LOD Russia – Enabling Russian National Knowledge with Scientific Open Data

Daniel Hladky

National Research University – Higher School of Economics (NRU HSE), Moscow/Russia
{dhladky}@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 paper is highlighting challenges of the life-cycle management of LOD data and provides a first view in use cases that will be realised based on the LOD data set.

"The work was performed with financial support from the Ministry of Education and Science of the Russian Federation"

Keywords: LOD, Linked Data, Scientific data, NLP, RDF

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 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 sets and have

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

adfa, p. 1, 2011.

different reports and analysis according the individual needs. Following picture depicts the various stakeholders.

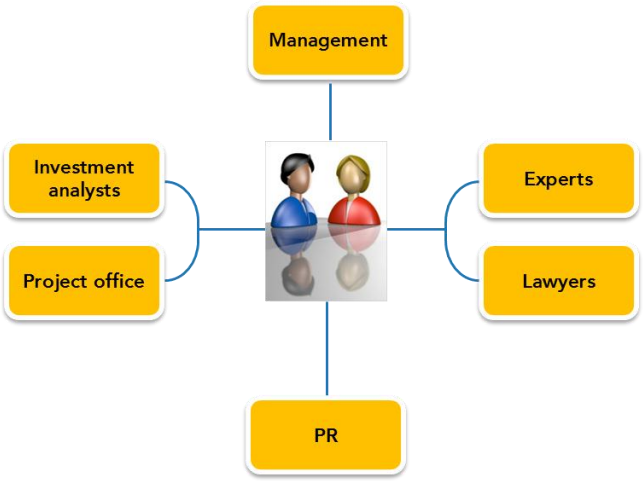

Figure 1 – Stake holders using LOD data

Within the research project we will address three main challenges that are described in the next section. Later in chapter three we will provide more details of the use case and the user interface related to the stakeholders needs.

2 Research challenges

The consortium consisting of NRU HSE, Avicomp Services[4] and AKSW University of Leipzig[5] have identified four major tasks. The first is dedicated to Data Management, second to Interlinking and Fusion, third is related to analyzing large amounts of unstructured data resources and the final task is to build the application use case which is described in chapter three.

2.1 Large-scale Data Management

First generations of RDF stores were either in memory only or used an external RDBMS, often MySQL, for persistent storage. Performance and scalability were understandably limited, especially when running complex queries. Within the project we explore the usage of new methods combining RDBMS and RDF stores in order to overcome disadvantages (Sidirourgos et al. 2008). The RDF store will have a quadruple indexing method that includes the subject, predicate, object and context.

[4] http://avicomp.ru/
[5] http://aksw.org/About

2.2 Interlinking and Fusion

The central idea of LOD is to link the different data sources[6]. Within the project we deal with the issue of generating automatic or semi-automatic RDF links based on unique identifiers (UID). Within the project we will explore other means of generating those links using similarity algorithms. Previous work of database research is used to build those methods. A specialty of the LOD Russia project is the usage of an Identification Knowledge Base (IKB). The following picture depicts the process of the IKB.

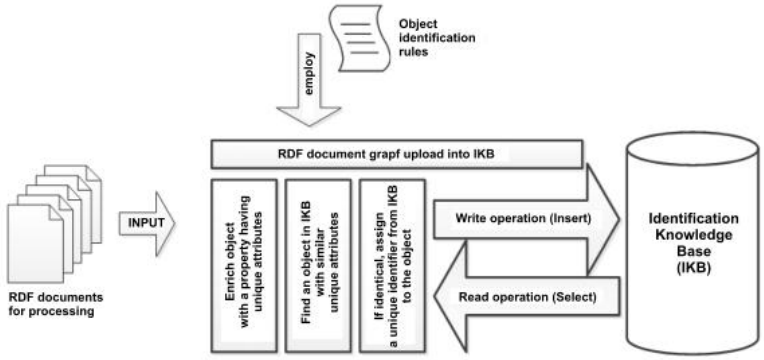

Figure 2 – IKB process diagram

The IKB main purpose is to create an UID for a named entity (NE) (Lim et al. 1993) and therefore support the process of data cleaning (Do et al. 2000) and data scrubbing (Widom 1995). 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. 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.

2.3 Analyzing large amounts of unstructured data resources

The third challenge of the research work is related to the process of natural language text using rule based and statistical methods (Maynard et al. 2008) within the natural language processing (NLP). The process involves also background knowledge such as the usage of LOD or existing embedded metadata like RDFa[7] or Microdata[8]. The aim is to improve quality of named entity recognition, semantic relationships, co-reference resolution and the recognition of relations between events.

[6] Linked Data – Design issues. http://www.w3.org/DesignIssues/LinkedData.html
[7] http://www.w3.org/2010/02/rdfa/
[8] http://www.w3.org/TR/microdata/

2.4 Summary

The overall objective is to create a sustainable and dynamically growing linked data pool and provide intelligent web services allowing integration into external platforms and web applications.

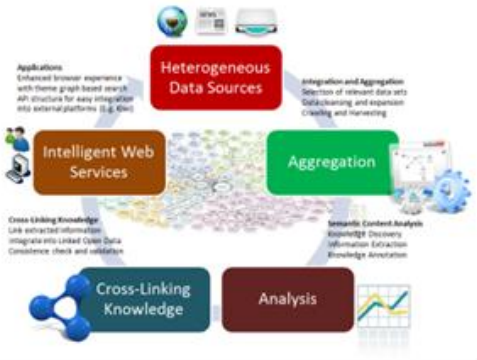

Figure 3 – LOD Russia integrated process

3 Use Case

The aim of the use case is to provide to various stakeholders (see Figure 1) an access to a knowledge portal which is connected to the LOD Russia data sets. Each of the stake holders 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.

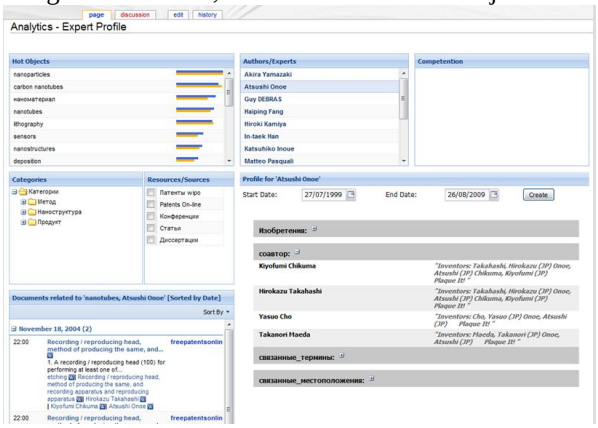

Figure 4 – Expert Search

The knowledge portal provide 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 5 right) or to investigate top publications by numbers in a specific environment (see figure 5 left).

Figure 5 – Left: Top publications by domain; Right: Trend on growing markets

4 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.

After the conclusion of the project the aim is to continue to add other scientific data using the experience gained from the project.

References

Do, H.H. et al., 2000. Data Cleaning : Problems and Current Approaches. *Informatica*, 23(4), pp.1-11.

Lim, E.-P. et al., 1993. Entity Identification in Database Integration R. Hutterer & W. W. Keil, eds. *Information Sciences*, 89(1), pp.294-301.

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.

Sidirourgos, L. et al., 2008. Column-store support for RDF data management: not all swans are white. *Proceedings of the VLDB Endowment*, 1(2), pp.1553-1563.

Widom, J., 1995. Research problems in data warehousing. *Proceedings of the fourth international conference on Information and knowledge management CIKM 95*, pp.25-30.