DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Designing a Linked Data Service across borders and timezones: the National Library Board's experience

Robin Dresel
National Library Board,
Singapore
robin_dresel@nlb.gov.sg

Abstract

This paper presents the implementation of a Linked Data Management System (LDMS) at the National Library Board Singapore (NLB), which aims to provide a unified view of bibliographic descriptions from diverse collections across the National Library, the Public Libraries, and the National Archives. The National Library Board worked with a vendor to convert metadata from multiple sources into entities in a triplestore for use in resource discovery. This paper will outline the challenges faced and lessons learnt in the production of Linked Data and improving data quality. The global pandemic forced the team to work remotely with an overseas vendor, which compounded the complexity of communicating relatively complex concepts and troubleshooting data issues. Great emphasis was placed on determining causes for data issues and correcting these. Challenges arose from the insertion of URIs into the source records to identify entities matching the string label as the reconciliation of entities extracted from different source systems is based on similarity of the name label and associated properties. The paper concludes with an outlook on the continuous refinement of data quality and the development of public interfaces to demonstrate the benefits of Linked Data to stakeholders. The development of this service inserts the discovery aspect directly into resources, demonstrating the potential of Linked Data to shape future services.

Keywords: Linked Data; Knowledge Graph; Entity Extraction; Schema.org; MARC21; BIBFRAME; Data Management; Data Quality; Data Refinement; Public Interfaces; Data Consistency; Entity Reconciliation; Data Transformation

1 Background

Libraries and other cultural institutions have a long history of making their resources available to the public through their catalogues (Lerner, 2006). However, the MARC (Machine-Readable Cataloguing) standard, which has been used by libraries for the past 60 years (McCallum, 2002), is unable to fully take advantage of the benefits of the World Wide Web such as modelling information as a network of entities based on relationships (Styles et al., 2008). This is because MARC was designed for card catalogues that manage library-owned resources, and not for sharing information online.

In contrast, the internet has enabled new ways of sharing information, including Google's Knowledge Graph. The Knowledge Graph leverages relationships between entities to relate information based on semantic context, allowing for better search results (Singhal, 2012). For example, a search for "Mona Lisa" can retrieve "Leonardo Da Vinci" as the creator, even if the search query did not mention the painter.

To take advantage of these benefits, libraries have begun using Linked Data to publish their resources. Linked Data is a data model that enables discovery by following relationships between entities. The Library of Congress has published the BIBFRAME ontology, which is a Linked Data model that allows institutions to create Linked Data catalogues. Other domains are also developing related ontologies to enable entity-based discovery.

https://doi.org/10.23106/dcmi.953335608

Unlike traditional catalogues, which require users to search for a known concept, a Linked Data Discovery Interface enables previously unknown relationships to be surfaced. Additionally, Linked Data supports resource sharing (Bizer et al., 2009). Instead of duplicating descriptions of entities, institutions can simply refer to them through a "same as" relationship, saving time and reducing redundancies.

Overall, the adoption of Discovery Interfaces based on Linked Data is a significant development in the cultural sector's efforts to make their resources more widely available, and to support collaboration and sharing across institutions.

2 Concept

2.1 Intent

Linked Data enables the display of entities from different data sources in a combined interface. The National Library Board Singapore (NLB) aims to bring together resources, both physical and digital, from its diverse systems across the National Library, National Archives, and Public Libraries into its Linked Data Management System (LDMS). To achieve this, a Discovery Interface was developed as an experimental Linked Data catalogue. The structure of the chosen solution allowed the data management interface, in which staff would be viewing and managing all entities with their respective nodes and edges, to duplicate the structure of the discovery layer.

To align the different vocabularies, Schema.org was chosen for all data sets to be mapped to. This ensures that the data, once exposed to the web, is in a format that search engines can easily understand. Since the library data available in MARC21 is transformed into BIBFRAME, both print and eBook collections are now available in both BIBFRAME and Schema.org in the system.

The double-conversion of MARC data into BIBFRAME and then into Schema.org was chosen to accommodate data sharing across institutions that are and will be working with BIBFRAME, while the conversion to Schema.org would assist search engines' understanding of the data exposed in such a way.

To achieve this, metaphacts GmbH from Germany was awarded the contract in 2021 to form a consortium together with Data Liberate (UK) and KewMann (Singapore) to convert metadata into entities in a knowledge graph and to develop and maintain the LDMS. While metaphacts was bringing the low-code, configurable platform, and the workflows to the table, Richard Wallis from Data Liberate was appointed as the project consultant. KewMann, located in Singapore, was charged with the business integration work, looking after infrastructure and implementation.

2.2 Scope

The metadata to be merged into the Knowledge Graph are spread over different source systems and are available in different data formats that need to be transformed to form a consistent graph based on entities. The source systems are:

- the Integrated Library System (ILS), which stores all MARC21 records of the physical and eBook collections across National and Public libraries,
- the Content Management System (CMS), which stores the Dublin Core (DC) records for the digital resources as well as the DC records converted from ISAD-G archival standard that describe the collections of the National Archives
- authority records used by cataloguers that are stored locally and exported to CSV-XML.

To harmonize the data sets, the vendor set up conversions across different pipelines as part of the entity extraction process. This included the MARC21 to BIBFRAME conversion using the MARC2BIBFRAME scripts published by the Library of Congress. The data model and the scripts were not adjusted for local purposes but taken as published to align with a widely adopted model that enables potential sharing across libraries. The output in BIBFRAME was further converted

https://doi.org/10.23106/dcmi.953335608

into Schema.org, which was chosen as the base vocabulary for the knowledge graph. The conversion was based on the Bibframe2Schema.org W3C community group, with updates planned to be published back there as well.

Bibframe 2.0	Schema.org Equivalent
WORK	
bf:MusicAudio bf:Monograph	schema:AudioObject + schema:CreativeWork
IF	
bf:content https://id.loc.gov/vocabulary/contentTypes/prm.html	schema:MusicRecording
IF	
bf:content 'content:https://id.loc.gov/vocabulary/contentTypes/spw.html	schema:AudioBook

FIG. 1. Example mapping of Bibframe 2.0 in Schema.org through the help of a table.

For the conversion of Dublin Core metadata to Schema.org, the vendor designed a custom mapping and conversion script, with the output reviewed by the librarians. Following the extraction of the records and the corresponding entities, a complex process was designed by the vendor and NLB's technology team to handle ingestion and deduplication of entities. This process had to account for the different source systems as well as the integrity of the Knowledge Graph (KG) over time. As the source systems were considered the ultimate source of truth, the structure was designed to allow for the KG to be updated automatically upon ingestion of updated data and manually, without losing data or accidentally reversing edits. Occurrences of duplicate entities derived from the various source records were merged based on the external "same as" URIs received from the respective source records. The matching criteria used for the strings was the name label and a matching second property such as birthdate. An <owl:sameAs> relationship was also considered a relevant second property.

3 Challenges

3.1 Working Across Borders

As the development work was scheduled for 2021 and 2022, the global pandemic saw the team working around restrictions surrounding onsite meetings. At that time, we thought that working with an overseas vendor with a local team to execute the essential manual data transfers, would enable NLB to source internationally for the best fit based on the established needs.

However, communicating complex concepts such as system structure, ingestion pipelines, mapping and troubleshooting data issues was a significant challenge to overcome online. The time difference compounded the complexity, as scheduling was limited around common timeframes in the late afternoon Singapore time, which forced the teams to work more independently than expected. Troubleshooting was usually done via screenshots, which were then sent to the vendor to assess. For the most complex issues, live sharing sessions were arranged.

The complexity was further increased through security measures that prevent overseas access to the production system. As a result, the vendor had to reproduce the issues reported in a staging environment, which became increasingly difficult with progressive data updates to the production environment.

3.2 Data Quality

Any service based on the LDMS can only be as good as the underlying data. If the data is inconsistent or inaccurate, any service that builds on it will inherit these issues. Therefore, great emphasis was placed on determining causes for data issues and correcting them. The types of problems can be largely categorized into two clusters based on their characteristics.

https://doi.org/10.23106/dcmi.953335608

3.2.1 Source Data

Source data problems refer to issues that already exist in the systems from where the data originates. The process of aggregating data to be displayed as entities increases the visibility of source data issues. For example, the omission of an article like "The" in the title of a specific item may go unnoticed in a traditional catalogue. But due to the nature of the data model, aggregation of such information will make this very apparent. For instance, "Lord of the Rings: Return of the King" vs. "The Return of the King".

Other issues arose from cataloguing standards being refined over time. This led to new items being catalogued following the new standard while the existing collection would retain records crafted under the old standard. In NLB, retrospective updates to cataloguing records are rarely practiced due to resource considerations. As the conversion scripts are generally based on the most recent version of the source format, older catalogue records may lead to errors when transformed.

Similarly, blanks, specifically at the end of a field, may lead to conversion problems. For example, the MARC tag 337 subfield a, where the label "rdamedia" is followed by a blank " " where subfield $a=audio. These cases would fail to transform using the MARC21 to BIBFRAME conversion scripts published by the Library of Congress.

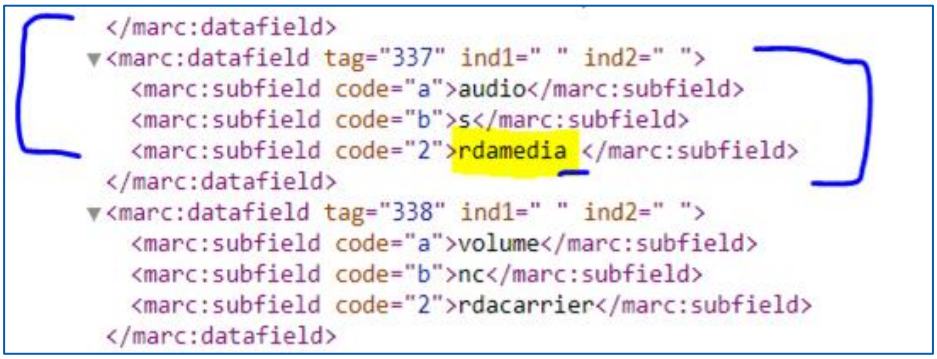

FIG. 2. A blank in the source data record causes the automated conversion script to fail.

A relatively common issue encountered is the generation of multiple entity types. A concept could end up being both a "person" type and an "organization" type, as not all metadata schemas identify an entity distinctly as a "person" or an "organization", but simply accept "agent" as a higher-level category. Since Schema.org as the chosen vocabulary for display does not cater to the type "agent", the system will assign both entity types and the issue has to be rectified through manual review.

Another strikingly incorrect display resulted from incorrectly referenced URIs from the source data. When the Library of Congress (LoC) "Person" URI is appended by the opensource tool MARCEdit to the end of a Person in a subject string with subdivisions, the Person LoC URI became associated with the whole subject string including the subdivisions. The ETL (Extract, Transform, Load) process used then resulted in Alternate Names identified and displayed together for example,

"Lee, Kuan Yew, 1923-2015--Political and social views",

"Lee, Kuan Yew, 1923-2015--Biography",

"Lee, Kuan Yew, 1923-2015--Interviews",

3.2.2 Aggregation

In the reconciliation process, entities from different source systems are merged based on similarities in their name labels and associated properties. One of the key advantages of Linked Data is the reduction of duplicate entities. Google deals with this issue in a similar fashion, by aggregating information across multiple sources for a given entity in their own knowledge graph. This is displayed in the knowledge panel on the right side of the search results.

https://doi.org/10.23106/dcmi.953335608

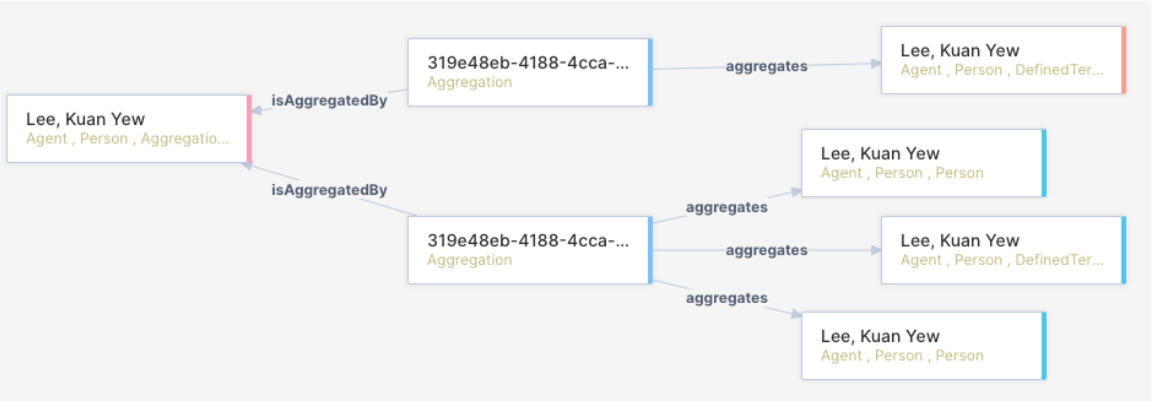

FIG. 3. Primary (left) and static (centre) aggregations based on entities extracted from source records (right)

The challenge is to find the right balance that merges as many duplicate entities together without merging entities that are distinct and should not be merged. This is an iterative process with different stages. Most of the reconciliation work is done through an algorithm, considering several factors, such as string matching and the use of additional matching properties. When this process reaches its limit, patterns can help the team to correct the remaining reconciliation issues. In a final step, manual verification will settle those entities that are not merged or are wrongly merged.

However, a significant challenge is presented when entities lack sufficient data, as they cannot be easily reconciled with other entities that carry the same name label. For instance, "Catherine Lim" is not a unique name, making it difficult to automate the disambiguation process for entities with identical name labels.

Additionally, issues arise when URIs are inserted into source records to identify related entities matching the string label. For example, a person entity's name could be merged with the description and photo of another entity. This leads to confusion when searching for entities, as in one case the system presented two "Lee Kuan Yew" entities instead of one, and "Goh Chok Tong" was not found in the search results. The underlying issue was traced to incorrect LoC URIs assigned in the source MARC21 record.

An overly enthusiastic algorithm initially merged too many entities, causing "Catherine Lim" to be merged with "Catherine, Duchess of Cambridge," "Catherine II, Empress of Russia," and others. In another case, the label "JJ Lin" caused problems due to its brevity. The team adjusted the algorithm to recognize and manage unique labels with less than three characters. Adjusting the reconciliation algorithm needs to be done to determine the optimal level of aggregation.

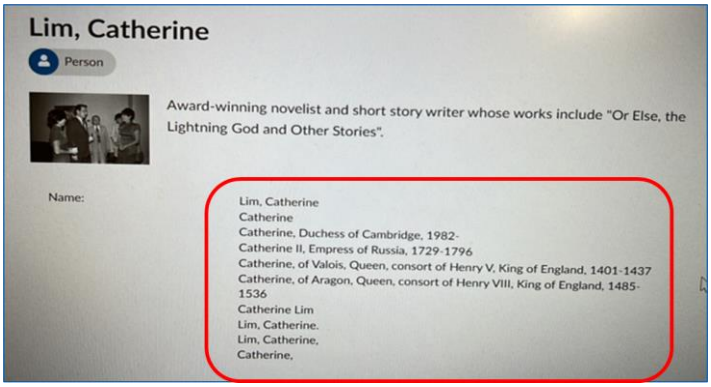

FIG. 4. Over-aggregation of an entity was corrected by introducing the honorary prefix into the algorithm.

Another process that saw some refinements was the way the MARC21 to BIBFRAME scripts assign URIs to entities found in source records. For instance, URIs would be assigned based on the position of a contributor, so the first contributor would receive URI#1, the second URI#2, etc.

https://doi.org/10.23106/dcmi.953335608

If in the process of updating the source records, the positions of the contributors were swapped, the script would then assign new URIs to these entities, as it would not recognize them due to the change in the order. A change in the algorithm is required to check the headings when assigning URIs to avoid creation of duplicate entities.

4 Outlook

The development of the system was completed in December 2022, allowing the team to focus on refining two key aspects to prepare for a public launch.

4.1 Data Quality

Following the launch of the minimum viable product, the team realized that the quality of the data needed to be improved to meet the high standards that NLB's customers have come to expect. This was particularly important for local entities such as politicians, authors, and other public figures, as well as places and organizations. It was essential to ensure that the information presented is accurate and consistent, as inaccuracies could damage the library's reputation as a trusted source of information.

Even while data quality work is an ongoing effort, the team decided to make a concerted attempt to enhance the consistency of the data. One proposed solution is to use exclusive sources to define entities. For local authorities, only the local authority file will be considered as the valid source, while for international entities, LoC URIs will be the single source of truth. This approach is expected to reduce conflicts resulting from multiple data sources, such as birthdates in differing date formats displayed for a person. The local authority control system would store the date with the month and day included, while the ILS would only have the birthyear available. Eliminating multiple sources in aggregating data would solve this issue.

This approach will also define the concepts to be matched against, making entity reconciliation through string matching easier by removing potential incorrect URIs. An aggregation from multiple sources could inherit the flaws of any data source and increase the inaccuracy of the data. In a later step, data from other sources can be introduced carefully to enhance the entities with the mentioned risks in mind.

4.2 Public Interfaces

The team also planned to enhance the public interface before launching them. The Linked Data catalogue is being developed to give a guided view into the entities and related resources. It can also be used to develop services that allow customers to be connected directly to resources from within the themed websites. Through such a service, an article can be linked to related resources based on the relationships that exist within the KG. For example, an article on "Stamford Raffles' career and contributions to Singapore" could be linked to resources about the subject "colonial administrators" and the person "Stamford Raffles". Additionally, a knowledge panel can be displayed, highlighting key dates and facts about Raffles as a person, such as birthdate, spouse, etc.

Developing this service together with the National Library as a proxy for the customer needs will insert the discovery aspect directly into the resources, allowing customers to benefit from the system without changing their behavior. It also illustrates the benefits of Linked Data to staff, generating ideas on how it can be used to shape future services. It is the first step in an ongoing journey towards a linked data discovery.

5 Conclusion

Embarking on a Linked Data journey is an endeavor that requires know-how and patience with data challenges that are numerous and complex.

https://doi.org/10.23106/dcmi.953335608

Having said that, the advantages of providing data in RDF have not been fully realized, with many collection owners yet to grasp the benefits of this new data service. Whereas the use of knowledge graphs has seen various enterprise users benefiting already from a linked data structure.

However, if we don't make a first step now, to create a sandbox that can be explored and understood, we are short-changing ourselves, as the benefits of such an approach will be seen in the long run through the connection of entities and thus resources, enabling discovery of relationships not known yet. Building the LDMS is our first step in this journey.

References

The Library of Congress. (n.d.). BIBFRAME 2.0 Vocabulary List View - LC Linked Data Service: Authorities and Vocabularies | Library of Congress. Retrieved August 16, 2023, from https://id.loc.gov/ontologies/bibframe.html.

Bizer, Christian, Tom Heath, Tim Berners-Lee. (2009). Linked Data: The Story so Far. International journal on Semantic Web and information systems, 2009, 1-22

dataliberate.com. Retrieved August 16, 2023, from https://www.dataliberate.com/.

KewMann | AI Software & Big Data Analytics Applications – KewMann. Retrieved August 16, 2023, from https://www.kewmann.com/.

Lerner, Fred. (2006). The Story of Libraries: From the Invention of Writing to the Computer Age. Retrieved August 16, 2023, from https://books.google.com.sg/books?id=YVhWu06J5j0C&pg=PA48.

McCallum, S. H. (2002). MARC: keystone for library automation. IEEE Annals of the History of Computing, vol. 24, no. 2, 2002, 34-49.

metaphacts - Decision Intelligence through Knowledge Democratization. Retrieved August 16, 2023, from https://metaphacts.com/.

Schema.org. Retrieved August 16, 2023, from https://schema.org/.

Singhal, Amit (2012). Introducing the Knowledge Graph: things, not strings. Retrieved August 16, 2023, from https://blog.google/products/search/introducing-knowledge-graph-things-not/.

Styles, R., Ayers, D., & Shabir, N. (2008). Semantic Marc, MARC21 and The Semantic Web. LDOW.

https://doi.org/10.23106/dcmi.953335608