases – examples of how LODGrefine can be used to clean data from different sources and domains and how to transform it into Linked Data.

3.1 100 best novels

In first example we will demonstrate how to convert data from a website first to a LODGrefine project, reconcile it, augment it additional columns from DBpedia and then export it as Linked Data.

In this example we will transform a list of 100 best novels from Modern library web page[3]. The list contains two rows for each novel - first row contains the title and the second one the author, but we need data in columns - one column for title and one for author. Fortunately, LODGrefine has an option to import line based text files and it can read text from clipboard. With some minor changes of default settings our data is imported in columns in few seconds instead of minutes or even hours. With GREL functions we convert titles from uppercase to titlecase and remove *'by'* preceding authors names.



Title	Author	book [/book/boo	country [/locatic	has abstract
Atlas Shrugged Choose new match	Ayn Rand Choose new match	Atlas Shrugged Choose new match	United States Choose new match	Ayn Rand was a Ru philosopher, playwri known for her two b Fountainhead and A a philosophical syste and educated in Rus States in 1926. She Hollywood and had a 1935–1936. After tw novels, she achieve Fountainhead. In 19 work, the philosophi Afterward she turne philosophy, publishir releasing several co

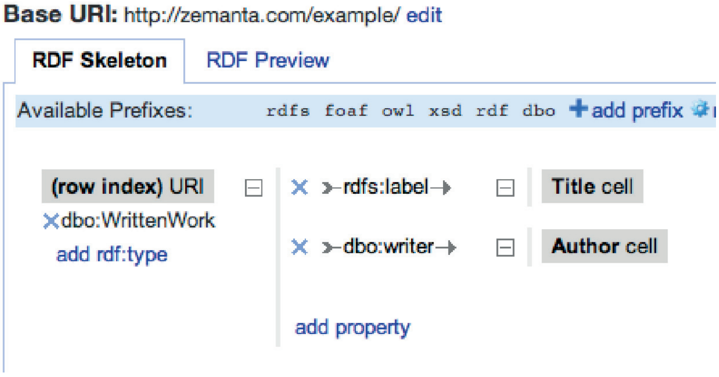
Fig. 2. Reconciled and extended data. Third and fourth column contain entities extracted from autor's biography in the last column obtained from DBpedia.

[3] http://www.modernlibrary.com/top-100/100-best-novels/

Next step is reconciling author names with DBpedia using RDF extension to entity type dbo:Person[4]. After reconciliation data is ready to be extended with *has abstract* property from DBpedia using Zemanta extension (Fig. 2).

The last step of converting online list of novels into Linked Data is configuring RDF schema alignment skeleton, with which we specify how RDF data will be generated (Fig. 3). At any time we can preview the Turtle representation of generated RDF data to see whether schema we defined produced expected results. After the schema has been configured, data can be exported into one of the RDF representations supported by LODGrefine - RDF/XML or Turtle (fig. 4).

In this example we demonstrated how easy it can be to transform data from a website into Linked Data using LODGrefine. We also demonstrated its most important functionalities.



Base URI: http://zemanta.com/example/

Available Prefixes: rdfs foaf owl xsd rdf dbo

(row index) URI
dbo:WrittenWork
rdfs:label → Title cell
dbo:writer → Author cell

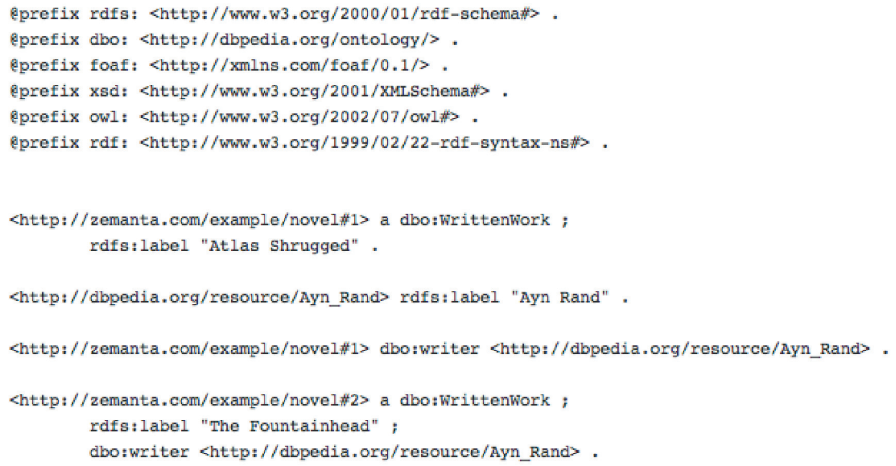
Fig. 3. RDF alignment schema for describing novels.

3.2 CKAN datasets

The Comprehensive Knowledge Archive Network (CKAN) is well known system for storage and distribution of data. It is widely used for storing government data and national registers as well as Data Hub, *the community-run catalogue of useful sets of data on the Internet*[5]. CKAN data is especially interesting for Linked Open Data community, but currently not all CKAN datasets in the Data Hub are provided as RDF (either SPARQL endpoint or RDF dump). A lot of datasets are provided as files with comma separated values (CSV), as Excel files or XML. Our goal is to show how it can be relatively easy transformed into triplets using LODGrefine in a similar way as described in previous example with novels.

[4] http://dbpedia.org/ontology/Person

[5] http://thedatahub.org/about



```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .


<http://zemanta.com/example/novel#1> a dbo:WrittenWork ;
        rdfs:label "Atlas Shrugged" .

<http://dbpedia.org/resource/Ayn_Rand> rdfs:label "Ayn Rand" .

<http://zemanta.com/example/novel#1> dbo:writer <http://dbpedia.org/resource/Ayn_Rand> .

<http://zemanta.com/example/novel#2> a dbo:WrittenWork ;
        rdfs:label "The Fountainhead" ;
        dbo:writer <http://dbpedia.org/resource/Ayn_Rand> .
```

Fig. 4. Turtle representation of first few rows.

3.3 Looking for entities in extracted links

In the last example we will demonstrate the full power of LODGrefine by using it to clean and filter links extracted from blog posts to obtain links, that could be considered as descriptions for entities.

Many times bloggers include links into their blog posts to point the reader to Wikipedia or any other web page, that can be considered as information resource (e.g. Google Maps, Crunchbase[6] - free database of technology companies, Amazon). Ideally, links that could be considered as entity candidates, have non-empty anchor text (entity surface form) and href attribute set to external URL, which directs to a page containing a description of concept/person/object (disambiguation page) mentioned in anchor text.

Unfortunately, blog posts also contain even more links that can be considered as noise or even spam, e.g. links with anchor text without semantic value (e.g. here, this, Read more). We used LODGrefine faceted browsing and filtering abilities to quickly identify patterns of occurring anchor texts or target links, which could be considered either as entity candidates or noise. We used GREL[7] expressions to simply extract features from columns containing anchor texts and target URL, e.g. number of words in anchor texts, flag whether first word in anchor text is capitalized or not, domain part of the target URL, path level of the target URL and more.

When applying faceted browsing on a large number of rows it is not always possible to display all unique values. LODGrefine offers the ability to display all

[6] www.crunchbase.com

[7] GREL - Google Refine Expression Language: http://code.google.com/p/google-refine/wiki/GRELFunctions

different values by choice count, which can be further used in mathematical expressions. For example, if only a few anchor texts appear 100x more frequently than the rest of the anchor texts, it is difficult to filter out anchor texts with occurrences between 20 and 35. In this case it is better to use logarithmic scale. It is worth mentioning, that filtering in LODGrefine works really fast even for 100 000 rows, where some other tools might start having problems.

After filtering we reconciled entity candidates against DBpedia and/or Freebase to link them to existing entities and then exported entity candidates in Turtle representation.

4 Conclusions

LOD-enabled version of Google Refine is one of the best open-source tools for cleaning and linking. With the examples we demonstrated its versatility and powerfulness for transforming tabular data to Linked Data for different problem domains.

5 Acknowledgments

This work was supported by a grant from European Union's 7th Framework Programme (2007-2013) provided for the project LOD2 (GA no. 257943).

References

1. S. Auer, J. Lehmann, and A.-C. N. Ngomo. Introduction to linked data and its lifecycle on the web. In *Reasoning Web*, pages 1–75, 2011.
2. C. Bizer, P. Boncz, M. L. Brodie, and O. Erling. The meaningful use of big data: four perspectives – four challenges. *SIGMOD Rec.*, 40(4):56–60, Jan. 2012.
3. R. Cyganiak and A. Jentzsch. Linking open data cloud diagram. http://lod-cloud.net/.
4. G. Inc. Google refine homepage. http://code.google.com/p/google-refine/.
5. F. Maali and R. Cyganiak. RDF Refine homepage. http://refine.deri.ie/.
6. F. Maali, R. Cyganiak, and V. Peristeras. Re-using cool uris:entity reconciliation against lod hubs. In *Linked Data on the Web*, volume 813. CEUR-WS, 2011.
7. T. Schandl and A. Blumauer. Poolparty: Skos thesaurus management utilizing linked data. In L. Aroyo, G. Antoniou, E. Hyvnen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, editors, *The Semantic Web: Research and Applications*, volume 6089 of *Lecture Notes in Computer Science*, pages 421–425. Springer, 2010.
8. M. Verlic. LodGrefine homepage. http://code.zemanta.com/sparkica/lodgrefine/.
9. Zemanta. Zemanta developers page. http://developer.zemanta.com/.

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 38-42, 2012.

Lightweight Transformation of Tabular Open Data to RDF

Miel Vander Sande, Laurens De Vocht, Davy Van Deursen,
Erik Mannens, and Rik Van de Walle

Ghent University - IBBT
Department of Electronics and Information Systems - Multimedia Lab
Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium
`firstname.lastname@ugent.be`

Abstract. Currently, most Open Government Data portals mainly offer data in tabular formats. These lack the benefits of Linked Data, expressed in RDF graphs. In this paper, we propose a fast and simple semi-automatic tabular-to-RDF mapping approach. We introduce an efficient transformation algorithm for finding optimal relations between columns based on ontology information. We deal with multilingual diversity between data sets by combining translators and thesauri to create context. Finally, we use efficient string and lexical matching approaches in a learning loop to ensure high performance with good precision.

1 Introduction

Every government in the world possesses a huge number of public data sets. Today, the idea of Open Government Data is uprising, where data sets are made freely available on the Web. This way, governments can boost their transparency, create economic value and facilitate governmental participation.

Most portals (e.g., data.gov.be) mainly offer downloadable spreadsheets or tabular records in an XML or CSV like format. Although these initiatives are considered Open Data, they do not reach their full potential. The content in many different sets overlaps, but there are no links identifying this. Also, the data may be machine readable, but not machine understandable, which makes data integration the responsibility of the developer. These obstructions can be tackled with Linked Open Data, where data sources are reformulated in the graph-structured RDF standard. However, current data sets are still structured in classic hierarchical schema's, while graph patterns introduce a whole new way of organising data. Hierarchical patterns can easily be translated into a simple graph, but this is not always the optimal choice. In this paper, we propose lightweight automatic mapping approach for optimally restructuring tabular data as RDF.

2 Related Work

A popular solution is the RDF extension for Google Refine[1]. The transformation structure can be edited, links with other datasets can automatically be resolved and the result is generated in RDF. Although the reconciliation results are very good, the restructuring of the data is fully manual. It does not use structural knowledge extracted from ontologies or other Linked Data sources to automatically suggest an optimal graph. Only supported by search in selected ontologies, the user needs to explicitly specify which Class or Property map each field, requiring many manual operations. Also, Refine is a pure offline application, decreasing accessibility and means to collaborative editing.

More related web-oriented approaches are database abstraction or extraction systems (e.g., D2RServer, Virtuoso, Triplify). However, these transformation methods are based on database schemas and manually defined mappings. The former requires the data source to be a relational database, or at least imported into one. The latter again requires full manual editing without automated assistance. Our approach is fully compliant with these systems, since it can serialize its output as mapping rules. Finally, simple converters like Any23 do not consider any ontology information and create triples using column labels defined in a CSV file. This results into RDF with little semantics.

The closest related work is done in the KARMA [1] framework. A CRF (Conditional Random Fields) machine learning model is trained using following process in an interactive loop. Each column in the dataset is assigned a semantic type from a selected ontology. This assignment is based on user input and formerly learned types. A graph is formed using these types and structural information described in the ontology. Finally, a Steiner tree algorithm is used to extract the right relations. Although this approach achieves very good results in type assignment and tree selection, high precision is only achieved after a considerable amount of manual input.

We propose a more lightweight iteration process that replaces the existing CRF model with string, lexical, data type and context analysis. By using this analysis, we can present a possible mapping before input by the user is required. Note that we accept low precision in the first iteration, to achieve high precision, by fully exploiting user input, in later iterations. The CRF approach is also very memory consuming and copies almost all the data. Our analysis is based on far lighter analysis, resulting in higher performance. Also, stored data are restricted to ontology concepts and data types, which is significantly less. Finally, KARMA learns learns semantic types based on the data format, which are only helpful for data sets in a similar domain. We cover a broader domain of data sets, since we rely more on generic information stored in the fields and ontology, refined by user input.

[1] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

3 Overall approach and basic assumptions

In Linked Data, ontologies define the way data should be structured. Therefore, we will base this approach on the defined concepts and relations. Considering Open Government Data, we make the following assumptions:

- Most data sources only provide column labels and data rows
- These sources cover a wide variety of domains, resulting in heterogeneous data between sets
- Ontologies are available, are well described in OWL. The considered ontologies are average in size
- Ontologies are written in English, while data sets are multilingual
- Since this approach is used in an interactive loop with the user, low precision is acceptable in the first iterations; only if it results into simpler and faster mapping (considering public servant workload).

This approach takes two inputs: a tabular data set and an ontology which we want to match. Serialised mapping rules are constructed in three steps: concept matching, context construction and tree extraction.

3.1 Concept matching

For each column header, we will extract matching concepts from the ontology using header and data type information, as shown in figure 1. During this process, we keep a candidate list containing each retrieved concept with a confidence score between 0 and 1.

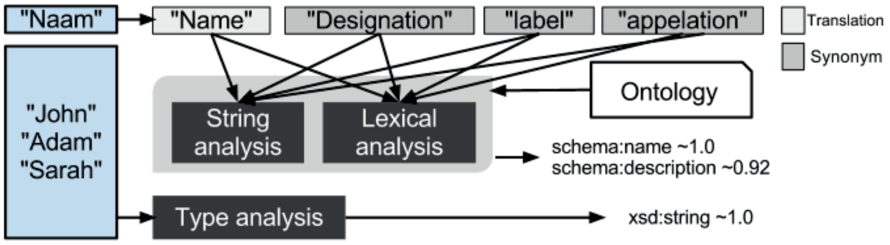


Fig. 1. Example of context extraction for a column 'naam'(dutch)

We start by translating our field label to English using the Microsoft Translator API and MyMemory[2]. The results are combined to form a unique result set. Each entry in this set will be the input of a thesaurus service, giving us a unique set of synonyms. Next, each translation or synonym is matched to the names of all classes and properties using the string and linguistic methods described in [2]. The former calculates the Jaro-Winkler distance with threshold

[2] http://www.microsofttranslator.com/dev/ http://mymemory.translated.net/

0.81 and the Smith-Waterman distance with threshold 10. The latter performs the Jiang-Conrath measure with threshold 1.0, which uses WordNet *Synsets* to score semantic relations between two nouns. The retrieved concept is then added to the candidate list, together with a normalised combination of these results as score. Then, we measure the compatibility of the column's data with the data types in the XMLSchema[3] name space. We sort them in order of compatibility and then look for the best match. The indices of this match and its preceding items indicate the compatibility score of each type.

3.2 Context graph construction

Based on the retrieved concepts, we construct a directed weighted labelled context graph in a three step process, as shown in figure 2. Firstly, we add the field label as a node and add all compatible (compatibility score greater than 0) data type properties as edges with their domain as source. The weight is calculated from their range's compatibility score and, if present in the candidate list, their similarity score. High compatibility and/or similarity result in a low edge weight. Secondly, we add all the classes from the candidate list. Since super classes might have a fitting relation to the field node, we copy each edge with a superclass as source, to the current class with a slightly higher weight (unless the current class is a candidate. E.g., *ex:PersonName* in figure 2). For completeness, we add a *subClassOf* relation between the super classes. Thirdly, we add object properties to connect the current nodes in the graph, based on their domain and range. When an object property is in the candidate list, we use its confidence score to determine the edge weight. If not, the edge gets a default weight.

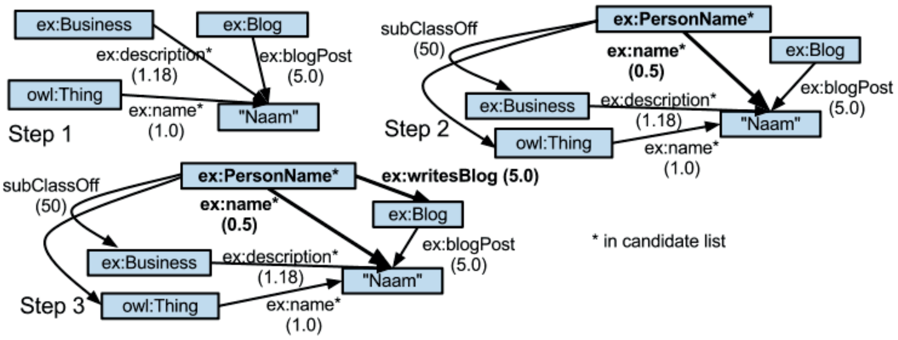


Fig. 2. Example of graph construction for a column 'name'

[3] http://www.w3.org/2001/XMLSchema

3.3 Tree extraction

Finally, we look for optimal paths between the different fields. First, we merge all the different context graphs into one and keeping the lowest edge weights. Second, we use a tree algorithm for finding paths in the graph. For extracting a tree, we use the *Steiner Tree algorithm* as described by Craig A. Knoblock et al. This algorithm finds the minimum-weight spanning tree in a graph between a subset of nodes, called Steiner Nodes. It extends the *minimal spanning tree* algorithm to dynamically add nodes if they provide a shorter path. The result, as shown in figure 3, determines our final mapping.

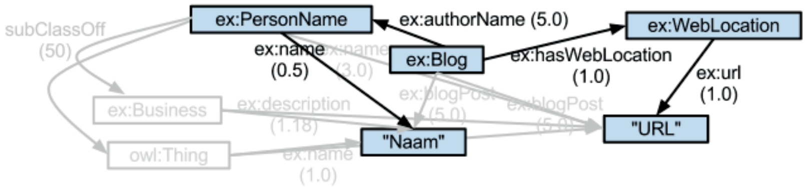


Fig. 3. Extracted steiner tree from merged graph

4 Discussion

We introduced a fast automatic mapping approach for transforming plain tabular data optimally into RDF. It uses a combined method of translation services and thesauri to deal with multilingual data sets. Furthermore, string, lexical and data type analyses is used to match ontology concepts to the different columns. For each column, a context graph is constructed based on the retrieved concepts. Finally, a possible mapping is selected by merging all context graphs and finding a tree connecting all columns.

In future work, an evaluation can be performed by comparing the result again manually defined mappings. This can be done in two ways: the different amount of manual operations and the precision/recall difference between the two. Furthermore, using this approach in a semi-automated loop, could dramatically increase the precision. Corrections by the user can be used to improve the tree selection process, or introduce machine learning for better concept matching.

References

1. Craig A. Knoblock, Pedro Szekely, Jos Luis Ambite, Shubham Gupta, Aman Goel, Maria Muslea, Kristina Lerman, and Parag Mallick. Interactively mapping data sources into the semantic web. In *Proceedings of the 1st Int. Workshop on Linked Science 2011 in Conjunction with the 10th Int. Semantic Web Conference*, 2011.
2. Feiyu Lin and Andrew Krizhanovsky. Multilingual ontology matching based on wiktionary data accessible via sparql endpoint. *CoRR*, abs/1109.0732, 2011.

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 43-47, 2012.

User Controlled Privacy for Filtering the Web of Data with a User-Friendly Manager*

Owen Sacco, Alexandre Passant, and John G. Breslin

Digital Enterprise Research Institute,
National University of Ireland, Galway, Ireland
{owen.sacco,alexandre.passant}@deri.org,{john.breslin}@nuigalway.ie

Abstract. Web of Data applications simplify the process for publishing information such as user's personal information. Although some provide privacy settings, these are limited and users require further options to restrict access to some parts of their data. In this demo paper, we present the extended Privacy Preference Manager (PPM) that enables users to (1) create privacy preferences using the Privacy Preference Ontology (PPO) and (2) grant or restrict access to their data to third-party users based on user profile features for example interests.

1 Introduction

Web of Data applications provide minimum privacy settings such as restricting access to private data to those who are in approved user lists. Yet, users require more complex privacy settings as some current systems do not meet their requirements for example in Social Networks [1]. Current work, for instance [5], use specific policy languages which require users to understand how to interpret privacy rules.

In addition to the full-paper from the research track [2], in this demo paper we describe the implementation and user interface of the extended Privacy Preference Manager (PPM). Moreover, we also demonstrate how it solves the above issues by allowing users to: (1) create fine-grained privacy preferences using the Privacy Preference Ontology (PPO) [3] for data residing in documents or SPARQL endpoints; and (2) let other users access this data, filtered according to these privacy preferences.

2 Overview of The Privacy Preference Ontology (PPO)

The Privacy Preference Ontology (PPO) [3] [2] – http://vocab.deri.ie/ppo# – is a light-weight vocabulary that allows people to describe fine-grained privacy preferences for granting or restricting access to specific Web of Data. Besides

* This work is funded by the Science Foundation Ireland (grant SFI/08/CE/I1380 (Líon 2)) and by an IRCSET scholarship co-funded by Cisco systems.

other use-cases, PPO can be used to grant part of a FOAF user profile only to users that have specific attributes. It provides a machine-readable way to define preferences for instance "Provide my work email address only to DERI colleagues" or "Grant write access to my blog only to my relatives"; assuming that the requester's attributes are trustworthy.

A privacy preference defines: (1) the resource, statement, named graph, dataset or context it applies to; (2) the conditions refining what to grant or restrict, including operators that connect several conditions; (3) the access control type; and (4) a SPARQL query, (`AccessSpace`) *i.e.* a graph pattern that must be satisfied by the user requesting information. The access control type is defined by using the extended Web Access Control (WAC)[1] vocabulary which defines the `Create`, `Read` and `Write` (which also includes `Update`, `Delete` and `Append`) access control privileges.

3 *The Privacy Preference Manager (PPM)*

The *Privacy Preference Manager (PPM)*[2], enforces privacy preferences for filtering data on the Web of Data. Although it is designed to work with any data on the Web of Data[3], we demonstrate how to define privacy preferences for FOAF profiles. Our aim is to illustrate how PPO can be applied to create privacy preferences and how personal information can be filtered based on those preferences.

The *PPM* allows users to manage their privacy preferences and also grants or denies access to user's information when requested by others. During this demo, users can (1) create their own Privacy Preference Manager instance; (2) authenticate to their instance and create privacy preferences for their data such as their FOAF profile; and (3) authenticate to other user's instance and access the filtered data, in this case filtered FOAF profile of these users.

The architecture of the system, illustrated in figure 1, consists of: (1) WebID Authenticator: handles user sign-on using the FOAF+SSL protocol [4]; (2) RDF Data Retriever and Parser: retrieves and parses RDF data such as FOAF profiles from SPARQL Endpoints or RDF documents; (3) Privacy Preferences Creator: defines privacy preferences using PPO; (4) Privacy Preferences Enforcer: queries the RDF data store to retrieve and enforce privacy preferences; (5) User Interface: provides users the environment to create privacy preferences and to view filtered RDF data; and (6) RDF Data store: an ARC2[4] RDF data store to store the privacy preferences[5].

[1] WAC — `http://www.w3.org/ns/auth/acl`

[2] Screencast online – `http://vmuss13.deri.ie/ppmv3/screencast/screencast.html`

[3] Currently Web of Data modelled as RDF

[4] ARC2 — `http://arc.semsol.org`

[5] Although ARC2 was used for the implementation of the Privacy Preference Manager, any RDF store can be used.

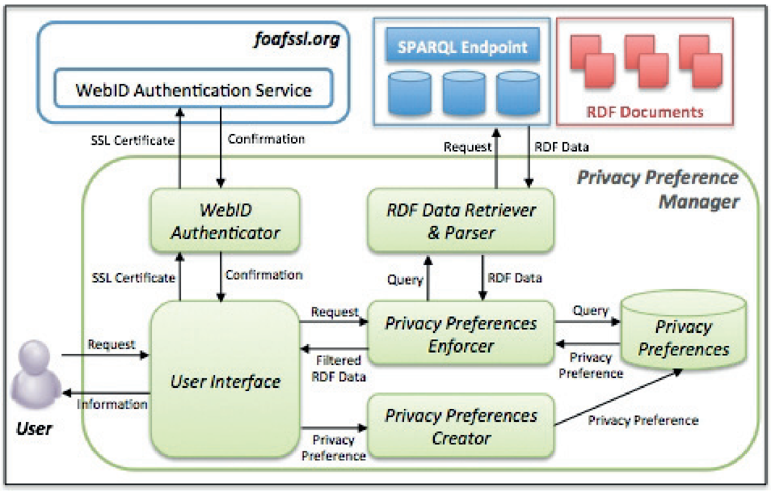


Fig. 1. Privacy Preference Manager

3.1 Creating Privacy Preferences

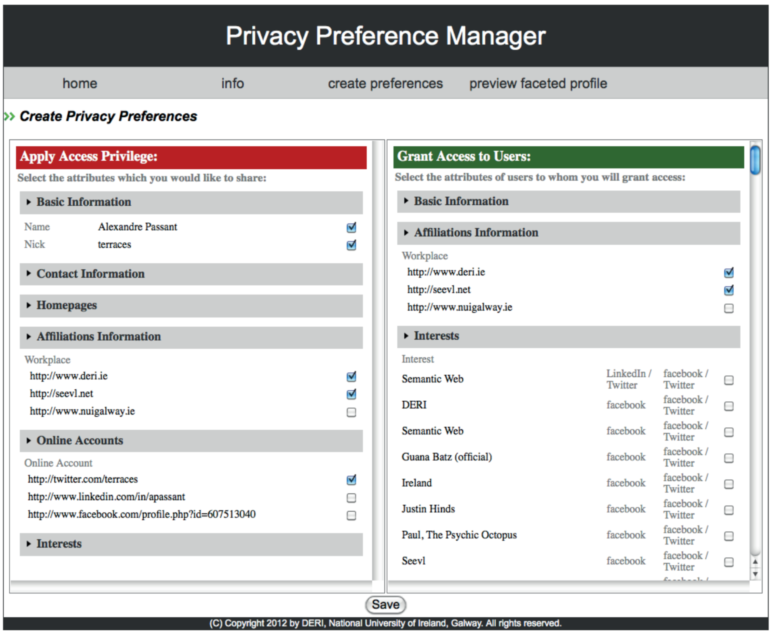
The system's interface provides users to create privacy preferences. Privacy preferences can be applied to triples retrieved from either a manually inputted SPARQL query; or a manually inputted URL; or from a users' FOAF profile extracted from the URL within a WebID certificate.

Once the triples are retrieved, the interface then consists of 2 columns representing: (1) on the left - the profile attributes which the user wants to share; (2) on the right - other attributes that a requester must satisfy to get access to the specific information. Once the choices are validated, the corresponding PPO preferences are created and stored in the system. For example, fig. 2 represents how a user can enable that his name, contact details, homepage and affiliation are available only to people working at DERI.

3.2 Requesting and Enforcing Privacy Preferences

The *PPM* has been extended to support both granting and restricting access; unlike previously that only supported granting access and assumed that everything is private by default unless defined otherwise.

The sequence in which privacy preferences are requested and enforced, consists of: (1) a requester authenticates to another user's manager instance using the WebID protocol so that the system can request the other user's FOAF profile; (2) the privacy preferences of the requested user's FOAF profile are queried to identify which preference applies; (3) the access space preferences are matched according to the requester's profile to test what the requester can access; (4) the requested information (in this case, FOAF data) is retrieved



Privacy Preference Manager

›› Create Privacy Preferences

Apply Access Privilege:

Select the attributes which you would like to share:

▸ Basic Information

Name	Alexandre Passant	☑
Nick	terraces	☑

▸ Contact Information

▸ Homepages

▸ Affiliations Information

Workplace

http://www.deri.ie	☑
http://seevl.net	☑
http://www.nuigalway.ie	□

▸ Online Accounts

Online Account

http://twitter.com/terraces	☑
http://www.linkedin.com/in/apassant	□
http://www.facebook.com/profile.php?id=607513040	□

▸ Interests

Grant Access to Users:

Select the attributes of users to whom you will grant access:

▸ Basic Information

▸ Affiliations Information

Workplace

http://www.deri.ie	☑
http://seevl.net	☑
http://www.nuigalway.ie	□

▸ Interests

Interest

Semantic Web	LinkedIn / Twitter	facebook / Twitter	□
DERI	facebook	facebook / Twitter	□
Semantic Web	facebook	facebook / Twitter	□
Guana Batz (official)	facebook	facebook / Twitter	□
Ireland	facebook	facebook / Twitter	□
Justin Hinds	facebook	facebook / Twitter	□
Paul, The Psychic Octopus	facebook	facebook / Twitter	□
Seevl	facebook	facebook / Twitter	□

(C) Copyright 2012 by DERI, National University of Ireland, Galway. All rights reserved.

Fig. 2. Creating privacy preferences in the Privacy Preference Manager

based on what can be accessed; and (5) the requester is provided with the filtered data she can access.

During the demo, the user is able to view how other users would see his/her profile based on the privacy preferences created which enables the user to validate that the privacy preferences created are the ones that were intended. Moreover, the user can also log into other user's instances to view the filtered information which they can access based on the instance owner's privacy preferences.

4 Conclusion

Privacy Preference Manager provides users to create privacy preferences for their data on the Web of Data and it filters data on the basis of these privacy preferences, which will be demonstrated during the demo session. Although an evaluation is still ongoing, we evaluated the system with 15 users who all confirmed that the filtered FOAF profile is what they expected after creating their preferences. As next steps, we will enhance the user interface based on users' feedback, notably to provide more options from which they can select from.

References

1. D. Boyd and E. Hargittai. Facebook privacy settings. Who cares? *First Monday*, 15(8), August 2010.
2. O. Sacco and J. Breslin. PPO & PPM 2.0: Extending the Privacy Preference Framework to provide finer-grained access control for the Web of Data. In *Proceedings of the 8th Int. Conference on Semantic Systems, I-SEMANTICS'12*, 2012.
3. O. Sacco and A. Passant. A Privacy Preference Ontology (PPO) for Linked Data. In *Proceedings of the Linked Data on the Web Workshop, LDOW2011*, 2011.
4. H. Story, B. Harbulot, I. Jacobi, and M. Jones. FOAF + SSL : RESTful Authentication for the Social Web. *Semantic Web Conference*, 2009.
5. J. Zeiss, R. Gabner, A. V. Zhdanova, and S. Bessler. A Semantic Policy Management Environment For End-Users For End-Users. In *Proceedings of International Conference on Semantic Systems, I-SEMANTICS'08*, 2008.

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 48-54, 2012.

SmartReality: Integrating the Web into Augmented Reality

Lyndon Nixon[1], Jens Grubert[2], Gerhard Reitmayr[2], and James Scicluna[3]
[1] STI International, Neubaugasse 10/15, 1070 Vienna, Austria
lyndon.nixon@sti2.org
[2] Graz University of Technology, Inffeldgasse 16c, 2. Floor, 8010 Graz, Austria
{grubert, reitmayr}@icg.tugraz.at
[3] Seekda GmbH, Grabenweg 68, 6020 Innsbruck, Austria
james.scicluna@seekda.com

Abstract. This poster and accompanying demo shows how Semantic Web and Linked Data technologies are incorporated into an Augmented Reality platform in the SmartReality project and form the basis for enhanced Augmented Reality mobile applications in which information and content in the user's surroundings can be presented in a more meaningful and useful manner. We describe how things of interest are described semantically and linked into the Linked Open Data cloud. Relevant metadata about things is collected and processed in order to identify and retrieve related content and services on the Web. Finally, the user sees this information intuitively in their Smart Reality view of the reality around them.

Keywords. Augmented Reality, mobile, Semantic Web, Linked Data, Web services, Web APIs, Web content, Things of Interest.

1 Introduction

SmartReality is a project which began in October 2010 as a nationally funded project in Austria. The participating partners are STI International, Technical University of Graz, Seekda GmbH and play.fm GmbH. Together they represent expertise and innovation in Augmented Reality (AR), semantic technology, Web services and online media. In the context of the project, they explore how people may be able to access relevant information, content and media about things of interest in their vicinity via AR. The AR is enhanced dynamically by the Web of data and the Web of services. This enhancement is made possible by a SmartReality platform which mediates between the client device and the Web-based data and services, focused on using available metadata to select the most appropriate content and services for display to the user. This poster and accompanying demonstrator will present the first results and prototypes of the project. We will explore and demonstrate how semantic technology and Linked Data can be integrated with AR in order to make a user's reality "smarter", based on a scenario with street club posters.

2 SmartReality Vision and Scenario

The value of the ideas of Semantic Web and Linked Data to the AR domain has been reflected in recent position papers, not only from SmartReality [1] but also other researchers [2]. A mobile application, ‘Mobile cultural heritage guide’ [3], also explored use of Linked Data and faceted browsing in combination with device GPS information and cultural artefact metadata. These apps make use of (often imprecise) GPS positioning and are designed to work in specific domains, with limited ability to make use of new data sources or adapt the presentation of resulting content around the objects in the AR focus. The vision of SmartReality is to leverage a richer description of points of interest (POIs) interlinked with the broad source of concept metadata found in Linked Data. This forms the basis of enabling dynamic and relevant content selection and presentation at the AR client. This results in a more useful, aware and flexible Augmented Reality experience.

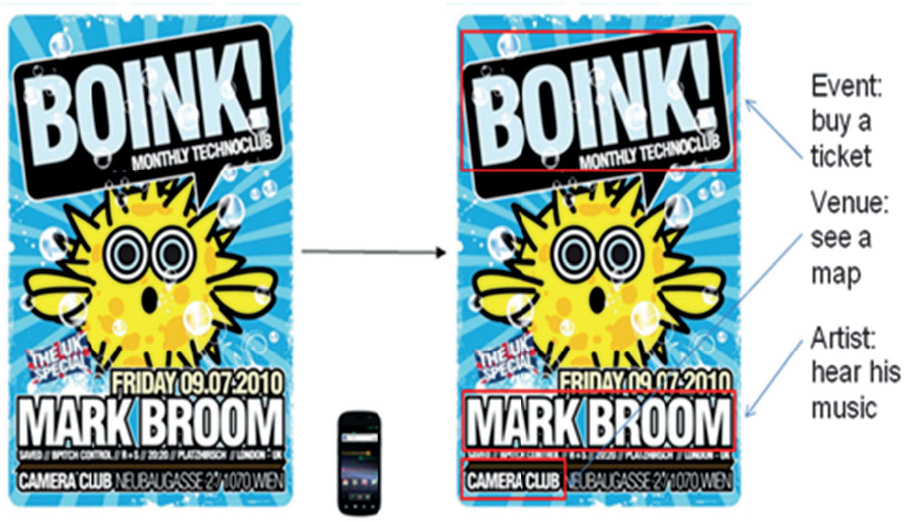


Fig. 1. SmartReality poster scenario

For our initial application domain, SmartReality is focusing on music. Music is a common and increasingly shared experience via the Internet. The user group most likely to be early adopters of Smart Reality solutions are young professionals who are typically interested in listening to music, discovering artists and attending concerts. Together with our partner Play.fm GmbH, whose web site and mobile apps offer access to more than 18 000 DJ mixes and live recordings to 150 000+ users a month, our goal is to use semantics and Linked Data to dynamically link references to music around us in our current reality to virtual information, content and services from the Internet which enhance our experience of music in our current reality. In the SmartReality use case, for example, we consider a music-conscious person wandering city streets and seeing the ubiquitous street posters advertising concerts and club nights. Now, at best, if this person is interested in the content they may have mobile Internet and can begin to search on aspects like the artist, event or venue. However, this presupposes they can identify these aspects from the poster and they still must gather and link together the information they are interested in across various Web searches, e.g. find out about the artist, find some audio to listen to from the artist, check where is the venue where the artist is playing, find out how to get there, find out how to get tickets for the concert. However, with Smart Reality they can view an enriched version of the poster with related content automatically overlaid over the poster references to artists, events, venues and other things. This is illustrated in Figure 1.

3 SmartReality implementation

A SmartReality server handles the interaction between the client and the Web of data and services. A simplified illustration of the steps taken in SmartReality is given below (Fig. 2). First, the object in the mobile devices camera view is identified via an image recognition service (we use Kooaba[1]), which returns an identifier for the object. This identifier is linked with a description of a "Thing of Interest" (TOI) in a datastore we term a "TOI Repository". The TOI description, enriched by links to concepts in the Web of Data, is processed in order to select the most appropriate content and services from the Web. The resulting content is packaged and sent to the client for display in the AR view.

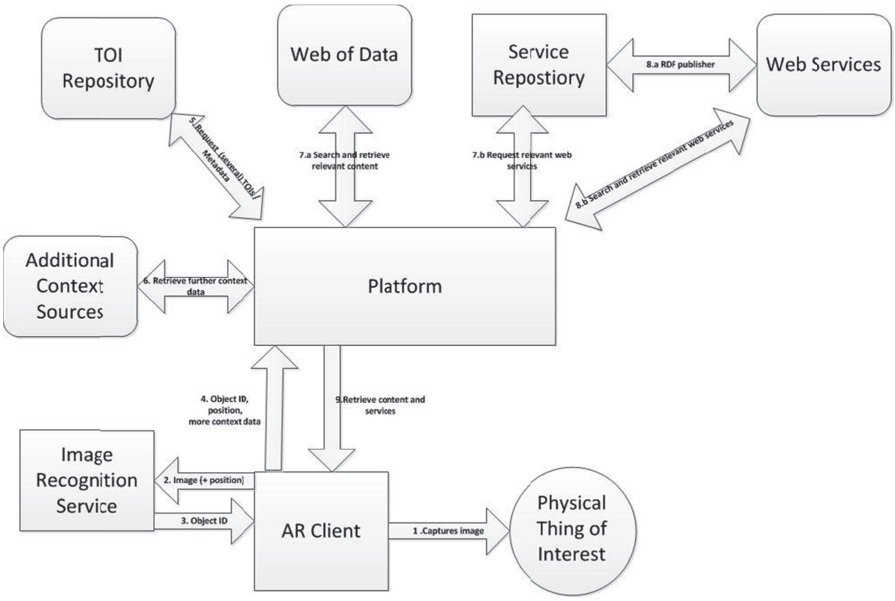
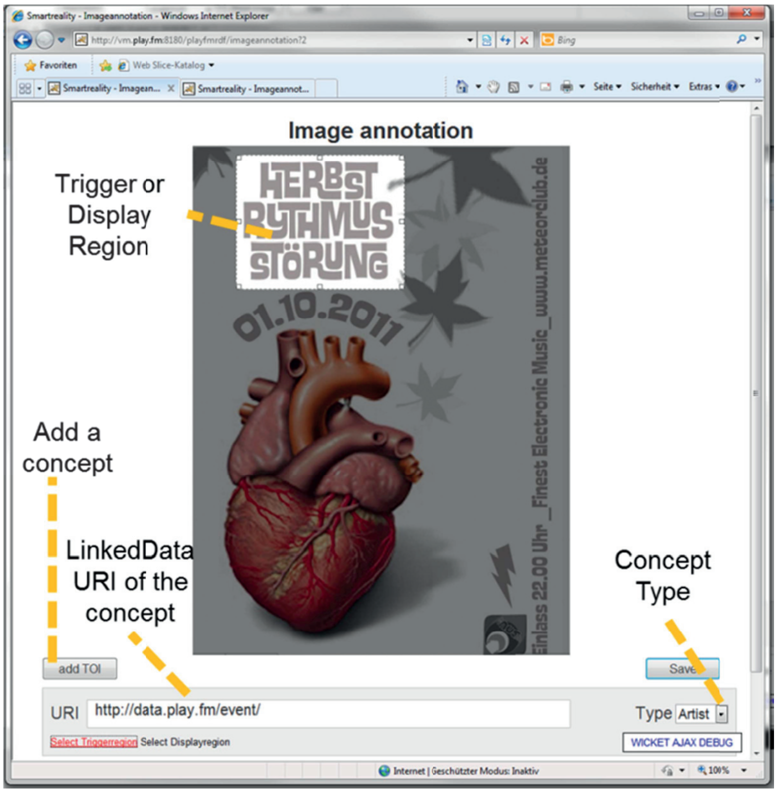

TOI Repository
Web of Data
Service Repostiory
Web Services
8.a RDF publisher
5.Request (several) TOIs/ Metadata
7.a Search and retrieve relevant content
7.b Request relevant web services
8.b Search and retrieve relevant web services
Additional Context Sources
6. Retrieve further context data
Platform
4. Object ID, position, more context data
9.Retrieve content and services
Image Recognition Service
2. Image (+ position)
3. Object ID
AR Client
1 .Captures image
Physical Thing of Interest

Fig. 2. SmartReality workflow illustration

[1] [1] http://www.kooaba.com

3.1 Annotation

To ease the process of generating the initial metadata about Things of Interest (TOI), we implemented a Web based annotation tool (Fig. 3). The tool currently only supports selecting street posters from play.fm's image database. Here, the user selects regions for triggering the appearance of content as well as regions where to actually display the content over the poster. Instead of adding a concrete link to content from the selected regions, users rather select a LinkedData URI representing a concept from an existing conceptual scheme. In this case we use the Linked Data identifiers for play.fm (artists, events, clubs) and support their addition by allowing free text entry and Ajax-based concept selection (automatically filling in the full URI of the concept). The user is also free to use a full URI from any other Linked Data source. When editing is finished, the annotation tool generates a Thing of Interest (TOI) annotation for the poster and stores it in a TOI repository. Additionally, the image of the event poster is uploaded to Kooaba to make it possible to identify the poster at runtime. The TOI data model has been created in RDF/S specifically for SmartReality and is published at http://smartreality.at/rdf/toi.



Smartreality - Imageannotation - Windows Internet Explorer

http://vm.play.fm:8180/playfmrdf/imageannotation?2

Image annotation

URI http://data.play.fm/event/

Type Artist

Bing
Favoriten
Web Slice-Katalog
Smartreality - Imagean...
Smartreality - Imageannot...
Seite
Sicherheit
Extras
Trigger or Display Region
HERBST RYTHMUS STÖRUNG
01.10.2011
Einlass 22.00 Uhr _Finest Electronic Music_ www.meteorclub.de
Add a concept
LinkedData URI of the concept
Concept Type
add TOI
Save
Select Triggerregion
Select Displayregion
WICKET AJAX DEBUG
Internet | Geschützter Modus: Inaktiv
100%

Fig. 3. SmartReality annotation tool.

3.2 Server

The server is developed as a set of components which interchange messages to realize the SmartReality functionality expressed in the above workflow (Fig. 2). The platform has been developed in Ruby on Rails and exposes a RESTful API to clients. The repositories and APIs used by the components to retrieve data into the workflow are separated from the code of the core components so that different storage and remote API solutions (including cloud) could be used as required. After parsing the TOI's metadata (the TOI being identified via the Kooaba identifier which is included in its description in the repository), it provides an initial response to the client which identifies the TOI's regions of interest for display in the AR view (see below, Fig. 4 left). Two further functional steps are realized on the server to provide the content bundle for the enrichment of the TOI's regions of interest in the AR view with links to content:

• Linked Data consumption. The (Linked Data) concepts used to annotate the TOI's regions and extracted from the TOI's metadata are crawled and further, related concepts extracted as defined in a set of Linked Data crawling rules. The rule syntax makes use of LDPath[2]. As a result, a local repository of relevant structured metadata about the concepts of interest in the TOI has been created. We use mainly play.fm Linked Data[3] in the current demo, while the approach is vocabulary-independent, i.e. any Linked Data could be used in the annotation and supported in this step. For the Linked Data step, we make use of the Linked Media Framework (LMF[4]), which provides on top of a RDF repository the means to directly cache Linked Data resources and synchronize their metadata with the description on the Web. This means for a new annotation the LMF will automatically use the locally cached resource metadata in available rather than repeatedly retrieve it from the Web which can lead to latency in the platform response.

• Service selection and execution for content retrieval. Based on this local metadata, a service description repository is queried. Services or APIs are described in terms of their conceptual inputs and outputs so that, for the given class of a concept in the annotation, appropriate services can be found to provide content for the enrichments of the TOI in the AR view. In the current demo, we use an API provided by play.fm to access audio streams of recordings by artists as well as an API provided by Seekda to link an event to a booking interface for nearby hotels with rooms available on the night of the event. Service execution will require querying the concept descriptions to extract the necessary input values – e.g. for the hotel booking interface, the service API needs the event's data and its location's longitude and latitude to be passed in the input request. Likewise, the service response needs to be parsed to return to the SmartReality platform the content inks which can be used to enrich the TOI with respect to the original concept. For this, we use the concept of "lowering" and "lifting" in the Linked Services approach [4] where the semantic concept is 'lowered' to datatype values for the input request to the service, and the datatype values from the service

[2] http://code.google.com/p/ldpath/wiki/PathLanguage

[3] http://data.play.fm

[4] http://code.google.com/p/lmf/

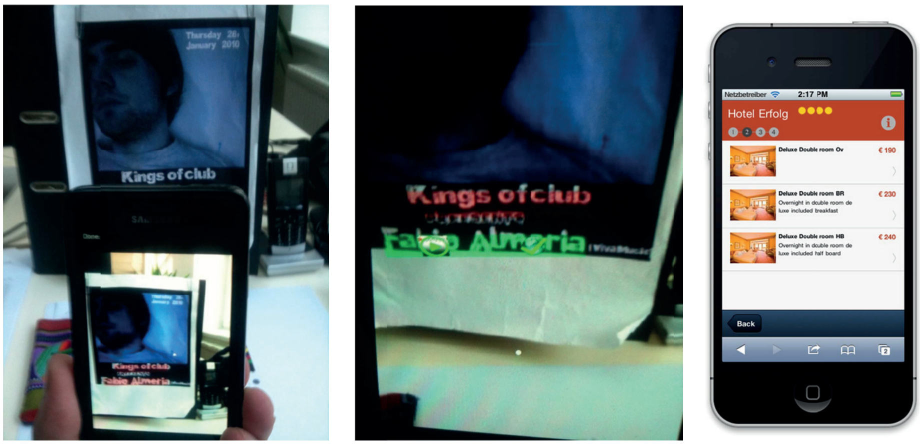
output response are 'lifted' to new semantic concepts (e.g. from a Place to images of Maps of the place).

The "lifted" responses are collected and sent as a content bundle to the client in JSON.

3.3 Client

We built a client application prototype for Android smartphones. It leverages an Augmented Reality interface visualizing relevant information that is spatially registered on the physical object via Natural Feature Tracking. A user points her smartphone on the physical thing of interest to initialize a TOI query. After successful initialization, segments containing relevant information are highlighted through an Augmented Reality interface on the physical object (Fig.4 below left). The user can now point towards individual segments and obtain detailed information (Fig.4 below right). The rendering of content that is spatially registered to segments on the physical object in 3D space is based on OpenSceneGraph[5].

The SmartReality demo will use two real club event posters with the installed client on Android smartphones to give visitors the experience of SmartReality for themselves.



Netzbetreiber 2:17 PM

Hotel Erfolg

1 2 3 4

Deluxe Double room Ov € 190

Deluxe Double room BR € 230
Overnight in double room de luxe included breakfast

Deluxe Double room HB € 240
Overnight in double room de luxe included half board

Back

Fig. 4. SmartReality view in the client. Left: poster is recognized and the regions with content are indicated. Middle: on pressing a content region, the available content items are shown. Here the artist Fabio Almeria is associated with an audio stream from play.fm and a web link to booking a hotel room for after his next concert from Seekda. Right: following the web link the user is at a hotel room booking screen (date and location is not input, as it is known from the event's description)

[5] http://www.openscenegraph.org

4 Future Work

The SmartReality project has focused on a proof of concept with club event posters and enrichment via LOD from mainly the play.fm database. The infrastructure developed has been deliberately designed to separate distinct data and content sources from the workflow which realizes a SmartReality experience, i.e. the use of other objects as "Things of Interest", the annotation with other LOD sources, or the linkage to content from other providers for display in the AR view, should be feasible as a configuration issue and not require any changes to the SmartReality platform or client.

5 Acknowledgements

This work has been performed in the Austrian project SmartReality (http://www.smartreality.at) which is funded by the FFG.

6 References

1. Nixon, L., Grubert, J. and Reithmayer, G.: Smart Reality and AR Standards, at the 2nd international Augmented Reality Standards meeting, Barcelona, Spain, 2011.
2. Reynolds, V., Hausenblas, M., Polleres, A., Hauswirth, M., Hegde, V.: Exploiting Linked Open Data for Mobile Augmented Reality. In: W3C Workshop on Augmented Reality on the Web, Barcelona, Spain, 2010.
3. Van Aart, C., Wielinga, B. and van Hage, W.: Mobile cultural heritage guide: location-aware semantic search. In: Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW '10), Lisbon, Portugal, 2010.
4. Pedrinaci, C., Liu, D., Maleshkova, M., Lambert, D., Kopecky, J., Domingue, J.: iServe: a linked services publishing platform. In: Ontology Repositories and Editors for the Semantic Web Workshop, Heraklion, Greece, 2010.

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 55-62, 2012.

ConnectME: Semantic Tools for Enriching Online Video with Web Content

Lyndon Nixon[1], Matthias Bauer[1], Cristian Bara[2], Thomas Kurz[3], and John Pereira[3]
[1] STI International, Neubaugasse 10/15, 1070 Vienna, Austria
{lyndon.nixon, matthias.bauer}@sti2.org
[2] Seekda GmbH, Neubaugasse 10/15, 1070 Vienna, Austria
cristian.bara@seekda.com
[3] Salzburg Research, Jakob Haringer Str. 5/3, 5020 Salzburg, Austria
{thomas.kurz, john.pereira}@salzburgresearch.at

Abstract. This poster and demo submission presents a set of tools and an extended framework with API for enabling the semantically empowered enrichment of online video with Web content. As audiovisual media is increasingly transmitted online, new services deriving added value from such material can be imagined. For example, combining it with other material elsewhere on the Web which is related to it or enhances it in a meaningful way, to the benefit of the owner of the original content, the providers of the content enhancing it and the end consumer who can access and interact with these new services. Since the services are built around providing new experiences through connecting different related media together, we consider such services to be Connected Media Experiences (ConnectME). This paper presents a toolset for ConnectME – an online annotation tool for video and a HTML5-based enriched video player – as well as the ConnectME framework which enables these media experiences to be generated on the server side with semantic technology.

Keywords: Hypervideo, clickable video, Web media, Linked Data, media linking, annotation, enrichment

1 Introduction

ConnectME is a project which began in June 2011 as a nationally funded project in Austria. The participating partners are STI International (research.sti2.org), Salzburg Research (www.salzburgresearch.at), Planet Digital (www.planet-digital.com) and Yoovis GmbH (www.yoovis.tv). The goal of ConnectME is to develop a hypervideo platform based on open Web standards for the delivery of interactive video experiences and Web services which support the conceptual annotation of video, Web-based linkage between concepts and content, and on-the-fly augmentation of video with content including aspects of personalisation and contextualisation. In this submission, we present the Web based annotation tool for video, which generates storable and sharable RDF based media annotations, the Web based hypervideo player, and the ConnectME framework, which extends an existing system known as the Linked Media Framework, which handles the server side processing from the media annotations to the final content for the enriched video.

2 Related Work

While hypervideo – the idea of hyperlinking to content from within video – has been around since the 1980s[1], the combination of online video, semantic annotation and Web linking in ConnectME is to the best of the authors' knowledge unique in the field. Online video is a clear trend in media consumption, yet the automated association of videos to related Web material is still a subject of technology demos like Mozilla's Popcorn[2] which uses textual tags associated to video to link into Wikipeda articles, maps and so on. Semantics could solve the inherent ambiguities of textual tagging. Work on semantic annotation of video has focused on using the rich metadata captured in improving multimedia indexing, search and retrieval, but the role it could play in enabling an enriched playout of the video is taken up anew in ConnectME. Traditionally, multimedia presentation systems [1] have indeed relied on formal knowledge about the multimedia but not agreed on a shared model for that knowledge. Earlier work on the Cuypers presentation engine [2] did explore use of RDF based knowledge models [3]. The emergence of Linked Data has meant semantic annotations can refer to freely accessible Web based metadata which can be reused in UIs for content selection and browsing, but work has gone not much further than the limited media linked to directly from Linked Data descriptions [4]. Automated linking from semantic annotations to online content related to the annotation needs to incorporate Multimedia Information Retrieval techniques [5] and benefit from increased publication of media metadata in a structured/Linked Data form [6]. The state of the art in Web hypervideo today does not have answers to these issues being addressed by research in ConnectME, and hence focuses on manual annotation and linkage to other content in the video (see Web based offers by companies such as WireWax, Videoclix, Overlay.TV or Klickable).

3 ConnectME workflow

The ConnectME workflow uses a set of executable Web services called from a server side platform which also provides for the workflow's data storage and retrieval in order to generate, from the starting point of a semantic annotation of an online video, a final set of content linked to spatial and temporal moments in the video that can be played out as a form of dynamic content enrichment in the ConnectME hypervideo player. Figure 1 provides a high level view of this workflow. The main steps in the workflow, printed on the left, are to identify objects in video, annotate them with

[1] Systems such as Hypersoap (www.media.mit.edu/hypersoap/) demonstrated the possibility of interactive product placement in a broadcast setting

[2] http://webmademovies.etherworks.ca/popcorndemo/

(Linked Data[3]) concepts and make use of this annotation to link the video objects to other Web content. From an implementation perspective, this means a hypervideo annotation tool (section 3) to help humans generate the semantic annotation of video, a ConnectME framework (section 4) which orchestrates the use of various components to support the steps of annotation, concept selection, content selection and packaging for playout on the client, and a hypervideo player (section 5) to give the client access to the enriched video.

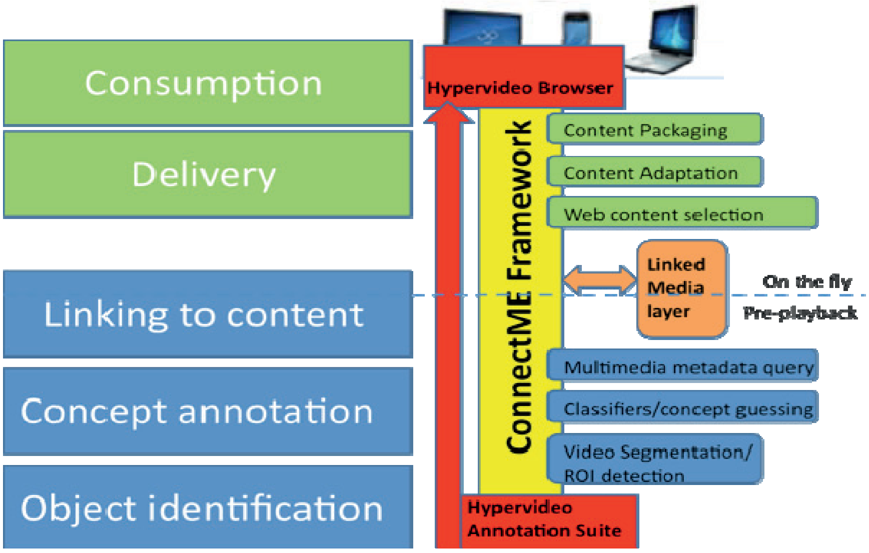


Fig. 1. ConnectME workflow

4 ConnectME Annotation Tool

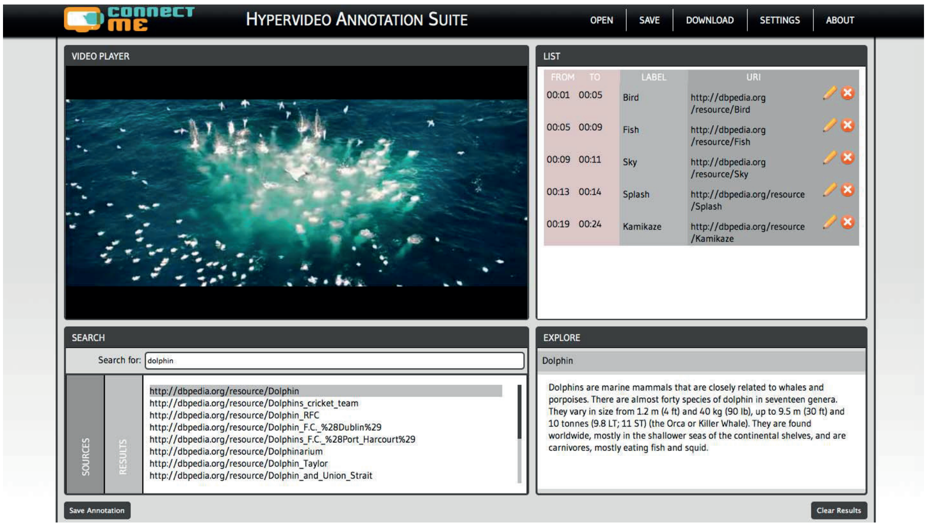
ConnectME has developed a Web-based hypervideo annotation tool in PHP. The user interface uses HTML5, the Video.js player and jQuery with extensions to provide for video loading and manipulation, such as selecting spatial and temporal parts of the video, and hence works across all latest versions of Web browsers.

Using the HTML 5 video tag for embedding video files allows playing videos without need for any Flash-based plug-in and the Video.js library provides several useful video control methods. Ajax powers as-you-type concept suggestion from the DBPedia[4] concept base to support annotators in quickly finding the right concept: also when a concept is selected, the bottom right hand area shows some explanatory text (the DBPedia concept's abstract) to help annotators be sure they choose the correctly

[3] http://linkeddata.org provides a Web based concept space with URI based metadata look-up for more information about concepts

[4] http://dbpedia.org publishes Wikipedia data in RDF, hence every Wikipedia page has a DBPedia Linked Data URI for its subject

intended concept. Furthermore, the annotation tool supports searching for geographic locations in Geonames[5]. The results are displayed by using the Google Maps JavaScript API V3. Locally the annotation being made is stored as JSON in a file on the server and the file itself gets referenced by using a Cookie, so that even if the browser page is reloaded or closed/opened again, the annotation task can be continued without loss of information. When the annotation is saved, the tool is directly connected to an instance of the ConnectME framework which stores the annotations and makes them available to the ConnectME workflow when the video is requested from the hypervideo player. For the video annotation schema a RDF based format has been selected which is based around the W3C Media Ontology with some extensions for enabling an annotation of the annotations (who made them, when, what are the rights for re-use) re-using the Open Annotation Model, and a backwards compatible ConnectME specific extension for describing how concepts are represented by the video object, which is leveraged in the framework.



FROM	TO	LABEL	URI
00:01	00:05	Bird	http://dbpedia.org /resource/Bird
00:05	00:09	Fish	http://dbpedia.org /resource/Fish
00:09	00:11	Sky	http://dbpedia.org /resource/Sky
00:13	00:14	Splash	http://dbpedia.org/resource /Splash
00:19	00:24	Kamikaze	http://dbpedia.org/resource /Kamikaze

Search for: dolphin

http://dbpedia.org/resource/Dolphin
http://dbpedia.org/resource/Dolphins_cricket_team
http://dbpedia.org/resource/Dolphin_RFC
http://dbpedia.org/resource/Dolphin_F.C._%28Dublin%29
http://dbpedia.org/resource/Dolphins_F.C._%28Port_Harcourt%29
http://dbpedia.org/resource/Dolphinarium
http://dbpedia.org/resource/Dolphin_Taylor
http://dbpedia.org/resource/Dolphin_and_Union_Strait

Dolphin

Dolphins are marine mammals that are closely related to whales and porpoises. There are almost forty species of dolphin in seventeen genera. They vary in size from 1.2 m (4 ft) and 40 kg (90 lb), up to 9.5 m (30 ft) and 10 tonnes (9.8 LT; 11 ST) (the Orca or Killer Whale). They are found worldwide, mostly in the shallower seas of the continental shelves, and are carnivores, mostly eating fish and squid.

Fig. 2. ConnectME annotation tool

5 ConnectME Framework

ConnectME draws on the concept of Linked Media to enable a Web based connection between concepts from the Linked Data concept space and Web content which, for the purposes of this linking, have been annotated in terms of Linked Data concepts. The key principles of Linked Media are:

Web media needs to be annotated in terms of its online parts

[5] http://www.geonames.org offers a geographical database which covers all countries and contains over eight million placenames all over the world.

Web media needs to be annotated with terms which represent a shared understanding of a domain or identification of a thing

Web media needs to be annotated using a media ontology which supports the above two issues

The expressed representation of different concepts by different media fragments in different ways shall be the basis to interlink media across the Web

The first three points are covered in the annotation tool (W3C Media Fragments specification, Linked Data as concept namespace, W3C Media Ontology and extensions as annotation scheme). The fourth point is part of a Linked Media implementation in our framework.

5.1 Linked Media Framework

The Linked Media Framework[6] is an easy-to-setup server application that bundles central Semantic Web technologies to offer advanced services. The Linked Media Framework consists of LMF Core and LMF Modules. The core component of the Linked Media Framework is a Linked Data Server that allows to expose data following the Linked Data Principles. The Linked Data Server implemented as part of the LMF goes beyond the Linked Data principles by extending them with Linked Data Updates and by integrating management of metadata and content and making both accessible in a uniform way. As extension for the LMF Core, we are working on a number of optional modules that can be used to extend the functionality of the Linked Media Server:

- LMF Permissions implements and extends the WebID and WebACL specifications for standards-conforming authentication and access control in the Linked Media Framework
- LMF Media Interlinking will implement support for multimedia interlinking based on the work in the W3C Multimedia Fragments WG and the W3C Multimedia Annotations WG
- LMF Reasoner implements a rule-based reasoner that allows to process Datalog-style rules over RDF triples; the LMF Reasoner will be based on the reasoning component developed in the KiWi project, the predecessor of the LMF
- LMF Versioning implements versioning of metadata updates; versioning itself is already carried out by LMF Core, but the management of versions will be carried out by this module
- LMF Enhancer offers semantic enhancement of content by analysing textual and media content; the LMF Enhancer will build upon UIMA, Apache Tika, and the semantic lifting engine of the Apache Stanbol framework[7]

[6] http://www.newmedialab.at/LMF

[7] http://incubator.apache.org/stanbol/

5.2 Connected Media Framework

In order to implement the Connected Media Framework, we chose to build upon the Linked Media Framework, which already offered out of the box much of the necessary basis functionality such as storage and retrieval of the semantic media annotations, as well as a means to access media or its metadata in a straightforward manner, following Linked Data principles. Since additional functionalities are plugged in via modules, ConnectME develops its own specific modules to turn the Linked Media Framework into a Connected Media Framework: The concept extraction module supports the video annotation tool by suggesting concepts to link to the video via textual analysis of available subtitling or transcript files for that video. For this, an instance of Apache Stanbol has been specifically trained to handle the particular corpus of concepts in ConnectME materials. The Linked Media engine is a specific component implementation which exposes multimedia object descriptions to the ConnectME workflow in a common structured metadata format. The provision of usable multimedia object descriptions on the Web as Linked Data is referred to in the project as the Linked Media layer. To find objects relevant for any concept in the video annotation, media repositories need to be queried and their responses provisioned as Linked Media. Hence the engine incorporates a semantic service middleware which brokers between ConnectME and heterogeneous media sources (Web APIs, SPARQL endpoints, etc.).

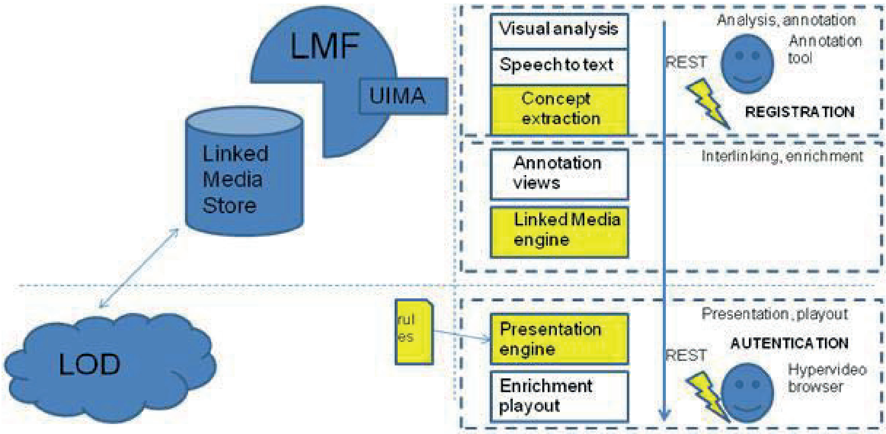
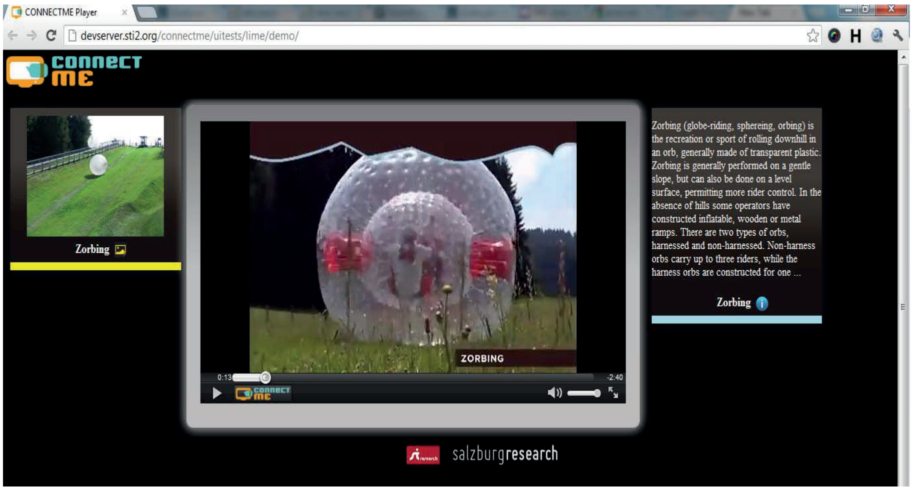


Fig. 3. ConnectME extending the Linked Media Framework.

6 ConnectME Hypervideo Player

The current implementation of the hypervideo player is HTML5 based (making us of the video tag, video.js, JQuery, CSS3, JSON2, backbone.js, underscore.js, RDFQuery and VIE library) and runs in any latest version of any of the main browsers. The player incorporates support for the W3C Media Fragment syntax that should allow video to be accessed not as an entire media resource but in terms of a temporal and/or spatial part thereof. As the video plays, a Javascript code checks for annotations on the next active video segment, and enables access to additional content when it is relevant to the concept annotating that segment via a plugin and widget architecture. Annotations refer to Linked Data resources and the ConnectME framework has collected links to content relevant to those resources. The Hypervideo player has a core that sustains the video playback mechanism and connects to the ConnectME Framework to retrieve the annotations in an initialization phase. A set of plugins is then attached to the core, each of which is specialized in recognizing a certain type of annotation resource. Plugins will retrieve and render relevant content for given resources and display them in the form of widgets. In Figure 4, a depiction plugin rendered a picture of Zorbing and created a widget, displayed on the left hand side; an abstract plugin fetched the abstract and concept label fields from DBPedia that describes Zorbing and composed a widget displayed on the right hand side. Since plugins can be configured for any Linked Data source, the player architecture is very flexible regarding the content selected and displayed in a widget.



Zorbing (globe-riding, sphereing, orbing) is the recreation or sport of rolling downhill in an orb, generally made of transparent plastic. Zorbing is generally performed on a gentle slope, but can also be done on a level surface, permitting more rider control. In the absence of hills some operators have constructed inflatable, wooden or metal ramps. There are two types of orbs, harnessed and non-harnessed. Non-harness orbs carry up to three riders, while the harness orbs are constructed for one ...

Zorbing

Fig. 4. ConnectME hypervideo player.

7 Conclusions

The ConnectME annotation tool, framework and player act as a proof of concept for semantics-based dynamic enrichment of videos based on Linked Data annotations. In the project we continue to explore how we can better automate and simplify the annotation step for the user, maximize the flexibility and relevance of the linkage to online media resources, and improve the intuitiveness of the user interaction with the hypervideo player in order to easily and effectively access and browse video enrichments, on desktop, tablet and SmartTV platforms.

8 Acknowledgements

This work has been performed in the Austrian project ConnectME (http://www.connectme.at) which is funded by the FFG.

9 References

1. Bordegoni, M., Faconti, G., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias, P., Wilson, M.: A Standard Reference Model for Intelligent Multimedia Presentation Systems. Computer Standards & Interfaces, 18:477-496, 1997.
2. Ossenbruggen, J. R., Cornelissen, F. J., Geurts, J. P., Rutledge, L. W., Hardman, L. Cuypers: a Semi-Automatic Hypermedia Generation System. Technical Report. UMI Order Number: INS-R0025., CWI (Centre for Mathematics and Computer Science), 2000.
3. Lindley, C. A., Davis, J. R., Nack, F., Rutledge, L. W.: The Application of Rhetorical Structure Theory to Interactive News Program Generation from Digital Archives. Technical Report. UMI Order Number: INS-R0101, CWI (Centre for Mathematics and Computer Science), 2001.
4. Schreiber, G., et al.: Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture, Demonstrator. Journal of Web Semantics (JWS), 6(4):243–249, 2008.
5. Hanjalic, A., Lienhart, R., Ma, W.-Y., Smith, J. R.: The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? In: Proceedings of the IEEE, 96(4):541–547, 2008.
6. Bürger, T., Simperl, E.: A Conceptual Model for Publishing Multimedia Content on the Semantic Web. In: Proceedings of the 4th International Conference on Semantic and Digital Media Technology (SAMT '09), 02.12-04.12., Graz, Austria, 2009

Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 63-67, 2012.

Making Ontology Documentation with LODE

Silvio Peroni[1], David Shotton[2], and Fabio Vitali[1]

[1] Department of Computer Science, University of Bologna (Italy)
[2] Department of Zoology, University of Oxford (UK)
essepuntato@cs.unibo.it, david.shotton@zoo.ox.ac.uk, fabio@cs.unibo.it

Abstract. In this demo paper we provide a brief overview of the main features of **LODE**, the *Live OWL Documentation Environment*. LODE is an online service that generates a human-readable description of any OWL ontology, taking into account both the ontological axioms and the annotations, and ordering these with the appearance and functionality of a W3C Recommendations web page, namely as an HTML page with embedded links for ease of browsing and navigation.

Keywords: OWL ontologies, Web tool, ontology documentation

1 Introduction

Understanding the aim and extent of an ontology is one of the most important task of the Semantic Web. To this end, users usually start by consulting its human-readable documentation. A large number of ontologies, especially those used in the Linked Data world, have good comprehensive Web pages describing their theoretical backgrounds and the features of their developed entities. However, problems arise when we look at under-developed models, since natural language documentation is usually only published once an ontology has become stable. This approach is justifiable: writing proper documentation costs effort, and re-writing it every time the developing ontology is modified is not practical. For this reason, tools for the automatic online generation of HTML documentation from ontologies – e.g. Parrot [5], the Ontology Browser[3], Neologism [1] – are critically important. Not only do they ease the task of creating effective ontology documentation, but they also enable this to be done earlier in the creation life-cycle.

In this paper we introduce the main features of *LODE*, the *Live OWL Documentation Environment*[4] [3], a tool we developed to address the issue of the automatic production of ontology documentation. LODE is an online service that takes any OWL ontology or, more generally, an RDF vocabulary, and generates a human-readable HTML page designed for browsing and navigation by means of embedded links.

[3] OWLDoc-based Ontology Browser: http://owl.cs.manchester.ac.uk/browser/.

[4] LODE, the Live OWL Documentation Environment: http://www.essepuntato.it/lode.

The rest of the paper is structured as follows. In Section 2 we present LODE, highlighting its main characteristics and features, and in Section 3, we conclude the paper by sketching out the future developments of our work.

2 The Live Documentation Environment

LODE, the*Live OWL Documentation Environment* [3], is an XSLT-powered on-line service that automatically generates a human-readable description of an OWL ontology (or, more generally, an RDF vocabulary), taking into account both ontological axioms and annotations, and presents these with the appearance and functionality of a W3C Recommendations document.

LODE automatically extracts classes, object properties, data properties, nam-ed individuals, annotation properties, meta-modelling (punning), general axioms, SWRL rules and namespace declarations from any OWL or OWL 2 ontology, and renders them as ordered lists, together with their textual and graphic definitions, in a single human-readable HTML page designed for easy browsing and navigation by means of embedded links. LODE can be invoked with a number of optional parameters so as to limit or extend the final documentation produced. For instance, it is possible to take into account all the entities in the ontology closure and/or the inferred axioms. The following pseudo-URL describes how to call LODE:

http://www.essepuntato.it/lode/**optional-parameters**/**ontology-url**

Fig. 1 illustrates the alternative ways to build the URL to call LODE and the related modules used[5]. The main features of LODE are introduced in the following paragraphs.

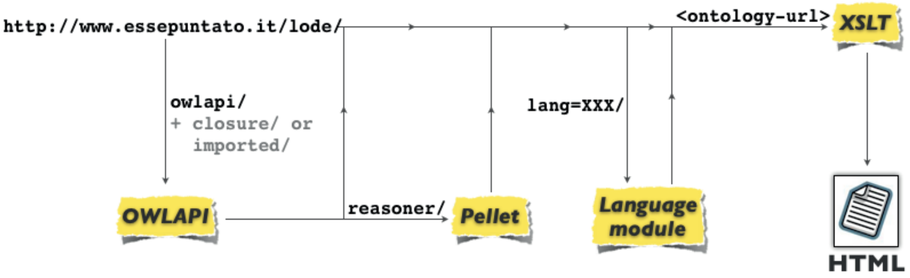


Fig. 1. All the possible ways, according to specific needs, for making a request to LODE.

Single page of documentation. The entire documentation is displayed in a single HTML page that starts with the title of the document (*dc:title* or

[5] Note that routine use of the optional parameter "owlapi" is strongly recommended.

rdfs:label) and then lists the metadata of the ontology, such as its IRI, authors (*dc:creator* annotations), date of publication (*dc:date*), and imported ontologies (*owl:imports* axioms). A brief abstract (*rdfs:comment*) and an automatically-built Table of Content follow. When *dc:description* assertions are specified for the ontology, the section "Introduction" is created as the first item after the Table of Content. Then, LODE renders all the classes, object, data and annotation properties, named individuals, general axioms and SWRL rules, and concludes the document with a list of the namespaces defined in the ontology.

Displays information about imported ontology. All the annotations and axioms defined in the imported ontology are fully rendered in the documentation when one of the parameters "imported" or "closure" is specified.

Displays of axioms inferred by a reasoner. All the axioms inferable by Pellet [4] are fully rendered in the documentation when the parameter "reasoner" is specified. However, this option has not yet been fully tested.

Axioms descriptions fully displayed. All the axioms and restrictions are rendered within each entity definition through the Manchester Syntax [2], as shown in Fig. 2a[6].

Cool URIs to call the service. Any LODE URL built as shown in Fig. 1 is compliant with Tim Berners-Lee's "Cool URI" definition[7]. This enables LODE to be used in conjunction with content-negotiation mechanisms, so as to produce HTML representations of ontologies and their entities automatically when accessing ontologies from a web browser.

Images can be included in the documentation. The annotation property *dc:description* can be used to refer to graphic files in addition to specifying simple text. In this case, LODE is able to render these graphic files as part of the documentation itself, as shown in Fig. 2a.

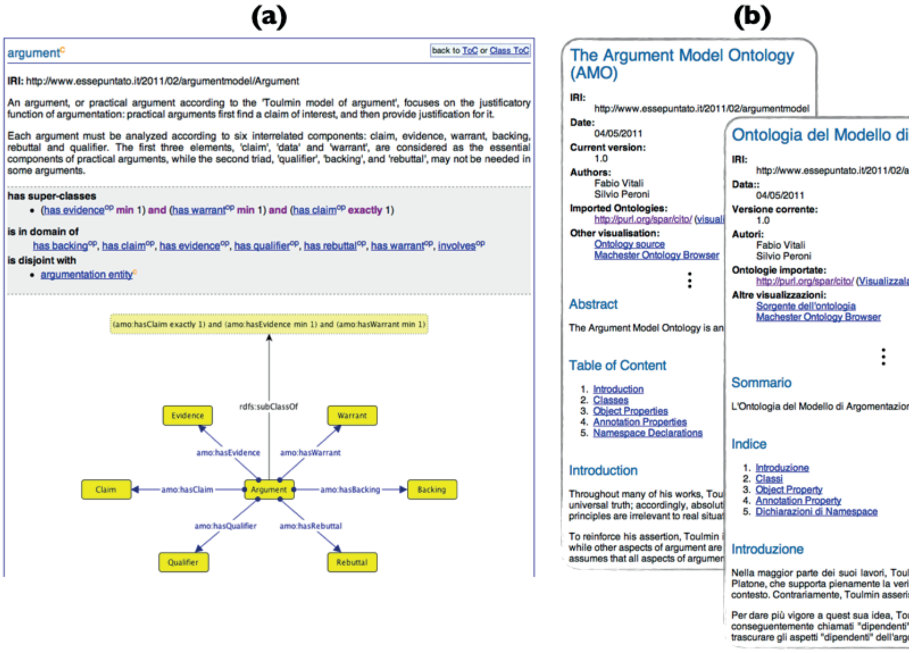
Permits choice of language in which to display the documentation. The text of the documentation can be rendered in a specific language when the parameter "lang" is specified. The selected language will be used as preferred language instead of English (the default) to render the documentation, as shown in Fig. 2b. Of course, this presupposes that appropriate language annotations are present in the ontology.

3 Conclusions

In this demo paper we briefly overview the main features of **LODE**, the *Live OWL Documentation Environment*, our tool that automatically produces HTML documentation of OWL ontologies. We recently performed two different user evaluations so as to assess the usability of LODE when users use it to deal with ontology understanding and navigation tasks. Preliminary results,

[6] The screenshots shown in Figure 2 of this paper refers to the documentation produced by LODE for the Argument Model Ontology (`http://www.essepuntato.it/2011/02/argumentmodel`).

[7] Cool URIs don't change: `http://www.w3.org/Provider/Style/URI`.



argument^{c} back to ToC or Class ToC

IRI: http://www.essepuntato.it/2011/02/argumentmodel/Argument

An argument, or practical argument according to the 'Toulmin model of argument', focuses on the justificatory function of argumentation: practical arguments first find a claim of interest, and then provide justification for it.

Each argument must be analyzed according to six interrelated components: claim, evidence, warrant, backing, rebuttal and qualifier. The first three elements, 'claim', 'data' and 'warrant', are considered as the essential components of practical arguments, while the second triad, 'qualifier', 'backing', and 'rebuttal', may not be needed in some arguments.

has super-classes
- (has evidence^{op} min 1) and (has warrant^{op} min 1) and (has claim^{op} exactly 1)

is in domain of
has backing^{op}, has claim^{op}, has evidence^{op}, has qualifier^{op}, has rebuttal^{op}, has warrant^{op}, involves^{op}

is disjoint with
- argumentation entity^{c}

The Argument Model Ontology (AMO)

IRI:
http://www.essepuntato.it/2011/02/argumentmodel
Date:
04/05/2011
Current version:
1.0
Authors:
Fabio Vitali
Silvio Peroni
Imported Ontologies:
http://purl.org/spar/cito/ (visuali
Other visualisation:
Ontology source
Machester Ontology Browser

⋮

Abstract

The Argument Model Ontology is an

Table of Content

1. Introduction
2. Classes
3. Object Properties
4. Annotation Properties
5. Namespace Declarations

Introduction

Throughout many of his works, Tou
universal truth; accordingly, absolut
principles are irrelevant to real situat

To reinforce his assertion, Toulmin i
while other aspects of argument are
assumes that all aspects of argumen

Ontologia del Modello di

IRI:
http://www.essepuntato.it/2011/02/a
Data::
04/05/2011
Versione corrente:
1.0
Autori:
Fabio Vitali
Silvio Peroni
Ontologie importate:
http://purl.org/spar/cito/ (Visualizzala
Altre visualizzazioni:
Sorgente dell'ontologia
Machester Ontology Browser

⋮

Sommario

L'Ontologia del Modello di Argomentazior

Indice

1. Introduzione
2. Classi
3. Object Property
4. Annotation Property
5. Dichiarazioni di Namespace

Introduzione

Nella maggior parte dei suoi lavori, Toul
Platone, che supporta pienamente la veri
contesto. Contrariamente, Toulmin asseri

Per dare più vigore a quest sua idea, To
conseguentemente chiamati "dipendenti"
trascurare gli aspetti "dipendenti" dell'arg

Fig. 2. Screenshots of the documentation produced by LODE. (a) The way the class "argument" is rendered by LODE: its IRI, a brief abstract (*rdfs:comment* assertion), its axioms and a graphic representation (*dc:description* assertion pointing to a PNG file) of the class itself. (b) The Argument Model Ontology rendered in English (parameter "lang=en") and Italian (parameter "lang=it").

presented in [3], show that LODE is perceived as a usable tool that favourably compares with similar applications such as Parrot [5] and the OWLDoc-based Ontology Browser[3]. In the future, we plan to widen the evaluation of the outcomes of the aforementioned user testing sessions, and to extend LODE features to include suggestions highlighted by users, such as adding a tree-browsing capability and a facetted free-text search box, thus significantly enhancing its usability for single ontologies in the one area in which LODE is presently lacking, search functions.

References

1. Basca, C., Corlosquet, S., Cyganiak, R., Fernández, S., Schandl, T. (2008). Neologism: Easy Vocabulary Publishing. In Proceedings of the 4th Workshop on Scripting for the Semantic Web. http://ceur-ws.org/Vol-368/paper10.pdf (last visited July 30, 2012).
2. Horridge, M., Patel-Schneider, P. (2009). OWL 2 Web Ontology Language: Manchester Syntax. W3C Working Group Note, 27 October 2009. World Wide Web

Consortium. http://www.w3.org/TR/owl2-manchester-syntax/ (last visited July 30, 2012).

3. Peroni, S., Shotton, D., Vitali, F. (2012). The Live OWL Documentation Environment: a tool for the automatic generation of ontology documentation. To appear in Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012). http://palindrom.es/phd/wp-content/uploads/2012/05/lode-ekaw2012.pdf (last visited July 30, 2012).
4. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. In Journal of Web Semantics, 5 (2): 51-53. DOI: 10.1016/j.websem.2007.03.004.
5. Tejo-Alonso, C., Berrueta, D., Polo, L., Fernandez, S. (2011). Metadata for Web Ontologies and Rules: Current Practices and Perspectives. In Proceedings of the 5th International Conference on Metadata and Semantic Research (MTSR 2011). DOI: 10.1007/978-3-642-24731-6_6.