Publishing TUWien Data as Linked Open Data

Lukas Baronyai

March 7, 2016

Contents

List of Figures

Abstract

Here comes the abstract

1 Introduction

While the pressure on governments and public organizations to release *Open Data (OD)* has significantly grown with the spread of information systems, there has been also a need for *linking* these data from various sources to understand the information as a whole.

Open Data includes non-privacy-restricted and non-confidential data. Therefore any restrictions in distribution are prohibited and data is funded only by public money. Janssen et al. [9] The application domain for Open Data providers is not restricted by its nature in any way and ranges from traffic, weather, statictics to budgeting in the public sector. Just the publication of Open Data seems not enough but in addition the implementation of a feedback loop result in *Open Government*. This has the advantage of a constant adaption to the citizen's needs instead of just visualizing former closed data.

Despite its wide adoption of Open Data it does restrict the published Data format in any way, thus complicating the integration of heterogeneous data sets. The World Wide Web has proven great success spreading knowledge of various data sources all over the world. The building block of the Web are documents and links connecting them to form a global information space. This can be seen as the key success factor in its nearly unconstrained growth [8]. Following these principles of publishing and connecting data on the Web is known as *Linked Data (LD)*. More technically, it refers to machine-readable data which is linked to external data sets and can in turn be linked from other data sets.

Berners-Lee [1] developed a five star rating scheme for classifying *Linked Open Data (LOD)*, which combines Linked Data and Open Data. The scheme ranges from one star describing Open Data only to five stars describing Open Data in a machine-readable format using open standards with links to other data sets.

Although Linked Open Data offer universities new opportunities for providing unprecedented insight into its core activities and ease application development, a major **problem** is that **Linked Open Data has not been widely adopted by universities yet**. Even tough there are a few examples [5] of publishing university related data as Linked Open Data, there has been little knowledge of using Linked Open Data for publishing university related information.

The remainder of this section states the addressed research question of this paper, describes the contributions in the course of investigating the research questions and gives an overview of the structure of this paper.

1.1 Research Questions

The fundamental research questions underlying this paper is:

> *How can Linked Open Data help to improve processes in university context and how can it be successfully applied?*

More concrete this paper concentrates on the following more concrete research questions:

Q1: What are best practices regarding the applicability of Linked Open Data in university settings? As of now, there are no established best practices for the use of Linked Open Data due to its little adoption in university contexts. For this very reason it is crucial to identify strengths and limitation from previous experiences [5] of using Linked Open Data as core technology.

Q2: What are major benefits and barriers for each stakeholder and what are useful use cases? We identified three different stakeholders in university context *Students*, *Researchers* and *Administration staff*. Since the success of any new technology highly depends on the acceptance of the stakeholders, the needs of each of the target groups needs to be examined. Furthermore, use cases are important to showcase profits and shortcomings to a non-technical audience.

Q3: What are major challenges for the implementation of a Linked Open Data solution? As the implementation of a Linked Open Data solution is a time consuming task, the knowledge of probable challenges from the technical perspective as well as from the management perspective is a key factor for the successful adoption.

Q4: How would a prototypical implementation of Linked Open Data look like? Among the various existing data sets available it needs to be investigated if a (semi-) automatic transformation is feasible or is the manual data provision enough. In addition, from an implementers perspective of view, critical factors regarding the storage and retrieval of Linked Open Data need to be identified.

1.2 Methodology

Finding an answer for the research questions above has lead to the following three methodologies:

A coordinated set of semi-structured interviews To answer research questions (RQ) two to four, we interviewed a selected set of stakeholders representing *Students*, *Researchers* and *Administration staff* respectively. Semi-structured interviews were selected as the means of data collection because they are well suited for exploring the impressions and interests of the interviewees as in a discussion while still following a defined structure.

Litrature Review Undertaking a litrature review to justify scientific contributions and making sound conclusions is an established practice in any scientific community. Since our scientific work targeted in particular to the Semantic Web community, we made some pre-assumptions of a basic understanding of the technologies and concepts regarding Linked Data. More specifically, the concept of an ontology and example knowledge descriptions languages describing these will not be covered in this paper.

Conceptual System Design The development of applications based on Linked Open Data requires a methodology which describes a common understanding of the overall system infrastructure. Therefore we designed a conceptual model of a prototypical implementation of a Linked Open Data solution.

1.3 Contributions

The work in this paper mainly contributes to different aspects wich need to be considered when designing and implementing a Linked Open Data application. More precisely, our contributions can be categorized into the following four areas:

1. Identifying best practices for Linked Open Data in university context. Due to the crowing complexity and the large amount of data information systems need to process, there has been the need to efficiently handle Linked Data as well. We gave a brief overview of the already published research work regarding Linked Open Data in university context. In particular, we compared the profits and shortcomings in existing Linked Open Data solutions.

2. Finding benefits/barriers with additional use cases for stakeholders. As with every software project the very first phase of the Software Development Lifecycle (SDLC) is the *Evaluation of the Requirements*. As a Linked Open Data solution has additional requirements to the structure of the data and due to its open nature, we investigated if the overhead compared to an established technology (e.g. a database based solution) is worth the effort. A set of selected participants from the areas Research, Student Affairs and Administration are interviewed at the University of Technology in Vienna and their benefits/barriers are compared. Additionally, we proposed several use cases emphasising their point of view.

3. Discovering possible obstacles for implementers of a Linked Open Data Solution. As the application domain for a Linked Data is limited to the university context, our work includes a defined set of Linked Open Data applications which were merged together from the conducted interviews. That use cases showcased probable shortcomings which might arise before, during or after the implementation.

4. Sketching a prototypical implementation of a Linked Open Data Solution. In consideration of the above mentioned obstacles of a possible Linked Open Data Solution, we gave an outline of a prototypical implementation. It begins by covering the whole process of data provision and ends by applications made for end users.

1.4 Structure of this Paper

%%%%tbd%%%%

2 Related work

2.1 Linked Universities

One of the most important university projects in the world of LD are the LinkedUniversities. They are "‘*an alliance of european universities engaged into exposing their public data as linked data*"’[1], providing help and knowledge for other universities who wants to implement LD-Systems in their infrastructure. Addressing the problem of connecting data and developing new sites by inexperienced universities, the alliance provide information so they don’t have to be re-learned. For this purpose the LinkedUniversities offering a portal as collaborative space with common vocabularies and practices for reusing, describing and sharing.

Their goals are:(d’Aquin et al. [6])

- "‘Identify, support and develop common linked data vocabularies, usable accross universities for common concepts such as courses, qualifications, educational material, etc."’
- "‘Describe reusable recipes, and share reusable tools, for exposing linked data in universities"’
- "‘Support, through experience sharing and reuse, initiatives towards exposing university data as linked data"’

The members of this alliance are:(d’Aquin et al. [5])

- The Open University, UK
- University of MÃ¼nster, Germany
- Aalto University, Finland
- University of Southampton
- Royal Institute of Technology (KTH) / MetaSolutions AB
- Aristotle University of Thessaloniki, Greece
- Ege University, Turkey
- Charles University in Prague
- Universitat Pompeu Fabra

2.1.1 Example: The Open University and the LUCERO Project

LUCERO (Linking University Content for Education and Research Online) was a project from the Open University, funded for 1 year by the JISC Information Environment 2011 Programme under the call Deposit of research outputs and Exposing digital content for education and research. Aim of the project was to "‘ *scope, prototype, pilot and evaluate reusable, cost-effective solutions relying on linked data for exposing and connecting educational and research content*"’.d’Aquin et al. [3] The projects connected with other organizations through LinkedUniversities.org to gather common issues and practices. They outcome was the first university linked data platform, `http://data.open.ac.uk/`, with much impact on The Open University and the education community.

%%%%TODO: section about Open Research Online and/or Open Science%%%%

2.2 Austrian Open Data

2.3 Linked Data for Libraries (LD4L)

%%%%tbd%%%%

2.4 Earlier studies

- Interlinking educational Resources and the Web of Data - a Survey of Challenges and Approaches (http://linkeduniversities.org/)
- a few more at http://linkeduniversities.org/lu/index.php/publications/index.html

[1] d’Aquin et al. [6]

3 Benefits and challenges of using LOD at TUWien

3.1 Methodology

In this study the data were acquired by a coordinated set of semi-structured interviews. As mentioned the stakeholders were classified into three groups (*administrative staff*, *students* and *researchers*) and therefore three different versions of the questionnaire but with joint parts for statistically evaluation were worked out. For each version exists an according paper, in this work only the category "‘researcher"’ will be described.

3.1.1 Design of questionnaire

The main purpose of the interviews were the collecting of the stakeholders thoughts, needs and knowledge, so the method of of a semi-structured interview was chosen. A *fully structured interview* would not be adequate because of it's strict character allowing only predefined answers and a *unstructured interview* would be to difficult to analyze.

After choosing the method, the questionnaire was defined. To allow a general, generic shared analyze of the interviewees the team decided to mix open questions from the semi-structured model with closed questions with fixed, predefined questions. The result had four parts:

1. General question about the interviewee for classification, about his/her work
2. General question about the interviewee's knowledge in general technical and LOD context. This part is the part for statistical evaluation.
3. Explanation of LD, followed by a specific set of questions targeting the thoughts and opinions of the interviewee about presented use cases and example application. Motivation of this part is to introduce the interviewee to LOD if it is an unknown topic and let him/her start to think about LOD to prepare the next part
4. Wide open Questions to explore and find use cases and existing data sources for LOD application at the university.

The examples from part 3 were LD in libraries (see 2.3) and an obvious source of research related data: the publication database.

3.1.2 Description of interviewed people

As mentioned the interviewees of this study were chosen according to the category "‘*Research*"’, so the interview partner were active researcher in various fields. Altogether four interviews were done. Because of the technical character of LOD the chosen people are all technically experienced so they are able to imagine use cases at the university. In future work there is a need of more less experienced researchers to understand their thoughts.

3.1.3 Data Validity and Quality

To ensure both a continuous conversation flow and a high quality recording of the spoken words, the interviews were held in teams, one speaker and one writer making notes. Additional all interviews were audio recorded. As result the data are available as interview notes and audio records.

3.2 Results

%%%%tbd%%%% The opportunity to develop new ideas based on an access to open data was very common welcome across all questions (though no concrete ideas came up).

Costs

3.2.1 General statistically evaluation

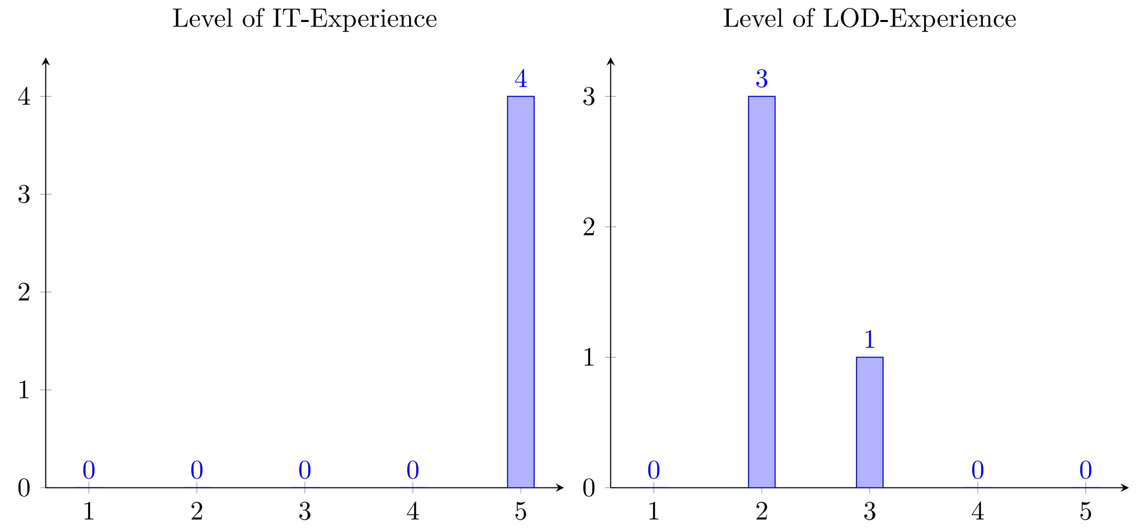


Figure 1: Level of Experience

3.2.2 Library data

Use case %%%%TODO: more interlinking with LD4L%%%% This question aimed for a use case similar to the project Linked Data Libraries (described in Section 2.3). The proposed scenarios was to publish the data or meta-data of the university library (and all of its specialized libraries) as LOD and provide an application to access the data. A further option of this scenario would be an interlink to other LOD data sets, e.g. from the publisher "‘Springer"’.

Statistically evaluation It can be seen in Figure 3 that the interviewed persons strongly agreed to the scenario (found it "‘extremely useful"’) and could imagine a similar project at the TU Vienna. Only one interviewee found it difficult to see advantages and therefore argued that he wouldn't use it. Another one found it indeed useful in a general context but not for his own work.

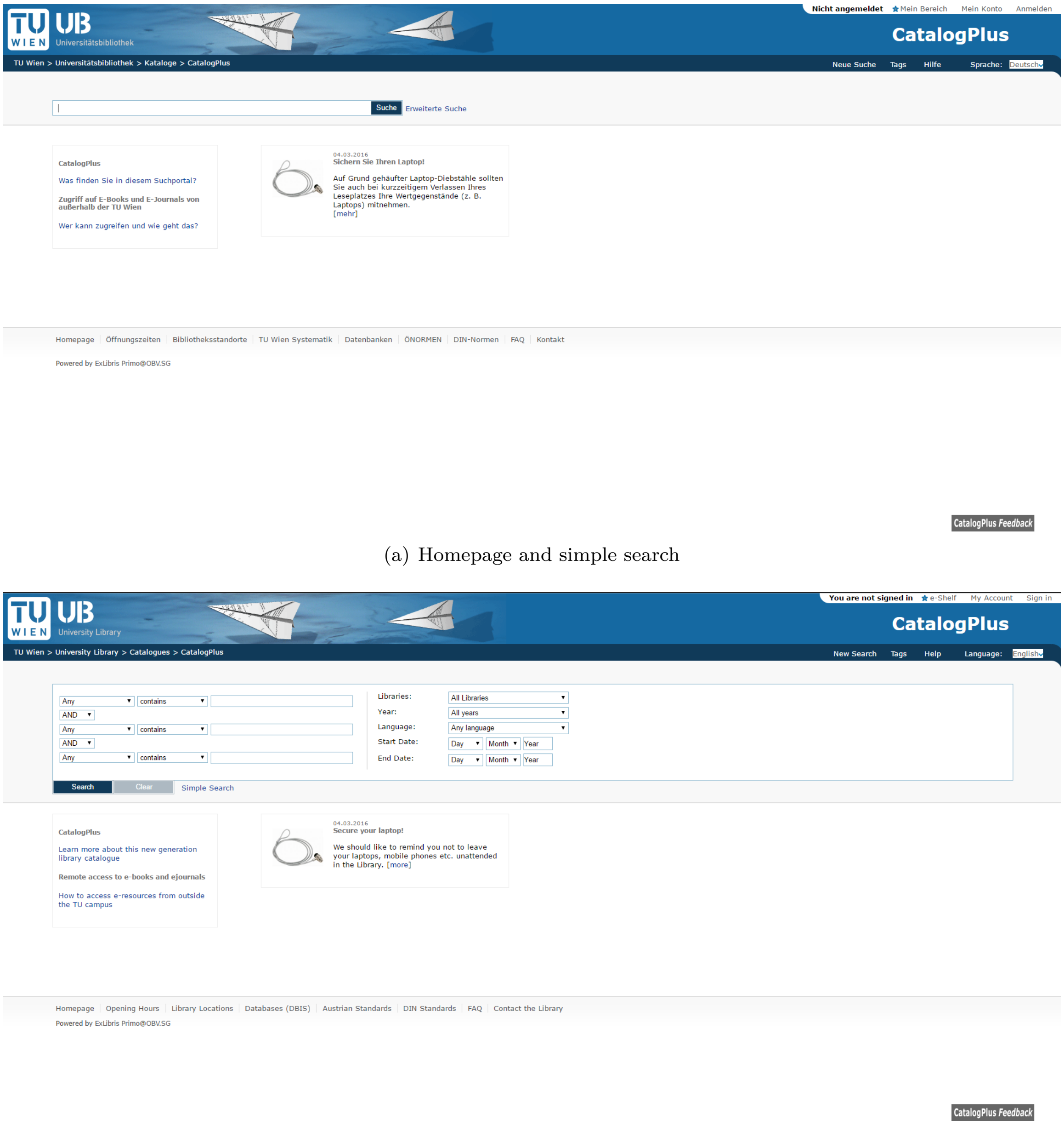
Needs, potential benefits As stated, all of the interviewed persons could imagine a similar project. One of the main reasons of the strong acceptance was the current interface of the library website [2], which only allows a search with only a few, specified parameters (see Figure 2). Also the physical search in the library itself was claimed due to a lack of orientation and knowledge about the position of e.g. a searched book. Both point of criticism are expected to vanish by an open access to the data and appropriate applications, which provides a detailed and personalizable search interface.

Furthermore an open access to the data was seen as a chance for everyone to interact with it and as opportunity to stimulate creativity of the people.

Barriers, potential disadvantages The library institution was very conservative perceived with skepticism, refusals and resentment against new ideas, especially in a technical term. It was stated by some interviewees, that this kind of project would only be seen as an extra amount of work in the library and not as an opportunity.

Beside the expected opposition there was also a real amount of work estimated by the interviewees. In particular there would be an effort to invest in digitizing books and keep this data up-to-date. This would be an important part of the project, otherwise the data would loose there value - not

[2] www.ub.tuwien.ac.at/



Nicht angemeldet ★Mein Bereich Mein Konto Anmelden

TU WIEN UB Universitätsbibliothek

CatalogPlus

TU Wien > Universitätsbibliothek > Kataloge > CatalogPlus

Neue Suche Tags Hilfe Sprache: Deutsch

| Suche Erweiterte Suche

CatalogPlus

Was finden Sie in diesem Suchportal?

Zugriff auf E-Books und E-Journals von außerhalb der TU Wien

Wer kann zugreifen und wie geht das?

04.03.2016
Sichern Sie Ihren Laptop!

Auf Grund gehäufter Laptop-Diebstähle sollten Sie auch bei kurzzeitigem Verlassen Ihres Leseplatzes Ihre Wertgegenstände (z. B. Laptops) mitnehmen.
[mehr]

Homepage | Öffnungszeiten | Bibliotheksstandorte | TU Wien Systematik | Datenbanken | ÖNORMEN | DIN-Normen | FAQ | Kontakt

Powered by ExLibris Primo@OBV.SG

CatalogPlus Feedback

(a) Homepage and simple search

You are not signed in ★e-Shelf My Account Sign in

TU WIEN UB University Library

CatalogPlus

TU Wien > University Library > Catalogues > CatalogPlus

New Search Tags Help Language: English

Any ▾ contains ▾
AND ▾
Any ▾ contains ▾
AND ▾
Any ▾ contains ▾

Libraries: All Libraries ▾
Year: All years ▾
Language: Any language ▾
Start Date: Day ▾ Month ▾ Year
End Date: Day ▾ Month ▾ Year

Search Clear Simple Search

CatalogPlus

Learn more about this new generation library catalogue

Remote access to e-books and ejournals

How to access e-resources from outside the TU campus

04.03.2016
Secure your laptop!

We should like to remind you not to leave your laptops, mobile phones etc. unattended in the Library. [more]

Homepage | Opening Hours | Library Locations | Databases (DBIS) | Austrian Standards | DIN Standards | FAQ | Contact the Library

Powered by ExLibris Primo@OBV.SG

CatalogPlus Feedback

(b) Extended search

Figure 2: Website of the TUVienna library

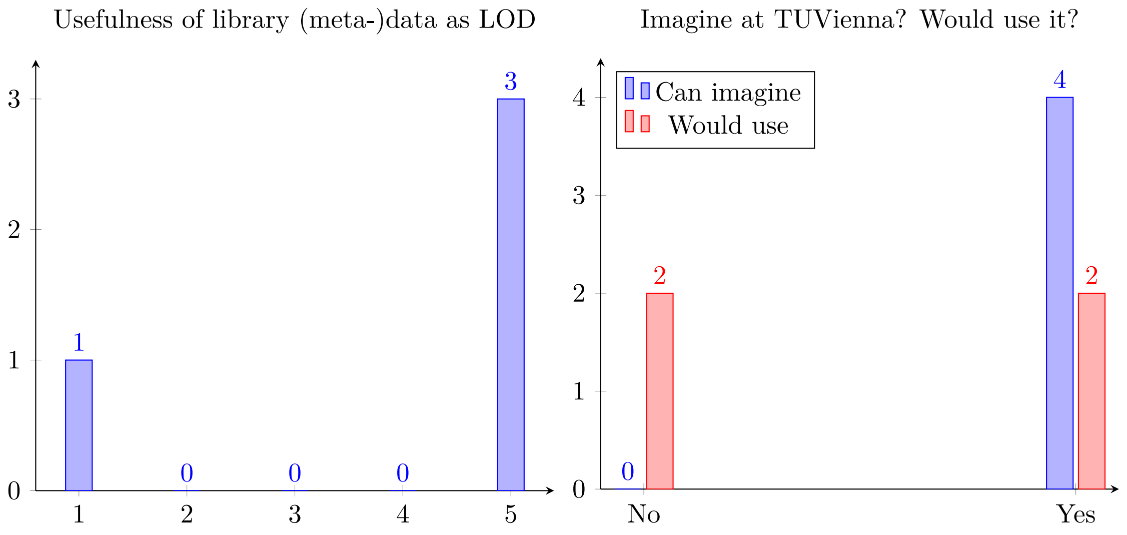
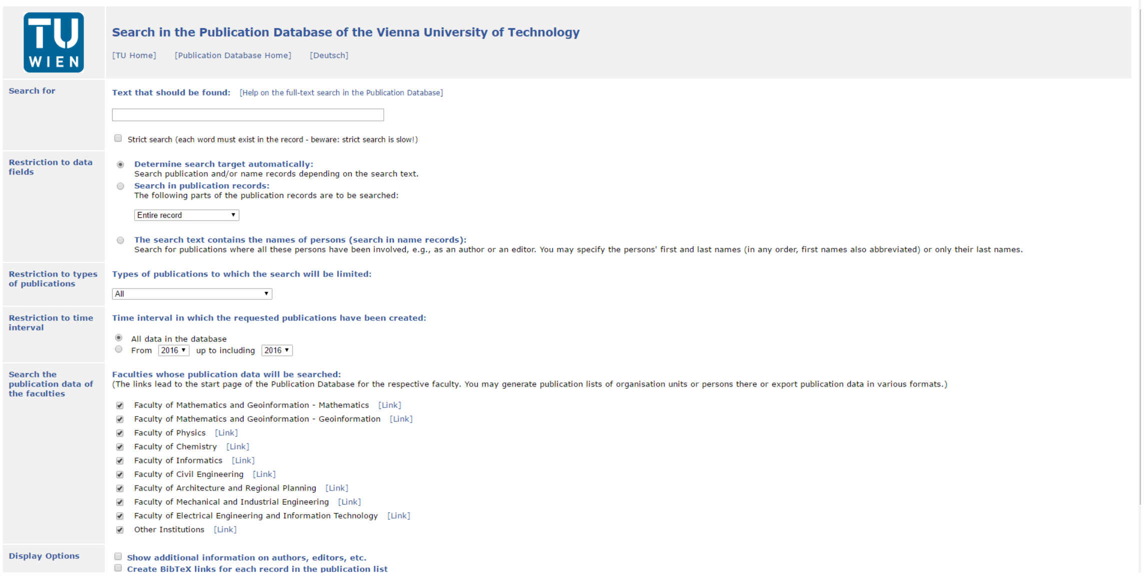


Figure 3: Usefulness of library (meta-)data as LOD

digitized data could not be distinguishable from not available data. Therefore it would be essential to have a complete and current database.
Another concern of the interviewed people were the copyright of the data. However, this problem was seen as handleable, because only meta-data would be open accessible.
At last, one interviewee declined the idea and usefulness of the LD4L and similar projects because of the existing platform Google Scholar [3]. He stated, that it already provides similar data and therefore everything he needs for his work. Further a additional platform would be too intricate to use.

3.2.3 Data from the Publication database



Search in the Publication Database of the Vienna University of Technology

[TU Home] [Publication Database Home] [Deutsch]

Search for

Text that should be found: [Help on the full-text search in the Publication Database]

□ Strict search (each word must exist in the record - beware: strict search is slow!)

Restriction to data fields

● Determine search target automatically:
Search publication and/or name records depending on the search text.
○ Search in publication records:
The following parts of the publication records are to be searched:
Entire record ▾

○ The search text contains the names of persons (search in name records):
Search for publications where all these persons have been involved, e.g., as an author or an editor. You may specify the persons' first and last names (in any order, first names also abbreviated) or only their last names.

Restriction to types of publications

Types of publications to which the search will be limited:
All ▾

Restriction to time interval

Time interval in which the requested publications have been created:
● All data in the database
○ From 2016 ▾ up to including 2016 ▾

Search the publication data of the faculties

Faculties whose publication data will be searched:
(The links lead to the start page of the Publication Database for the respective faculty. You may generate publication lists of organisation units or persons there or export publication data in various formats.)

☑ Faculty of Mathematics and Geoinformation - Mathematics [Link]
☑ Faculty of Mathematics and Geoinformation - Geoinformation [Link]
☑ Faculty of Physics [Link]
☑ Faculty of Chemistry [Link]
☑ Faculty of Informatics [Link]
☑ Faculty of Civil Engineering [Link]
☑ Faculty of Architecture and Regional Planning [Link]
☑ Faculty of Mechanical and Industrial Engineering [Link]
☑ Faculty of Electrical Engineering and Information Technology [Link]
☑ Other Institutions [Link]

Display Options

□ Show additional information on authors, editors, etc.
□ Create BibTeX links for each record in the publication list

Figure 4: Publication database

Use case %%%%TODO: interlinking with open science and/or Open Research Online%%%% The proposed use case applied to the already existing publication database of the university. In the current state the database can be accessed via its website [4], where a interface for search exists. For other search parameters a request to the administration has to be made. The introduced LOD approach provides an LOD interface to the existing database, so the data can be accessed over e.g. SPARQL.

Statistically evaluation Similar to the library use case the interviewees liked the idea and found it all "'extremely useful"'. Everyone could imagine such a project at TUVienna and would also use it.

Needs, potential benefits In contrast to the previous use case,the work amount of work was significant lesser stated, because the data are already there, just as the need of being up-to-date of the database.
Further some of the interviewed persons (independent to each other) come up with the idea of building an application based on the Linked Open Data interface to provide an overview of a researcher references as a widget or similar for a personal website. The data could be interlinked e.g. with the LOD interface of the Springer publisher [5] and complemented by including the Journal Impact Factor (JIF).

Barriers, potential disadvantages The interviewees identified uniformly a problem with the data ownership. Although there is an administration of the database, the *ownership* of the data itself is very unclear and therefore a contact and responsible person would be hard to find for such

[3] www.scholar.google.at/
[4] www.publik.tuwien.ac.at/pubstart.php
[5] www.lod.springer.com/

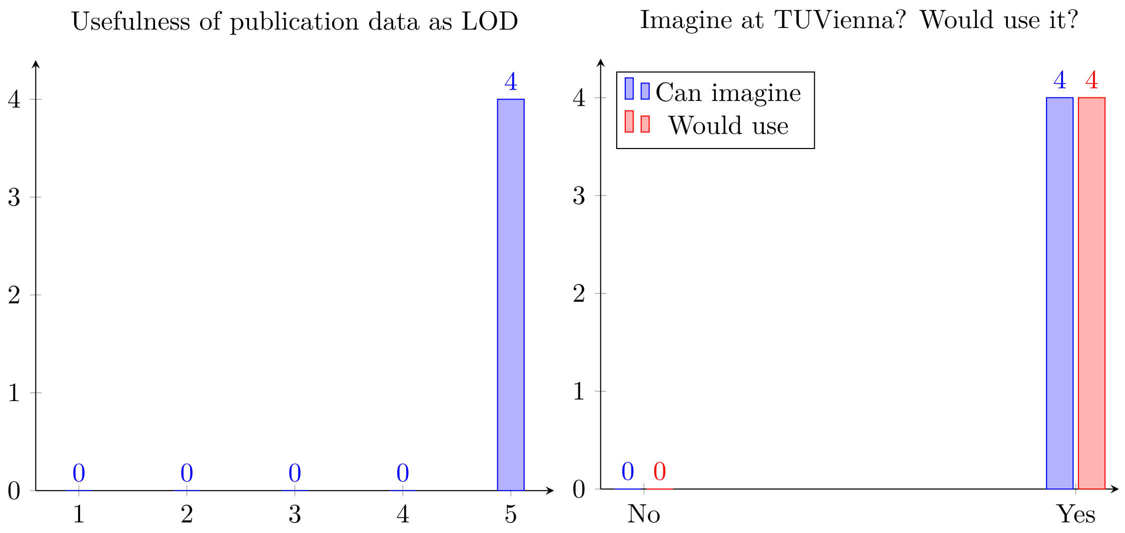


Figure 5: Usefulness of publication data as LOD

a project - there needs to be more investigation.
Another point, some of the interviewed persons were concerned of, was quality assurance - to use the data, they need to be synchronized with the original database. To taking care of this point, the implementation of a similar project has to ensure this.

3.2.4 Other data

%%%%tbd%%%%

Needs, potential benefits, use-cases

Challenges

4 Proposed technical architecture and challenges

As an organization covering areas like research, student affairs and administration a university has to manage a significant amount of knowledge, adding new information on a daily basis. Such an application domain is complex and includes areas like management of an academic library and provision of educational resources which have to be conform with stakeholders requirements. Traditionally the *Service Oriented Architectures (SOA)* have been used to meet these needs. However, as the application domain grows many small and similar services tend to emerge. That phenomena can not only be observed at the Vienna University of Technology, but also at the Open University[6] Zablith et al. [12].

A major problem of evolving similar, independent services are diverging data formats and service owners. Thus, knowledge and administrative information that has been collected by multiple services can not be easily interlinked. An example for such isolated services is the e-learning platform called TUWEL[7], combining moodle[8] and the central information system called TISS[9]. These services provide course information and material, but are intended for different purposes. Whereas TISS focuses mainly on administrative functionality, TUWEL supports the interaction between teacher and student. Adding additional services which, for example, synchronize deadlines and registration dates is costly due to the fact that the information is separated over different isolated sources and not easily accessible.

4.1 The big picture

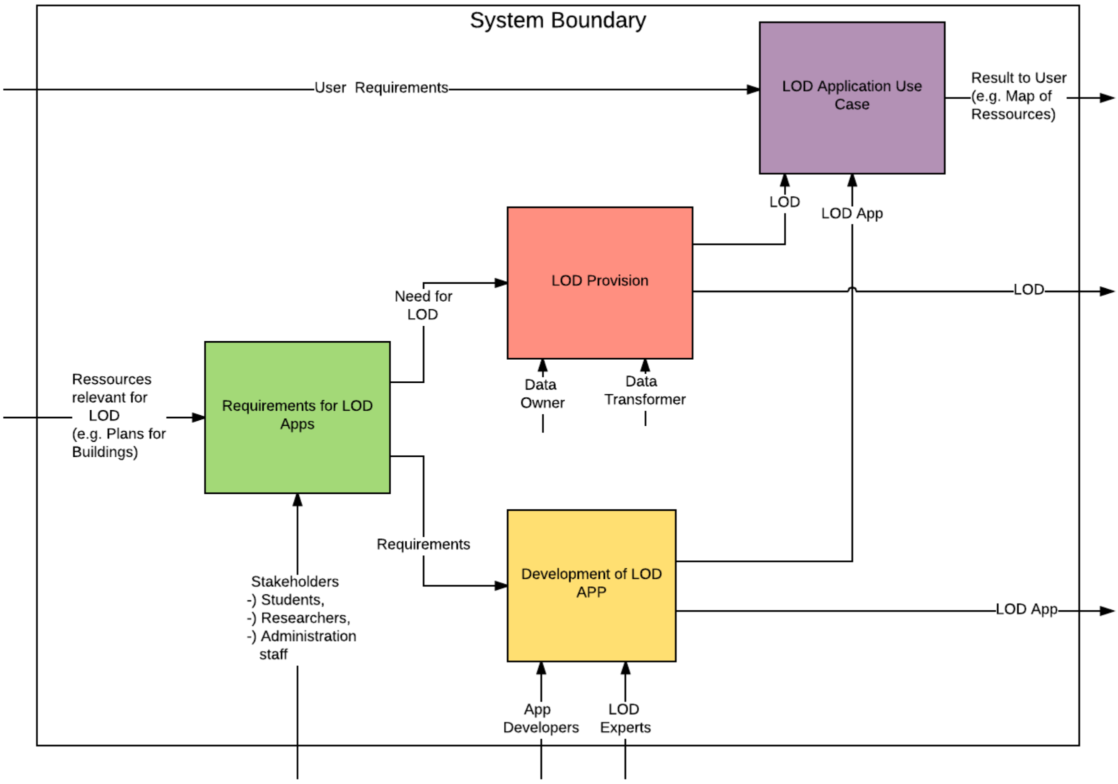


Figure 6: High level framework for LOD publishing in IDEF0 notation

4.1.1 Parts

In this sub-section we provide an overview of a high level architecture, illustrated in Figure 6, for a university wide publication framework which includes:

[6] http://www.open.ac.uk/

[7] https://tuwel.tuwien.ac.at/

[8] https://moodle.org/

[9] https://tiss.tuwien.ac.at/

- **Requirements for LOD-Applications**
  At the beginning stand the existing resources and the needs of the stakeholder. Combining these leads to the requirements for an application, the first step of the process. According to this requirement a decision has to be made whether they are realizable depending on the cost-value ratio and whether the solution has to be a LOD application. These process has to actively involve all stakeholders.

- **Data Provision**
  After defining the requirements, the existing data (e.g. a publication database) must be transformed in a appropriate, machine readable LOD format. These can happen in a manually (only for small data sets) or a (semi-)automated (for big data set) way. Ideally there is a automated transformer based on an existing, well maintained and up-to-date database so there has to be less cared about the of the data (for a more technical description see Section 4.2.1). Key roles for this process are the original data owner and the data transformer.

- **Development of an Application**
  Based on the requirements definitions from the previous step a proper application (e.g. a browser of publication data) now can be constructed considering the stakeholder needs and existing resources (transformed to LOD). To support this process and to not obtaining all knowledge from zero it is recommended to access the knowledge of LOD experts (e.g. provided by the LinkedUniversities [10]). The development can simultaneously be done with the data provision if proper interface between the application and its data are made.

- **Application Use Case**
  Combining the data, transformed in LOD, the LOD application and user requirements (not the application requirements from the first step) result in the actual application use case, representing the environment or the domain.

4.1.2 In- and outbound interfaces

There were several indirectly interfaces mentioned above - we define them now in a more formal way and divide them in inbound (arrows pointing from outside the system boundary) and outbound interfaces (arrows pointing from inside the system boundary).

The inbound interfaces are:

- **User Requirements**
  Understanding user requirements is an essential part of the software development process. Due to the open nature of Linked Open Data privacy concerns and legal issues should be already considered in the requirements as misunderstandings are hard to fix in later phases.

- **Existing Resources**
  The starting point of every LOD application are resources, ideally already existing (to reduce the amount of work). It can be everything from relational databases, simple Excel files or other Linked (Open) Data sets. For more details see section 4.2.1.

- **Stakeholders**
  Stakeholders are defined as *"‘a person, group or organization that has interest or concern in an organization. Stakeholders can affect or be affected by the organization's actions, objectives and policies."’* Post et al. [10] Their needs and demands have to be considered in a LOD project as similar as in every other software project.

- **Application Developers**

- **LOD Experts**
  To avoid unnecessary redundancy in acquiring knowledge of LOD implementation, it is highly recommended to involve either LOD experts in the development or access their accumulated know-how e.g. by platforms like LinkedUniversities [11] (see section 2.1)

- **Data Owner**
  Every data set has its owner, therefore this role has to be considered in the development process to avoid organizational conflicts and copyright issues. Ideally he is directly involved to access his specific know-how about the data set.

[10] www.linkeduniversities.org/

[11] www.linkeduniversities.org/

- **Data Transformer**
  To use a data set in a LOD approach, the data have to be transformed either manually or (semi-)automatically into a proper format (see Berners-Lee [1] for the 5 star model of data format and section 4.2.1 for details about transformations).

The outbound interfaces are:

- **Linked Open Data**
  As a result of the LOD provision the actual data are provided e.g. as SPARQL endpoint, so others can easily access and use them in other applications. For technical details about endpoints see section 4.2.4 or as example the SPARQL endpoint of The Open University `www.data.open.ac.uk/query`.
- **Application**
  The outcome of the development phase are applications for end-users to access and interact with the data e.g. in form of an web platform. For technical details about endpoints see section 4.2.4 or as example the application list of The Open University `www.data.open.ac.uk/applications.html`.
- **End-User Result**
  Finally the main result is the actual interaction of the users with the applications and endpoints, (hoepfully) acting in the boundaries of the defined use cases.

4.2 Proposal of a technical architecture

In this section we will briefly describe a proposal technical architecture for publishing (existing) data as Linked Open Data . The proposal is mainly inspired by the toolkit "'Tabloid"' ("'Toolkit ABout Linked Open Institutional Data"') by the LUCERO Project [12] - a collection of tools, examples and documentions. The general principle is a system, which extract RDF data from existing data sources (section 4.2.1), load them into a triple store (section 4.2.3) and finally expose them to the web (section 4.2.4). For illustration see the LUCERO workflow in figure 8 (in this part only the generic parts are described). In this workflow there are way more components than we will talk about (e.g. mechanism of detecting data changing) than shown in the figure. Additional a more generic architecture can be seen in figure 7.

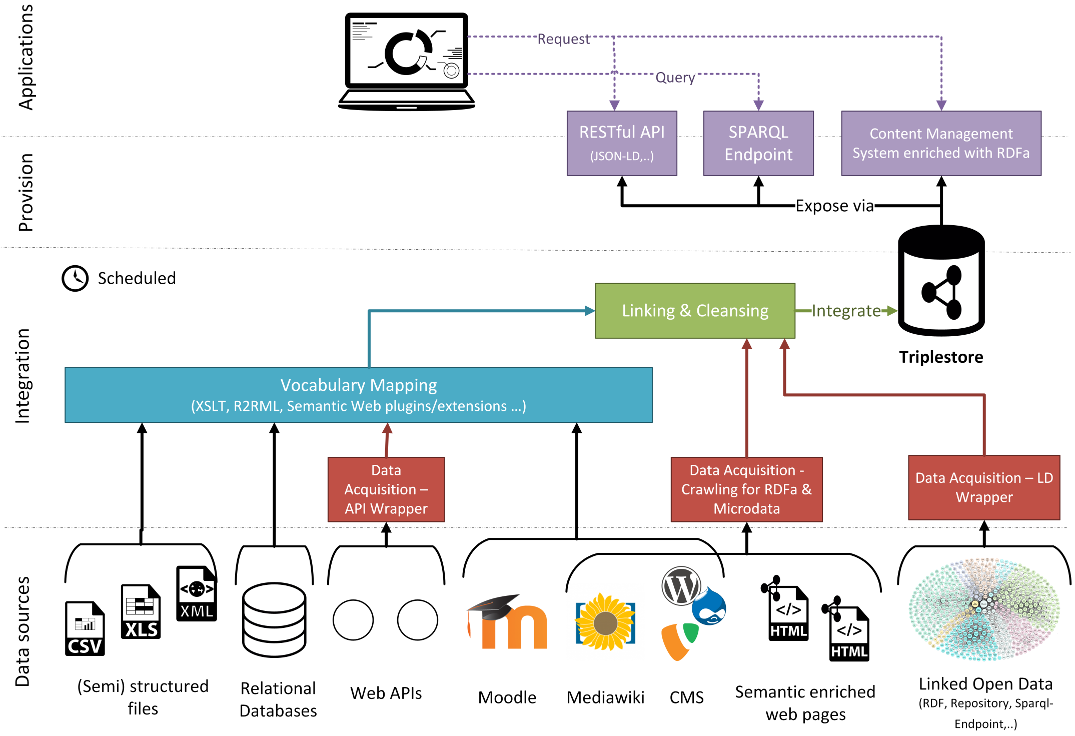


Figure 7: Generic high level technical architecture of a LOD system //TODO: source

4.2.1 Collect & extract data from sources

The first step must be to collect and extract the data from the source and transform them into a proper LOD format, e.g. RDF. By now there are basically for every data format tools for transforming them into RDF, LinkedUniversities made a collection of them [14], here are the main tools:

- **From Relational Databases**
  - **Triplify** [15] is a tool which use SQL queries to generate RDF data from a relational database.
  - **D2RQ** [16] is a tool which also use SQL queries but with the use of a mapping that relates the structure of a database to RDF triples. It transformers SPARQL queries at run-time into SQL queries using this mapping.
- **From XML and RSS**
  In terms of syntax, XML, RSS and RDF sharing the same base, so there have to be fewer effort to be done to transform them syntactic, commonly using XSLT. W3C recommended the GRDDL [17] language for this purpose.

[12] d'Aquin et al. [4]

[14] http://linkeduniversities.org/lu/index.php/tools/index.html

[15] http://triplify.org/

[16] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[17] http://www.w3.org/TR/grddl/

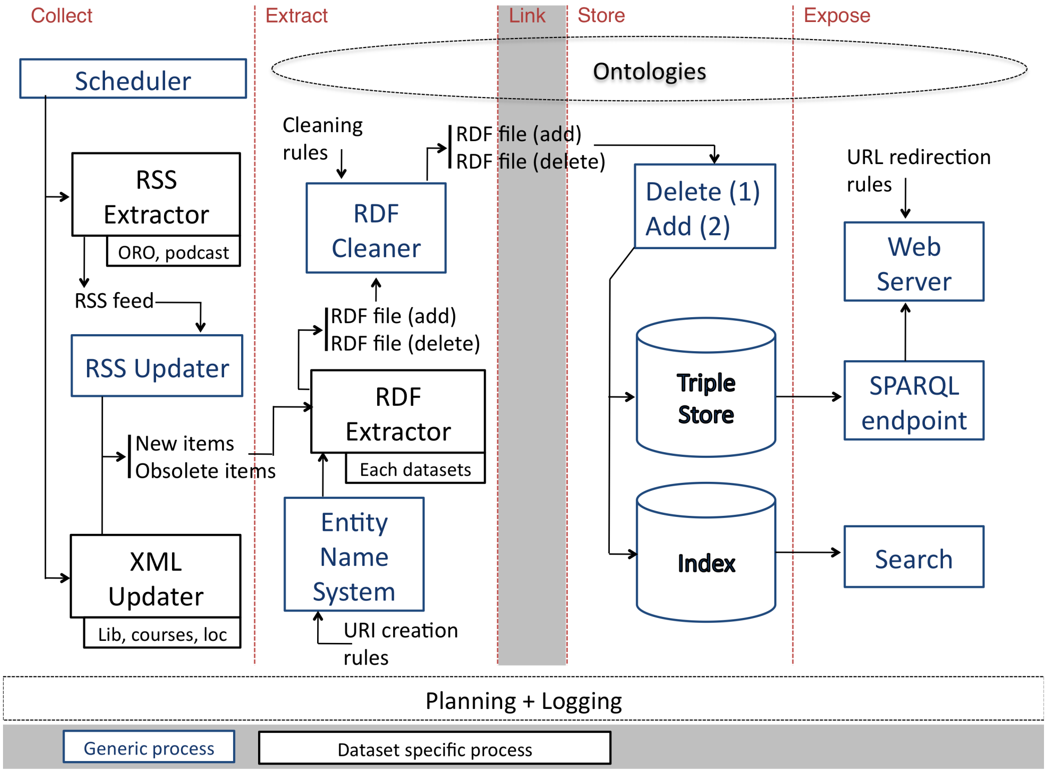


Figure 8: Example: the LUCERO workflow[13]

- **From Tables and Spreadsheets**
  - Google Refine [18] with the RDF Extension [19] is an easy way to clean, transform and explore data in a tabular format including MS Excel, Google Spreadsheet and CSV.
  - Other tools like Any23 [20] or QUIDICRC [21] providing simple transformation from CSV to RDF

After extracting the data as RDF from there source, they need to be cleaned and interlinked with themselves and other data.

4.2.2 Vocabularies and Ontologies

To represent and store the collected data, they must be mapped into an ontology and a vocabulary.

Gruber [7] defines an *ontology* as "'*an explicit specification of a conceptualization*"'. *Conceptualizations* are objects, entities or concepts that may or may not exist in the universe. In addition to that, the *vocabulary* defines the relationships between those objects. In other words, the vocabulary defines the conceptual model of what can be represented. Bock et al. [2] describe ontologies from a more practical point of view defining the three conceptual components of an ontology - *classes*, *instances* and *properties*. Regardless of what concrete implementation of an ontology is used graphical representations (e.g. graph) are preferred over textual ones to give a high level overview of the concepts used in an ontology.

Again, LinkedUniversities listed a lot of useful vocabularies and ontologies on their website [22]. As an example for such an ontology let's take a look at the use case "'publication database"' from section 3.2.3. To mapping the data from this database a bibliographic ontology is needed. LinkedUniversities recommend the BIBO (Bibliographic Ontology) [23]. It can be used as a citation ontology, as a document classification ontology, or simply as a way to describe any kind of document in RDF, so it is ideally for this purpose. You can see the BIBO graph in figure 9.

[18] http://code.google.com/p/google-refine/
[19] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/
[20] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/ (at the moment of the work unavailable)
[21] http://any23.org/(at the moment of the work unavailable)
[22] http://linkeduniversities.org/lu/index.php/vocabularies/index.html
[23] http://bibliontology.com/

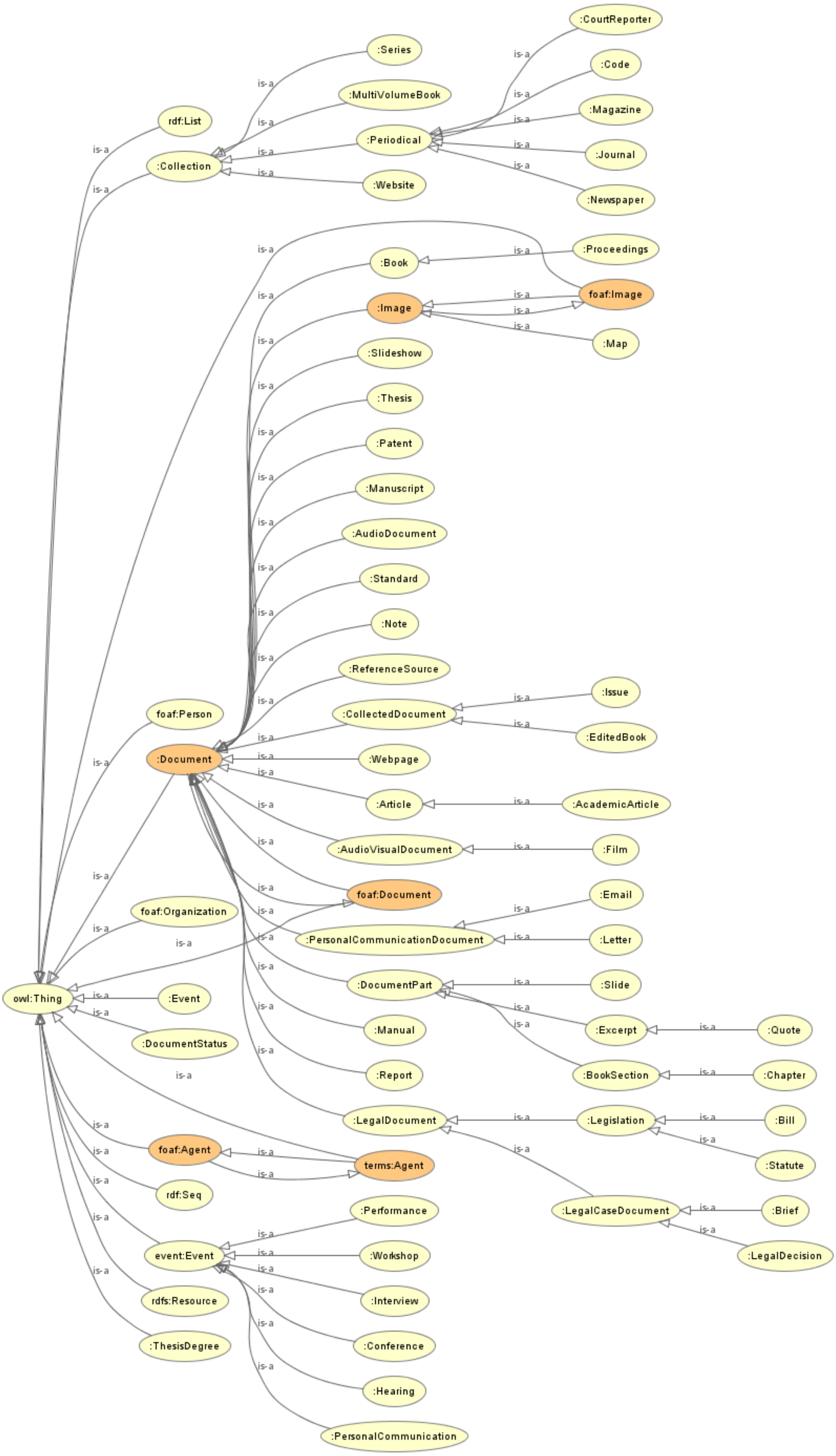


Figure 9: The BIBO graph

4.2.3 Store the data

Considering the performance of such a system and to avoid too much communication with the original data source, the center of the system is a triple store, where the extracted data are stored and from there exposed to the web. A triple store (or RDF store) is a purpose-built database, "‘*designed to store and retrieve identities that are constructed from triplex collections of strings (sequences of letters). These triplex collections represent a subject-predicate-object relationship that more or less corresponds to the definition put forth by the RDF standard.*"’ Rusher [11]. There are a lot of implemenations of such a triple store, like Sesame [24], Jena TDB [25], 4Store [26] or SwiftOWLIM [27]. The LUCERO project settled for the SwiftOWLIM, because it is "‘*free, scalable and efficient, and includes limited reasoning capabilities, which might end up being useful in the future*"’ d'Aquin et al. [4].

4.2.4 Expose the data to the Web

The last step is obviously to expose the data from the triple store to the web. For this purpose a SPARQL endpoint is commonly used, but there are also other way, e.g. resolvable URIs, the linked data API [28] or an Open RESTful API. All of them make the data available so others can query and access them to use it for their application or to serve themselves as resource for linking it with other data.

4.3 Challenges

4.3.1 Data ownership

4.3.2 Data quality

[24] `http://rdf4j.org/`

[25] `http://openjena.org/TDB/`

[26] `http://4store.org/`

[27] `http://www.ontotext.com/owlim/`

[28] `http://code.google.com/p/linked-data-api/`

5 Conclusions and future work

Revisit each research question and give a condensed answer derived from possibly multiple methods (e.g., benefits: as identified in interviews, as seen at other universities)

6 Acknowledgments

The authors would like to thank...

References

[1] Tim Berners-Lee. Linked data - design issues. *W3C*, (09/20), 2006. URL `http://www.w3.org/DesignIssues/LinkedData.html`.

[2] Conrad Bock, Peter Haase, Rinke Hoekstra, Ian Horrocks, Alan Ruttenberg, Uli Sattler, and Mike Smith. OWL 2 web ontology language structural specification and functional-style syntax. *W3C Recommendation, 2nd edn.(December 11, 2012)*, 2008.

[3] Mathieu d'Aquin, Fouad Zablith, Enrico Motta, Owen Stephens, Stuart Brown, Salman Elahi, and Richard Nurse. LUCERO - Aims, Objectives and Final Outputs of the Project. `http://lucero-project.info/lb/2010/06/lucero-aims-objectives-and-final-outputs-of-the-project/index.html`, 2010. [accessed 29-February-2016].

[4] Mathieu d'Aquin, Fouad Zablith, Enrico Motta, Owen Stephens, Stuart Brown, Salman Elahi, and Richard Nurse. The LUCERO Project: Tabloid. `http://lucero-project.info/lb/tabloid/index.html`, 2010. [accessed 07-March-2016].

[5] Mathieu d'Aquin, Carsten Kessler, and Tomi Kauppinen. Members of Linked Universities. `http://linkeduniversities.org/lu/index.php/members/index.html`, 2014. [accessed 24-February-2016].

[6] Mathieu d'Aquin, Carsten Kessler, and Tomi Kauppinen. Linked universities. `http://linkeduniversities.org/index.html`, 2014. [accessed 24-February-2016].

[7] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.

[8] Ian Jacobs and Norman Walsh. Architecture of the World Wide Web, Volume One. W3c:rec, W3C, 15 dec 2004. URL `http://www.w3.org/TR/2004/REC-webarch-20041215/`. http://www.w3.org/TR/2004/REC-webarch-20041215/.

[9] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4):258–268, 2012.

[10] James E. Post, Lee Preston, and Sybille Sachs. *Redefining the Corporation - Stakeholder Management and Organizational Wealth.* Stanford University Press, Stanford, new. edition, 2002. ISBN 978-0-804-74310-5.

[11] Jack Rusher. Triplestore, semantic web advanced development for europe (swad-europe), workshop on semantic web storage and retrieval - position papers. `https://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/rusher.html`, 2003. [accessed 07-March-2016].

[12] Fouad Zablith, Mathieu d'Aquin, Stuart Brown, and Liam Green-Hughes. Consuming Linked Data within a Large Educational Organization. Bonn, Germany, 2011.

.1 Questionaire

.2 Contact persons for further investigations