Final Report: Smithsonian Minority Awards Program Internship

Applying standardized identifiers to library data and practicing digital curation at Smithsonian Libraries

Allyson Ota

University of Hawaii, Manoa

August 10, 2016

Author note:

Allyson Ota, MLISc candidate, University of Hawaii, Manoa

All projects described in this report were completed under the supervision of Joel Richard, Head of Web Services for Smithsonian Libraries' Digital Programs and Initiatives Division, and was funded by the Smithsonian Minority Awards Program. The 10-week internship ran from June 6th through August 12th, 2016.
For additional questions, the author can be contacted directly at: allyson@hawaii.edu

Table of Contents

Introduction

I'm currently a second year graduate student in the Library and Information Science program at the University of Hawaii at Manoa. Born and raised in Hawaii, it was a dream come true to be able to do an internship with the Smithsonian Institution in a setting so relevant to my future goals within the field of librarianship. I chose to leave a career in IT for one in librarianship because I wanted to serve in a more community-service oriented profession. I'm a huge fan of lifelong learning and firmly believe that intellectual freedom is a right all citizens are entitled to. The projects I participated in this summer with Smithsonian Libraries have been invaluable, since I know I could not have gained this type of experience in school. It was also an opportunity to work with large amounts of data, alongside dealing with collections issues for a large institution.

Project 1: Creating a workflow for batch DOI registration to Smithsonian publications

My first project was the most challenging and also the most rewarding. Smithsonian Libraries maintains a digital library with Books Online, the Smithsonian Research Online portal, and DSpace—their document repository. These valuable resources hold citations or digital copies of publications for researchers and the general public to use. However, how easy is it for information to get lost in the vastness of the Internet? How permanent is a URL? While a hyperlink might work today, there are no guarantees it will still work a year from now. Using a direct URL is not the most reliable way to ensure people can find bibliographic sources in citations.

For these reasons, Digital Object Identifiers (DOIs) fill the role of tying publications to a unique and permanent identifier. You can type a DOI into a browser, and it will redirect you to the object you seek. DOIs also hold information about the publications they link to, e.g. title, author, year, etc. Most researchers encounter DOIs in citations for digital articles or books they've read. Publishers assign DOIs to digital publications in order to make their works findable and accessible for use. A DOI consists of a prefix and a suffix. The prefix, always starts with a "10" and is followed by an identifier. The suffix can be created by the publisher as long as it's unique. So a Smithsonian DOI might look something like: 10.5479/sil.12345.98033892. Smithsonian Libraries uses Crossref as its DOI registration agency, and it was my task to develop a workflow that could be used in order to register batches of DOIs for online publications currently lacking one.

Process:

I was given a database export from The Libraries' catalog for publications requiring DOIs. There were over 2,000 books and 24 serials total. The spreadsheet needed to be converted into an XML (Extensible Markup Language) file. XML is a markup language that is both human and machine readable commonly used to pass information between systems. The structure and values of your XML tell the receiving system where the information contained within each tag belongs, and separates individual objects. The title of a book might look like this:

```
<title>On the Origin of Species</title>
```

My XML file had to contain the tags Crossref requires for data being passed. Proper formatting, along with element tags, attributes, and the values for these fields could be registered through properly formatted XML. In order to make this happen I followed these steps:

1) Imported the database export spreadsheet into OpenRefine and cleaned the data and created some required fields, e.g. DOI, using General Refine Expression Language (GREL). Changed column headings to match element/tag names we'd need to pass our data to Crossref, etc.
2) Exported the properly formatted data back to excel from OpenRefine
3) Generated an XML document containing all metadata about each publication using an MS Excel 2003 Add-In, which enabled schema generation and XML export ability
4) Opened the XML file in oXygen XML Editor, which allowed us to use XSLT (Extensible Stylesheet Language Transformations) code co-written with Joel Richard, head of Web services for the Libraries which would further transform the data, to generate a second XML file that was properly formatted for Crossref's batch upload process.

Lessons learned:

Throughout the process there were data issues that arose and in some cases it turned out we had to fix them in the catalog. As a result, this project also helped Smithsonian Libraries perform data clean-up on the catalog's database. We found some information for books authored by corporations weren't pulling the entire corporation name into the database, so some of the authors displayed as just "United States" as opposed to "United States Department of Agriculture." Other data issues included human error when items were input in the catalog, e.g. typos, etc.

Outcomes:

I was able to successfully register DOIs for 24 articles from the *Herpetological Information Service* journal, and 2,668 digital books with Crossref. I submitted workflow documentation to Bess Missell, the Systems Librarian, that detailed all the software requirements, instructions for cleaning and creating data in OpenRefine, all of the GREL expressions I used to create fields in OpenRefine, and lessons learned throughout the process. Missell, should be able to refer to my workflow documentation for future uploads.

Project 2: Linking library data to VIAF

Another project that tied into the first, involved the Virtual International Authority File (VIAF). Libraries keep lists of authors, geographic locations, corporations, etc. which are referred to as their authority files. VIAF collects authority files from national libraries (and some other organizations) from countries around the world. For the U.S., the Library of Congress contributes its authority files. By linking all these authority files together, VIAF creates a "super authority" file for an entry. For this project, I was asked to link our authors (which consists of persons and organizations/corporations) to these super authority files found in VIAF. We wanted to see if this could be accomplished because we knew OpenRefine had the ability to link to external data sources, so we wanted to find a reconciliation process that would connect to VIAF and bring back VIAF IDs for authors of works held within our catalog.

Richard hopes to use the ability to link to open data sources in future projects in order to connect through APIs (Application Program Interface) to external data sources such as Wikipedia or DBPedia moving forward. The thought is that if VIAF links to other data sources (and vice-versa), we can use the VIAF ID to connect external data sources to our library data and pull information in via APIs in order to create a richer experience for researchers using our online resources. In the future if this can be implemented, it would make datasets more navigable for when searching through our online catalog and repositories.

Process:

I was provided a database export of authors, and tested 2 different open source reconciliation services found online. The first I tried was written by a developer named Roderic D.M. Page, who I found through OpenRefine's support pages, and was posted as a blog post in 2013. This service did not support reconciliation for corporations. I found a second reconciliation service and determined Jeff Chiu's refine_viaf service is still being maintained as I can see activity on the GitHub page showing updates to files. The service was easy to install and involved downloading a .jar file and running it from the command line. It also required Java version 1.7 or higher. I documented the process of downloading and running the service from within OpenRefine and was able to connect to VIAF. I tested the service on persons, sorting by last name (in order to catch possible duplicates) and ran through the first 50 names. After reconciling the list, I was then able to import VIAF IDs for the authors we received a match for.

Outcomes:

Of the 50 persons I attempted reconciliation with, I was able to retrieve VIAF IDs for 36 of them. This is not a purely conclusive test, since data issues would cause additional problems and I only tested on a small portion of the data. I discovered an issue with OpenRefine where the first column has to be populated with data. Rows beginning with blank cells ended up being appended to the previous row, so I adjusted the documentation and procedure to highlight this fact and mention persons with only a surname need to be reconciled separately, and to also reconcile corporations separately.

I discovered there was still a lot of manual work involved when it comes to verifying the matches made by the reconciliation service, since matches made by the service were not always accurate. A librarian should verify our authors are being linked to the correct authority file for each author in order to ensure accuracy. The project was successful, since it was done as an experiment to see what we could do with OpenRefine and VIAF. Richard hopes that linking library data to VIAF IDs can enable access to other linked data sources, for example, seeing that Wikipedia is also using VIAF IDs in their Authority Control could allow linkages to Wikipedia or DBPedia data to authors in the Smithsonian Libraries catalog, creating a richer experience for users of the collection.

Projects 3 & 4: Digital curation

The Galaxy of Images hosts images from digitized publications. It's utilized in generating educational materials by Smithsonian Libraries' staff working to create educational outreach materials, and can also be used by the general public, since all images are in the public domain. I think of the Galaxy of Images as an important tool of engagement that can draw users into The Libraries collections. Seeing a pretty or interesting image online can motivate someone to look for the book it came from, enticing them to further explore The Libraries' digital collections.

Weeding images from the Galaxy of Images:

In librarianship, the term weeding describes the process of removing items from your collection. Generally weeding is done in an effort to free up space for more desirable items to be added to a collection, and analyzing hit counts allowed us to see which images were getting the least use from viewers. The Galaxy of Images (GoI) holds over 16,000 images and is going to be migrated from ColdFusion to Drupal in order to add functionality and modernize the current site. In preparation for the migration, it was determined some weeding should be done to help prepare for the upgrade.

Process:

Metadata Librarian, Douglas Dunlop, provided criteria to evaluate individual images found in the collection that had received 4 or less views in the past three years, according to reports from Google Analytics.[1] The criteria consisted of a scoring system to be applied to each image, with the final number of total points determining the probability of an image being saved as either: 1) low, 2) moderate, or 3) high.

4 Categories ranked on a scale of 1-10
- Image quality
- Usability
- Uniqueness
- Visual Interest

A total score of 1-19 ranked low, 20-29, was moderate, and 30 and above was considered high for the probability of the image to be saved. In addition to scoring, additional factors taken into account were whether or not they had been featured in an online exhibit, whether they were grey-scale or color, and if there was a high text-to-image ratio.

Outcomes:

While some images were indeed beautiful and scored high points which earned them a "high" probability of being saved, there were instances of images containing a high text-to-image ratio, or lower-quality images that I scored a "low" probability to keep them around in the GoI. Dunlop will double-check and have the final say, but this preliminary groundwork will aid him since there were still thousands of other images to go through. I assessed a total of 10,205 images for Dunlop.

Selecting Images for inclusion in the Galaxy of Images:

The opposite of weeding is called selection. Selectors add items to a library's collection(s). This was more fun. Richard and Dunlop walked me through the process and explained that I would have to use my judgement on rating the interestingness of an image. After the weeding process, they felt I could apply some of the same criteria to these images, e.g. not selecting images with a high text-to-image ratio, but also taking into account uniqueness and rarity, etc.

For selection, I used Macaw, a program written by Richard, which he explained "is a tool meant to collect and organize page images from scanned or digitized books, collect metadata about the pages, and upload them to the Internet Archive for inclusion in the Biodiversity Heritage Library (BHL). It is also used to select page images to be delivered by Macaw to Smithsonian Libraries online image library."

Process:

Images were examined using Macaw, which allowed me to perform a facet based search to display only the pages with images based on metadata that was included when the book was scanned. I could select the ones I thought should be added to the Galaxy of Images and submit them to be reviewed. Throughout the selection process, I tried to ensure duplicate images were not selected (in the case of advertisements in journals), and also made note of criteria I used in the weeding process. A lot of this process was actually pretty subjective, and I deferred to being more inclusive than exclusive in my selection.

Outcomes:

I selected images from 98 books and submitted them for consideration to be added to the Galaxy of Images. I sent a listing of all the works I completed to Dunlop, who will further review my selections. Using Macaw

---

[1] Google Analytics provides an estimate of these page views based on a statistical sample of the data. It does not return exact actual values for three years' worth of data.

allowed me to interact with software being used by the department in their everyday workflow and gave me a feel for the experience of selecting images for use in an online collection.

Personal and Professional Development

During my internship, aside from working on projects, I participated in special tours and events for interns, sat in on a few departmental meetings, and also had the opportunity to attend the NISO Virtual Conference: BIBFRAME and Real World Applications of Linked Bibliographic Data. I also decided to independently ask multiple Smithsonian Libraries employees to do interviews with me in order to explain the duties of their current position at SI, (as well as any previous library positions they've held), required skills to perform their jobs, and finally asked for advice on resources or tips for job searching and skills I should try to obtain before graduating in May 2017. The willingness of staff to give me great advice was phenomenal. I spoke to 14 staff members and want to mention them by name here (in alphabetical order):

> Grace Costantino, Bianca Crowley, Douglas Dunlop, Alvin Hutchinson, Martin Kalfatovic. Monique Libby, Bess Missell, Richard Naples, Lesley Parilla, Suzanne Pilsk, Alex Reigle, Joel Richard, Erin Rushing, Keri Thompson.

Conclusion

Accessibility and engagement are huge aspects of librarianship, and all of my projects this summer were focused towards these ends. As the world becomes more and more technologically advanced and digital content is being created at exponential rates, the amounts of information can be difficult to wade through. The Internet, with its vast resources and seemingly endless search possibilities can become a place where information gets lost or is difficult to find. Librarianship has continually had to adjust to meet the new demands of users and rapidly evolving technology. This makes digital librarianship especially relevant moving forward, and I'm thankful to have had experiences I would not have had in school.

When I applied for the Minority Awards Program, I expressed in my essay that I had wanted to learn more about digital librarianship and/or archives, and this summer I was definitely able to accomplish that thanks to the multiple projects developed by my supervisor, Joel Richard. It has been a unique and invaluable experience to intern with the Digital Programs and Initiatives Division at Smithsonian Libraries. I would never have been able to afford to do this internship without the support of the Minority Awards Program, and am grateful for the financial support I received. I am incredibly thankful for the opportunities I've had here, and will never forget this experience. I will always consider the summer of 2016 one of the highlights of my entire adult life, and know the knowledge I've gained will help me move forward along the path I've chosen for this second career in my life. I will definitely apply what I've learned here towards the communities and organizations I serve in the future.