Enabling Russian National Knowledge with Linked Open Data (LOD Russia)

Daniel Hladky, Victor Klintsov, Grigory Drobyazko

National Research University – Higher School of Economics (NRU HSE),
Moscow/Russia
{dhladky, vklintsov, gdrobyazko}@hse.ru

Abstract. The LOD Russia research project funded by the Ministry of Education aims to create a first Linked Open Data Set in Russia enabling scientists, researchers and commercial users to share, access, analyse and reuse knowledge related to scientific data. The position paper is highlighting challenges of the life-cycle management of LOD data, especially focuses on the process of entity linking and the creation of a unique identifier (UID) based on the concept of the Identification Knowledge Base (IKB).
The publication is prepared with the financial support of the Ministry of Education and Science of the Russian Federation within the framework of the State Contract № 07.524.11.4005 of October 20, 2011.

Keywords: L**inked Data, NLP, RDF, UID disambiguation**

1 Introduction

The World Wide Web has enabled the creation of a global information space compromising linked documents. Under the umbrella of the Russian National Knowledge several Russian ministries have enabled projects that pursuit the desire to access scientific data not currently available on the Web or bound up in hypertext documents. Linked Data provides a publishing paradigm in which not only documents, but also data, can be first class citizen of the Web, thereby enabling the extension of the Web with a global data space based on open standards promoted by the W3C[1]. Based on first Russian experiences of e-Arena[2] and DANTE[3] the LOD Russia project is designed to build use cases for scientific data related to nanotechnology and mathematics. The research project has a runtime of 600 man-days and the project will finish by June 2013. The goals are to semantize various scientific papers, patents, research projects and create Linked Data sets based on a domain knowledge driven vocabulary. The access to the data should not only provide a better search but the use case shall also support analytical functions in order to make better decisions. Various stakeholders like scientists, researchers and lawyers shall have the possibility to use the data

[1] http://www.w3.org/
[2] http://en.e-arena.ru/
[3] http://www.dante.net/

sets and have different reports and analysis according the individual needs. This position paper will focus mainly on the process of creating unique identifiers (UID) and the merging of the UID using the Identification Knowledge Base (IKB).

2 LOD project

The consortium consisting of NRU HSE, Avicomp Services[4] and AKSW University of Leipzig[5] have identified different tasks within the process of creating the linked data sets. The simplified view can be described as such:

- Crawl/Harvest sources and create plain text
- Large-scale information extraction from unstructured, semi-structured heterogeneous sources using rule based NLP, statistical methods, background knowledge and bootstrapping.
- Process objects using the Identification Knowledge Base (IKB) for named entity disambiguation and the creation of unique identifier (UID).
- Aggregating the data into a scalable RDF store applying methods for interlinking named entities with the LOD.

The size of the test corpus consists of 15'000 documents related to mathematics and 100'000 documents related to nanotechnology resulting in approximately 2.5 million triples for the RDF store. Based on the current ontology we expect to create links to another 20 data sets on the Web[6]. The focus of the position paper is on the automatic creation of an UID and the merging of same objects (named entities) as shown in Figure 1.

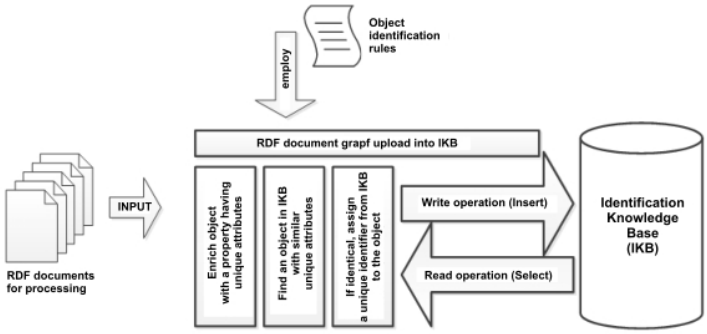

Figure 1 – Simplified process of UID creation using the IKB

[4] http://avicomp.ru/
[5] http://aksw.org/About
[6] Datasets (excerpt) incl.: The Open Library, British National Bibliography, CiteSeer, DBLP, RISKA Digest, ACM and DBpedia

2.1 Disambiguation and Fusion using IKB

The IKB main purpose is to create an UID for a named entity (NE) [3] and therefore support the process of data cleaning [1] and data scrubbing [6]. Within an IKB an object is automatically populated from the Natural Language Process (NLP) with diverse attributes (e.g. person: gender, first/middle/last name, date of birth, etc.) and relations to other objects [4]. From the list of those attributes a set of keys can be created called Data Driven Identifiers (DDI). Those DDI are used for the creation and merging of UID.

2.2 IKB-based approach for identification

Native attributes of an object as well as a context surrounding it are characteristics enabling the identification of any specific object. A context may be either values of other objects interlinked with the one being identified, or text words surrounding a given object. Therefore, an object can be treated as a point in the multidimensional space of its characteristics [2,5]. To identify an instance from the whole set of object characteristics belonging to a certain class, subsets of identifying characteristics are selected that constitute a kind of "keys" for each instance. Objects obtained from different documents are considered "identical" subject to the "nearness" of one of the sets of identifying characteristics. The IKB process involves the indexing of identifying characteristics of an object. The resulting index is used to search exactly matching or "similar" objects. This information may be adapted and corrected by human interaction. The IKB object is used for automatically identifying objects similarity and hence creates an UID.

3 Use Case

The aim of the use case is to provide to various stakeholders an access to a knowledge portal which is connected to the LOD Russia data sets. Each of the stakeholders has different expectations on how to search for data and on how to analyze the data. For the most part a user will be able to find experts and leaders in a specific domain of their interest, as well as to look for shadow groups of people and institutions working in the domain, based on thesauri and objects of interest. The knowledge portal provides different views to the LOD sets and allows simple filtering using thesauri, sources and filtering by the extracted named entities and semantic relations. Besides the search another use case is related to provide analyzes over the data sets. This process allows for better decision making, especially under the point of view of a patent lawyer or an investment analyst. For example create a report of existing patents related to a domain or on the other hand identify trends of research and link that information to expert groups. Other examples of analysis could be to identify trends in a specific domain (see figure 2 right) or to investigate top publications by numbers in a specific environment (see figure 2 left).

Figure 2 – Left: Top publications by domain; Right: Trend on growing markets

4 Discussion and Outlook

The approach sketched above for the UID creation and merging is a straight forward method comparing DDI Keys. In the remaining time of the project we envisage to enhance the existing approach using a weighting of DDI keys allowing the creation of relevancy of each DDI key.

5 Conclusion

At the end of the research project we expect to have the following impact:

- Better leverage of knowledge from unstructured sources applying new NLP methods such as rule based, statistical and background knowledge;
- Increase information sharing through cross-linking of knowledge using the fusion and data link to other LOD sets;
- Foster various stakeholders through use cases that allow better search and analysis of the data.

6 References

1. Do, H.H. et al., 2000. Data Cleaning : Problems and Current Approaches. *Informatica*, 23(4), pp.1-11.
2. Li, C., Jin, L., Mehrotra, S. Supporting record linkage in large data sets. *Proceeding of eighth International Conference on Database Systems for Advanced Applications (2003)*
3. Lim, E.-P. et al., 1993. Entity Identification in Database Integration R. Hutterer & W. W. Keil, eds. *Information Sciences*, 89(1), pp.294-301.
4. Maynard, D., Li, Y. & Peters, W., 2008. NLP Techniques for Term Extraction and Ontology Population. *Proceeding of the 2008 conference on Ontology Learning and Population Bridging the Gap between Text and Knowledge*, pp.107-127.
5. Tejada, D, Knoblock, C.A., Minton, S. Learning object identification rules for information integration. *Information Systems Journal (2001), pp.607-633*.
6. Widom, J., 1995. Research problems in data warehousing. Proceedings of the fourth international conference on Information and knowledge management CIKM 95, pp.25-30.