Exploring Linked Data

Eva Bolkovac
Assistant Catalog Management Librarian
Submitting on behalf of the Metadata Management Team, Resource Discovery Services, Technical Services
Metadata Management Team members: Youn Noh, Timothy Thompson, Dominique Bourassa, Yukari Sugiyama, Sarah Coe, Marena Fisher
344 Winchester Avenue, PO Box 208203, New Haven, CT 06511
(203) 432-7659
eva.bolkovac@yale.edu

Purpose and Expected Outcome

“Linked Data refers to a set of best practices for publishing and connecting structured data on the Web.”¹ Over the past few decades,² academic and research libraries have been exploring the role of linked data in sharing metadata about resources in their collections. These efforts have gained momentum with initiatives such as BIBFRAME,³ Linked Data for Libraries,⁴ and the SHARE project sponsored by Casalini.⁵ As a participant in SHARE, Yale University Library (YUL) is having all of its MARC bibliographic and authority records converted to BIBFRAME. The converted data will be shared among project partners, including several of Yale’s peers, in a data warehouse maintained by Casalini.

Many librarians at YUL are eager to acquire hands on experience with linked data and to understand its impact on library services and workflows. It is anticipated that linked data will enable a more collaborative, modular approach to resource description but will require the reconciliation or accommodation of metadata from disparate sources.⁶

In this project, we propose to use Blazegraph, an open-source graph database platform, to explore datasets about film from the Film Study Center and the Internet Movie Database (IMDb). The Film Study Center datasets will be derived from MARC records in Orbis and from the Film Study Center’s OpenBiblio database. These datasets describe the same resources but use different standards. The IMDb dataset² is publicly available for research via Amazon Web Services (AWS) and encompasses metadata available for searching on its site.

At the conclusion of the project, we expect to be able to provide the following deliverables:
1. Enhanced linked dataset for use in training and research
2. Linked data workshop

¹ http://linkeddata.org/faq
² http://www.dlib.org/dlib/may98/miller/05miller.html
³ https://www.loc.gov/bibframe/
⁴ https://www.ld4l.org/
⁵ http://share-vde.org/
⁶ https://www.loc.gov/aba/pcc/documents/LinkedDataInfrastructureModels.pdf
⁷ http://www.imdb.com/interfaces/
Assessment of MARC to BIBFRAME conversion technologies and approaches

Evidence-based FAQ on linked data

We believe that the lessons learned from the project will help us make informed contributions to discussions on linked data at YUL, as well as open up opportunities that both enhance and further improve services at YUL, and within the information science community.

Methodology and Timeline

Participants and Roles

The project participants will consist of staff reporting to Eva Bolkovac in Resource Discovery Services, Technical Services (TS). Participants will take responsibility for acquiring and sharing expertise in the following areas with other team members:

1. BIBFRAME
2. RDF, OWL, and SPARQL
3. Triple stores and graph database technologies
4. Initial consultation and assessment with the Film Study Center - Michael Kerbel, Director, Film Study Center (1-hour interview with Michael agreed to in an email)
5. Consultation and training in TS

Milestone 1: Convert, ingest, and clean up datasets. (3 months)

For the Film Study Center data, we will use a subset of the BIBFRAME dataset converted by Casalini. We will also use marc2bibframe2\(^8\) developed by the Library of Congress to convert MARC records from scratch. For the IMDb data and the Film Study Center data in OpenBiblio, we will convert the tab-delimited files to linked data using domain-specific ontologies, such as the Movie Ontology\(^9\) developed by researchers at the Free University of Bolzen-Bolzano, while also exploring additional relevant ontologies.

Converting MARC records to BIBFRAME from scratch will allow us to evaluate conversion tools, Casalini’s conversion process, and BIBFRAME itself. Converting data to other standards will give us a frame of reference for understanding BIBFRAME and will allow us to explore interoperability and integration issues.

Patrick Stone, Workflow Analyst/Programmer, LIT, whose time is split between LIT and TS, 50-50, will be the Library IT contact for the project and will coordinate LIT resources as needed, including the installation of Blazegraph and assistance with the ingestion of the datasets (confirmed by Marty Kurth, TS AUL, and Ray Frohlich, LIT, Director of Enterprise Systems and Architecture). We expect the conversion process to be iterative, as we clean up our data and refine our mappings.

Milestone 2: Learn SPARQL and graph database technologies. (3 months)

Once the data has been ingested, we will use the train-the-trainer model to learn SPARQL and to learn how to use Blazegraph to work with our datasets. A few team participants will lead

---

\(^8\) https://github.com/lcnetdev/marc2bibframe2

\(^9\) https://github.com/ontop/ontop/wiki/Example_MovieOntology
the team through selected chapters of Bob DuCharme’s *Learning SPARQL*,\(^{10}\) or a similar resource, applying what we have learned to our datasets along the way. A few participants will work with Eric James, Software Engineer, Yale Center for British Art, (up to 2 days of time commitment confirmed by Eric via email) to learn how to administer Blazegraph and to provide database training and guidance with its ongoing support. We will not require the use and support of existing library servers, but plan to purchase/use Amazon Web Services, Software as a Service (SaaS) options.

**Milestone 3: Apply inference. Enhance data. (2 months)**

Once we are comfortable using SPARQL and Blazegraph, we will look for queries we can run to add inferences to our asserted statements about Film Study Center resources. Inference\(^{11}\) is frequently cited as a major benefit of the Semantic Web. We believe our datasets are rich, multi-dimensional, and interconnected enough to allow interesting inferences to be drawn. Through this project, we expect to learn more about using inference for resource discovery.

**Milestone 4: Assess data conversion and enhancement. (2 months)**

Once we have made our final improvements to our datasets, we will consult with staff at the Film Study Center to devise a set of test questions that we should expect to be able to answer with our datasets. We will assess the data conversion and enhancement process based upon the quality of the answers to the test questions. While user interface development is out of scope for our project, we expect to be able to run SPARQL queries using Blazegraph’s web interface, interactively with our domain experts from the Film Study Center, to assess the information content of our data sources.

**Milestone 5: Create training materials. Provide workshop. Create FAQ. (2 months)**

We will use our experiences from Milestone 2 to create training materials for use in a workshop. The intended audience for the workshop will be staff in Technical Services and other interested members of the YUL community; no prior knowledge will be assumed. Training materials will be made available online after the workshop. We will also create an evidence-based FAQ to share what we learned from the data conversion, enhancement, and assessment process.

**Expenses**

<table>
<thead>
<tr>
<th>Expense</th>
<th>Amount</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Web Services (AWS), Software as a Service platform</td>
<td>$1,347.84</td>
<td>Estimated amount is based on an assumption of a 24/7 on-demand usage rate of $0.156/hour for a basic \texttt{i3.large} compute</td>
</tr>
</tbody>
</table>

\(^{10}\) http://search.library.yale.edu/catalog/11716309

\(^{11}\) https://www.w3.org/standards/semanticweb/inference
instance. See https://aws.amazon.com/marketplace/pp/B01JBL2M0O?qid=1517582187922 for pricing verification. Based on the needs of the project, different instance options or schedules may be appropriate (e.g., it may not be necessary to run the instance during non-work hours, in which case a more expensive and computationally powerful instance could be selected under the same budget).

<table>
<thead>
<tr>
<th>AWS General Purpose Storage</th>
<th>$180.00</th>
<th>150 GB of working storage (needed for running the AWS instance) at $0.10 per GB-month of General Purpose SSD (gp2) provisioned storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage@Yale</td>
<td>$43.20</td>
<td>One year of Archive Tier storage of the project data following completion of the grant period at a rate of $3.60/month. See <a href="https://its.yale.edu/services/communication-and-collaboration/document-sharing-and-team-sites/storagyale">https://its.yale.edu/services/communication-and-collaboration/document-sharing-and-team-sites/storagyale</a> for pricing verification.</td>
</tr>
</tbody>
</table>

TOTAL $1571.04

The dataset created as part of this project is intended for internal usage, training, and exploration. The steps used to generate it will be documented in order to facilitate replication of the data creation process. There is no long-term commitment to preserving the data indefinitely following the end of the grant period; however, the budget does include the cost of one year of archival storage on Storage@Yale so that the dataset will be available for follow-up consultation and evaluation after the immediate end of the project.

Benefits

Benefit 1: Enhanced linked dataset for use in training and research

The enhanced linked dataset created through this project will be made available online for use at YUL (and not for public consumption) in ongoing linked data training efforts and guided self-learning. It will also be suitable for use as a dataset for developing, testing, and implementing linked data technologies and applications.

Benefit 2: Linked data workshop

The linked data workshop offered in Technical Services will reflect the lessons learned over the year of the grant and will be based on practical experience. We believe that linked data
technologies are sufficiently mature for us to use a train-the-trainer model to share knowledge within our team and Technical Services at large and that such training will help establish a common base of knowledge from which to engage with external experts and consultants.

**Benefit 3: Assessment of MARC to BIBFRAME conversion technologies and approaches**

MARC is so widely entrenched that migrating from MARC to BIBFRAME would be a huge undertaking for YUL and the library and information science community. Many within the community have expressed reservations about BIBFRAME. This project will put us in a better position to make informed contributions to the discussion and, potentially, to contribute to a path for migration, as well as serve as a model.

**Benefit 4: Evidence-based FAQ on linked data**

In committee and departmental meetings, many staff members in Technical Services have asked about the practical implications of adopting linked data. We believe that the lessons learned through this project will help us provide evidence-based answers to these questions. We expect to reach our largest audience through this FAQ. As part of reaching out to staff participating in this grant, Eric James indicated an interest in forming a Working Group, after the project is completed, with two members of the Yale Linked Data Interest Group, Emmanuelle Delmas-Glass and Michael Appleby. This would allow for larger participation not only within the Yale community, but also for soliciting wider collaborations from other communities.