Libraries in the Semantic Web Era.

Marek Kopel[1] and Aleksander Zgrzywa[2]

[1]Wrocław University of Technology, Poland
[2]Wrocław University of Technology, Poland

Abstract: This paper focuses on problems with exposing the digital libraries supporting OAI-PMH to the Semantic Web services. The most wanted thing for OAI-PMH metadata is being querable via SPARQL. This means the OAI-PMH formats must be converted and served as RDF. Another important thing is supporting the fourth rule of linked data, which is about interlinking relevant resources. The interlinking can be done by mapping resources across different SPARQL endpoints. But problem arises when the fields to be mapped are not a perfect match. The solution may be some similarity metric with an established threshold.

Keywords: Linked Data, OAI-PMH, Semantic Web, SPARQL

1. Introduction

The upcoming era of Semantic Web (SW) domination gives the promise for autonomous machine reasoning. But before the promise becomes reality there are a lot of efforts to make public Web data semantically enriched. The semantic enrichment is believed to allow the Web be readable not only by humans, but also by the machines. As the author of the WWW idea envisioned in Berners-Lee, Hendler and Lassila (2001) the existence of SW would allow the existence of intelligent agents that can process information from the Web autonomously. Autonomously means that unlike today Web browsers these agents would only need a high level of abstraction commands. For example, getting a command “Find a drugstore that is closest to my way home” the agent would estimate optimal way home based on current GPS read, match it to search results for local drugstores and check whether they are open at the time of coming home. It may optionally create an event in user personal calendar with a reminder set after the scheduled work time. That automatic reasoning needs inferencing based on Web data. To make the Web machine readable the data should be described by easy accessible semantic metadata.

As far as libraries are concerned, metadata have always been a part of cataloging. Today in digital libraries it is common to represent metadata using standard ontologies, such as Dublin Core (DC). There is also the idea of exposing the metadata for harvesting in order to create one central database, which would allow to search all the available metadata at once. The protocol for that purpose is Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which is described, along with the whole initiative, in Lagoze and

Van de Sompel (2001). The OAI-PMH is RESTful, which means it fulfils Representational State Transfer (REST) principles, as described in Fielding (2000). But even with that architecture, the idea of duplicating the harvested metadata and a central storage is in a way with the SW idea. For the purpose of interoperability of SW services Berners-Lee created and described in Berners-Lee (2007) four rules of linking data on the Web. The main idea of linked data is to make each piece of information identified by dereferenceable URI and available via HTTP. When fulfilling the rules the data can be easily accessed an processed by a query language. The query language used widely for this purpose is SPARQL.

2. Exposing the Data to SW Services

Almost every information system today is using relational database as a storage. Relational database stores data in tables and can be queried using SQL. This is not what SW services can use. SW assumes the data are RDF triples which can be queried via SPARQL. So the simple idea to help SW services gaining traction is to make a transformation interface which would serve the relational stored data as linked data. A tool for publishing the content of a relational database to the Semantic Web is D2R Server described in Bizer and Cyganiak (2006). To describe the mapping between the relational database schemata and the OWL/RDFS ontologies D2R Server uses a declarative language D2RQ. The mapping description is used by the D2RQ Engine, which makes the transformation on the fly and wraps the local relational database into a virtual RDF graph. The graph is then exposed by the server to the RDF browsers, which are enabled to navigate the content of the database. The server also creates a SPARQL endpoint via which SW services can make structured queries. The results can then be retrieved in XML or JSON serialization.

3. Linking the Data

Beside giving object a dereferenceable URI, the other important thing about exposing information as linked data is linking to other resources. The fourth rule of linked data is: “Include links to other URIs, so that they (people) can discover more things”. To support that rule D2R Server includes an rdf:seeAlso triple with every resource description that point to an RDF document containing links to other resources. These triples serve as “breadcrumbs” to RDF crawlers and browsers.

However, sometimes the information about other resources that may be useful in context of the retrieved resource is not stored in the relational database or is not available. This is the case with OAI-PMH. One can retrieve metadata records, but there is hardly any information about other useful records. This is because OAI-PMH is about exposing the data to the Web, but it is not the linked data. In Haslhofer and Schandl (2008) authors deal with converting the OAI-PMH endpoint into the SPARQL endpoint. They present the OAI2LOD

Server which is build on the D2R server, but it enables to expose the linked data without accessing directly the relational database. This means the SPARQL endpoint can be set up by anyone, because it only needs an OAI-PMH endpoint as the data source. To satisfy the fours rule of linked data OAI2LOD Server may search and map resources to other linked data resources based on some similarity metric. For example: places or concepts may be linked by name to dbpedia – the linked data version of Wikipedia articles. If a matching entry is found then an owl:sameAs property may be added to the metadata record. In order to make this work the server administrator needs to specify:

1. target OAI2LOD Servers for linking,
2. pairs of source and target fields to be analyzed,
3. a similarity threshold for each pair.

4. The Matching Problem

Using the OAI2LOD Server to expose the DBC digital library metadata records as linked data an unexpected problem emerged . The requirement of linked data fourth rule was supposed to be fulfilled by associating library OAI-PMH records (*http://www.dbc.wroc.pl/dlibra/oai-pmh-repository.xml*) with D2R Server publishing the DBLP Bibliography Database (*http://dblp.l3s.de/d2r/*). The association was supposed to be made by simple matching author names. In this way, when looking up a DBC library record, agent is given a link to other works of the same author in the DBLP database.

It turned out that the exposed OAI-PMH records do not serve author names in any consistent way. Therefore direct matching on the author names could not give positive results. Table 1 shows some examples of author names in DBC library.

Daniłowicz, Czesław [dr hab. inż.], promotor
Dankowska-Żołnierowicz, Bogna
Darboux, Gaston (1842-1917). Wyd.
Daubrée, Auguste (1814-1896)
Dec, Ignacy (1944-)
Dec, Ignacy (1944-). Promotor
Dittrich
Dowojna - Sylwestrowicz, Mieczysław. Red.
Dudziński, Włodzimierz Jan [dr hab. Inż.], promotor
Dudziński, Włodzimierz Jan. Promotor
Dufour, Piotr. Wyd.
Dumas, Jean Baptiste André (1800-1884)

Tab. 1. Example of inconsistency in author names in DBC

On the other hand DBLP database also showed some inconsistency: sometimes giving full first names, sometimes initials and sometimes a combination of the two. In one case the name field was containing a nickname put in quotation marks. Other problems with author field inconsistency concern character encoding, inversing the order of first and last name and occasional notes introduced in parentheses or square brackets. Since the inconsistency is not deterministic cleaning the data is not a solution.

Facing the above problems matching the names with any similarity metric with threshold set to 1 would not give expected results. Thus two questions arise:

1. Which similarity metric to use?
2. To what value set the threshold?

In order to answer these questions an experiment have been carried out.

5. Experiment

For the purpose of an experiment only the author names staring with "D" were taken into account. DBC containing 1700 records stores 56 authors with names starting with "D" in 62 distinct fields. The names were retrieved using SPARQL endpoint with the following query:

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT DISTINCT ?uri ?name WHERE {{
 ?uri dc:creator ?name .
 FILTER(regex(?name, "^D")) }
UNION {
 ?uri dc:contributor ?name .
 FILTER(regex(?name, "^D")) }}
ORDER BY ?name
```

In DBC authors are described using DC ontology with two properties: dc:creator and dc:contributor. Authors from DBLP were retrieved in an analogical manner. The DBLP database contains more than 950 000 articles and 570 000 authors. The retrieved authors with name staring with "D" gave 25 876 items stored in a foaf:name and rdfs:label properties.

The author names retrieved from DBLP where tested against each author name retrieved from DBC. This gives over 1.6 million tests for each metric. The tests were based on the following similarity metrics:

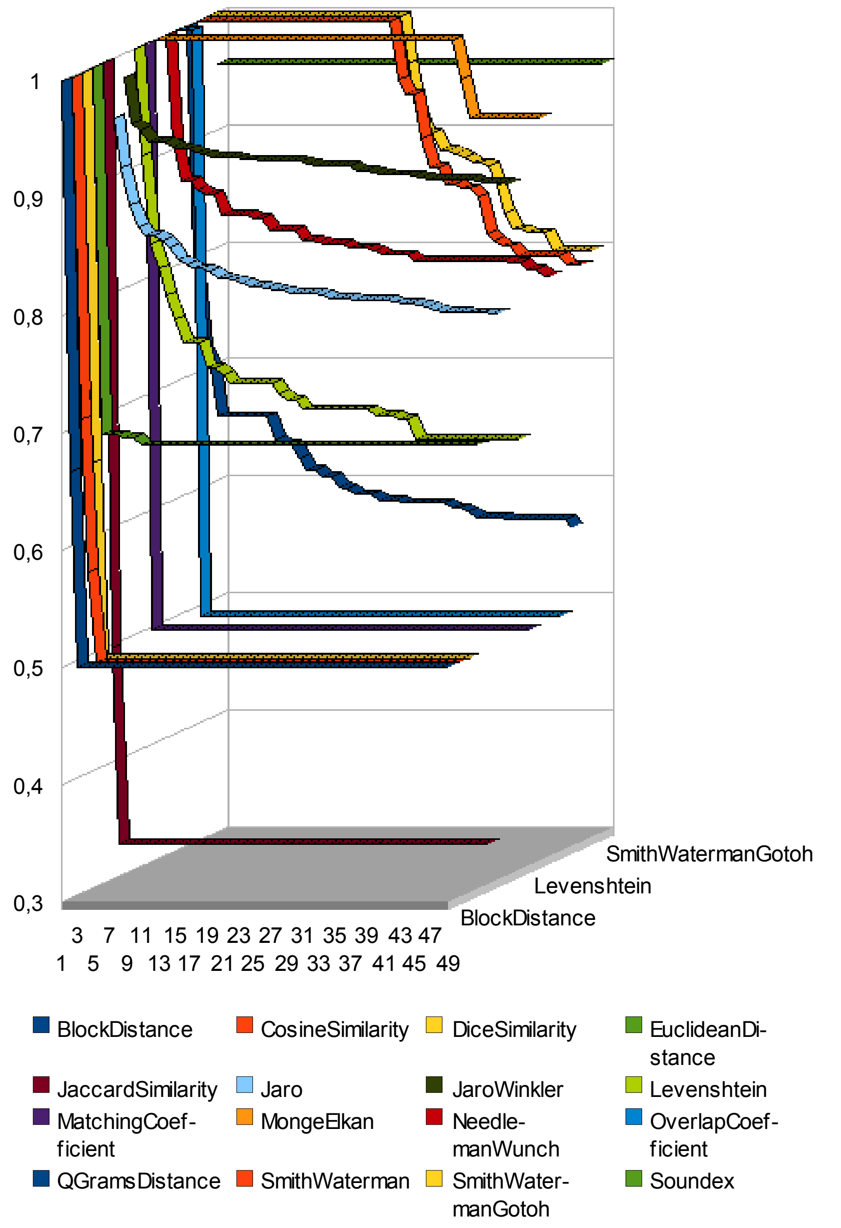
1. Block distance or L1 distance or City block distance
2. Cosine similarity
3. Dice's Coefficient
4. Euclidean distance or L2 distance
5. Jaccard Similarity or Jaccard Coefficient or Tanimoto coefficient – Jaccard (1912)
6. Jaro distance – Jaro (1989)
7. Jaro Winkler distance – Winkler (1999)

8. Levenshtein distance
9. Matching Coefficient
10. Monge Elkan distance – Monge and Elkan (1996)
11. Needleman-Wunch distance or Sellers Algorithm
12. Overlap Coefficient
13. Q-gram distance – Gravano et al. (2001)
14. Smith-Waterman distance – Smitha and Waterman (1981)
15. Gotoh Distance or Smith-Waterman-Gotoh distance – Gotoh (1982)
16. SoundEx distance

The open source java implementations of the metrics are available at *http://www.dcs.shef.ac.uk/~sam/simmetrics.html.* They are normalized, which mean when testing 2 strings they return a value from 0 to 1. Value 1 means a perfect match.

	Dziedzic, Andrzej (1957-)	Dziedzic, Andrzej [dr hab. inż.], promotor	Dziedzic, Andrzej. Promotor	Dziedzic, Janusz
BlockDistance	0,67	0,50	0,40	0,50
CosineSimilarity	0,71	0,58	0,41	0,50
DiceSimilarity	0,67	0,50	0,40	0,50
EuclideanDistance	0,68	0,68	0,52	0,50
JaccardSimilarity	0,50	0,33	0,25	0,33
Jaro	0,00	0,00	0,00	0,00
JaroWinkler	0,60	0,60	0,60	0,60
Levenshtein	0,65	0,40	0,63	0,65
MatchingCoefficient	0,50	0,33	0,33	0,50
MongeElkan	0,50	0,54	0,72	0,63
NeedlemanWunch	0,65	0,51	0,63	0,76
OverlapCoefficient	1,00	1,00	0,50	0,50
QGramsDistance	0,72	0,54	0,71	0,54
SmithWaterman	1,00	1,00	1,00	0,63
SmithWatermanGotoh	1,00	1,00	1,00	0,66
Soundex	0,96	0,96	0,96	0,96

Tab. 2. Similarity metrics' scores for testing name "Dziedzic, Andrzej" against 4 examples



Fig, 1. Monotonicity of top 50 scores for 16 similarity metrics applied to matching author names

6. Discussion

Direct comparison of the names from DBC and from DBLP established only one association. Most of the metrics scored this mapping as 1.0. But, as can be seen on figure 1, three metrics: 6, 7 and 16 didn't evaluated those two names with highest score. Even though they had tested two exactly the same strings. Testing "*Dziedzic, Andrzej*" against "*Dziedzic, Andrzej (1957-)*" and "*Dziedzic, Andrzej [dr hab. inż.], promotor*" got into top 5 scores using metrics: 1-5 and 13. Metrics 12, 14, 15 scored these tests as 1.0. As shown in table 2 metrics' scores in these cases are similar to those from a false positive case: "*Dziedzic, Janusz*".

For overall test results, false positives were given most frequently by metrics: 10 and 16. A true positive "*Danilowicz, Czeslaw*" against "*Daniłowicz, Czesław [dr hab. inż.], promotor*" was found in top 100 scores only for measures: 14 and 15.

The example from table 2 shows global characteristic: false positives with some metrics get better scores than true positives and sometimes there is no difference..

Figure 1 show that the range for top scores is quite discrete for metrics: 1-5, 9 and 12. For others, especially: 6, 7, 8, 11 and 13 the score range is more regular.

7. Conclusions

This paper covers problems with publishing digital libraries metadata to the Semantic Web in the form of linked data. Since the relational databases are widely used as a storage the idea is to transform the data and expose them as a SPARQL endpoint. When access to relational database is not possible, library metadata may be accessed vie OAI-PMH endpoint and then transformed to RDF graph. Described in this paper tools for that task are D2R Server and OAI2LOD server. Recently a new tool was developed to make exposing RDF graph even easier for Web developers. The tool is triplify (*http://triplify.org*). The main reason for facilitating users creating Web of data and not only Web of documents is to help SW services gain traction.

Another issue about exposing linked data is interlinking relevant resources. In order to find useful resources to be connected, property values matching must be carried out. Described in the paper experiment verifies matching based on similarity metrics. Experiment results reveal that using one metric does not guarantee positive effects. Using a combination of different similarity metrics for interlinking Semantic Web resources should be further investigated.

On the other hand the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) is the new standard for extending the OAI-PMH with the functionality of creating and sharing aggregations. When widely applied, this may be a better source of connections to relevant resources. Again, reusing the existing connections when exposing metadata records as linked data should be more efficient than staring from scratch with similarity metrics.

References

Berners-Lee, T., Hendler, J. and Lassila, O., (2001). *The semantic Web*. Scientific American, 284(5), 28-37.

Berners-Lee, T., (2007). *Linked Data - Design Issues*. Available at: http://www.w3.org/DesignIssues/LinkedData.html

Fielding, R. T., (2000), *Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine

Lagoze, C. and Van de Sompel, H., (2001). *The Open Archives Initiative: Building a low-barrier interoperability framework*. Proceedings of the *1st ACM/IEEE-CS joint conference on Digital libraries*. ACM Press New York, 54-62.

Bizer, C. & Cyganiak, R., (2006). *D2R Server - Publishing Relational Databases on the Semantic Web.* Proceedings of the *5th International Semantic Web Conference*.

Gotoh, O., 1982. *An improved algorithm for matching biological sequences.* Journal of Molecular Biology, 162(3), 705.

Gravano, L. et al., (2001). *Using q-grams in a DBMS for approximate string processing.* IEEE Data Engineering Bulletin, 24(4), 28-34.

Haslhofer, B. & Schandl, B., (2008). *The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data.* Available at: http://events.linkeddata.org/ldow2008/papers/03-haslhofer-schandl-oai2lod-server.pdf.

Jaccard, P., (1912). *The distribution of the flora in the alpine zone.* New Phytologist, 37-50.

Jaro, M.A., (1989). *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.* Journal of the American Statistical Association, 414-420.

Monge, A.E. and Elkan, C., (1996). *The field matching problem: Algorithms and applications.* Proceedings of the *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. 267-270.

Smith, T.F. and Waterman, M.S., (1981). *Identification of common molecular subsequences.* Journal of Molecular Biology, 147, 195-197.

Winkler, W.E., (1999). *The state of record linkage and current research problems.* Statistical Research Division, US Bureau of the Census, Washington, DC.