kOre: Using Linked Data for OpenScience Information Integration

Ivan Ermilov Konrad Höffner Jens Lehmann

University of Leipzig, Institute of Computer Science, AKSW Group
Augustusplatz 10, D-04109 Leipzig, Germany
{iermilov,hoeffner,lehmann}@informatik.uni-leipzig.de

Dmitry Mouromtsev
ITMO University
49 Kronverksky Ave., St.Petersburg, 197101, Russia
d.muromtsev@gmail.com

ABSTRACT

While the amount of data on the Web grows at 57 % per year, the Web of Science maintains a considerable amount of inertia, as yearly growth varies between 1.6 % and 14 %. On the other hand, the Web of Science consists of high quality information created and reviewed by the international community of researchers. While it is a complicated process to switch from traditional publishing methods to methods, which enable data publishing in machine-readable formats, the situation can be improved by at least exposing metadata about the scientific publications in machine-readable format. In this paper we aim at metadata, hidden inside universities' internal databases, reports and other hard to discover sources. We extend the VIVO ontology and create the VIVO+ ontology. We define and describe a framework for automatic conversion of university data to RDF. We showcase the VIVO+ ontology and the framework using the example of the ITMO university.

1. INTRODUCTION

While the amount of data on the Web grows at 57 % per year [4], the Web of Science maintains a considerable amount of inertia, where growth varies between 1.6 % and 14 % [7] depending on the type of publication and research area. The share of the Web of Science inside the whole Web is small. For example, DBLP lists only 2 892 316 publications up to the date of writing [1]. The Library of Congress with over than 26 million books only consumes up to 10 terabytes, while the size of the Web is measured in exabytes (i.e. millions of terabytes).

On the other hand, the Web of Science consists of high quality information created and reviewed by the international community of researchers. Moreover, by publishing research contributions using traditional methods, which are targeted at print and screen media (i.e. PDF documents), the data has become inaccessible for automatic processing. In the new era of Big Data and the Web of Data, the scientific community started developing new publication methods, which reflect the 3V Concept (i.e. volume, velocity, variety), such as nanopublications [5]. While it is a complicated process to switch from traditional publishing methods to methods that enable data publishing in machine-readable formats, the situation can be improved by at least exposing metadata about the scientific publications in machine-readable format. In this paper we aim at metadata, which is hidden inside universities' internal databases, reports and other hard to discover sources. Unlocking this hidden metadata will facilitate integration of the Web of Science with the new scientific data publication media (i.e. RDF datasets, nanopublications etc.), resulting in a Data Web of Science.

The Data Web of Science will enable new ways of data access based on the data semantics. In particular, complex search queries based on metadata will be available. Therefore, the availability and discovery for research papers will be increased. Given the availability of annotated PDF documents it will be possible to measure key performance indicators across universities as well as browse university data world wide. Also, the Data Web of Science will facilitate search for specialists and researchers given a particular research field, thus enabling new collaborations.

In this paper we present a framework to expose metadata about the research institutions according to the Linked Data principles [2]. In particular our contributions are as follows:

- We extend the VIVO ontology and create the VIVO+ ontology to tackle several deficiencies which we identified when using it.
- We define and describe a framework for automatic conversion of university data to RDF.
- We made the implementation of our approach freely available (on GitHub).
- We showcase how academical RDF data can be utilized and integrated with other data sources in an efficient way using the example of the ITMO university in St. Petersburg.

.

2. ONTOLOGY MODELLING

Using an established ontology for a particular domain reduces the development effort and leads to higher quality and better integration. In our usage scenario, we aim to publish organizational university data within a joint project of the University of Leipzig and ITMO. According to our research the best-fitting candidate ontology for such a task is VIVO [3]. However, it contains a lot of gaps left which we address with the VIVO+ ontology built upon VIVO and described in this section. VIVO+ adds missing key concepts, models and relationships. Moreover, it provides a more flexible and general modelling of relationships regarding academic degrees, which provides a better coverage, e.g. by allowing to represent Russian particularities.

Missing Concepts.

VIVO lacks many key concepts while others are modelled in insufficient detail. For example, it contains a concept for a PhD student but this concept does not reference any other concept of this domain except its super class. While a human user can understand the concept using its label, the more fine grained modelling of VIVO+ with related concepts and their connections allows easier extension, higher expressiveness, better querying and automatic processing.

The following example shows VIVO+ modelling of the student-degree-qualification domain.

```
:Alice a foaf:Person;
:academicQualification :AliceQualification;
:aspiresDegree :PhDDegree.

:AliceQualification
:academicDegree :MasterDegree;
:academicSubject :Chemistry.
```

Use of Established Vocabularies.

Besides the core ontologies RDF, RDFS, OWL, DCMI Metadata Terms, XSD and FOAF, VIVO+ uses the LinkedGeoData and GeoNames ontologies to specify locations.

In order to categorize the research topic of a laboratory, we use the 6 digit subject matter identifiers of the Russian State Index of Scientific and Technical Information (SISTI)[1].

Design Principles.

VIVO+ was designed with the following design principles in mind, following [6]. **Clarity** is provided by labels and comments in both English and Russian for all described classes and properties. URIs have speaking names. Class and property definitions are formally stated using consistent OWL restrictions as well as domain and range definitions, which ensures the **coherence** of datasets that conform to VIVO+. Future **extendibility** of the ontology and related datasets is provided through the usage of established vocabulary when fitting along with general definitions that are not restricted to the presented use cases. Using VIVO+ on top of VIVO presents **minimal ontological commitment** as it merely extends the latter ontology. **Encoding bias** was minimized through designing the ontology according to the presented use cases.

[1]Translated, in the Russian original: Государственный рубрикатор научно-технической информации (ГРНТИ), see `http://grnti.ru/`

Peculiarities of the Russian Education System.

We aim to provide an ontology that is sufficiently generic to be usable internationally. Different countries have different educational systems, however and VIVO is modelled according to the system of the USA. Adapting VIVO to ITMO University poses three challenges: (1) identifying the peculiarities of the Russian system (2) modelling the resulting additional concepts and (3) appropriately linking them to existing concepts. One difference between Russia and most of the world is that it has two different doctoral degrees: the Candidate of Sciences (кандидат наук, kandidat nauk) and the Doctor of Sciences (доктор наук, doktor nauk). The Candidate of Sciences is equivalent to the PhD while the Doctor of Sciences can be earned after a period of further study following the award of the Candidate of Sciences degree and requires five to fifteen years beyond the award of the Candidate of Sciences. VIVO+ contains classes for both degrees and relates them to existing concepts.

3. KORE: AUTOMATIZED MAPPING FRAMEWORK

In this section, we describe the transformation of the university data from RDBMS to RDF modeled according to the VIVO+ ontology. To perform such a transformation, we developed **kOre**[2] framework on top of the Sparqlify SPARQL-SQL rewriter[3], depicted in Figure 1. The framework is evaluated using ITMO university infrastructure (i.e. Oracle RDBMS) in section 4.

The main components of the kOre framework are:

- The **Mappings Repository** contains mappings in SML (Sparqlify Mapping Language). Mappings are used by the **Sparqlify SPAQRL-SQL rewriter** to convert the data from RDBMS to RDF. Mappings have to be maintained and updated by the framework maintainer in order to reflect university RDBMS schema changes and tackle new data inside RDBMS.
- The **Sparqlify SPARQL-SQL rewriter** establishes the connection to RDBMS and converts the data to RDF using mappings from the **Mappings Repository**.
- The **Triple store** stores RDF data and executes queries over it.
- The **SPARQL Endpoint** publishes RDF data on the Web thus enabling various web applications which are capable of using the W3C semantic web standards.

We define interactions between components with four basic operations:

1. *Poll.* As a university RDBMS is a live system updated with new information several times a day, **Sparqlify SPARQL-SQL rewriter** *polls* the **Mappings Repository** periodically to check for changes and perform transformations.
2. *Push.* After successful transformation of the data from the RDBMS the data has to be *pushed* to the **Triple store**. The *Push* operation adds/deletes/updates RDF data inside the **Triple store**.

[2]kOre: Using Linked Data for OpenScience Information Integration.

[3]`http://aksw.org/Projects/Sparqlify.html`

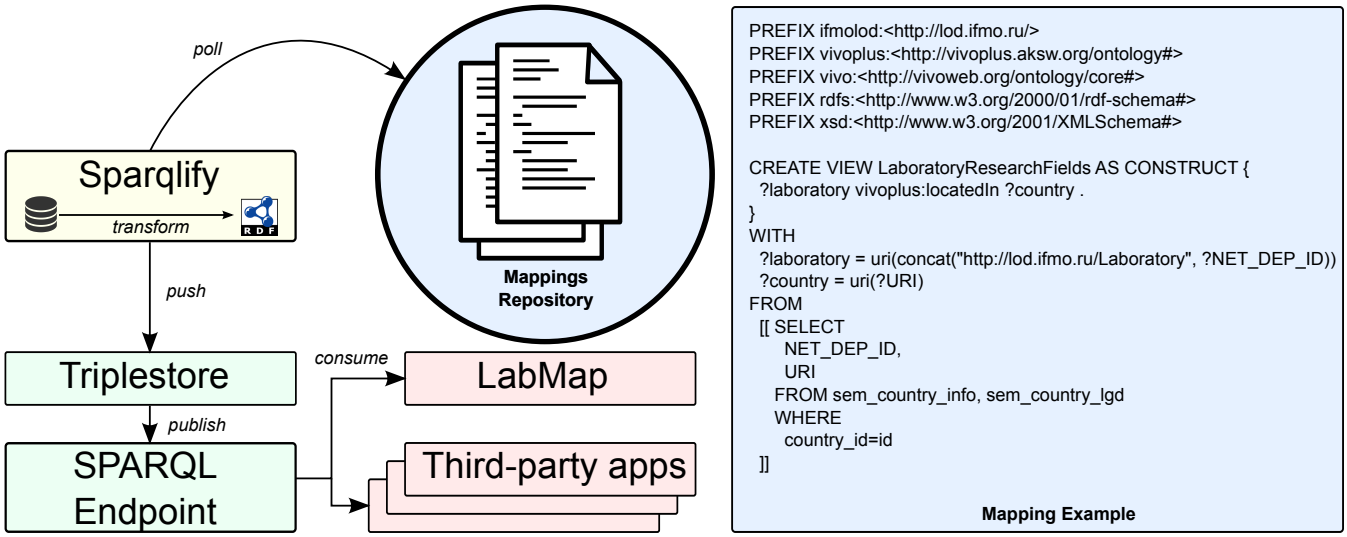

Figure 1: kOre: architecture overview.

3. *Publish.* The **Triple store** *publishes* the data by providing a **SPARQL Endpoint** thereby making it accessible on the Web.

4. *Consume.* Web applications can consume RDF via the **SPARQL Endpoint**, which provides a set of interfaces (e.g. RDF/JSON interface).

The presented framework provides a persistent layer on top of RDBMS infrastructure. Schema changes of underlying RDBMS are reflected with SML mappings, therefore publicly available data provided by the framework is consistent over time. Thus data consumers can rely on data schema provided by the kOre framework and save development time required for tackling schema changes inside applications.

4. MAPPING FRAMEWORK DEPLOYMENT AT ITMO

We implement and deploy the kOre framework using infrastructure of ITMO (Saint Petersburg State University of Information Technologies, Mechanics and Optics) University.[4] ITMO university uses an Oracle RDBMS to store the data about laboratories, people, publications among others. The database is only accessible from the university network, therefore a VPN connection is necessary.

For deployment we set up an Ubuntu 14.04 server with 6 GB of RAM and 35 GB of disk space. The **Mappings Repository** on the deployed framework corresponds to a system folder. **Sparqlify** is installed for all users and available through command-line interface. As a **Triple store** we use Virtuoso Open Source, which provides **SPARQL Endpoint** with RDF/JSON interface out of the box. *Poll* and *push* operations are configured using shell scripts and scheduled with crontab. *Publish* operation is performed by Virtuoso Open Source internally. *Consume* operation is possible through the **SPARQL Endpoint** using POST requests.[5]

[4]The implementation is freely available on GitHub: `https://github.com/AKSW/itmolod`

[5]The SPARQL endpoint for ITMO LOD project is located at `http://lod.ifmo.ru/sparql`

URI	`http://vivoplus.aksw.org/ontology#`
Version Date	2014-11-01
Version Number	1.0
License	PDDL 1.0[7]
Triples	15 572
SPARQL Endpoint	`http://lod.ifmo.ru/sparql`
N-Triples Dump	`http://lod.ifmo.ru/data.itmo.nt`

Table 1: Technical Details of the ITMO LOD dataset

At the moment of writing, the ITMO LOD dataset contains information about 43 laboratories, 188 research areas and 1001 persons. The dump of the dataset is available online[8]. Table 1 provides statistics and links to services such as the SPARQL endpoint. The dataset is openly licensed under the PDDL 1.0. in accordance with the open definition[9].

5. APPLICATIONS

In this section we describe how the *consume* operation can be utilized by developers and end-users.

For end-users we published the data from ITMO LOD SPARQL endpoint under lod.ifmo.ru domain using the Pubby[10] Linked Data interface. Pubby makes dataset URIs dereferenceable and enables navigation between linked entities inside the dataset with the easy-to-use web interface. For example, in Figure 2 we show, that a user is able to navigate through persons and research areas for a particular laboratory.

Developers can *consume* the data through SPARQL endpoint. Here we showcase how the ITMO LOD dataset can be utilized in such a way by implementing LabMap application[11] (the code is published in the GitHub repository), which shows

[8]`http://lod.ifmo.ru/data/itmo.nt`

[9]`http://opendefinition.org/`

[10]`http://wifo5-03.informatik.uni-mannheim.de/pubby/`

[11]`http://lod.ifmo.ru/usecases/heatmap.html`

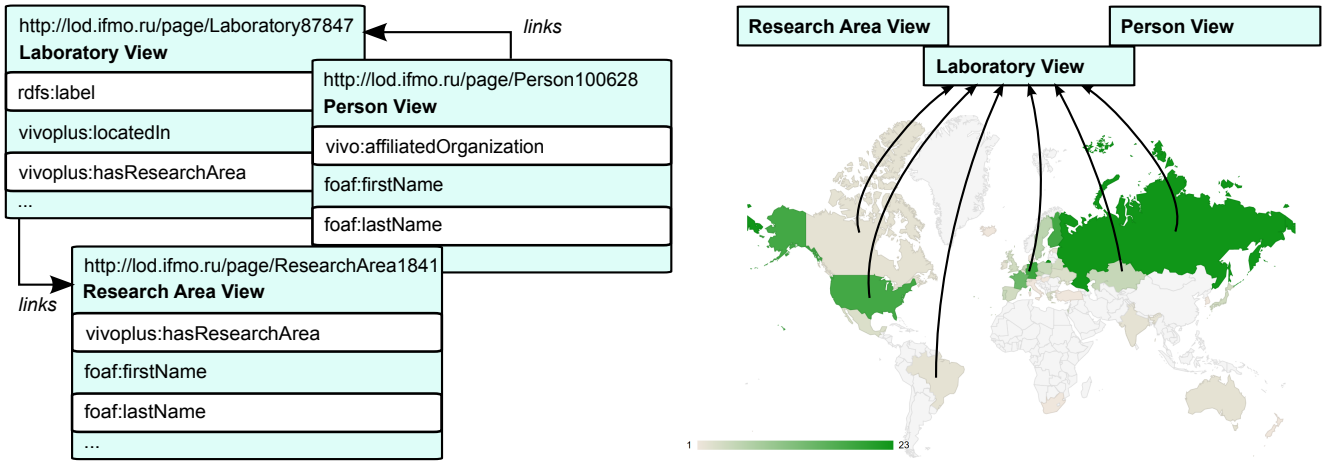

Figure 2: Linked Data Interface enables browsing between linked entities (on the left). LabMap application shows the laboratories of ITMO university filtered by collaborating countries (on the right).

laboratories of ITMO university by collaborating countries. User is able to see the list of laboratories per country as well as persons working in particular laboratory (see Figure 2).

6. CONCLUSIONS AND FUTURE WORK

We implemented the the kOre framework, which facilitates data publishing for universities. To support our framework we extended the VIVO ontology resulting in VIVO+. To showcase the applicability of the kOre framework we deployed Linked Data interface for end-users implemented simple web application as an example of data consumption for developers.

Unlocking the hidden metadata for universities is a step forward in the integration effort between the original Web of Science and the Data Web of Science. In the future we plan to deploy the kOre framework for University of Leipzig. We plan to support ongoing effort of ITMO university to expose more data publicly. Also, we plan to involve scientific personnel and students into the development of interesting applications using available data.

7. ACKNOWLEDGMENTS

We want to say thank you to our colleagues in AKSW and ITMO university, whom supported ITMO LOD project:

- Claus Stadler for adapting Sparqlify to Oracle SQL on our request.
- Maxim Kolchin for administrating the server environment.
- Denis Varenikov for helping getting access to ITMO RDBMS and exposing data for us.

This work was partially financially supported by Government of Russian Federation, Grant 074-U01.

References

[1] DBLP publications per year. `http://dblp.uni-trier.de/statistics/publicationsperyear`, accessed: 11-06-2015

[2] Linked Data Design Issues. `http://www.w3.org/DesignIssues/LinkedData.html`, accessed: 11-06-2015

[3] VIVO-ISF Ontology. `https://wiki.duraspace.org/display/VIVO/VIVO-ISF+Ontology`, accessed: 11-06-2015

[4] Gantz, J.F., Reinsel, D.: The expanding digital universe: A forecast of worldwide information growth through 2010. IDC (2007), `http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf`

[5] Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use 30(1), 51–56 (2010)

[6] Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. International journal of human-computer studies 43(5), 907–928 (1995)

[7] Larsen, P.O., Von Ins, M.: The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84(3), 575–603 (2010)