Publishing TUWien Data as Linked Open Data

Lukas Baronyai

March 4, 2016

Contents

1 Introduction **4**
1.1 Research Questions . . . 4
1.2 Methodology . . . 5
1.3 Contributions . . . 5
1.4 Structure of this Paper . . . 6

2 Related work **7**
2.1 Linked Universities . . . 7
2.1.1 Example: The Open University and the LUCERO Project . . . 7
2.2 Austrian Open Data . . . 7
2.3 Linked Data for Libraries (LD4L) . . . 7
2.4 Earlier studies . . . 7

3 Benefits and challenges of using LOD at TUWien **8**
3.1 Methodology . . . 8
3.1.1 Design of questionnaire . . . 8
3.1.2 Description of interviewed people . . . 8
3.1.3 Data Validity and Quality . . . 8
3.2 Results . . . 9
3.2.1 General statistically evaluation . . . 9
3.2.2 Library data . . . 9
3.2.3 Data from the Publication database . . . 11
3.2.4 Other data . . . 12

4 Proposed technical architecture **13**
4.1 The big picture . . . 13
4.2 Stakeholder specific issues (datasets, application types) . . . 13
4.3 (Technical) Challenges . . . 13

5 Conclusions and future work **14**

6 Acknowledgments **15**
Bibliography16

List of Figures

1 Level of Experience 9
2 Website of the TUVienna library 10
3 Usefulness of library (meta-)data 11
4 Publication database 11
5 Usefulness of publication data 12

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. [6] 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 [5]. 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 [3] 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 [3] 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:[4]

- "'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:[3]

- 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*"'.[2] 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.

2.2 Austrian Open Data

2.3 Linked Data for Libraries (LD4L)

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] [4]

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

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

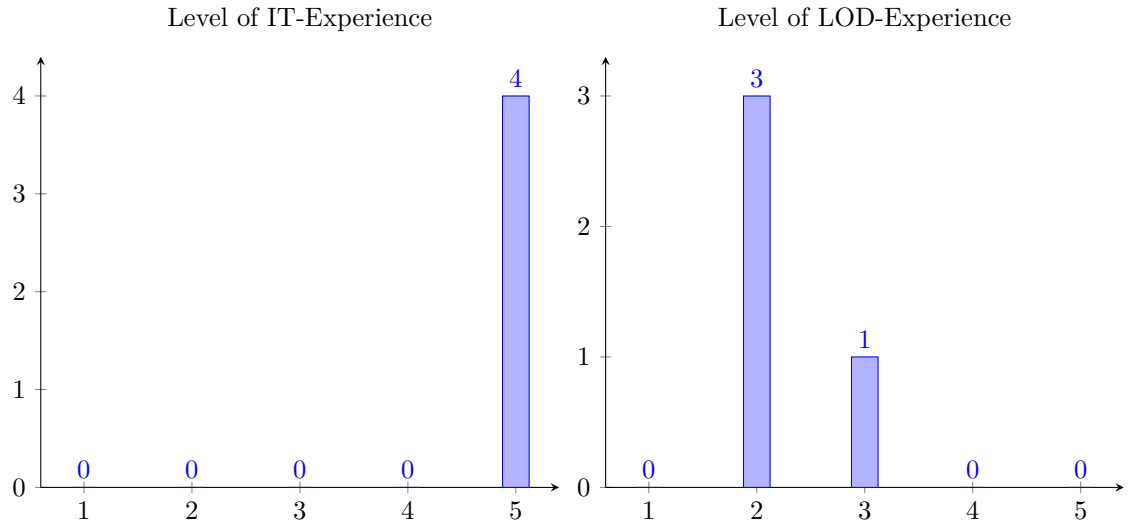


Figure 1: Level of Experience

3.2.2 Library data

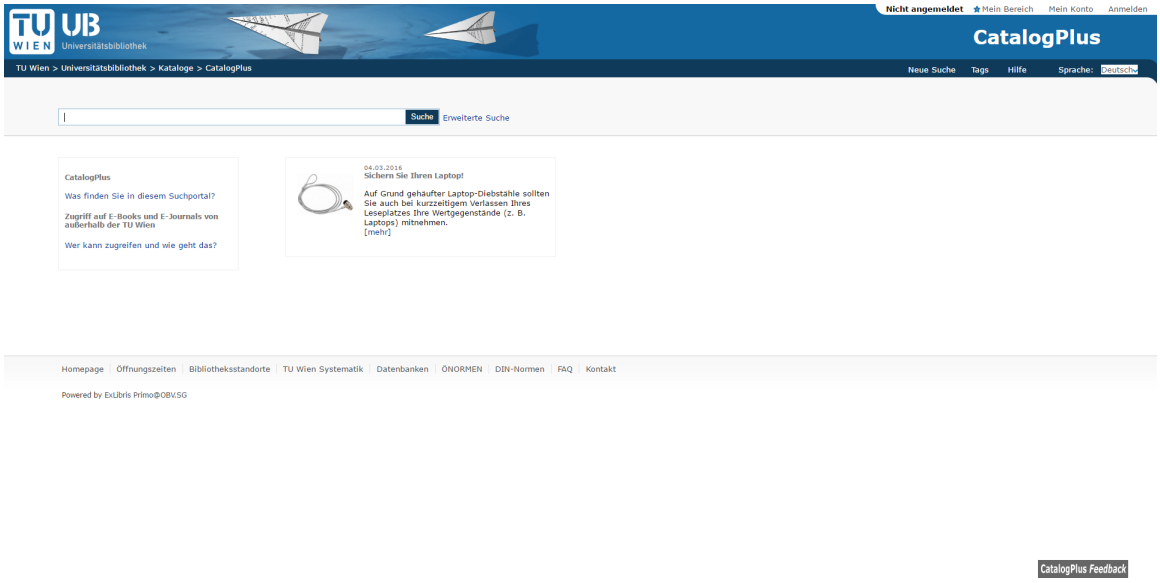
Use case This question aimed for a use case similar to the project Linked Data Libraries (described in Section 2.3). The proposed scenarios was to publish the data or meta-data of the university library (and all of its specialized libraries) as LOD and provide an application to access the data. A further 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.

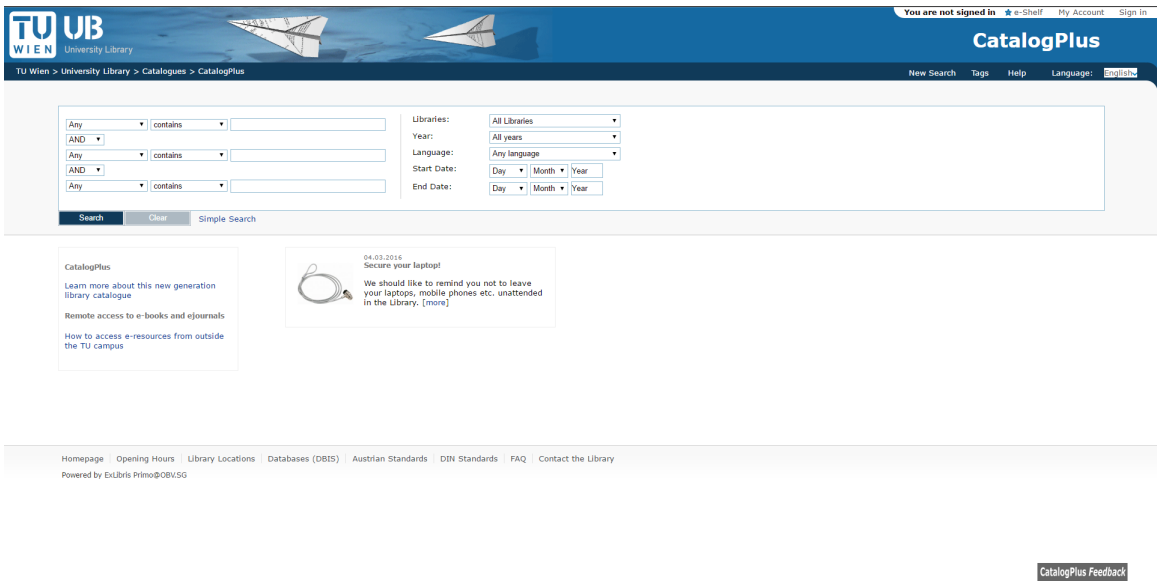
Needs, potential benefits As stated, all of the interviewed persons could imagine a similar project. One of the main reasons of the strong acceptance was the current interface of the library website [2], which only allows a search with only a few, specified parameters (see Figure 2). Also the physical 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.

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



(a) Homepage and simple search



(b) Extended search

Figure 2: Website of the TUVienna library

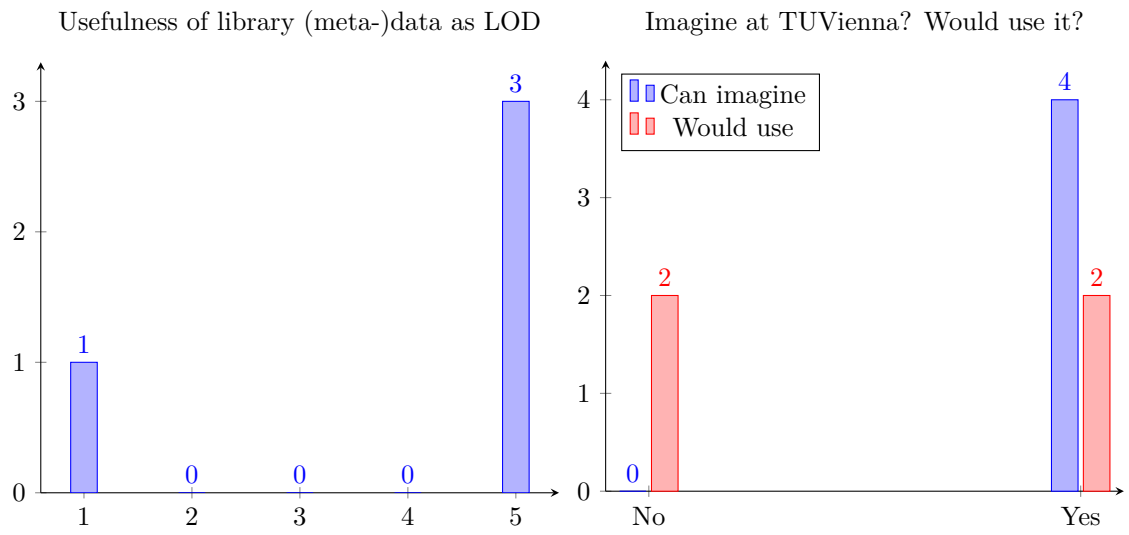


Figure 3: Usefulness of library (meta-)data as LOD

This would be an important part of the project, otherwise the data would loose there value - not 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

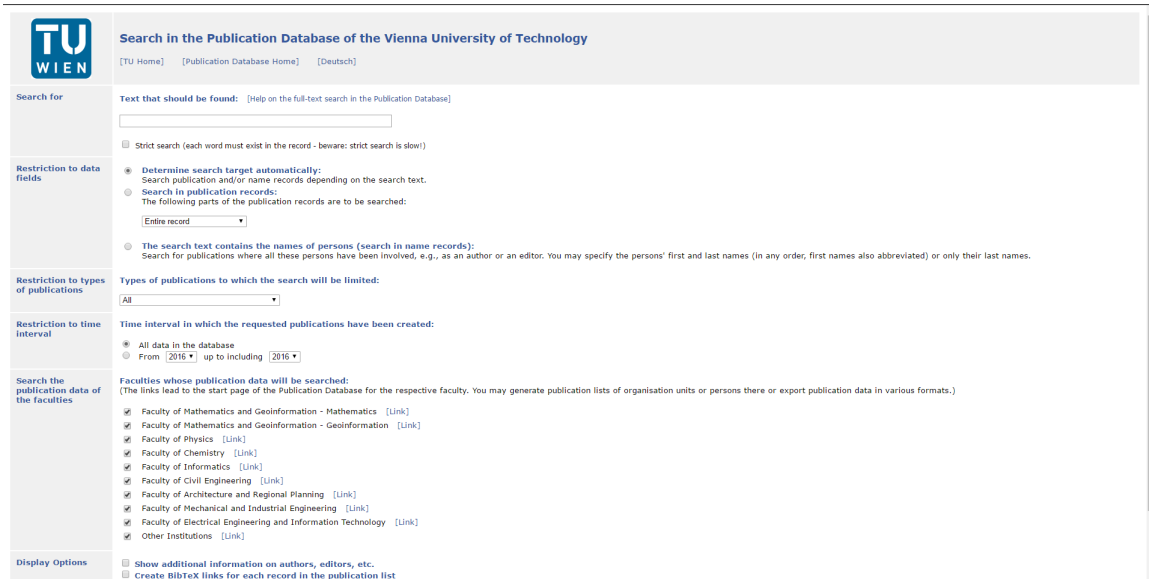


Figure 4: Publication database

Use case 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

[3] www.scholar.google.at/
[4] www.publik.tuwien.ac.at/pubstart.php

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.

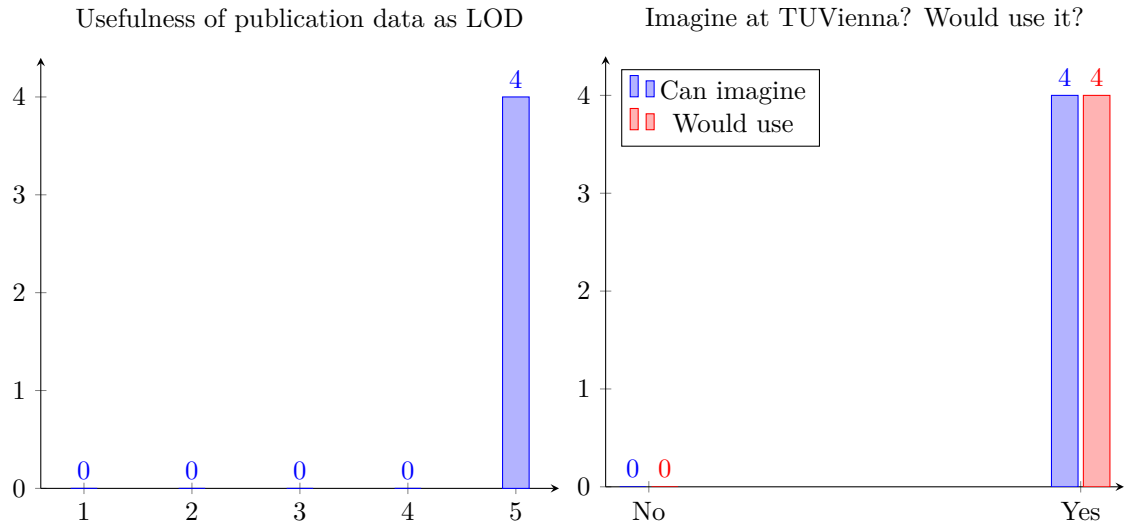


Figure 5: Usefulness of publication data as LOD

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 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

Needs, potential benefits, use-cases

Challenges

[5] `www.lod.springer.com/`

4 Proposed technical architecture

4.1 The big picture

4.2 Stakeholder specific issues (datasets, application types)

4.3 (Technical) Challenges

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

References

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

[2] 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].

[3] 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].

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

[5] 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/.

[6] 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.

Appendices

QUESTIONNAIRE
LINKED OPEN DATA

Name:_____

Organization:_____

Position (Role):_____

General Questions

1. **What are is your area of responsibility? How would you characterize your *daily* work tasks?**

_____

_____

_____

2. **How would you say classify your level of experience with Information Systems?**

Fundamental	Novice	Intermediate	Advanced	Expert
☐	☐	☐	☐	☐

3. **How would you classify your level of expertise with Linked Open Data?**

never heard	heard but never used	used in small example	used in practice	Expert in LOD
☐	☐	☐	☐	☐
1	2	3	4	5

LOD in Research

4. How useful do you find the improvements with Linked Data in libraries? E.g. in context of an use case of my presentation

not useful		useful		extremely useful
☐	☐	☐	☐	☐
1	2	3	4	5

Why?

_____

_____

5. Can you imagine a similar project at the Vienna University of Technology?

Yes ☐
No, because:

_____

_____

What benefits do you see?

_____

_____

What disadvantages or barriers do you see? E.g. copyright?

_____

_____

Would you use it?

Yes ☐
No, because:

Can you imagine/recommend other roles/persons?

6. Do you think that publishing the scientific publications or their metadata of TU as LOD might be useful?

not useful		useful		extremely useful
☐	☐	☐	☐	☐
1	2	3	4	5

Why?

_____

_____

7. Can you imagine a similar project at the Vienna University of Technology?

Yes ☐
No, because:

_____

_____

What benefits do you see?

_____

_____

What disadvantages or barriers do you see?

_____

_____

Would you use it?

Yes ☐
No, because:

_____

Can you imagine/recommend other roles/persons?

LOD Applications

8. Which kinds of applications based on LOD could be useful in a research context (e.g., single researcher, research workgroup, faculty, scientific community)?

What benefits do you see?

What disadvantages or barriers do you see?

Would you use it?

Yes ☐
No, because:

Can you imagine/recommend other roles/persons?

9. Which kinds of applications based on LOD could be useful in a general university context (e.g., teaching, administration, orientation)?

What benefits do you see?

What disadvantages or barriers do you see?

Would you use it?

Yes ☐
No, because:

Can you imagine/recommend other roles/persons?

LOD Data

10. Which kinds of data could be useful to publish as LOD in a research context (e.g., single researcher, research workgroup, faculty, scientific community)? Do you have any data that may be useful?

What benefits do you see?

What disadvantages or barriers do you see?

Would you use it?

Yes ☐
No, because:

Can you imagine/recommend other roles/persons?

11. **Which kinds of data could be useful to publish as LOD in a general university context (e.g., teaching, administration, orientation)?** Do you have any data that may be useful?

_____

_____

_____

_____

What benefits do you see?

_____

_____

What disadvantages or barriers do you see?

_____

_____

Would you use it?

Yes ☐
No, because:

_____

_____

Can you imagine/recommend other roles/persons?

_____

_____

End

12. Do you know any other person who might be relevant for this interview?

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

13. Do you want the final report of the outcome of the study?

Yes	☐
No	☐