

# Linked Data Services for Theses and Dissertations

**Thomas Johnson**

Oregon State University  
thomas.johnson@oregonstate.edu

**Michael Boock**

Oregon State University  
michael.boock@oregonstate.edu

## ABSTRACT

Linked Data presents new opportunities to expand services surrounding theses and dissertations. By creating open datasets, ETD systems can be built to interoperate with other institutional data as well as with outside metadata sources. However, much foundational work must be done before these advantages can be fully realized.

This paper details work at Oregon State University to create a Linked Dataset covering the University's theses and dissertations. Using data from existing MARC and Qualified Dublin Core records, we have established a process and model for crosswalking data from existing records into a variety of Semantic Web vocabularies. Our approach is to create basic services on a dedicated thesis and dissertation interface, incrementally extending those available through our institutional repository. We describe services implemented, those in progress and plans for continued work. We also address the limitations of our existing metadata and resulting challenges in crosswalking and interoperability.

While Linked Data has great promise, implementation must target specific services that can be implemented today. We plan continued work to improve our data models and to utilize new data from other linked data sources as they emerge.

## Keywords

metadata, linked data, SKOS, RDF, Dublin Core, authority control, discovery

## INTRODUCTION

Linked Data has been a growing topic of discussion in libraries in recent years. Recent reports from the Working Group of the Future of Bibliographic Control (2011), the W3C Library Linked Data (LLD) Incubator Group (Baker et al. 2011) indicate increasing recognition of the value this approach can have for libraries. This "sharable, extensible, and reusable" way of expressing data improves the utility and interoperability of metadata, and connects our information more effectively to other web resources and services (Baker et al., 2011).

The key mechanism of Linked Data is the use of HTTP Uniform Resource identifiers (URIs) to identify non-web things and concepts in metadata. URIs provide a globally unique way of referencing a specific concept and a standard mechanism for finding information about it (by visiting the URI). In conjunction with semantic web standards for representing (RDF) and querying (SPARQL) data, this idea forms the basis for a universally understandable "web of data".

In libraries, progress joining that "web of data" has been strongest in the area of developing element sets and value vocabularies for expressing library concepts as Linked Data. The proliferation of vocabularies opens doors for the creation of rich, interoperable datasets for

library resources. However, as noted by the LLD Incubator Group, “[r]elatively few bibliographic datasets have been made available as Linked Data, and even less metadata has been produced for journal articles, citations, or circulation data” (Baker et al., 2011). Yet, it is this category of data that drives existing library services.

With this in mind, work on Linked Data at Oregon State University (OSU) Libraries has focused on building end-user services, improving workflows and solving metadata problems based on datasets generated from existing metadata. By building tangible services, we create a strong impetus for maintaining and expanding datasets in practical ways, targeting our institutional needs. In the process we produce an open dataset that is naturally situated as a part of the growing LLD cluster.

This paper describes our ongoing work to apply this service-centric approach to thesis and dissertation metadata as an initial Linked Data initiative for OSU Libraries.

### **THESES AND DISSERTATIONS AT OSU**

Beginning in July 2005, individual colleges at Oregon State University began to require the deposit of theses and dissertations to the ScholarsArchive@OSU (then DSpace@OSU) open access repository. By January 2007, all graduate students were required to deposit their theses to the Electronic Theses and Dissertations (ETD) Collection in the repository. As of August 2012, the university has produced a total of 22,942 theses and dissertations associated with a final degree. Since the mandate in 2007, students have deposited over 2,500 theses and dissertations as part of their degree requirements.

As part of the ScholarsArchive@OSU deposit process, the author is responsible for assigning the bulk of the metadata associated with the thesis or dissertation. The student author enters author, title, abstract, committee members, advisor, degree name, degree level, college, copyright date, graduation date and keywords metadata. A total of 24 descriptive and administrative metadata elements, most of which are recommended by the Networked Digital Library of Theses and Dissertations ETD metadata standard (ETD-MS), are assigned to incoming theses and dissertations as part of the deposit process.

After the Graduate School approves a thesis or dissertation, a library technician reviews the student-submitted metadata, makes revisions, and “uses the MarcEdit cataloging utility to map the DSpace Qualified Dublin Core metadata to MARC and export the MARC metadata to the library online catalog and to OCLC WorldCat.” [Boock/Kunda] This process substantially improves the efficiency with which catalog records are created for theses and dissertations.

Shortly after the mandate was approved, OSU Libraries began a large-scale digitization project to scan all of the University’s print theses from prior to 2007, accounting for roughly 20,000 theses and dissertations. Aside from a couple hundred theses and dissertations deposited by students from individual colleges between July 2005 and December 2006, the Libraries had two physical copies of every thesis and dissertation ever produced prior to the mandate; one circulating copy on the shelves and one in remote storage, available for digitization. The same data dictionary that is used for new items deposited by graduate students is used for the digitized theses. The library has MARC records for every thesis ever produced at the University. As part of the digitization workflow, student scanners copy and paste information from the MARC record in the online catalog to a DSpace Dublin Core record.

An initial motivation for creating a Linked Dataset for our theses and dissertations was to convert metadata from the disparate MARC and Dublin Core records into a single data model. Since each record is mapped directly to the RDF model, it is possible to automatically crosswalk data from each source to the other. We expect to further streamline the ETD workflows described above based on this automation.

### AN ETD DATA MODEL

To convert MARC and Qualified Dublin Core metadata into linked data, we established an RDF data model formalizing the expression of key data points for theses and dissertations. In keeping with practice recommended by Bizer, Cyganiak, and Heath (2007), the terms used are selected from well known used vocabularies. Where possible, we've chosen to use general purpose vocabularies relying on terms specific to bibliographic and library data where they add meaning. An outline is presented in Figure 1; see the example data in the appendix for a more complete treatment.

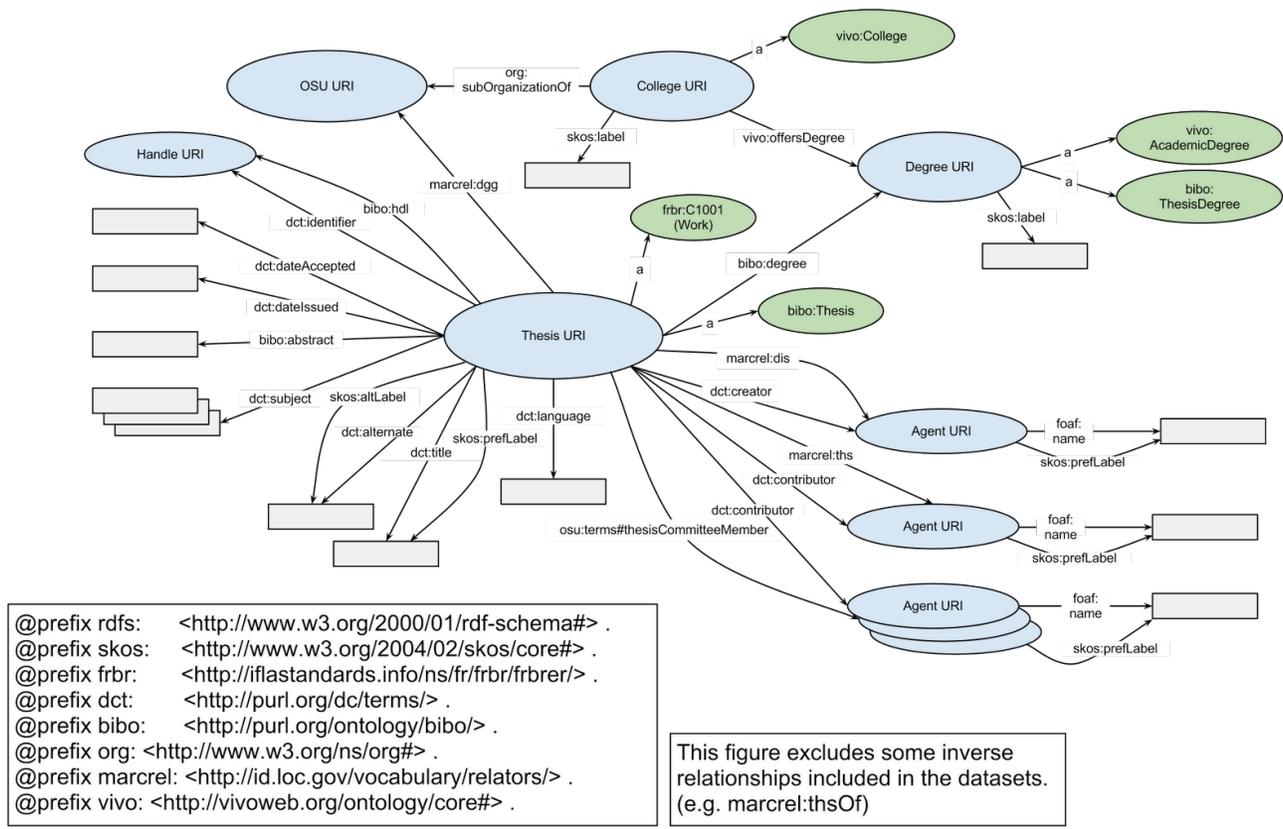


Figure 1. The basic thesis data model

This model is not conceived as either static or complete. Rather, it is designed as a flexible foundation we expect to build on as new needs and community practices emerge. Linked Data offers a fundamentally extensible, "open world" view of metadata. This allows not only for linking with external data, but also for continuous expansion when new data can be added. In contrast to the larger "open world" graph, our model can be seen as a combination of what was achievable given the limitations of existing ETD metadata at OSU and what was needed to support the "closed world" requirements of our target services.

In terms of data richness, the primary advantage of the linked data model over our Dublin Core and MARC records is the representation of people, academic departments, and degrees as independent resources. These concepts were previously represented either as flat text entries or, in the case of some names in MARC records, as name headings. They are now assigned URIs and can be the subject of metadata statements in their own right.

Degrees, for example, have metadata specifying their type, labels, degree level, and relationships to colleges. What was previously represented as a string (e.g. “Master of Science (M.S.) in Electrical and Computer Engineering”) is now a complex data object.

The theses themselves are specified as “works” within the FRBR entity-relationship model. The dataset, therefore, deals specifically with the abstract concept of the thesis, rather than any particular expression or document. The decision to invoke the FRBR model is based on analysis by Rochkind (2012), work by Westrum, Rekkavik and Tallerås (2012), and our desire to keep our data extensible.

Future work on linked data theses should address gaps in the available vocabularies. We weren’t able, for instance, to identify an existing term for committee members. To express those relationships, we defined a term in our own namespace (osuterm:thesisCommitteeMember). This is a suitable solution to support our applications, but it isn’t ideal since this term is not widely known or used and not part of a larger vocabulary; other datasets and applications will be less likely to use or understand its meaning.

## **PUBLISHING AND MAINTAINING THE DATASET**

### **Crosswalking**

The process of converting existing data into the RDF data model is handled by a series of Python scripts. The two core scripts work with MARC and Dublin Core data respectively, reading each thesis record in turn. If an existing URI is found, the metadata in the record is processed and written as RDF, overwriting existing data where appropriate. If there is no match, a new thesis URI is created along with accompanying data. These scripts work safely against the existing dataset, allowing easy updates when data in the catalog or repository changes.

Separate scripts manage URI assignment for resources other than the theses themselves. These are run prior to the main thesis scripts, allowing the MARC and Dublin Core crosswalks to add appropriate relationships between theses and people, departments, and degrees.

Using a process based on custom scripts rather than a one-to-one mapping allows us to adjust more flexibly to the specifics of our metadata implementations. Information implicit in local cataloging but not explicit in formal term definitions can be included to the extent that the script maintainer is aware of it. This is especially important in the case of Qualified Dublin Core records, where the use of qualifiers has been shaped by the needs of the repository software.

We are aware of several downsides to this approach. First, it translates less easily to the creation of datasets for other collections. When generating new datasets, we will need to recreate the process in large part. Second, the mapping itself is not easily readable for most librarians and staff. This shortcoming can be assuaged through good documentation practices, but the script maintainer must have knowledge of the collections and metadata involved.

## Namespace Management

URIs assigned by OSU Libraries are created within a general purpose namespace at <http://data.library.oregonstate.edu/>. Several sub-namespaces are used to simplify namespace management. The sub-namespaces are:

- <http://data.library.oregonstate.edu/thesis/>
- <http://data.library.oregonstate.edu/person/>
- <http://data.library.oregonstate.edu/degree/>

URIs for individual resources are assigned by appending a random, opaque extension to the relevant sub-namespace. The extensions are modeled after the California Digital Library's NOID identifier specification (NOID, 2006). Organizational units within the University are an exception to this; their URIs are assigned from the base namespace and take a more human readable form (e.g. [http://data.library.oregonstate.edu/OSU\\_Libraries](http://data.library.oregonstate.edu/OSU_Libraries)).

## Publishing the Data

The complete dataset is maintained as a set of RDF graphs housed in an RDF triplestore running on the 4store platform. The graphs are organized as portions of the overall dataset. By splitting the data in this way, specific applications are able to work with the data they need, bypassing the rest. The entire dataset--as a union graph--can be queried via a SPARQL endpoint provided by the triplestore software.

A Python based web application (Djubby, <http://code.google.com/p/djubby/>) supplies an interface for URI resolution and HTTP content negotiation, supporting retrieval of both human and machine-readable representations of resources within the <http://data.library.oregonstate.edu/> namespace. Bulk download of the entire dataset and of component graphs will also be available.

## USE CASES

To identify potential use cases, we examined existing implementations from the library domain, as well as the LLD Incubator Group's report on use cases (Vila Suero, 2011). We have drawn particular inspiration from the French national thesis dataset developed by ABES at <http://theses.fr> ("A Propos de theses.fr", n.d.).

In keeping with the service-building philosophy of our project, we have tried to identify use cases with immediate practical benefits for OSU Libraries.

## Discovery

Like many library resources, thesis discovery is spread across multiple data systems and interfaces. The primary points of access at OSU are the online catalog and the institutional repository. Neither of these systems supports thesis discovery through properties unique to theses. Building a thesis specific semantic search interface based on linked data would allow discovery of theses by advisor, committee member, department, degree, and graduation date.

As the outside data sources are linked to resources in our dataset, this functionality could expand to include, as examples, browsing other works by the author and advisor, or searching for related thesis work from other universities.

## Data Unification and Crosswalking

Creating a single, unified record for a given item has been cited as a core reason for a move to linked data by Singer (2009) and as a general use case by the LLD Incubator Group (Vila Suero, 2011). Singer (2009) cites, in particular, the resulting “workflow inefficiencies, unnecessary duplication of data” and the lack of a “standard means to provide relationships between the discrete entities” as shortcomings of the status quo.

These problems are evident at OSU in the MARC and Dublin Core data silos discussed previously. Though we will continue to use the “siloes” infrastructure, automatic crosswalks can smooth over workflow problems and limit the negative effects of data duplication.

### **Linking Theses to Research Data**

OSU Libraries is piloting a project to help students manage and preserve research data collected or used in their thesis research. It is unclear how often and under what conditions these datasets will be made available by the library, or when they will be housed by funding organizations or disciplinary repositories. In either case, there is a need to track datasets and their relationship to theses in our collection.

Linked Data will allow us to describe those relationships in simple ways (eg. `rdfs:seeAlso`) or using complex semantics, with interoperability toward both internally and externally maintained datasets. On the service side, this would make it possible to incorporate links to—and descriptions of—external datasets into thesis displays.

### **CURRENT PROJECTS**

In addition to the data crosswalking work detailed above, we have undertaken two initial projects.

#### **Faceted Thesis Search**

As our primary publicly visible service, we are prototyping a thesis-specific search system to meet our discovery use case. The system indexes each thesis in the dataset, including all metadata statements associated with the theses and related entities. Metadata terms explicitly included in our model are also given stronger weighting in “keyword” search results. Specific terms from the model are mapped to title, person, and subject searches.

We have opted to use a faceted search approach that takes emphasis off of advanced search functionality in favor of a “drill down” mechanism. While we have only a few search types, each common data element can be included as a facet allowing users to limit their initial results.

Because the interface needs are limited to theses, results pages display metadata not typical to other search applications. Advisor, degree, college, and graduation date all appear alongside title and author in the search result listings. Each thesis also has a record page tailored to the thesis data type to a greater degree than the one in the institutional repository.

Record pages also exist for people, degrees, and colleges, each with its own browse functionality. We expect these features to be valuable to internal University users; students can easily access past theses associated with their degree program, faculty can get a listing of past advisees and committee membership, and college administrators can find projects for degrees granted in a particular academic year.

#### **SKOS Name Authority**

One major shortcoming of institutional repository software is the lack of authority control. Salo (2009) describes this problem and its effects on usability in depth, noting “[s]hould a user arrive at a specific item and desire to see more items by the same author, clicking on the author’s name will lead only to results for that particular name spelling or variant, though possibly case-independently.”

Repository systems, including DSpace, have created interfaces for adding authority keys to metadata, but this still requires an externally maintained name authority record as a source for keys and name variants. Since most thesis authors and many other repository submitters do not appear in major library name systems, these solutions are of limited help. What is needed is a locally maintained name database.

Because our data model assigns a URI for each person in the dataset, we have been able to develop tools for internal name authority based on the Simple Knowledge Organization System (SKOS) vocabulary. SKOS provides a broad framework for organizing concepts into “thesauri, classification schemes, subject heading lists and taxonomies” for the Semantic Web (“SKOS Simple Knowledge Organization System - Home Page”, n.d.).

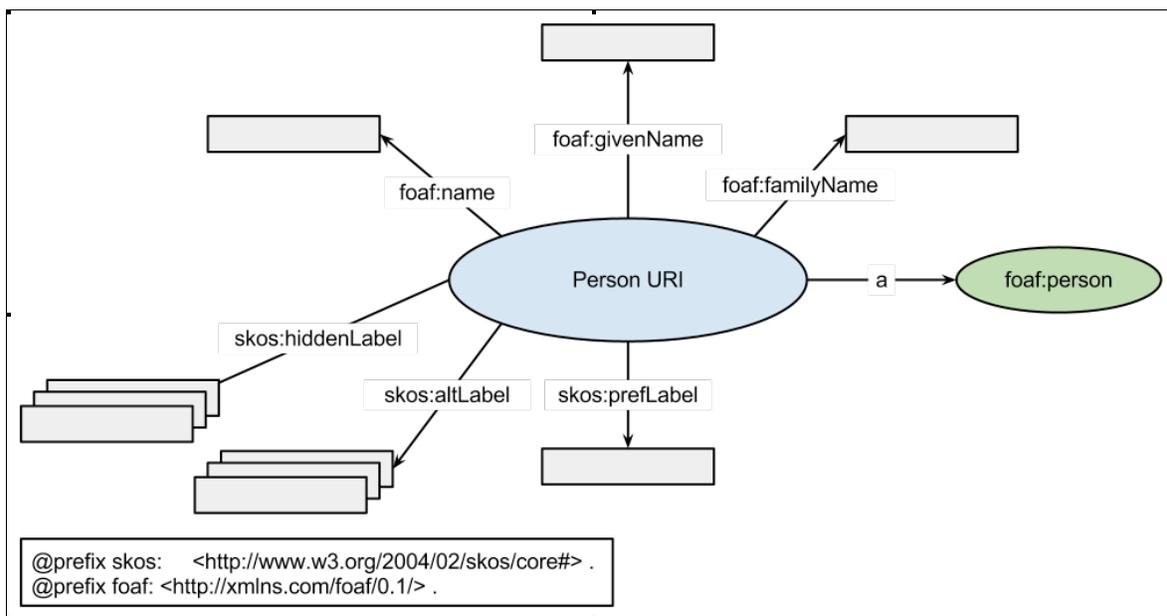


Figure 2. SKOS and FOAF properties for name authority

Harper (2006) has previously explored the use of SKOS for library authorities, and implementations are in place at the Library of Congress, as described by Summers, et. al. (2008). Our authority system mirrors the simplicity of their approach, modeling names as either skos:prefLabel, skos:altLabel, or skos:hiddenLabel. For greater interoperability with other Linked Data descriptions of people, name terms from the broadly used Friend of a Friend (FOAF) Ontology are included as well (“Friend of a Friend Project”, n.d.).

The data needed to construct this model is added by the same script that assigns person URIs in the thesis crosswalking process. The script attempts a rudimentary search for names, adding the new name as a skos:altLabel when a record is found. The automated name matching is subject to error (especially when the quality of existing data is poor) but even when

the SKOS records are incomplete or inaccurate the use of authority keys represents an improvement for search. In addition, staff can manage name data for each person through a Python web application we created for editing, splitting, and merging records.

We are in the process of implementing this authority system on a broad scale in our institutional repository. URIs serve as authority keys within the DSpace metadata system and a SPARQL client handles data interchange between the repository and the triplestore. A name review step will be inserted into the review process for new items, gradually correcting errors from the automated process. The initial names are sourced from the thesis dataset (authors, advisors, and committee members), but we are currently working to add names from other repository collections.

## **FUTURE DIRECTIONS**

### **Linking to External Data Sources**

Linking with other datasets allows our services to take advantage of interoperability with other participants in the linked data web. Some unrealized opportunities for linking already exist in the form of datasets for Library of Congress Subject Headings and Authorities, the Virtual International Authority File, open bibliographic datasets, and WorldCat. Non-library data sources like DBpedia and Freebase may also be high value targets.

### **Improved Name Authority**

Though our SKOS-based name authority structure addresses our needs for name search and cross-referencing in the institutional repository, its simplicity limits interoperability with MARC Authority Records and the discovery systems using them. We have considered implementing MADS/RDF as a second, SKOS compatible, layer. Since MADS/RDF is designed to encode MARC Authorities in RDF, such an implementation could allow us to convert between MARC and linked data authorities.

### **Standardization of a Thesis Data Model**

There is a need for community agreement about representing theses in linked data. Our model serves well to support the use cases discussed above and, perhaps, provide a starting point for future work. However, it won't represent a solution for the community at large.

A standardized thesis data model should be developed to represent the requirements of existing metadata formats and provide guidelines for community practice. It should be possible to develop complete mappings for common data formats such as ETD-MS and MARC while still emphasizing interoperability with common Semantic Web vocabularies.

### **Additional Datasets**

Lastly, we expect to use the expertise gained and the infrastructure built in this process to create new datasets based on other resources at OSU Libraries. In the short term, our focus will remain on unique collections with clear internal use cases.

## **CONCLUSIONS**

Linked data is an important trend in libraries and the groundwork is largely laid for us to begin implementing this technology. What remains is to create datasets covering library resources and link them together. As they do this, libraries need to be cognizant of use cases and build

services based on the value that RDF and data linking can add to our metadata. Many of these use cases can be implemented effectively today.

## REFERENCES

1. A Propos de Theses.fr. n.d. Available <http://www.theses.fr/apropos.html>
2. Baker, T., Bermès, E., Coyle, K., Dunsire, G., Isaac, A., Murray, P., ...Zeng, M., 2011. Library Linked Data Incubator Group Final Report. Available <http://www.w3.org/2005/Incubator/ld/XGR-ld-20111025/>
3. Bizer, C., Cyganiak, R., and Heath, T., 2007. How to Publish Linked Data on the Web. Available <http://www4.wiwiw.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
4. Boock, M. and Kunda, S., 2009. Electronic Thesis and Dissertation Metadata Workflow at Oregon State University Libraries, *Cataloging & Classification Quarterly*, 47, 297–308. doi:[10.1080/01639370902737323](https://doi.org/10.1080/01639370902737323)
5. Friend of a Friend Project, n.d. Available <http://www.foaf-project.org/>
6. Harper, C., 2006. Authority Control for the Semantic Web. Encoding Library of Congress Subject Headings, *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2006. Available <http://dcpapers.dublincore.org/index.php/pubs/article/view/842/838>
7. Rochkind, J. (2012, February 15) Yet Another Defense of FRBR in a Linked Data World. Available <http://bibwild.wordpress.com/2012/02/15/yet-another-defense-of-frbr-in-a-linked-data-world/>
8. Salo, D., 2009. Name Authority Control in Institutional Repositories, *Cataloging & Classification Quarterly*, 47, 249-261. doi:[10.1080/01639370902737232](https://doi.org/10.1080/01639370902737232)
9. Singer, R., 2009. Linked Library Data Now!, *Journal of Electronic Resources Librarianship*, 21, 114-126. doi:[10.1080/19411260903035809](https://doi.org/10.1080/19411260903035809)
10. SKOS Simple Knowledge Organization System - Home Page, n.d. Available <http://www.w3.org/2004/02/skos/>
11. Westrum, A., Rekkavik, A., and Tøllerås, K., 2012. Improving the presentation of library data using FRBR and Linked data, *Code4Lib Journal*, 16. Available <http://journal.code4lib.org/articles/6424>
12. Vila Suero, D., 2011. Library Linked Data Incubator Group: Use Cases. Available <http://www.w3.org/2005/Incubator/ld/XGR-ld-usecase-20111025/>

## APPENDIX: EXAMPLE DATA

This example data is serialized in Notation3 format.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix marcrel: <http://id.loc.gov/vocabulary/relators/> .
@prefix vivo: <http://vivoweb.org/ontology/core#> .
@prefix osuterm: <http://data.library.oregonstate.edu/terms#> .

<http://data.library.oregonstate.edu/thesis/498k> dcterms:title "Hybrid electric vehicle converter harmonics";
    skos:prefLabel "Hybrid electric vehicle converter harmonics";
    rdfs:type <http://iflstandards.info/ns/fr/frbr/frbrer/C1001>.
```

```
    bibo:Thesis;
dcterms:creator <http://data.library.oregonstate.edu/person/vmvp>;
marcrel:dis <http://data.library.oregonstate.edu/person/vmvp>;
marcrel:dgg <http://data.library.oregonstate.edu/OSU>;
bibio:degree <http://data.library.oregonstate.edu/degree/499p>;
marcrel:ths <http://data.library.oregonstate.edu/person/1npd>;
osuterms:thesisCommitteeMember <http://data.library.oregonstate.edu/person/82dn>,
    <http://data.library.oregonstate.edu/person/zxt>;
dcterms:contributor <http://data.library.oregonstate.edu/person/82dn>,
    <http://data.library.oregonstate.edu/person/zxt>,
    <http://data.library.oregonstate.edu/person/1npd>;
bibio:abstract "Hybrid electric vehicles (HEVs) are a very important part of today's
transportation system.";
dcterms:language "en";
dcterms:dateAccepted "2005";
dcterms:dateIssued "2004-06-08";
dcterms:subject "Hybrid electric vehicles",
    "Electric current converters -- Design and construction",
    "Voltage regulators";
dcterms:identifier <http://hdl.handle.net/1957/10462>;
bibio:handle <info:hdl/1957/10462> .

<http://data.library.oregonstate.edu/person/vmvp> rdfs:type foaf:Person;
skos:prefLabel "Bowers, Waylon, T.";
foaf:name "Bowers, Waylon, T.";
foaf:givenName "Waylon";
foaf:familyName "Bowers";
marcrel:disOf <http://data.library.oregonstate.edu/thesis/498k> .

<http://data.library.oregonstate.edu/person/1npd> rdfs:type foaf:Person;
skos:prefLabel "von Jouanne, Annette";
skos:altLabel "von Jouanne, Annette R.",
    "Von Jouanne, Annette R.";
foaf:name "von Jouanne, Annette";
foaf:givenName "Annette";
foaf:familyName "von Jouanne";
marcrel:thsOf <http://data.library.oregonstate.edu/thesis/498k>,
    <http://data.library.oregonstate.edu/thesis/1wjg>,
    <http://data.library.oregonstate.edu/thesis/3kqf> .

<http://data.library.oregonstate.edu/degree/499p> rdfs:type bibo:degrees/ms,
    vivo:AcademicDegree;
vivo:degreeOfferedBy <http://data.library.oregonstate.edu/College_of_Engineering>
skos:label "Master of Science (M.S.) in Electrical and Computer Engineering";

<http://data.library.oregonstate.edu/College_of_Engineering> rdfs:type vivo:College;
skos:label "College of Engineering; Oregon State University";
skos:prefLabel "College of Engineering";
vivo:offersDegree <http://data.library.oregonstate.edu/degree/499p>;
org:subOrganizationOf <http://data.library.oregonstate.edu/OSU> .
```