Good data, bad data: getting ready for linked data

Abstract

Linked data in libraries provides opportunities for increased discoverability and reuse and repurposing of content for new applications. The benefits extend beyond libraries, and are being seen increasingly in the cultural, academic, government, and health sectors. This presentation outlines the steps taken to prepare bibliographic records for linked data, and identifies other opportunities for libraries in developing linked data.

Libraries have many years of experience in creating and managing data. Legacy data is a valuable resource, but does not integrate well with new standards. To be successful in the linked data environment, metadata must differentiate between content and carrier, and must enable appropriate links to unambiguously identify entities to expose and leverage correct and relevant relationships. The implementation of RDA and the release of Library of Congress RDA authority records with the recent inclusion of URIs in 024 fields are significant and exciting developments.

The University of Sydney Library has approximately three million bibliographic records; a large percentage of these are AARC2 and some are AACR1. There are over one million Library of Congress authority records, but not all bibliographic records have been through authority control. In order to fill the gaps and to get the data fit for purpose, we decided to kick off by identifying significant and unique categories of resources to put through onsite quality checks followed by authority control and RDA conversion by our offsite vendor. The immediate benefit is more authorised access points and references providing helpful information delivered 24/7 to library clients. More than that, data cleaning and record enhancement will build the backbone of reliable links with better discoverability for the linked data future.

Libraries can play other roles to support linked data. They can promote the take up of persistent identifiers such as ORCID, publish their own institutional data in a linked data format for reuse, and ensure linked data capabilities are considered in the development of new library and institutional systems.

Introduction

Libraries now are expected to be accessible around the clock to a global audience. To support this, libraries are increasingly investing in digitizing content from unique legacy collections along with investment in born digital content. The value of library collections is widely understood and appreciated. Librarians not only build significant collections, they also create vast amounts of metadata that describe and provide access to individual resources. The creation and maintenance of metadata takes significant intellectual and financial input, making it a valuable asset. However in its current form it is often not discoverable by the search engines that library clients use. Linked data is a means to lift our content out of the Deep Web by enabling library metadata to be discoverable by search engines, making library resources more accessible.

Furthermore, with linked data the metadata will be actionable; one resource will be able to generate many new relationships and will allow a human or machine to dereference terms in the Semantic Web. Extending resource description to place a publication within a knowledge context will make the library catalogue much more than a list of largely known and uninformative items. It will enable serendipitious discovery of the unknown (Coyle, 2016). The benefits to be gained from exposing library collections as dynamic, information-rich resources are obvious. The British National Bibliography as Linked Open Data (http://bnb.data.bl.uk/) has 3.1 million records representing the publishing activity of the United Kingdom and the Republic of Ireland.

Linked data in the information community

Librarians' experience with and commitment to data quality places us well to lead linked data initiatives in the metadata community. Colleagues in the cultural and heritage institutions have been making use of linked data to expose content from their unique collections and implement new ways of facilitating access to their content, e.g. through 'generous interfaces' that are rich and browsable, revealing the scale and complexity of digital collections (Whitelaw, 2015). That this content is being made available through linked data mechanisms from institutions across the GLAM sector means increasing opportunities to draw connections across institutions, and reunite content in a digital form. As institutions publish more linked data sets, the opportunity to create a fuller picture about items in our collective collections increases.

Large academic institutions such as the University of Sydney are uniquely placed to collaborate in developing and sharing linked data. The University of Sydney Library has a library catalogue, electronic resources, rare books and special collections, digital content, institutional repository, and research data registry. In addition to the Library, the

University of Sydney has museums, an art gallery, and the University Archives. If these entities were able to enrich their metadata and expose at least some of their content as linked data, there would be opportunities for students, researchers, and a global audience to build a more complete picture of an item. For example, an artefact from the Nicholson Museum could be linked to published items held in the catalogue (or globally in other catalogues), to related publications or data in the institutional repository, and to similar items held in other galleries or museums. However this ideal state would require a significicant amount of time, effort, and resources.

The challenge

The challenge for libraries is that linked data is still in development and many aspects are uncertain. Major developments so far include the release of VIAF (Virtual International Authority File http://viaf.org/) and LC (Library of Congress http://id.loc.gov/) authorities as linked data in 2009. In 2011 LC announced the Bibliographic Framework Initiative (https://www.loc.gov/bibframe/) as a linked data alternative to MARC. In 2012 OCLC released WorldCat as linked data using the Schema.org vocabulary. RDA (Resource Discovery and Access) was widely implemented in March 2013. LC has started releasing authority records with URIs (Uniform Resource Identifiers) in multiple 024 fields that link out to VIAF, DBpedia, Wikidata, and the IMDb to name a few.

Some libraries have published their data as linked data, but the barriers for many libraries to do so are prohibitive because of the required resources and LMS capabilities. This should not stop us from taking action as failure to do so will affect our ongoing capacity to interoperate with other communities as there will be an increasing reliance on the richness of our metadata (Sheih & Reece, 2015). Furthermore, URIs will facilitate the migration of library data to other formats such as BIBFRAME.

There is much that we can do to prepare our data to be fit for purpose. Firstly, we need to ensure that the data is present and correct. MARC coding is still being developed but the changes are not reflected in older records. Non-compliant RDA legacy data has to be re-examined and retro-converted. Secondly, we can use standard http URIs to transform strings of text into explicit references to structured information, and to encode meaning using ontologies to define types and relationships (Krafft, 2016). The benefit for library staff will be simplified workflows; updates will be instant through machine-to-machine processing of bibliographic and authority data.

Most importantly we need our people to be ready to take the first steps. The University of Sydney Library implemented RDA in March 2013 with customised bibliographic training provided by Sydney TAFE. RDA authority control training was done in-house. Building our knowledge and understanding of linked data concepts and tracking developments is ongoing. Our membership of the OCLC Research Partnership and the OCLC Metadata Managers' Focus Group gives us access to an international pool of experts and input into the metadata management issues undertaken by OCLC Research. Our goal is to upgrade and enrich the metadata of specific categories of resources using the processes, tools and people at our disposal to build the foundation for future development. Our capacity to invest staff time is limited; our actions have to be strategic and within budget.

Making a start: data remediation

The University of Sydney Library has approximately 3 million bibliographic records with over 1 million Library of Congress authority records that are supplied and maintained by an authority control provider. Our metadata remediation workflow is focused on updating legacy data and enriching string-based access points with URIs in subfield 0. Our basic tool for metadata remediation is Sierra, our LMS (Library Management System). However, the bulk of the remediation work is done by our authority control provider as RDA conversion is a free service with authorities processing. It is an established, streamlined, cost-effective and automated process that can convert large quantities of bibliographic data in a short space of time. This has drastically reduced our workload and our reliance on tools like MarcEdit and OpenRefine.

Data remediation started in 2012 with the first release of RDA compliant authority records. Approximately 120,000 revised authority records were loaded resulting in thousands of bibliographic differences that had to be resolved. After each load Sierra's AACP (Automated Authority Control Processing) flipped bibliographic access points matching authority records on a see from tracing (4XX). Thousands of straightforward changes were made, for example abbreviations such as Dept. to Department and arr. to arranged. Those that could not be changed automatically were done category by category using Global Update. Wrongly formulated and complicated access points (often music and religious works) had to be done record by record.

We are now focussing on personal names. RDA does not permit undifferentiated personal names; the metadata must uniquely identify an entity in order to expose correct and relevant relationships. Disambiguation and reconciling variant forms of a name requires human intervention. Our authority control provider does ongoing checks for all access points that do not match an authority to ensure that we receive newly released authority records. RDA and the resulting work on disambiguaton is resulting in an increasing number of new match authority records. In our recent monthly file load we received 1,661 new match authority records.

The next step in authority control is to contribute our locally created authorities to the LC Name Authority File. Two specialist staff have commenced NACO (Name Authority Cooperative Program) training to contribute locally created authorities through the NACO CJK (Chinese, Japanese & Korean) Funnel. Local authority records are created if references are needed in the catalogue, particularly for University of Sydney staff and corporate bodies. In NAF these authorities will have URIs that we can capture in our bibliographic records as linked data. Authority remediation is progressing well.

However, our bibliographic data is not in such good shape. The University of Sydney Library has a large and old collection and metadata remediation presents many challenges. A large percentage of our records are AARC2, some are brief non-MARC, and many resources have card catalogue entries only. Not all bibliographic records have been through authority control. A major project is underway to convert legacy data. Significant or unique categories of resources have been identified to put through onsite checks followed by authority control and RDA conversion by our authority control provider. Onsite data clean-up targets problems that an automated process cannot do. For example, we check that every record has an item, order or checkin attached and is not wrongly coded for deletion. We input 006 and 007 fields in older records and input separate preferred title fields for translated works. Relationship designators are input wherever possible, largely on an ad hoc basis.

Results so far

So far 151,786 bibliographic records requiring authority control or RDA conversion have been put through this process. Only 774 records (0.45%) had no changes. The RDA conversion includes replacing obsolete 440 series with a 490#1/8XX combination, updating title, imprint and description fields, and creating appropriate content, media and carrier fields. Data clean-up includes updating obsolete country codes, standardising relator terms with RDA relationship designators, correcting non-filing indicators in the title field and removing initial articles from preferred title fields. The bibliographic data included 605,054 access points; 67% newly matched an LC authority record, 6% required modification and 27% were unrecognised. As a result we received 40,352 new LC name/title authorities and 6,810 new subject authorities. The work we are doing on disambiguating personal names will result in a higher success rate for name authorities.

Items in the card catalogue

We have three processes underway to create brief records for resources in Rare Books and Special Collections that have catalogue cards only.

- Cards are scanned and the data is converted into MARC using OpenRefine and MarcEdit.
- Staff working at the Rare Books service point copy and paste data from scanned cards directly into a bibliographic record template in Sierra during quiet times.
- Cards for the detective fiction collection have been made available through a web site and we have invited the community to input data through crowdsourcing.

The resulting brief records are suitable for inventory control, are keyword searchable in the catalogue and are placeholders for a full record. A very practical decision was made that temporary brief data is better than no data.

Linked data in the catalogue

The next step in our preparation for linked data is to enrich bibliographic records with URIs for all access points that are associated with an LC authority. LC is a trusted source of data in the global environment; the LC Linked Data Service is the source reported as most consumed by the respondents to the Linked Data Survey (Smith-Yoshimura, 2014). Inserting the URI representing the authority record in subfield 0 of bibliographic records is a free service with authority control by our authority control provider. Sierra has the capacity to store $0 URIs, but AACP is not yet able to flip variant access points to the valid form when authority records are loaded. Once this is resolved we will capture $0 URIs and suppress them from the public view to be stored for future LMS infrastructure development and functionality.

Planning for linked data has galvanised us into action to address deficiencies in our data through bulk, efficient and cost-effective data remediation. We know that we cannot remediate all legacy data and that not all access points will match an LC authority, but this will not prevent us from doing what we can. The goal is to enable our significant or unique resources to become more discoverable and interoperable. “From a single work, we can extract relationships from co-authors, citations, geo-location, dates, named entities, subject classification, institution affiliations, publishers and historical circulation information. From these relationships, we can connect to other works, people, patents, events, etc.” (Teets & Goldner, 2013).

The authority work librarians have done over many decades to establish unique identities is now proving its worth; it is fundamental to the creation of linked data entities. The immediate benefit for us is more reliable data, more authorised access points and more references providing helpful information delivered 24/7 to library clients through the catalogue. More than that, it is preparedness through data cleaning and enhancement to build the foundation of reliable links and better discoverability for the linked data future.

Linked data for research

As an academic institution, the University of Sydney is also interested in how linked data can benefit the research space. The current Australian research environment is strongly focused on cross-disciplinary and translational research as a means to find answers for some of the big issues of our time. Major funders in Australia, such as the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC) are now expecting to see more outcomes from research they've funded – both in terms of Open Access publications and open data. The publication component has been removed from the annual HERDC collection from 2017, while ERA will look more towards the *impact* of Australian research (Australian Research Council, 2015). With the push for research impact, a need to find collaborators, and having your work discovered by researchers in other disciplines, it's increasingly important for researchers to be able to show how their research has been used, by whom, and how.

Research funded by the ARC and NHMRC is assigned grant IDs and these are listed in IR records with a persistent URL (PURL) generated for the ID and harvested by Trove. Increasingly research outputs such as journal articles are being assigned Digital Object Identifiers (DOIs), and with a DOI for a publication and the PURL for a grant it's easier to link publications to their respective grants. The RD-Switchboard is aiming to connect datasets across multiple registries, by making use of information such as authorship, publications and grants (RD-Switchboard, 2016).

The ARC and NHMRC also make use of the Australian and New Zealand Standard Research Classification (ANZSRC), meaning that Field of Research (FOR) codes assigned to a research output have the possibility of being linked to other outputs from the same disciplines. The CSIRO Linked Data Registry (CSIRO, 2016) contains FOR codes with stable HTTP URIs, which could be used in an institutional repository as a step towards linked data.

In 2015 the ARC and NHMRC released a joint statement encouraging the use of ORCID identifiers by researchers in applying for funding (National Health and Medical Research Council, 2015). Author name disambiguation continues to be a major challenge for institutional repositories and ORCID IDs can play some role in trying to resolve this issue. It also means that researchers who have moved across multiple institutions have a better way to try to pull together their research housed in disparate systems through a single identifier. From a repository perspective, use of an ORCID ID for authors provides an HTTP URI for future linked data possibilities.

Within this research environment, the University of Sydney Library promotes and preserves research outputs, such as journal articles and research data, through the institutional repository, Sydney eScholarship (SES), and the Sydney Research Data Registry. Linked data in the research space provides opportunities for discoverability of content, as well as ways to repurpose data to show interconnectedness (e.g. the RD-Switchboard) and for researchers to be able to demonstrate how they have collaborated outside of their discipline. The University of Sydney is reviewing its existing repository model and will be implementing a new repository framework to support research outputs – including journal articles, non traditional research outputs (NTROs) and research data. This provides an opportunity to consider how to approach Linked data in this repository framework, but could also be considered by those who want to publish their repository data as linked data.

To sum up

Linked data is a useful tool to capture the value locked up in our metadata. Search engines will find and create a multipurpose infrastructure to build new options for collating data where new questions can be asked and new connections can be made. We will not only expose our significant resources, but we will also be able to gather geographically disparate items into a single global collection or exhibition. We need to expose the many other resources we provide, such as lectures, research data sets, multimedia collaboration spaces and research skills training. Building relationships with other data producers will better manage our physical and intellectual assets across the University’s faculties, institutions, museums and galleries. We must advocate for change and share the lessons we have learned to be central to the development of vocabularies, the preservation of linked data, and publishing data sets. This will make us central to the university’s scholarly endeavours and output. The intellectual and financial investment in these assets will ensure that the library will thrive into the future.

References

1. Australian Research Council (2015). *ARC welcomes new measures to boost innovation*. Retrieved from http://www.arc.gov.au/news-media/media-releases/arc-welcomes-new-measures-boost-innovation

2. Coyle, Karen (2016, February 16**).** More is more. *Coyle’s InFormation*. Retrieved from http://kcoyle.blogspot.com.au/2016/02/more-is-more.html

3. CSIRO (2016). *CSIRO Linked data Registry*. Retrieved 19 August 2016, from http://registry.it.csiro.au/def/keyword/anzsrc

4. Krafft, Dean B. (Last modified by Lynette Rayle June 01 2016). *Why Linked data?* Linked data For Libraries (LD4L). https://wiki.duraspace.org/pages/viewpage.action?pageId=43910411

5. National Health and Medical Research Council (2015). *NHMRC and ARC Statement on Open Researcher and Contributor ID (ORCID)*. Retrieved from

https://www.nhmrc.gov.au/grants-funding/policy/nhmrc-and-arc-statement-open-researcher-and-contributor-id-orcid

6. *RD-Swithboard* (2016). Retrieved 19 August 2016, from http://www.rd-switchboard.org/

7. Sheih, Jackie & Reese, Terry (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* 15(3-4), 208-226. doi: 10.1080/19386389.2015.1099981

8. Smith-Yoshimura, Karen (2014). *Linked data Survey Part 1*. Retrieved from OCLC Linked data Research http://www.oclc.org/research/themes/data-science/linkeddata.html

9. Teets, Michael & Goldner, Matthew (2013). Libraries' fole in curating and exposing big data. *Future internet* 5(3), 429-438.

10. Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, 9(1)