

A Semantic Knowledge Graph Aggregation of Library Resources

Richard Wallis

Data Liberate, 10 Yessell Lane, Hinton-on-the-green, Evesham, Worcestershire, WR11 2RE, United Kingdom

Abstract

In December 2022 the National Library Board Singapore (NLB) launched a continuously updated, Linked Data based, semantic Knowledge Graph (KG) to manage and aggregate resources extracted from their library management, authority management, National Archives, and, content management systems. Since then, developments have proceeded to utilise the data and functionality from the KG for sharing across the web and embedding in other NLB hosted services. Additionally, processes have been implemented to use external authority services, such as the Library of Congress, to enrich and improve the data quality of KG entities.

Keywords

Linked Data, Knowledge Graph, Digital Libraries, Authority Control, Web, User Interfaces

1. Introduction

This paper describes the background to, architecture, development, and operation of a Linked Data based semantic Knowledge Graph (KG) implemented by the National Library Board Singapore (NLB)².

In 2021 NLB issued a tender document for the development of a “Linked Data Management System and Discovery Interface”. After a competitive process the contract was awarded to a group of three organisations: metaphacts³, Kewmann⁴, Data Liberate⁵.

Over the following two years, in close cooperation with data teams at NLB, the system was developed to regularly import data from established operational source systems supporting library services such as the library management system, authority management system and National Archives. The linked data system (KG) was launched in December 2022.

In daily operation since that time data curator members of NLB’s team have utilized its data management interface to refine and enhance the entity descriptions and relationships within the KG. These having been automatically created from the daily imports from source systems.

* Corresponding author.

✉ richard.wallis@dataliberate.com (R. J. Wallis)

ORCID 0000-0001-8099-5359 (R. J. Wallis)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

² National Library Board Singapore <https://www.nlb.gov.sg/>

³ Metaphacts GmbH <https://metaphacts.com>

⁴ KewMann Pte Ltd <https://kewmann.com>

⁵ Richard Walls independent consultant trading as Data Liberate <https://dataliberate.com>

Having established a foundational context, by briefly describing the system's history, initial development, and operation; this paper goes on to focus on further developments beyond the basic KG.

2. History

The issuing of the invitation to tender for a linked data system came at the end of a period of several years of investigation, experimentation development and analysis of linked data techniques and potential, carried out by the Discovery Services Team at NLB, these included several projects such as:

2.1. Bibliographic Triplestore

An implementation of a linked data RDF-based triplestore[1], using an early release of BIBFRAME[2] as a format, populated by records converted from Marc data exported from NLB's Civica library management system.

Utilising a bespoke interface into the triplestore and read-only SPARQL[3] queries, Discovery Services Team members gained experience and understanding as to the linked data descriptions and relationships that can be derived from Marc format records.

2.2. Schema.org Training and Experience

Training courses, associated materials, and consultancy advice, provided by the author, introduced the Discovery Services Team to the use and potential of the Schema.org[4] vocabulary for the description of bibliographic resources and their sharing in the Web in a way that enables the consumption by search engines such as Google.

2.3. Schema.org Publishing

The development of a static website representation of the records within NLB's library system, public user interface (OPAC).

For each record in the OPAC a simplistic single web page was created including a link to the associated page in the OPAC.

Embedded in the html markup of each page was a schema.org description of the record for use by search engine crawlers.

The website did not include a user search facility, meaning users arriving at an individual page would only do so by following a link from another site or a search engine result.

Having used standard tools to invite Google to crawl the site, the website was hosted for several years. Tracking traffic and activity on the site indicated, from many millions of hits, and hundreds of thousands of those that resulted in onward clicks to the OPAC, that this is a potentially effective way to drive user traffic to bibliographically focused web interfaces.

2.4. Proposed Data Model for Web and Library Focused Linked Data

Consultancy report delivered, by the author, produced in cooperation with Discovery Services Team members. The report proposed a RDF-based linked data based data model for describing bibliographic resources utilizing a combination of two 'standard' vocabularies:

- BIBFRAME 2.0 [5]- A widely adopted RDF vocabulary issued by the Library of Congress with open source tools available for the conversion of Marc records into BIBFRAME format.
- Schema.org – The *de facto* vocabulary for the sharing of data on the Web in a manner that is easily consumed by search engine crawlers.

3. Requirements for a Linked Data System

Based upon the experience of these and other projects, and the environment of user accessible services provided by NLB, the Discovery Services Team produced the invitation to tender for a linked data system.

Explicit and implicit in the tender document are several technical and operational requirements that shaped the resulting KG system.

3.1. Combined Open Data Model

A data model based upon the open vocabularies BIBFRAME 2.0 (to capture, where appropriate, fine bibliographic description data potentially derived from source Marc data) and Schema.org (to capture generic data for all entities regardless of source for potential sharing in a web friendly form).

3.2. Automatic Ingest of Data from Multiple Sources

The system will be required to import and convert into a standard data model, the data from several operational source systems:

3.2.1. Library Management System (ILS)

MARC-XML formatted files describing records held in the library management system hosted by NLB. An initial dump of all records held in the ILS to be processed, followed by regular ‘delta’ update dumps containing changes (additions, updates, deletes). Regular delta dumps would be received daily and often contain many thousands of records.

3.2.2. Authority Control System (TTE)

CSV formatted output from the authority control system which maintains the NLB authority file of Singaporean related entities – persons, organisations, places, etc. The rate of change of data in this system is such that an export of the current state would be required monthly.

3.2.3. Content Management System (CMS)

NLB hosts several public webservices serving several needs:

- Infopedia – Singaporean-focused encyclopedia
- Singapore Arts
- Music SG – Singapore-focused music
- History SG – Singaporean history
- BiblioAsia – Ejournal

- Picture SG – Singapore related images

All these sites are hosted and managed by the CMS system, which are exported in Dublin Core (DC) formatted files on a monthly basis.

3.2.4. National Archives (NAS)

Dublin Core (DC) files, derived from ISAD-G stored data, exported on a monthly basis.

3.3. Source System Cataloging

The cataloguing processes and practices for each source system will remain in place. Those systems will remain the source of truth for the description of NLB's resources.

For example, Marc based cataloguing is, and will continue to be, the primary method of record management in the ILS. Changes made in the ILS should be reflected in the linked data system as those changes are described in delta update files.

3.4. Entity Reconciliation and Consolidation

Duplicate references to a single real-world entity created from records imported within, and between, data sources should be reconciled into a single entity representation. For example, individual person entities may be created for the same real-world person as ILS records for several works for which that person is the author. Equally the same person may be the subject of an authority record in the Singapore Authority Control system (TTE), and or the subject of an article in the CMS system.

Such duplicate source entities should be reconciled together to produce a single 'primary entity' for presentation in user interfaces.

The attributes of entities reconciled together should be consolidated to provide an aggregate view of all known attributes, for display and search indexing.

3.5. Linked Data Management Interface (DMI)

An interface for data curators to use to override reconciliation decisions made by system processes. Identifying individual entities that should or should not have been reconciled together, and correcting.

Additionally enabling the adjustment or correction of entity attribute values and or contents. Also providing an option for curators to suppress individual entity attributes, or entire entity descriptions, from search indexing and display.

Actions to suppress or change data or relationships should not be overridden by conflicting records being reimported via delta update.

3.6. Discovery Interface (DI)

A discovery interface designed for potential public users to search for and follow relationships between reconciled Work, Person, Organisation, Place and Subject entities.

Search should be refinable to specific entity types or across all types. Intuitive navigation paths, wherever possible inter-entity links enabled as clickable links. For example, Work to

Person via author, Person to Work via Works by or Works about, Place to Work via Works about.

3.7. A Cloud-based System

The developed system should not require specific hardware to be installed. The solution should be installed in the Singaporean Government Compute Cloud (GCC) – a secure implementation of Amazon Web Services (AWS) [6].

4. System Development

Working from the requirements, an 18 month development project was undertaken by the three contracted organisations working closely with the NLB Discovery Services Team in an agile way.

Based on those broad requirements, data models, processes, and user interfaces, were evolved and developed to satisfy NLB's needs both functionally and operationally. [7]

5. Delivered System

The system was created using metaphactory [8] an open standards based Knowledge Graph Platform, utilizing the Ontotext GraphDb [9] Semantic RDF database.

In December 2022 the developed system passed its acceptance tests and went into live operation. Daily, weekly and monthly updates from source systems commenced. Data curation began, managing entity reconciliation (merging/splitting as required).

Viewing data in a different system inevitably identified issues from source cataloguing. Data curators corrected these using the DMI or, more often communicated the issue to the source system cataloguing teams. Thus, improving data quality in both the KG and source systems.

A simple example of this being a Person entity having multiple and different date of birth values, derived from the consolidation of attributes from multiple source entities. Identifying the source entity containing such an anomaly resulted in communication with the responsible cataloguing team for them to fix their record. Such an update would then be reflected in following delta updates and the anomaly would resolve itself in the KG.

The Discovery Interface (DI) passed performance load, and usability tests and was accepted. It was used by the Discovery Services Team as a tool demonstrating the characteristics and benefits of a linked data service, advising future design and deployment proposals.

6. Continued Developments

As stable operation of the KG was established, future enhancements and developments were considered and commissioned. These focused on three main areas:

1. Enhancements to the DMI to improve its capabilities and usefulness in managing entities.
2. The provision of interfaces to the KG, and the data insights it provides, to two communities of potential users: librarians and similar bibliographic data consumers both within NLB and globally and users already using NLB web services such as

Infopedia offering recommendations, and intuitive navigation paths through NLB resources.

3. Enrichment and quality improvements to entity descriptions in the knowledge graph. Specifically, the use of internal, and publicly available authority, data to increase the searchability and usefulness of the display of entities for user consumption.

Several projects have since been commissioned which include the following:

6.1. Entity Data Service (EDS)

Having established and curated the KG over a two year period one next step was to share it with the wider global linked data community by ensuring the URIs used within its data were resolvable as true Linked Open Data [LOD] identifiers. To that end the development of the EDS was commissioned.

Based upon the initially delivered Discovery Interface a publicly accessible entity display interface was produced. This has been designed with librarians, cataloguers, and other data consumers in mind. The design took its inspiration from the data display interfaces of other systems, such as VIAF [10] and the Library of Congress Linked Data Service [11], designed to satisfy the needs of data focused consumers, such as librarians and not general public searching users.

It enables navigation through the data based upon the URI of the entity to be displayed. Specifically, this means a user from anywhere on the web following a URI link to an entity within the KG (for example, in the form <https://eresources.nlb.gov.sg/linkeddata/primary-entity/work/10cf19af-6d30-40ff-a862-cd3e6b2dca37>) will be presented with a representation of the data associated with that entity. An example use case being the clicking on a link to an NLB entity referenced within a Wikidata record.

The EDS facilitates navigation to the display of associated entities via clickable links within a user interface designed to support data interested users.

In common with Open Linked Data principles, data consumers are presented with options to access entity data in ways most useful to them. The web page display offers download options for common RDF serialization formats such as RDFXML & JSON-LD. Using http content-negotiation techniques, download in common RDF serializations is also accessible. Finally, the html markup of entity pages include an embedded Schema.org representation of the entity.

6.2. Sidebar API

Analysis of the data descriptions of, and more especially the relationships between, entities in the knowledge graph identified that they could add significant value to users' navigation through and consumption of NLB's already established online resources. Introducing new intuitive navigation paths within those systems to relevant resources both within the system and available in other NLB systems. For example, providing a link to a book held in the library system that is 'about' the author of an article in the Infopedia information system. Introducing navigable entity relationships that are not derivable from the system's source data.

One way of realising this value was to surface these data and links in a knowledge panel, or side-bar, in currently hosted systems focused on specific areas of NLB resources.

To simplify the embedding of such a panel in a service, and enable its potential embedding in other services, an API to access the KG independently from user interfaces was developed.

Embedding within the html page of the target service is enabled by using JavaScript code to call the API and format the returned results.

The call to the API includes the internal identifier for the item (eg. Infopedia article) on display. Upon receipt of this identifier the API looks up the entity associated with that item and then returns information for entities with an ‘about’ relationship with it.

For example, called on a page for an article describing a place, the API will return descriptions of Work entities that are ‘about’ that place. Such descriptions may for example include linking information for items in the ILS system.

6.2.1. Article Mentions

A parallel project conducted by NLB “Metadata Enrichment with Named Entity Recognition using GPT-4” used Named Entity Recognition (NER) techniques to identify, within the text of articles, references to entities in the KG. These were imported into the KG, programmatically introducing a schema:mentions relationship between the article and the relevant entity.

The sidebar API was enhanced to include these mentions relationships in its lookup process. The result being enhanced intuitive recommendations, based on data beyond that which was catalogued in the source systems.

6.3. Name Authority Lookup and Ingest

One of the goals of the curators of the KG is to wherever possible improve the quality and consistency of its data. Most data ingested into the KG was created at source by human cataloguing processes. Despite the application of standards, such as MARC21 and Dublin Core, the human element inevitably introduces inconsistencies, especially with attributes such as names. There have been several initiatives, such as VIAF and the Library of Congress Name Authority File (LCNAF) [12], to provide open authoritative name resources, to address this problem in the global library community.

[Marc] Data ingested from the ILS system often includes name authority references (URIs) to LCNAF. These are derived from \$1 or \$0 values for contributing authors.

Analysis of the Person entities derived from such data, when reconciled and consolidated with descriptions from other sources, indicate a wide variance in data quality resulting in multiple similar names for individual persons. This further results in confusing display for users and inconsistent search results.

Further analysis indicated that in the majority of cases it would be desirable for the published LCNAF data for name, data of birth, etc., to take precedence over descriptions derived from other sources.

To that end, an automatic process was implemented to identify such LCNAF identified entities, when added or updated. This process then imports from an external representation of the LCNAF file an authoritative description of the named entity.

The consolidation process was then enhanced to enable this imported external data to take precedence and, replace using the already implemented suppression mechanism for existing data.

As no data set is perfect, curators can if required, manually adjust this process for individual entities or attributes.

Further enhancements are under development to expand this approach to include lookup against the LCNAF using strings, in addition to URIs, thus enabling data quality improvements for a much broader subset of entities than those specifically catalogued with LCNAF URIs. Because of the less exact nature of string matching, this also involves human validation of potential matches, generated by automatic processes, before publishing into the KG.

7. Summary and Key Highlights

The narrative presented in this paper covers a period in excess of three years comprising many technical and operational aspects of the Linked Data System. A long successful, and still on going, project. Its success in no small measure, can be attributed to the commitment, enthusiasm and cooperation between the NLB staff members and those in the vendor team at metaphacts, KewMann and Data Liberate, working together in an agile way to develop, establish and evolve a system to satisfy NLB's needs.

Highlights from the project worth singling out:

- A Linked Data Based Semantic Knowledge Graph system – underpinned by an RDF triplestore.
- Independent from established cataloguing practices and processes – the system takes data from four separate systems, each utilizing their own data formats and cataloguing tools. No changes to these systems, or cataloguing practices, were required to enable the implementation of the KG.
- An aggregation of all NLB online and physical resources – delivering a linked data view across NLB's services providing equality of prominence between authority, bibliographic, archive, and web services resources.
- Entity Data Service – the integration of KG data into the global web of data. Directly accessible via KG entity URIs for viewing by other librarians and data scientists.
- Sidebar API – programable access to KG data, resources and functionality, delivering recommendations and enhanced navigation into existing web services.
- LCNAF lookups – using external authority services to enrich, and improve the data quality of, KG entities.

8. References

- [1] RDF-based triplestore “What is an RDF Triplestore?”
<https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-triplestore/>
- [2] BIBFRAME <https://en.wikipedia.org/wiki/BIBFRAME>
- [3] SPARQL <https://en.wikipedia.org/wiki/SPARQL>
- [4] Schema.org <https://schema.org>
- [5] BIBFRAME 2.0 <https://www.loc.gov/bibframe/docs/index.html>
- [6] AWS – Amazon Web Services <https://aws.amazon.com>

- [7] Dresel, R. (2024, January 31). Designing a Linked Data Service across borders and timezones: the National Library Board's experience. National Library Board, Singapore. DOI: 10.23106/dcmi.953335608.
- [8] metaphactory <https://metaphacts.com/product/metaphactory-overview>
- [9] GraphDB <https://www.ontotext.com/products/graphdb>
- [10] VIAF <https://viaf.org/>
- [11] Library of Congress Linked Data Service <https://id.loc.gov/>
- [12] LCNAF – Library of Congress Name Authority File <https://id.loc.gov/authorities/names.html>