Qualitative and Quantitative Methods in Libraries (QQML) 7: 595-606, 2018

Crafting Linked Open Data to Enhance the Discoverability of Institutional Repositories on the Web

Qiang Jin[1] and Jane Sandberg[2]

[1] Qiang Jin, Authority Control Team Leader, Associate Professor, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[2] Jane Sandberg, Electronic Resources Librarian, Linn-Benton Community College, Albany, Oregon, USA

Abstract: Institutional repositories are archives for collecting and disseminating digital copies of the intellectual output of institutions. Linked open data is to expose and connect pieces of data, information, and knowledge on the Semantic Web. This paper studies how BIBFRAME 2.0 can be used to describe objects in institutional repositories, with the goal of bringing together efforts within two communities devoted to openness. We examine a sample of mappings and conversions from Dublin Core to BIBRAME 2.0 ontology to see if BIBFRAME 2.0 will increase visibility of local digital collections on the Web.

Keywords: open standards, open access, institutional repositories, linked data, metadata crosswalks, BIBFRAME 2.0

1. Introduction

Institutional repositories are archives for collecting and disseminating digital copies of the intellectual output of institutions. At the University of Illinois at Urbana-Champaign, the institutional repository is named the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS). Like many institutional repositories, IDEALS collects the digital copies of research output by the faculty, staff, and students at the University of Illinois at Urbana-Champaign.

Linked data is a method of publishing structured data so that it can be interlinked. It uses the Web to connect pieces of data, information, and knowledge on the Semantic Web using URIs and RDF. (1) According to Berners-Lee, linked open data (LOD) is linked data which is released under an open license, which does not impede its reuse for free. (2)

Received: 31.7.2017 Accepted: 21.11.2017 ISSN 2241-1925

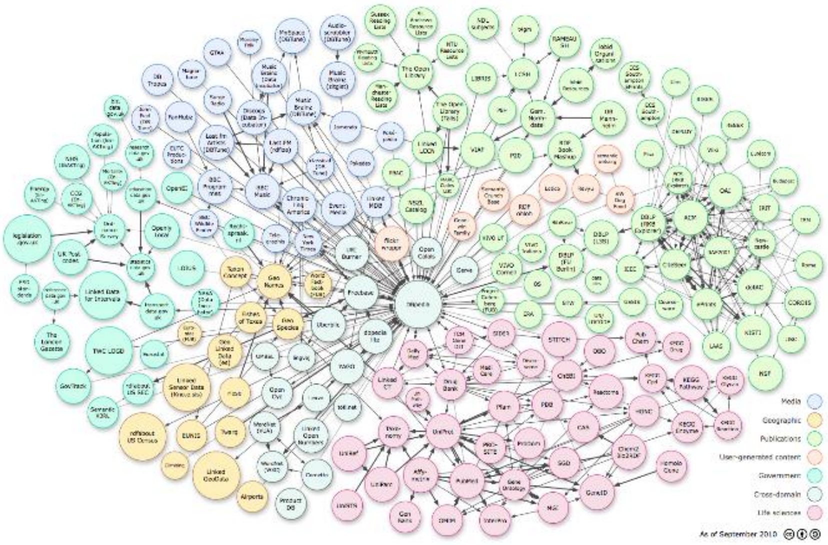


Figure 1. Linking Open Data Cloud Diagram (3)

In some ways, the missions of institutional repositories and linked data are very much linked -- both value making information more accessible to users outside an immediate college, university, or library context. Some institutional repositories, particularly those based on Fedora, are already making linked data fundamental to their metadata.

One way to bring cataloging linked data efforts and institutional repository linked data effort into better alignment is to use the same linked data ontologies, such as BIBFRAME 2.0. (4) BIBFRAME (Bibliographic Framework) is a data model for bibliographic description. It was designed to replace the MARC standards, which can accommodate a broader user community such as museum, archives, and publishers to use linked data principles to make bibliographic data more useful both within and outside the library community. BIBFRAME and linked library data enable libraries to publish bibliographic resources in a way that the Web understands, so that users from various communities will be able to find them during their first searches on the Web. However, there has been little research on how effective BIBFRAME 2.0 might be in describing objects in institutional repositories.

Like many libraries, catalogers at the University of Illinois at Urbana-Champaign provide double descriptions for theses and dissertations -- a MARC one for the online library catalog, and a Dublin Core (5) one for the institutional repository. This makes it particularly appealing to use a shared description for both systems. If we could do just one BIBFRAME 2.0 *Work* description with two linked *items*, it would not only save staff time, but would also help display our local data on the Web. Structured data like BIBFRAME 2.0 can be interlinked and become more useful through semantic queries.

2. Bibframe 2.0

Many libraries in the United States including Library of Congress (LC), and some foreign national libraries tested BIBFRAME between 2013 and 2015. In April 2016, LC published the BIBFRAME 2.0 model. Afterwards, LC issued a call encouraging libraries to test the BIBFRAME 2.0 model. We answered LC's call by conducting this research with the hope to contribute to the transformation from Dublin Core to BIBFRAME 2.0 ontology to find out if BIBFRAME 2.0 will increase visibility of local digital collections on the Web.

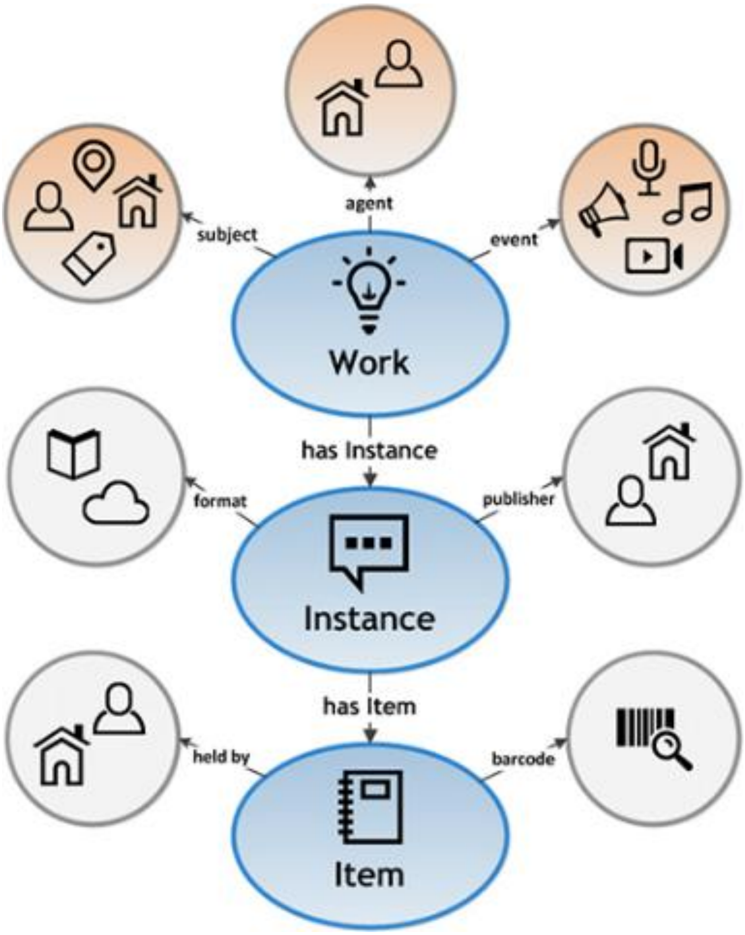


Figure 2. The BIBFRAME 2.0 Model (6)

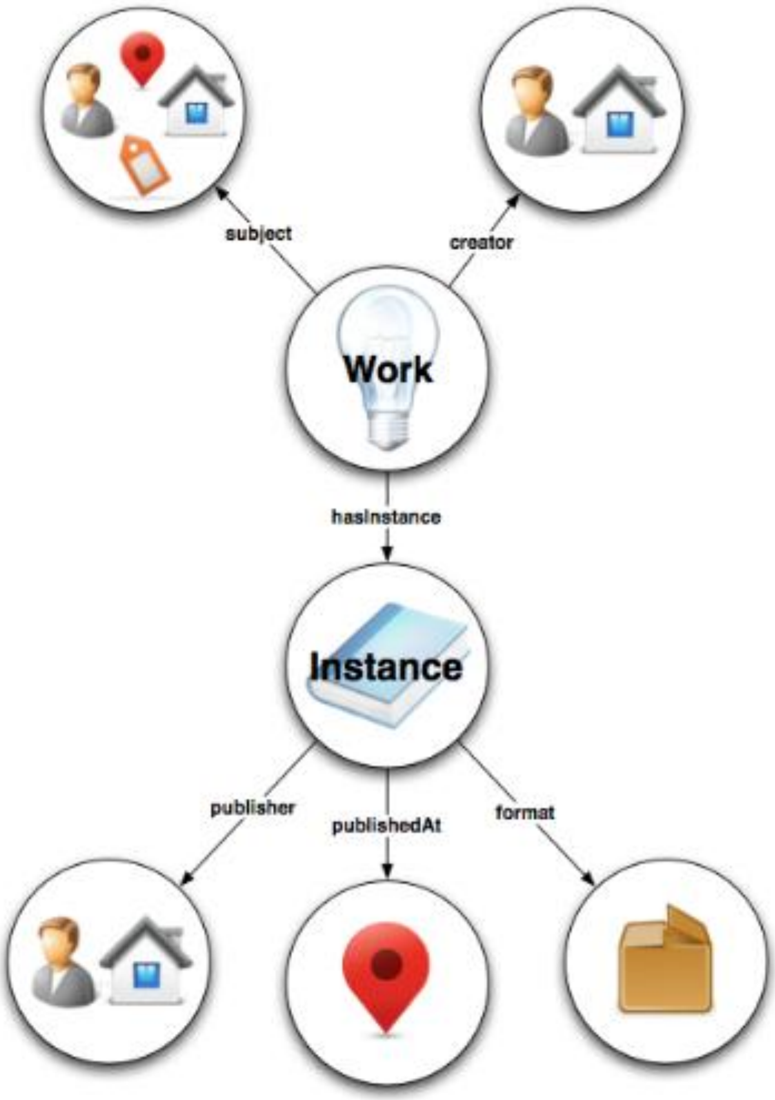


Figure 3. The BIBFRAME 1.0 Model (7)

The BIBFRAME 2.0 model descends from the BIBFRAME 1.0 model, but changes it significantly. BIBFRAME 2.0 focuses on three core levels of abstraction: *work*, *instance*, and *item*. (8) The BIBFRAME entity *work* is the highest level of abstraction. (9) The BIBFRAME entity *instance* is an individual, material embodiment of a work. (10) The BIBFRAME entity *item* is an actual copy (physical or electronic) of an *instance*. (11) BIBFRAME 2.0

introduces the entity *item*, but eliminates Authority class. Subject and event entities can be related to a *work* entities, and Agent entities can be related to work, instance, or item entities. Persons, organizations, places (things with names) are represented as real world objects rather than identified by name. (12) It is designed to integrate with and engage in the wider information community and still serve the very specific needs of libraries. (13)

This paper studies how to provide enhanced discovery of similar sets of content in the institutional repository described based on Dublin Core with the BIBFRAME 2.0 model, and enrich the BIBFRAME 2.0 model with linked data that connected to other open linked data projects with the goal of bringing together efforts within two communities devoted to openness. We are mapping a collection of our senior theses in IDEALS from Dublin Core to BIBFRAME 2.0 to test if BIBFRAME 2.0 could help increase the Web visibility of our senior theses. Few researchers have tested Dublin Core against BIBFRAME 2.0, so our research is unique.

3. Bibframe transformation and the linked data enrichment process

We selected a collection of 459 senior engineering theses in the University of Illinois at Urbana-Champaign's institutional repository, IDEALS. In order to retrieve the metadata for these 459 theses, we created a python script, which fetched the Dublin Core from the HTML in IDEALS. We then crosswalked and mapped Dublin Core metadata terms to MARC. We also enriched the data with links to linked data authorities via LC Linked Data Service https://id.loc.gov/, which was accessible for human and machines processing. This BIBFRAME transformation and linked data enrichment process involved several steps. First, we used an Extensible Stylesheet Language Transformations (XSLT) stylesheet based on LC's DC2MARC21slim.xsl stylesheet to convert the Dublin Core data to MARC. Here are a few changes we made:

- Our stylesheet processed a file containing several Dublin Core records, rather than just one record per file.
- Our stylesheet used a more specific 264 statement as a mapping for dcterms:issued and dcterms:available.
- Our stylesheet put subjects in 650 fields, rather than 653 fields, so that they can be linked to linked data sources in the next step.
- Our stylesheet added the specific Library of Congress Genre/Form Terms (lcgft) Academic theses in 655.
- Our stylesheet added dcterms:abstract to the 520.
- Our stylesheet added language codes to the 041.
- Our stylesheet added 040$e local.
- Our stylesheet added 33x for RDA.
- Our stylesheet added degree, institution, department, and discipline info.

After running this conversion, we used MarcEdit to add linked data URIs in a manner similar to the method documented by Shieh and Reese (2015). (14) For any 650s that did not receive links – indicating that they were not authorized LCSH terms – we moved them to 653 fields. Finally, we used the Library of Congress BIBFRAME Converter software to create BIBFRAME 2.0 records.

Authority Modeling

BIBFRAME 2.0 authority is associated with a *work* or *instance* through roles such as author, editor, artist, photographer, composer, illustrator, etc. We searched through the Virtual International Authority File (VIAF), Open Researcher and Contributor ID (ORCID), and International Standard Name Identifier (ISNI), and we were not able to find any of our authors. We believed that it might be because these theses were unpublished, so no authority records had been created for them by the University of Illinois Library catalogers. It might also be that authors of these 2017 senior theses had not written any articles or books yet to have been included in these databases. To complicate searching for these authors, all of the author names were undifferentiated by dates or other potentially identifying characteristics.

BIBFRAME 2.0 subject is a work that might be about one or more concepts. Such a concept is to be subject of the work. Concepts may include topics, places, temporal expressions, events, works, instances, items, agents, etc. (15) We chose to link subject headings to LC Linked Data Service, which provided URIs for many LCSH in our theses collection.

An example linking to LCSH to https://id.loc.gov/

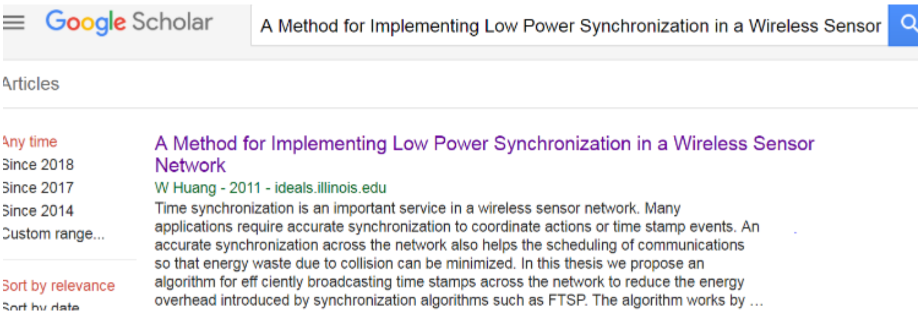
```
 <bf:Topic>
<bf:Topic:rdf:about="">
  <bf:authorizedAccessPoint>wireless                              sensor
networks</bfLauthorizedAcessPoint>
  <bf:label>wireless sensor networks</bf:label>
  <bf:hasAuthority>
  <madsrdf:Authority>
  <rdf:type                                                         rdf:
   resource=”http://id.loc.gov/authorities/subjects/sh2008004547">
       <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Topic"/>
       <networks</madsrdf:authoritativeLabel>
<madsrdf:Authority>
</bf:hasAuthority
```

Work Modeling

To locate a *Work* identifier for these theses, we considered several sources, such as OCLC, the Internet Archives, and Open Library since they are all open source resources, but we were not able to find any of our theses in these databases. We finally found our theses in Google Scholar. Therefore, we chose Google Scholar link as our *Work* identifier as it was an excellent way for users to find our theses on the Web. Google Scholar is a freely accessible Web search engine that indexes the full text or metadata of scholarly literature across an array of publishing formats and disciplines, which meets our goal.

An example link to a Google Scholar as *Work* identifier and an image below

<bf:Work rdf: about=
“https://scholar.google.com/scholar?hl=en&as_sdt=0%2C14&q=A+Method+for+Implementing+Low+Power+Synchronization+in+a+Wireless+Sensor+Network+&btnG=”/>



A Method for Implementing Low Power Synchronization in a Wireless Sensor Network

W Huang - 2011 - ideals.illinois.edu

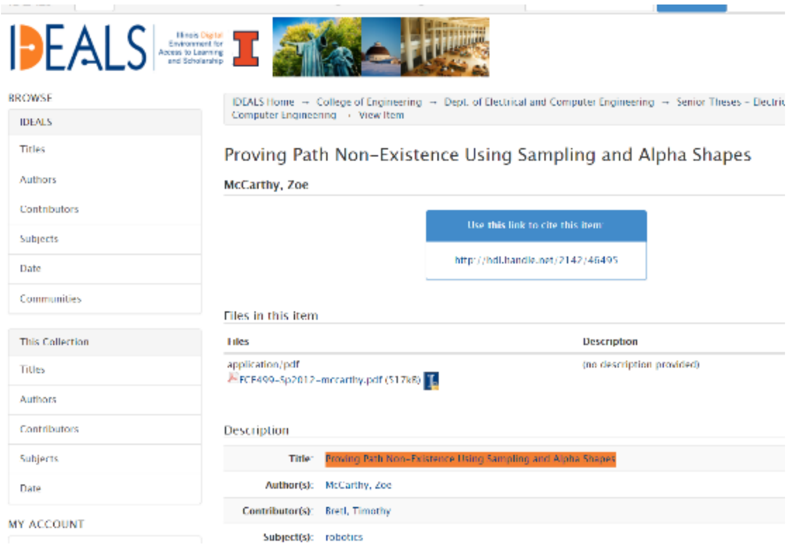
Time synchronization is an important service in a wireless sensor network. Many applications require accurate synchronization to coordinate actions or time stamp events. An accurate synchronization across the network also helps the scheduling of communications so that energy waste due to collision can be minimized. In this thesis we propose an algorithm for eff ciently broadcasting time stamps across the network to reduce the energy overhead introduced by synchronization algorithms such as FTSP. The algorithm works by …

Instance Modeling

For BIBFRAME 2.0 *Instance* identifier, we chose IDEALS’s link to a thesis. Our *instance* is expressed by its relationship with *work* via the properties bf:hasInstance and bf:instanceOF.

An example link to IDEALS as *Instance* identifier and an image below

<bf:hasInstance
rdf:resource=“https://www.ideals.illinois.edu/handle/2142/46495”/>



IDEALS Home → College of Engineering → Dept. of Electrical and Computer Engineering → Senior Theses – Electric
Computer Engineering → View Item

Proving Path Non-Existence Using Sampling and Alpha Shapes

McCarthy, Zoe

Use this link to cite this item:

http://hdl.handle.net/2142/46495

Files in this item

Files	Description
application/pdf ECE499-Sp2012-mccarthy.pdf (517kB)	(no description provided)

Description

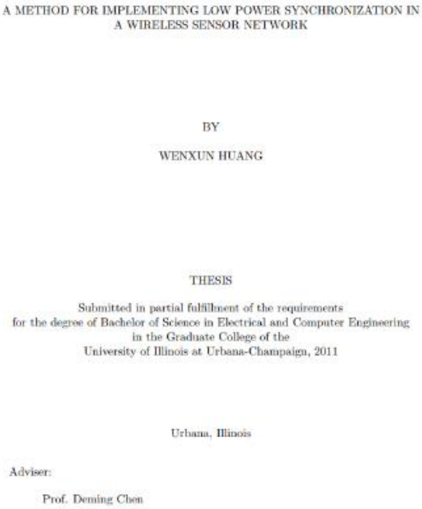
Title:	Proving Path Non-Existence Using Sampling and Alpha Shapes
Author(s):	McCarthy, Zoe
Contributor(s):	Bretl, Timothy
Subject(s):	robotics

Item Modeling

BIBFRAME 2.0 introduces the entity *item*, which was represented as annotation in BIBFRAME 1.0. (16) To locate BIBFRAME 2.0 *Item* identifier for these theses, we decided to use the pdf link to a thesis.

An example link to pdf of the thesis as *Item* identifier and an image below

<bf:hasItem
rdf:resource="https://www.ideals.illinois.edu/bitstream/handle/2142/46495/ECE499-Sp2012-mccarthy.pdf?sequence=2&isAllowed=y"/>



A METHOD FOR IMPLEMENTING LOW POWER SYNCHRONIZATION IN
A WIRELESS SENSOR NETWORK

BY

WENXUN HUANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Adviser:

Prof. Deming Chen

4. DISCUSSSION

BIBFRAME Transformation and the Linked Data Enrichment Process

We have learned several lessons during the transformation of a collection of digital theses in our institutional repository from Dublin Core to BIBFRAME 2.0, and linked data enrichment process.

[1] Lack of Authority Data for Authors

At the University of Illinois at Urbana-Champaign, senior theses had been kept by academic departments for decades. For the last several years, the library started collecting them, describing them using Dublin Core, and depositing them to IDEALS. Like many other libraries, catalogers at the University of Illinois at Urbana-Champaign do not create authority records for authors in its institutional repository. As a result, we were not able to find any links for our authors to the Virtual International Authority File (VIAF) or other authority databases such as ISNI. It will be harder for users to find these theses because of the lack of authority control. Users will find it very difficult to find theses for authors with non-English names, whose names have changed, or have undifferentiated names. We believe that providing consistency in the form of access points used to identify authors is one of key issues for linked data. Efforts such as opaque namespace at Oregon State University and University of Oregon is a great step forward in this work, since it creates URIs and simple RDF authority records for people not represented in the LC Name Authority Files. (17) Those URIs are linked to repository descriptions, and eventually could be conceivably linked to BIBFRAME descriptions too.

[2] Issue with *Work* identifier

Another issue we encountered was that these senior theses were not part of OCLC, Internet Archives, or Open Library. To locate a proper *Work* identifier for these theses, we considered several options before we decided to use Google Scholar https://scholar.google.com/, an open source. Google Scholar is part of a large online database connecting works across the world of scholarly research.

[3] Low Metadata Quality

Our Dublin Core records in IDEALS were of low metadata quality because we do not have adequate staff members to devote to the work. Currently, one metadata librarian and a graduate student are responsible for IDEALS. As a result, it is not possible for us to follow the standards and create quality metadata. Problematic metadata in Dublin Core or BIBFRAME affects search results, which leads to the invisibility of resources with a digital repository.

[4] Issues with transformation of Dublin Core to BIBFRAME 2.0

From the experience of a previous project mapping BIBFRAME 1.0 to MARC records, we believe that BIBFRAME 2.0 should map rather well with MARC records. BIBFRAME 2.0 does not map well with non-MARC records. We believe that using BIBFRAME 2.0 in a mix of other ontologies might be a good way to go. This is an approach that is becoming increasingly common in the Samvera Institutional Repository community: using BIBFRAME as one of many ontologies that are useful in describing repository contents. This might have also helped us with mapping some terms where it is hard to find a specific BIBFRAME attribute, like “Department” and “Discipline.”

Another problem we had was if we wanted to add non-BIBFRAME terms, it made the conversion more complicated, since we could not just accept the output of the BIBFRAME Transformation Service.

To summarize, we extracted a collection of 459 senior theses from our institutional repository, IDEALS. We converted its Dublin Core to MARCXML using MarcEdit. We then added links to the Library of Congress subject headings <http://id.loc.gov/authorities/subjects.html>

Next, we used the Library of Congress MARCXML to BIBFRAME Transformation software <http://id.loc.gov/authorities/subjects.html> to convert our MARCXML to BIBFRAME 2.0. We added links to entities *work*, *instance*, and *item*, so that they were connected.

We believe our work in enriching data for our senior theses is very instructive for future projects in the University of Illinois at Urbana-Champaign Library, and applies to library data work across institutions. We have learned new skills throughout the challenging work of converting and mapping from Dublin Core to MARCXML to BIBFRAME 2.0, and linked data enrichment process. We have also learned that it takes much longer than you expect to transform a digital collection from Dublin Core to BIBFRAME 2.0, and add links to enrich data because of various issues we encountered discussed above.

Finally, our next step is to design a search interface for the newly generated BIBFRAME 2.0 metadata. Once we have finished this interface, we can assess the data's utility in supporting discovery success, display understanding, ability of repository metadata to operate on the open Web, and ability to support FRBR user tasks.

5. Conclusion

Our study is based on the idea that openness can bring people together. An open standard like BIBFRAME 2.0 can be the spark that brings together catalogers, metadata librarians, software developers, scholarly communications librarians, repository librarians, and others to design better search and discovery experiences for our users. Our project suggests that the BIBFRAME 2.0 model and vocabulary are suitable for the basic modeling of digital theses in our institutional repository. Therefore, we have contributed to the evaluation of the BIBFRAME 2.0 model related to institutional repositories.

The cataloging world is in transition. BIBFRAME 2.0 is a profound step for the library community, which uses linked data to make discoverable library bibliographic resources on the Web. We believe that BIBFRAME 2.0 will help our users discover library resources across the Web and beyond the classic catalog paradigm.

References

1. Linked Data, Wikipedia, accessed April 26, 2018
https://en.wikipedia.org/wiki/Linked_data

2. Tim Berners-Lee, Linked Data
https://www.w3.org/DesignIssues/LinkedData.html

3. The Linking Open Data Cloud Diagram, accessed April 26, 2018
https://www.google.com/search?q=linked+data+diagram&tbm=isch&source=iu&ictx=1&fir=t-26qPFvugpHIM%253A%252CyZJ0Ehmo6HpNUM%252C_&usg=__m_RtewUFopCyz6JmBp05-xVjSsY%3D&sa=X&ved=0ahUKEwiL_MeT1tDaAhUr74MKHbzfCQUQ9QEIMTAD#imgrc=65OXV6pAizxcuM

4. Overview of the BIBFRAME 2.0 Model, Library of Congress, accessed April 26, 2018,
https://www.loc.gov/bibframe/docs/bibframe2-model.html

5. Dublin Core, Wikipedia, accessed April 26, 2018 https://en.wikipedia.org/wiki/Dublin_Core

6. Figure 2. The BIBFRAME 2.0 Model, Overview of the BIBFRAME 2.0 Model, Library of Congress, accessed April 26, 2018, https://www.loc.gov/bibframe/docs/bibframe2-model.html

7. Figure 3. The BIBFRAME 1.0 Model, Library of Congress, "Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services," accessed April 26, 2018 www.loc.gov/bibframe/pdf/marcld-report-11- 21-2012.pdf

8. Overview of the BIBFRAME 2.0 Model, Library of Congress, accessed April 26, 2018, https://www.loc.gov/bibframe/docs/bibframe2-model.html

9. Ibid.

10. Ibid.

11. Ibid.

12. What's New in BIBFRAME 2.0, accessed April 26, 2018 https://www.loc.gov/bibframe/docs/bibframe2-whatsnew.html

13. BIBFRAME Frequently Asked Questions, Library of Congress, accessed April 26, 2018, http://www.loc.gov/bibframe/faqs/#q01

14. Shieh, J. and Reese, T., (2015). The importance of identifiers in the new web environment and using the Uniform Resource Identifier (URI) in subfield zero ($0): a small step that is actually a big step, *Journal of library metadata*, Vol.15, No. 3/4, 208-226.

15. Overview of the BIBFRAME 2.0 Model, Library of Congress, accessed April 26, 2018, https://www.loc.gov/bibframe/docs/bibframe2-model.html

16. What's new in BIBFRAME 2.0, Library of Congress, accessed April 26, 2018, https://www.loc.gov/bibframe/docs/bibframe2-whatsnew.html

17. Simic, J. and Seymore, S., (2016). From Silos to Opaquenamespace: Oregon Digital's Migration to Linked Open Data in Hydra, Art Documentation: *Journal of the Art Libraries Society of North America*, Vol. 35, No. 2, pp. 305-320.