kOre: Using Linked Data for OpenScience Information Integration

Ivan Ermilov[1], Konrad Höffner[1], Jens Lehmann[1], and Dmitry Mouromtsev[2]

[1] University of Leipzig, Institute of Computer Science, AKSW Group
Augustusplatz 10, D-04109 Leipzig, Germany
{iermilov,hoeffner,lehmann}@informatik.uni-leipzig.de
[2] 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 nearly 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 [2], the Web of Science maintains a considerable amount of inertia, were 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[3]. 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 [3].

[3] http://dblp.uni-trier.de/statistics/publicationsperyear

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, 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 and thus preventing the original Web of Science from becoming a part of scientific heritage.

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[4]. 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 Related Work

In this section we provide an overview over existing Linked Data initiatives for universities, frameworks for publishing the university data and argue about existing ontologies for research data.

Open Universities Universities in Macedonia [9], Turkey [4], Germany [5], China [8] and the UK [10], among others[5], are joining the trend of making scholarly data available to the general public according to the Linked Data Principles. While publishing the data in such a way offers great benefits for developers and end-users, the publishing process itself is a demanding task in university environment. This is mainly attributed to heterogeneous infrastructure

[4] http://www.w3.org/DesignIssues/LinkedData.html

[5] http://linkeduniversities.org/

of university. For instance, in [9] Mitrevski et. al describe publishing of relational data for their faculty as RDF. Halac et. al [4] discuss a framework that utilizes Linked Data Principles to integrate four different software systems (data is stored as relational data and XML) in order to support decision-making processes inside Ege university. The LODUM project of University of Muenster [5] aggregates RDF data from the various custom triple extractors, each one of those dedicated to a particular part of university infrastructure system. Linked Data initiative for Tsinghua University in China [8] converts not only structured information about campus, faculties etc., but also unstructured information extracted by crawling the university web portal. The Open University in the UK [10] publishes video and audio content with RDF metadata. As we can see universities have different requirements and thus require different approaches for the data publishing on the university scale. However, in particular cases, such as publishing of relational data, the workflow can be generalized, which we try to achieve in this paper.

Scientific Data Publication Frameworks The state of the art scientific data frameworks are represented by open-source systems such as VIVO [1, 6] and CKAN[6]. Open-source systems can be deployed inside a university and customized according to specific requirements. While those systems store their data in non-semantically rich formats (mostly relational database management systems, RDBMS), it is possible to enrich the data by using various D2R mapping tools. On the other hand online web applications exist such as Research Gate[7]. Dependence on web applications results in the data lock-in and losing control over the data.

VIVO is a technology stack for building an interdisciplinary network. It stands out among other systems because of its wide adoption in the research community and usage of semantic web technologies. It was initially built for the data integration inside Cornell University in 2006. It supports flexible data integration where the underlying data schema can be extended on demand. The VIVO project was extended in 2009 to support cross university data integration [6]. The outcome of the project is Linked Data interfaces for seven universities inside the U.S.[8]. The corresponing VIVO ontology is described in the next subsection.

CKAN exposes metadata about datasets in a catalog and allows to publish, share, find and use the registered datasets. CKAN provides means for users and developers to easily access the published datasets. The registered datasets can be explored by end-users through free-text and faceted search based on various attributes, dataset groups and tagging. The CKAN API provides programmatic access to the metadata stored about datasets in a CKAN instance.

[6] `http://ckan.org/`

[7] `https://www.researchgate.net/`

[8] `http://vivoweb.org/about`

Research Gate is an online application for international researchers community. It supports better dissemination of publications in the de-facto PDF format by providing access to them for more than 2 000 000 scienitists[9].

Ontologies for Research Data Ontologies are one of the pillars of the Semantic Web. They structure the domain knowledge and provide a vocabulary to describe it. Using an established ontology for a domain saves development effort and leads to higher quality and better integration. Table 1 identifies ontologies related to the scientific and educational data along with their size, focus and usage intensity.

The following gives a short overview of each one, categorized by their focus.

Name	Domain	Triples	LOV score
VIVO	organization	6781	0.773
MLO	learning opportunites	130	0
XCRI-CAP	courses	110	0
Bowlogna	organization	934	0
TEACH	courses	225	0
AIISO	organization	319	0.773
org	organization	319	0.773

Table 1. University ontologies and their domain, size in triples and usage score as given by Linked Open Vocabularies (LOV) (`http://lov.okfn.org`) at 2014-08-26.

Course Ontologies

- **Meta data for Learning Opportunities** MLO[10] is a small ontology solely to represent learning opportunities such as presentations and reports. There are many concepts that it does not cover and it is not connected to any other ontology besides RDF, RDFS and OWL.
- **XCRI Course Advertising Profile** XCRI-CAP[11] is a small and narrow ontology for course advertising. It lacks classes for many important concepts such as subject matters and course equivalency for interlinking. Additionally, there is no recorded usage by the LOV.
- **Teaching Core Vocabulary** TEACH[12] is a teaching-centered ontology that covers the organizational aspects (room, building, teacher, student) but contains no means to model the content of a lecture subject matter.

[9] https://explore.researchgate.net/display/news/2012/09/19/Two+million+members,+two+million+stories

[10] http://svn.cetis.ac.uk/xcri/trunk/bindings/rdf/mlo_rdfs.xml

[11] http://svn.cetis.ac.uk/xcri/trunk/bindings/rdf/xrci_rdfs.xml

[12] http://linkedscience.org/teach/ns/

Organization Ontologies

- **Academic Institution Internal Structure Ontology** AIISO[13] models the internal organizational structure of an academic institution. With over 30000 occurrences in LOD datasets, using it would result in having many datasets with the same ontology available which makes interlinking easier.
- **Bowlogna Ontology** The Bowlogna Ontology[14] describes the process of studying according to the EU Bologna process, which standardized the study structure, for instance by unifying the different diplomas to Bachelor and Master. The LOV score of 0 shows that it is not significantly used, however.
- **Organization Ontology** The Organization Ontology[15] is a general ontology to model organizations. As such, it only covers concepts that universities have in common with other organizations. For the ones it does cover, it is well-suited, however, as it has a high usage and thus facilitates integrating between universities and other organizations.
- **VIVO Ontology** The extensive VIVO ontology models a university as an organization, which includes persons, areas, buildings and documents. While it is well suited for administrative purposes, there is only rudimentary modelling for academic, course and learning concepts where it suffers from many critical omissions.[16] The ontology is well integrated with other popular ontologies such as FOAF[17] and vCard[18].

Conclusion While the VIVO Ontology was identified as the most promising candidate for our requirements, its modelling is incomplete and not detailed enough, which hinders data integration between departments and universities.

3 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. While the above described VIVO Ontology is an obvious candidate for this task, it contains a lot of gaps left which we address with the VIVO+ ontology which builds upon VIVO and which is 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.

[13] `http://purl.org/vocab/aiiso/schema#`

[14] `http://diuf.unifr.ch/main/xi/bowlogna`

[15] `http://www.w3.org/ns/org#`

[16] For example, it does not have a concept for an academic subject or research area.

[17] `http://xmlns.com/foaf/spec/`

[18] `http://www.w3.org/TR/vcard-rdf/`

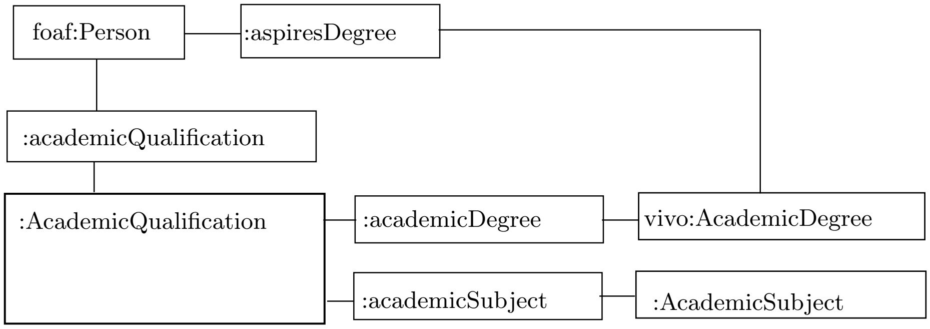


Fig. 1. Excerpt from the VIVO+ ontology structure

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 superclass. 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, whose structure is detailed in fig. 1.

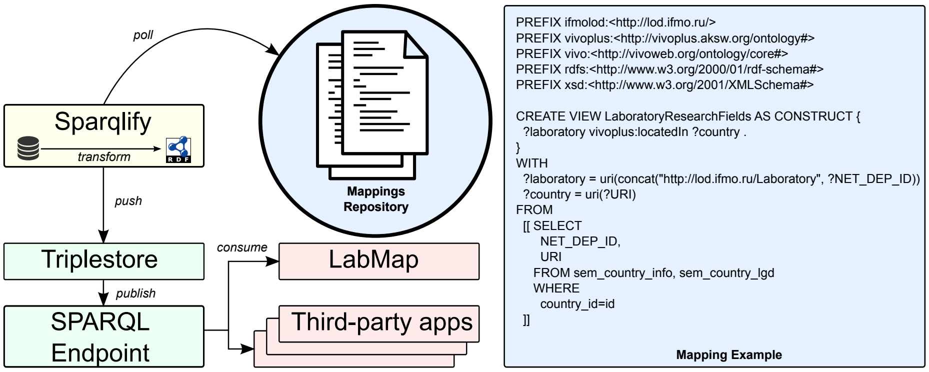
```
:Alice  a  foaf:Person;
           :academicQualification  :AliceQualification;
           :aspiresDegree            :PhDDegree.

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

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.

4 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**[19] framework on top of the Sparqlify SPARQL-SQL rewriter[20], depicted in Figure 2. The framework is evaluated using ITMO university infrastructure (i.e. Oracle RDBMS) in section 5.



```
PREFIX ifmolod:<http://lod.ifmo.ru/>
PREFIX vivoplus:<http://vivoplus.aksw.org/ontology#>
PREFIX vivo:<http://vivoweb.org/ontology/core#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>

CREATE VIEW LaboratoryResearchFields AS CONSTRUCT {
  ?laboratory vivoplus:locatedIn ?country .
}
WITH
  ?laboratory = uri(concat("http://lod.ifmo.ru/Laboratory", ?NET_DEP_ID))
  ?country = uri(?URI)
FROM
  [[ SELECT
       NET_DEP_ID,
       URI
     FROM sem_country_info, sem_country_lgd
     WHERE
       country_id=id
  ]]
```

Fig. 2. kOre: architecture overview.

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.

[19] kOre: Using Lin<u>k</u>ed Data for <u>O</u>penScience Info<u>r</u>mation Int<u>e</u>gration.

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

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**.
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.

5 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.[21] 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.[22]

At the moment of writing, the ITMO LOD dataset contains information about 43 laboratories, 188 research areas and 1001 persons. Overall it contains 12 970 triples, 2143 distinct entities and 123 distinct properties. The dump of the dataset is available online[23].

6 Applications

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

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

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

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

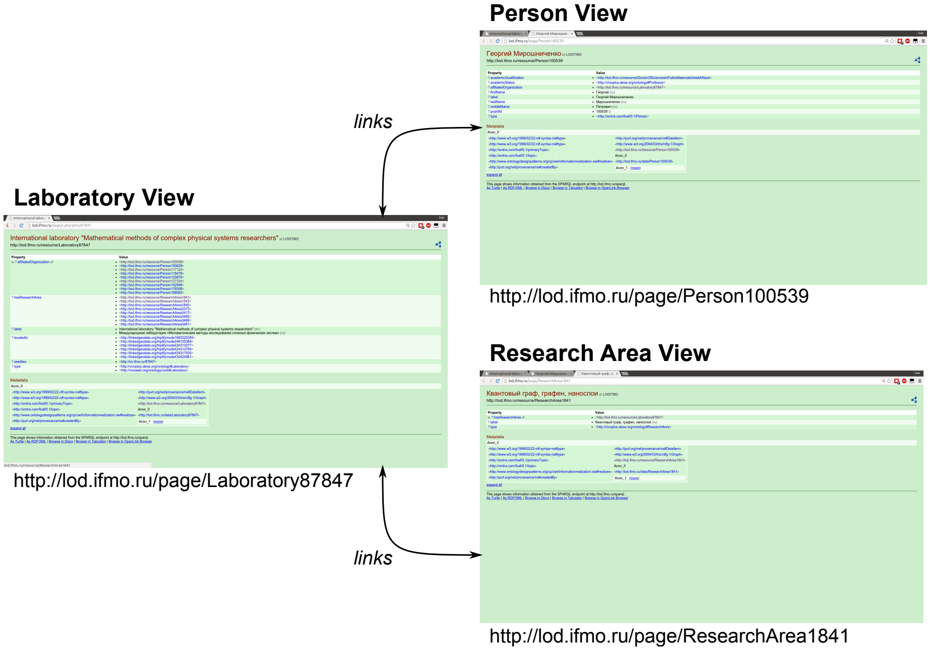

Person View
links
Laboratory View
http://lod.ifmo.ru/page/Person100539
Research Area View
http://lod.ifmo.ru/page/Laboratory87847
links
http://lod.ifmo.ru/page/ResearchArea1841

Fig. 3. Linked Data Interface enables browsing between linked entities.

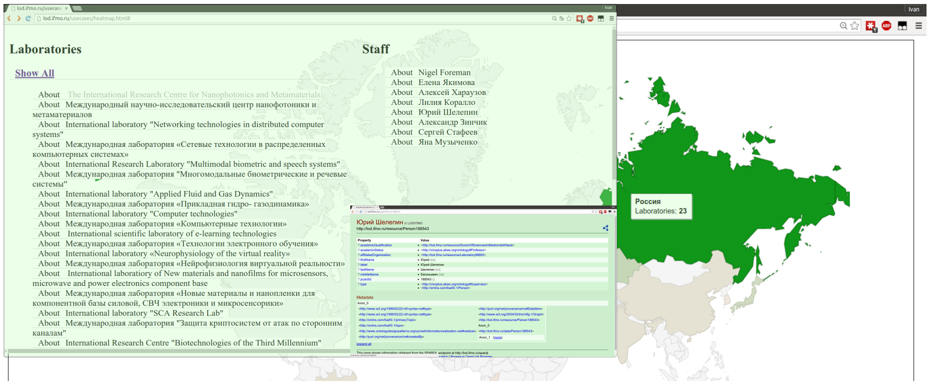
For end-users we published the data from ITMO LOD SPARQL endpoint under lod.ifmo.ru domain using the Pubby[24] Linked Data interface. Pubby makes dataset URIs dereferenceable and enables navigation between linked entities inside the dataset with the easy-to-use web interface. For example, in Figure 3 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[25] (the code is published in the GitHub repository), which shows 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 4).

7 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

[24] `http://wifo5-03.informatik.uni-mannheim.de/pubby/`
[25] http://lod.ifmo.ru/usecases/heatmap.html


lod.ifmo.ru/usecases/heatmap.html#
Ivan
Россия
Laboratories: 23
lod.ifmo.ru/page/Person188543

Laboratories

Show All

About The International Research Centre for Nanophotonics and Metamaterials
About Международный научно-исследовательский центр нанофотоники и метаматериалов
About International laboratory "Networking technologies in distributed computer systems"
About Международная лаборатория «Сетевые технологии в распределенных компьютерных системах»
About International Research Laboratory "Multimodal biometric and speech systems"
About Международная лаборатория "Многомодальные биометрические и речевые системы"
About International laboratory "Applied Fluid and Gas Dynamics"
About Международная лаборатория «Прикладная гидро- газодинамика»
About International laboratory "Computer technologies"
About Международная лаборатория «Компьютерные технологии»
About International scientific laboratory of e-learning technologies
About Международная лаборатория «Технологии электронного обучения»
About International laboratory «Neurophysiology of the virtual reality»
About Международная лаборатория «Нейрофизиология виртуальной реальности»
About International laboratiory of New materials and nanofilms for microsensors, microwave and power electronics component base
About Международная лаборатория «Новые материалы и нанопленки для компонентной базы силовой, СВЧ электроники и микросенсорики»
About International laboratory "SCA Research Lab"
About Международная лаборатория "Защита криптосистем от атак по сторонним каналам"
About International Research Centre "Biotechnologies of the Third Millennium"

Staff

About Nigel Foreman
About Елена Якимова
About Алексей Хараузов
About Лилия Коралло
About Юрий Шелепин
About Александр Зинчик
About Сергей Стафеев
About Яна Музыченко

Юрий Шелепин at LODITMO
http://lod.ifmo.ru/resource/Person188543

Property	Value
?:academicQualification	• <http://lod.ifmo.ru/resource/DoctorOfSciencesInMedicinskihNauk>
?:academicStatus	• <http://vivoplus.aksw.org/ontology#Professor>
?:affiliatedOrganization	• <http://lod.ifmo.ru/resource/Laboratory88693>
?:firstName	• Юрий (ru)
?:label	• Юрий Шелепин
?:lastName	• Шелепин (ru)
?:middleName	• Евгеньевич (ru)
?:pcardId	• 188543 ()
?:type	• <http://vivoplus.aksw.org/ontology#Supervisor> • <http://xmlns.com/foaf/0.1/Person>

Metadata

Anon_0

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://purl.org/net/provenance/ns#DataItem>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.w3.org/2004/03/trix/rdfg-1/Graph>
<http://xmlns.com/foaf/0.1/primaryTopic>	<http://lod.ifmo.ru/resource/Person188543>
<http://xmlns.com/foaf/0.1/topic>	Anon_0
<http://www.ontologydesignpatterns.org/cp/owl/informationrealization.owl#realizes>	<http://lod.ifmo.ru/data/Person188543>
<http://purl.org/net/provenance/ns#createdBy>	Anon_1 (more)

expand all

This page shows information obtained from the SPARQL endpoint at http://lod.ifmo.ru/sparql.

Fig. 4. LabMap application shows the laboratories of ITMO university filtered by collaborating countries.

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, which possibly can prevent the former from becoming a part of scientific heritage. 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.

8 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.

Bibliography

[1] Devare, M., Corson-Rikert, J., Caruso, B., Lowe, B., Chiang, K., McCue, J.: Connecting people, creating a virtual life sciences community. D-Lib Magazine 13(7/8), 1082–9873 (2007)

[2] 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

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

[4] Halaç, T.G., Erden, B., Inan, E., Oguz, D., Gocebe, P., Dikenelli, O.: Publishing and linking university data considering the dynamism of data-sources. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 140–145. I-SEMANTICS '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2506182.2506202

[5] Kessler, C., Kauppinen, T.: Linked Open Data University of Münster – infrastructure and applications. In: Demos at the 9th Extended Semantic Web Conference (ESWC2012). Heraklion, Greece (May 2012)

[6] Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J., et al.: VIVO: Enabling national networking of scientists. In: Proceedings of the Web Science Conference. vol. 2010, pp. 1310–1313 (2010)

[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)

[8] Ma, Y., Xu, B., Bai, Y., Li, Z.: Building Linked Open University Data: Tsinghua university Open Data as a showcase. In: The Semantic Web, pp. 385–393. Springer, Berlin Heidelberg, Germany (2012)

[9] Mitrevski, M., Jovanovik, M., Stojanov, R., Trajanov, D.: Open university data. In: Proceeding from the 9th Conference for Informatics and Information Technology (2012)

[10] Zablith, F., d'Aquin, M., Brown, S., Green-Hughes, L.: Consuming Linked Data within a large educational organization. In: Proceedings of the Second International Workshop on Consuming Linked Data (COLD) at the 10th International Semantic Web Conference (ISWC 2011) (2011)