Linked Data @NLB

Hanna Hussein
National Library Board
Singapore

Abstract
In early 2014, the National Library Board Singapore (NLB) started a Linked Data project with the aim of increasing the utilisation of NLB and its partners' resources through connecting related data, locally and with other datasets on the Internet. The project deliverables include the implementation of a Linked Data Management System (LDMS), developing a core Data Model and extended service vocabulary, transforming approximately 260,000 records (authorities, vocabularies and bibliographic data) from NLB, National Archives of Singapore (NAS) and National Heritage Board (NHB) into RDF, and developing a Linked Data service consisting of a web feature and a widget for the delivery of the transformed datasets, which has since been launched internally to staff and is currently under review. This paper will provide an overview of the main project deliverables, with a detailed focus on the Linked Data service.

Keywords
Linked Data, NLB, National Library Board, Web Feature, Widget, LDMS, NLB Data Model,

Introduction

Linked Data offers a unique opportunity for the National Library Board Singapore (NLB) to explore new ways in improving the information-seeking experience of its users as it allows interoperability of library resources with related data from archives, museums, other agencies, organisations, and external datasets available on the Web. This could potentially increase the visibility and the value of NLB resources.

"Linked Data has much to offer libraries if they can find ways to leverage this technology for their own uses" (Gonzales, 2014). The NLB Linked Data project thus aims to use Linked Data to enable this sharing of resources, increase the utilisation of NLB content through contextual linkages and semantic relationships, and explore new ways for NLB users to discover and access related information.

There are 4 key deliverables in this initial phase of the project:

1. Implementing a Linked Data Management System (LDMS) platform and URI Registry;
2. Developing an NLB data model (NLB DM);
3. Transforming selected data from NLB (bibliographic, authorities and vocabularies) and National Archives of Singapore (NAS) into RDF;
4. Developing a linked data service to showcase the transformed data.

1. Linked Data Management System (LDMS) and URI Registry

A key characteristic of Linked Data projects in comparison to other similar projects is in the use of RDF (Resource Description Framework) and URI (Unique Resource Identifier).
RDF is a "standard model for data interchange on the Web" and it "allows structured and semi-structured data to be mixed, exposed and shared across different applications" (RDF Working Group, 2014).

In RDF, each statement is 'deconstructed' into triples. Each RDF triple connects a subject to an object through a relationship (predicate). URIs are used "to name the relationship between things as well as the two ends of the link" or "triples" (RDF Working Group, 2014). Such a simple statement allows for maximum flexibility as any resource in the world (subject) can have a specific relationship (predicate) to any other resource in the world (object), thus allowing for limitless connections among the resources (Hooland & Verborgh, 2014).

The first deliverable that was put in place in this project is LDMS, which is a platform for the ingestion and transformation of data into RDF. LDMS also catered for the management and maintenance of the RDF'ised data, ontology, vocabularies and URIs.

URI "permits resources anywhere in the universe to be given a unique identification" and thus allows look-up of definitions for each resources and vocabularies while the RDF model allows "URIs for subjects, predicates, and objects" (Hooland & Verborgh, 2014). Hence, with RDF and URIs, computers can understand the meaning of content, rather than simply matching on strings of text and the data can be "easily aggregated and queried" and be "cross-referenced with related data" (Byrne & Goddard, 2010).

The URI Registry in LDMS thus ensures that individual pieces of data transformed from both structured and unstructured data could be uniquely identified using the http URIs, either assigned by the Registry if the entity has no URI available or to check existing entities and return the URIs (if already available).

Initially, the LDMS was implemented with basic system requirements and functionalities ("vanilla version"). However, it has evolved into a more robust platform as custom requirements were established and implemented to take into account more data ingested and possibly more services that will make use of the Linked Data.

2. NLB Data Model (NLB DM)

A data model defines the rules of structuring data for storage and retrieval. Thus, it has to be extensible, flexible, and interoperable to support varied datasets and future applications.
The NLB data model (NLB DM) is based on the Bibliographic Framework or 'BIBFRAME', which recognises "entities, attributes, and relationships between entities" and leverages the RDF modelling practice of uniquely identifying as Web resources "all entities (resources), attributes, and relationships between entities (properties)" (Library of Congress, 2012).

BIBFRAME vocabularies support a variety of data standards that NLB adheres to, such as RDA and RDF (https://www.w3.org/RDF/). It also supports vocabularies from Schema.org (https://schema.org/) which were re-used where possible. Some of the concepts of NLB DM were also mapped to ISAD-G (General International Standard Archival Description) which defines the elements that should be in the archival finding aid (International Council on

Archives, 2009), and CIDOC Conceptual Reference Model (CIDOC-CRM) which provides "definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation" (International Council of Museums, 2014).

Use cases were created to provide scenarios for how users would access information from multiple sources, include information outside of NLB that would better answer their queries. After the initial requirements for the Linked Data service were confirmed, a gap analysis exercise was carried out between the NLB DM and the service requirements, and a number of gaps in the NLB DM were noted. For example, in the NLB DM, the class 'Place' has no property for previous or historical names of a location or place in Singapore. Hence, the Linked Data service would not be able to connect to this information in its query response.

The team was advised to create an extension to the NLB DM, rather than adding on to the NLB DM which was designed as a core model and must remain so for easier management and version control. The Event Service Vocabulary (ESV) was developed as an extension to the NLB DM. It is structurally similar to the NLB DM but with properties required to address such specific service requirements. Hence, in the example given, 'Historical Name' is defined in the ESV as an additional property for the class 'Place', and the NLB DM 101 classes, 136 properties and relationships remain unchanged.

3. Data

Constraint in the project budget dictated that only a small subset (250,000 records) of NLB's and partners' data could be transformed for this initial phase. Hence, the team had to carefully consider the factors in the selection of data. Firstly, the small data subset should represent NLB's bibliographic and non-bibliographic data (NLB's Vocabularies and Authorities). Secondly, the data must be relevant to the envisaged Linked Data service. Lastly, the different datasets must have a high probability of convergence to ensure a higher chance of linking between the datasets.

The team decided on Singapore-related resources from the NLB MARC records (Singapore National Bibliographic and Singapore-related titles), NLB Dublin Core (DC) digital records for collections such as Singapore Heritage and Singapore Memory Project (SMP), and representative records from the National Archives Singapore (NAS) and the National Heritage Board (NHB) as the initial set of data for the transformation to RDF.

File formats from the various sources of raw data were also determined and standardised. Each dataset was ingested separately in the LDMS staging environment. This allowed the data team to verify the accuracy of each of the transformation processes and address any anomalies effectively. However, the verification process was not easy and posed a major challenge to the team due to the quality of the data which was not always consistent or complete and varied from agency to agency.

In straightforward instances of data errors which surfaced during data verification, back-end data cleansing at source was performed and the data was re-ingested into LDMS. If anomalies are due to internal practices and do not conform or differ from standards, the team would liaise with the source data team, discuss possible solutions and, when ready, re-ingest the data into LDMS. Once verification is completed, the data is moved to LDMS production and made available in the Linked Data service.

In order to realise the full potential of Linked Data, the NLB transformed data should also be linked to other linked datasets available in the open, semantic web. The same two criteria in selecting the internal / partner's data (relevant to the envisaged Linked Data service as well as the probability of convergence) were used as basis for selection. Another additional criterion included was the level of comprehensiveness of the dataset such that the data would complement and expand NLB's data corpus in areas where information is sparse. Thus, LCSH (Library of Congress Subject Headings), VIAF (Virtual International Authority File) and Wikipedia were selected for linking in this phase.

4. Linked Data Service

The main challenge in the conceptualisation of the Linked Data Service is how to show the value of Linked Data that is above and beyond what 'ordinary un-linked' data can do. The team realised that the uniqueness of linked data in enabling serendipitous discovery should be the focus of this service, where users can explore contents not known, and probably not discovered or retrieved, if they had gone through the usual route of getting information.

The Linked Data Service was developed based on the need to have an interface that displays 'human-friendly' Linked Data. It should also allow for multi-dimensional entry points for the exploration of entities and for contextual browsing and exploration of the varied resources and data. In this way, users can explore fully regardless of where the data is from, in a single interface, and they are able to retrieve the source data where available.

For example, if there are photographs, clicking on the photographs would bring user to the NAS Archives Online or PictureSG. If there are related MARC records, clicking on the link would bring users to the NLB OPAC for physical resources that are written by the entity (i.e. person). Likewise, if there are other resources available about the entity, clicking on the article titles in 'About' (for example) will bring users to the digital article curated by NLB staff (Fig 1).

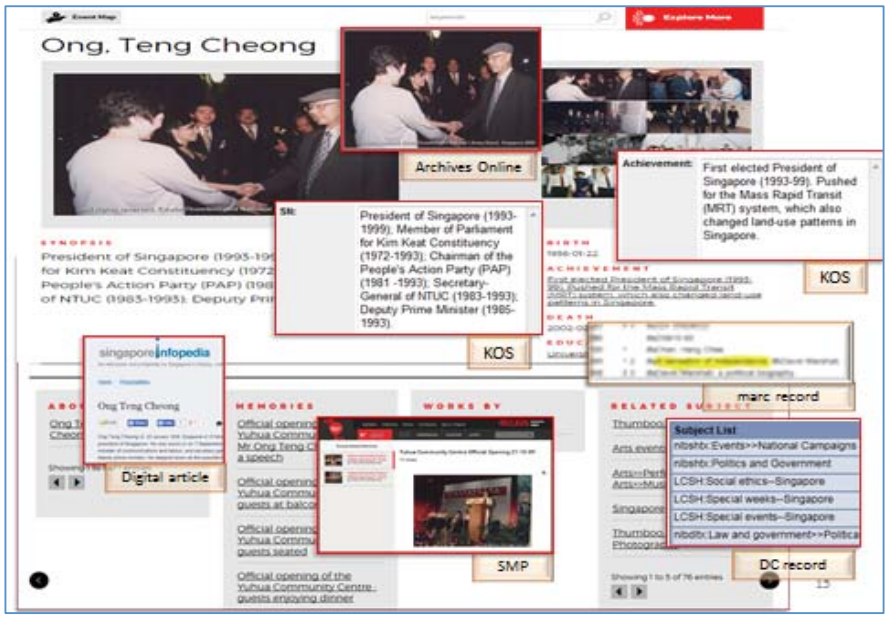


Fig 1: Entity page of the late Mr Ong Teng Cheong, displaying resources from NAS Archives Online (photographs), KOS (synopsis and attributes), digital article from NLB Infopedia, memories (public contributions via Singapore Memory Project (SMP)) and metadata from Dublin Core (DC) digital records and bibliographic (MARC) records

Lastly, the Linked Data service must also be able to display the varied resources with minimal 'clicks' and 'scrolling' of the page, and allow searching and display that is easy for the users.

4a.) Linked Data Web Feature

The Linked Data Web Feature was developed to address the above service requirements in 4 key sections, namely:

i. Entity page
ii. Entity Relationships page
iii. Entity Resources page
iv. SPARQL Search

i) Entity page

There are four classes (People, Places, Organisations or Events) featured in the Web Feature. Each entity in LDMS is grouped according to which class it belongs to, and each class has an entity page. For example, Fig 2 shows the entity page of the class 'People'.

Fig 2: Entity page of Singapore's 6th President, Mr S. R. Nathan displaying the photographs (top main section), Entity attributes (bottom main section), Event Map (top left) and 'Explore More' (top right)

Each Entity page will display the following:

Photographs:
Photographs from NLB's PictureSG website (http://eresources.nlb.gov.sg/pictures/) and NAS Archives Online (http://www.nas.gov.sg/archivesonline/) are shown if available. Clicking on the photograph leads users to the detailed page of the photograph at the respective portal (Fig 3), which would be useful if more information is required about the photographs. Hence, the 'link-out' to the other portals and websites aids in a seamless information-seeking experience for the user.

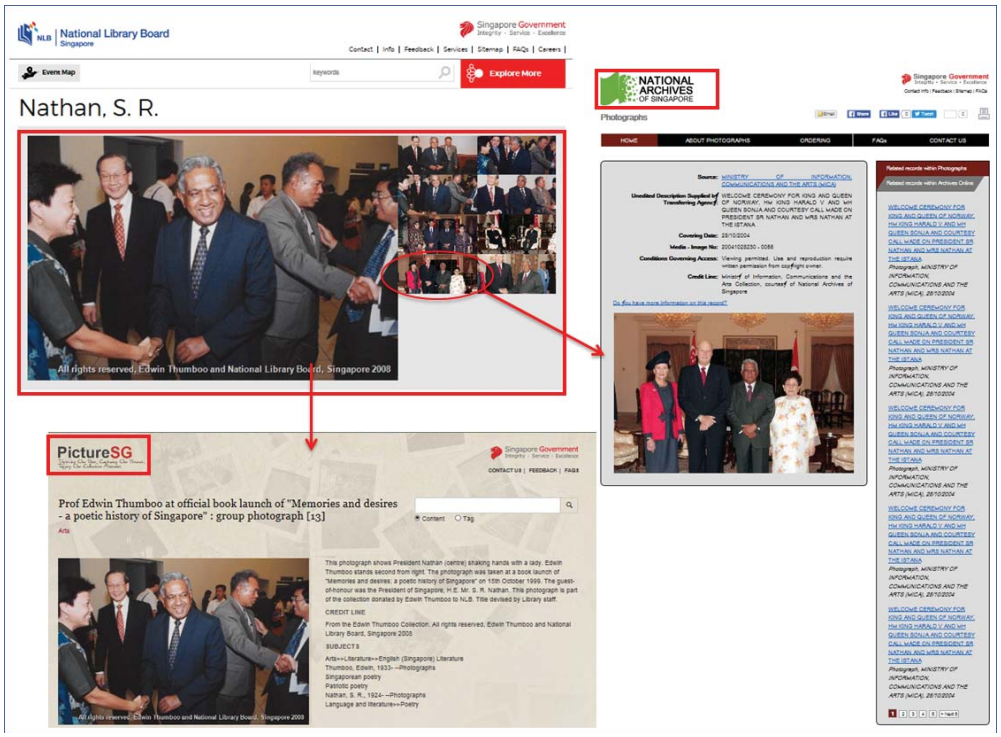


Fig 3: Related photographs of Mr S. R. Nathan from available resources

Entity Attributes:

This section displays more information about the entity. The main difference between the four types of entity pages is in the attributes or properties of the classes being displayed, as each class has a distinct and separate set of attributes. For example, attributes for 'People' would include 'Achievement', 'Award', and 'Education' (Fig 4), while entities belonging for 'Place' would display 'Historical Name' and 'Postal Code'.

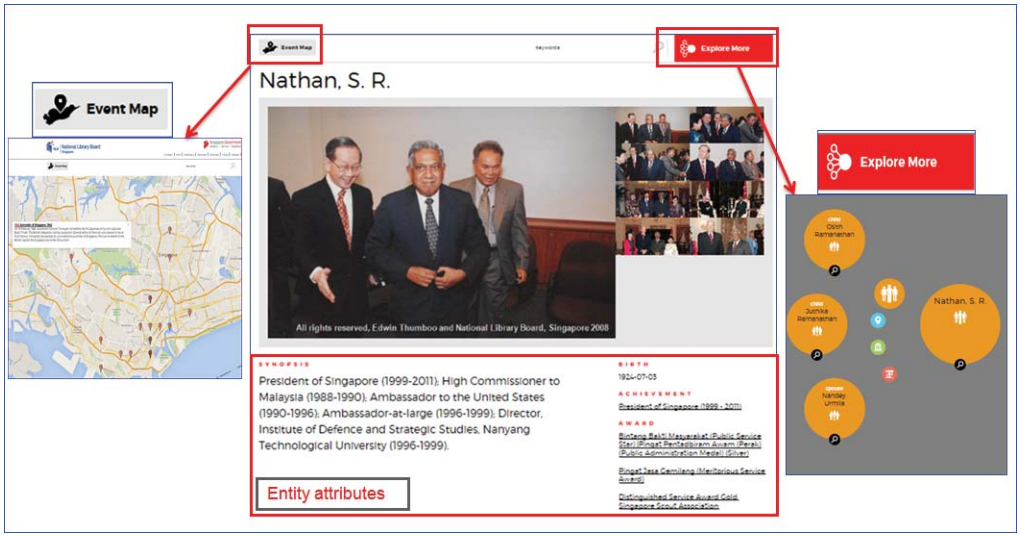


Fig 4: 'Event Map', 'Explore More' and 'Entity Attributes' at the Entity page

Event Map:
Clicking on 'Event Map' would open up the Singapore map which allows users to browse via timeline and topics. Historical events are displayed according to 12 categories (types of events), location (where the events occurred), and period (when the events happened) (Fig 5). There are currently 72 significant events in the NLB Historical Events vocabularies, such as the Bukit Ho Swee fire of 1961, Hock Lee Bus riots of 1955, Maria Hertogh riots of 1950 and the historical parliamentary general election of Singapore in 1968. The categories or types of events are listed in the right panel. Examples of the categories are Accidents, Crime, National Campaigns, and Singapore's First.

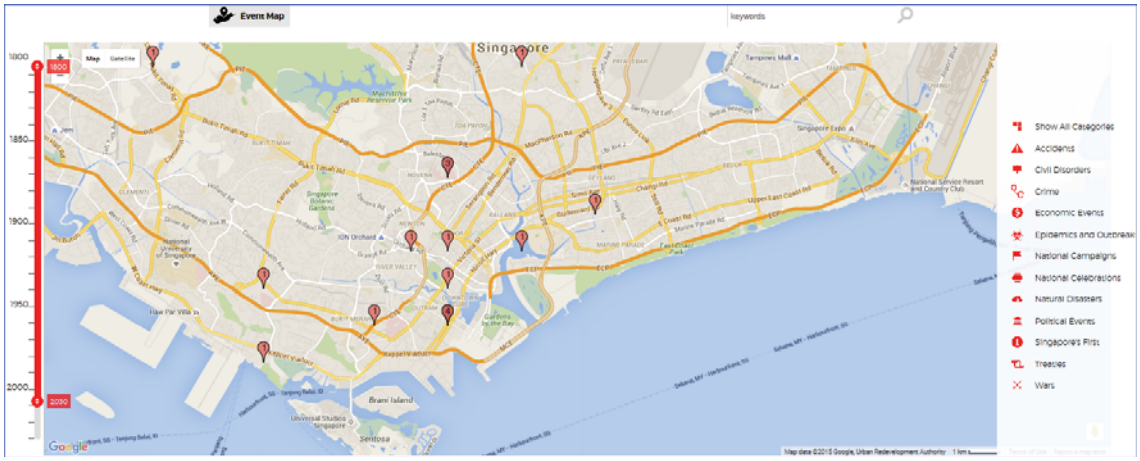


Fig 5: Events displayed on the Singapore map and the 12 categories (listed on the right panel)

The events are shown as pins on the Singapore map. The historical events would be selectively displayed on the map based on the categories selected (Fig 6). The number in the pin indicates the number of events that occurred in that location.

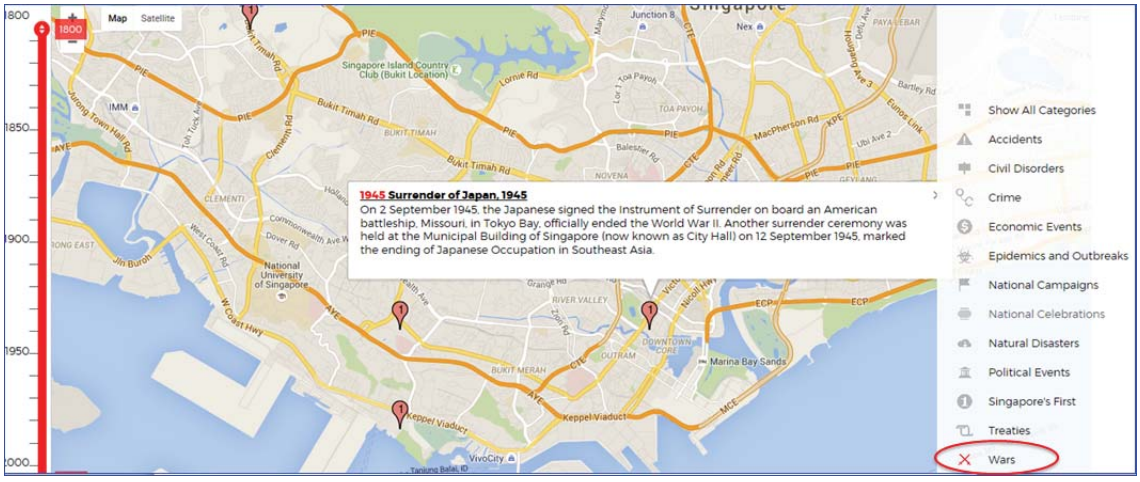


Fig 6: Events shown on the map based on the categories selected, e.g. Wars

Clicking on the event title would bring up the Entity Event page, and users can continue to navigate through the entity page for additional information.

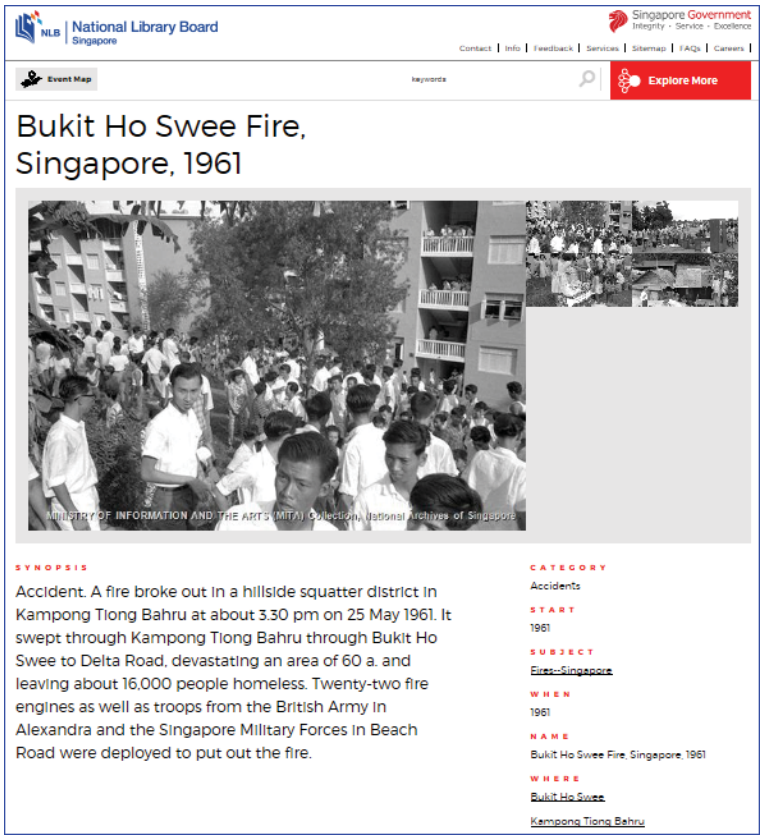
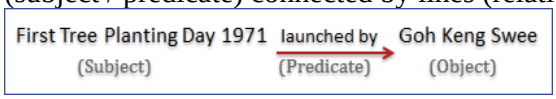


Fig 7: Entity page for historical event 'Bukit Ho Swee Fire, 1961'

Explore More:
This option brings users to the Entity Relationships page as detailed in the section below.

ii) Entity Relationships page

Clicking on 'Explore More' will open up the 'entity relationship' overlay where granular relationships between the entities in the 4 classes (People, Places, Organisations, Events) are pulled. During the initial discussions, the team attempted to visualise the triples as bubbles (subject / predicate) connected by lines (relationships). For example,



was visualised as shown below in the draft mock-up (Fig 8):

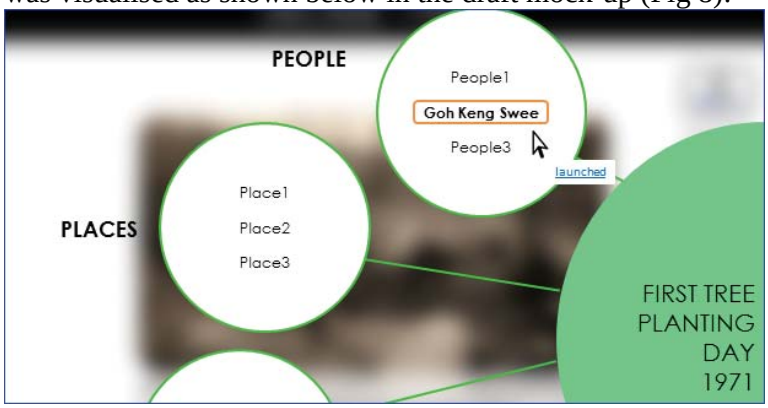


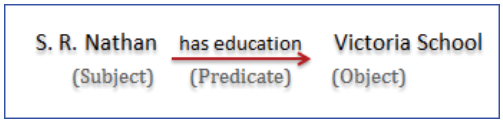
Fig 8: First attempt in visualising a triple with bubbles (representing entities) and lines (relationships)

However, there was a concern with many 'lines' in cases where the entities have multiple relationships, and this might inadvertently make the display too 'cluttered' and thus confusing for users. The team discussed and decided to remove the use of 'lines' and thus, decided that the triple such as this:

is now shown as below (Fig 9) where the relationship is embedded within each entity.

Fig 9: Visualisation of the triple for the late Mr Lee Kuan Yew and his wife, the late, Mdm Kwa Geok Choo, where the relationship 'spouse' is found within the bubble.

The web feature is also able to display an entity connected to another entity via a certain relationship that is also established with other entities. For example, Mr S. R. Nathan was educated in Victoria School:



However, Victoria School is also the place of education for Mr S. Dhanabalan (a former politician), the late Mr Devan Nair (Singapore's third president) and Mr David Lim Kim San, a 1979 Cultural Medallion recipient, to name a few. These are reflected automatically via the triples and thus, provided an interesting 'connectedness' which would not have been apparent otherwise (Fig 10).

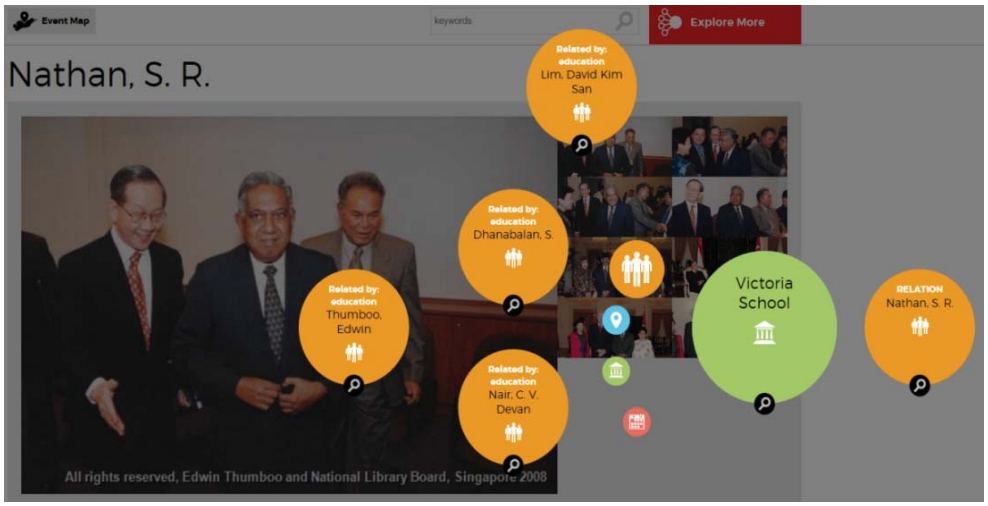


Fig 10: Mr S. R. Nathan (People) link to Victoria School (Organisation), which also has links to other entities (People)

iii) Entity Resources page

The Entity Resources page is located below the Entity Attributes section (Fig 11).

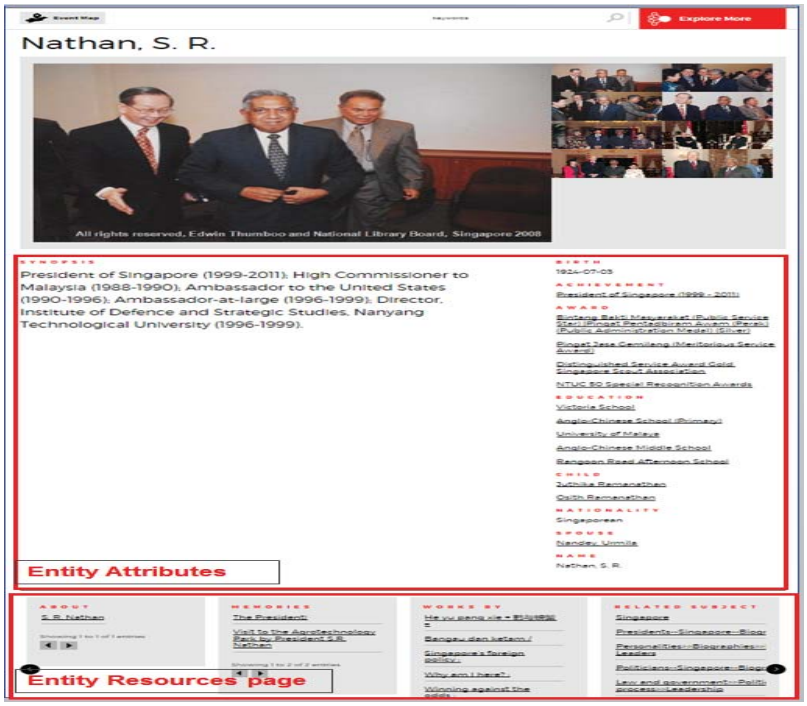


Fig 11: Entity Resources page for Mr S. R. Nathan

It lists all resources related to the entity and clusters them into 5 ‘containers’ (Fig 12) - About, Memories, Works By, Related Subject, and Related Resources.

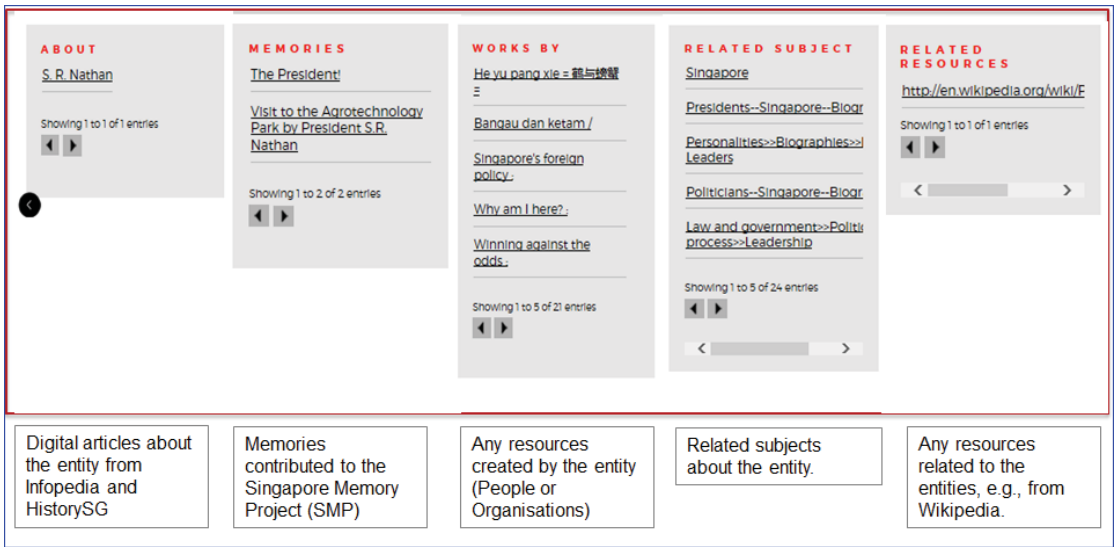


Fig 12: Entity Resources page showing the 5 'containers' and their brief descriptions

The links listed in each of the containers will bring users to the various resources such as the digital article itself listed in 'About', NLB catalogue records or NAS Archives Online in 'Works By'.

Clicking on the links of the subject headings in 'Related Subject' will bring up a list of resources with the same subject headings. For example, clicking on "Personalities>>Biographies>>Political Leaders", will display the below list (Fig 13), and users can select the entities to explore further.

Fig 13: List of entities related by subject

iv) SPARQL Search

"SPARQL Protocol and RDF Query Language" or SPARQL is the W3C recommended query language for RDF data (W3C SPARQL Working Group, 2013).

The inclusion of the SPARQL query is to show a distinction in the web feature to that of a usual website. However, a SPARQL query is not easy to construct and the results returned are not user-friendly. Hence, the team decided to include predefined SPARQL queries in the web feature (Fig 14). The predefined query acts like a natural language query but with configurable parameters for users to select. New SPARQL queries could also be configured backend in LDMS.

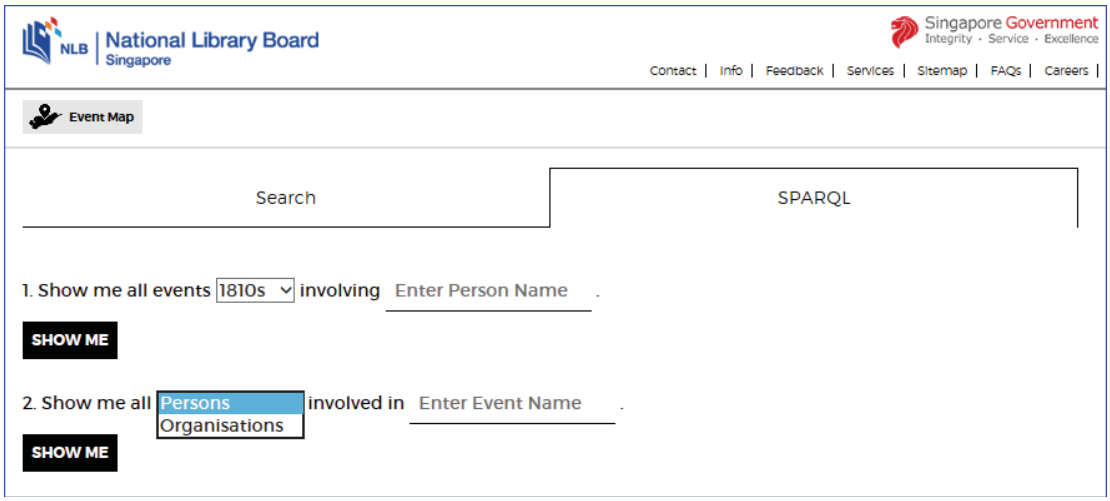


Fig 14: SPARQL query page

For example, users can select "Show me all 'Persons' involved in 'Hock Lee Bus Riot, Singapore, 1955'". This would trigger a keyword search against the LDMS triple store, and return the search results in a user-friendly format (Fig 15). Users can select any of the names listed and explore further from the Entity page.

Fig 15: Example of results from the SPARQL query

4b.) Linked Data Widget

As the service development progressed, it was apparent that the Web Feature, made available as a link from NLB websites, was not enough on its own. The challenge was how to increase the utilisation and visibility of the Linked Data. The Linked Data Widget was thus conceptualised after many rounds of discussions and consultations with user representatives, the NLB Linked Data Steering Committee, and the project vendor.

The primary objective of the widget was to enrich online text articles with the linked data entities. This is achieved by embedding the widget in the online article page, with links back to the Web Feature for the mentioned entities for further exploration.

HistorySG (http://eresources.nlb.gov.sg/history) was the first NLB website chosen for the Linked Data widget. It is a portal containing approximately 500 articles chronicling Singapore's history from 1299 to the present, including video and audio clips, photographs and newspaper articles from NLB and NAS (National Archives of Singapore) collections. Information can be searched and browsed through themes, timeline and categories (Fig 16).

Fig 16: NLB HistorySG portal

Based on the same 4 classes of People, Places, Organisations and Events, entities extracted from the text which matched the entities in LDMS are listed under its appropriate class. Users could 'click' on any of the displayed entities, and this would lead them to the LDMS Web Feature where they could further discover more information about the entities (Fig 17 and 18).

Fig 17: Linked Data Widget in a HistorySG article page shown under 'Explore Further'

Fig 18: 'People' entities as found in the text which matched the LDMS entities and listed in the Widget under 'People'

Potential of Linked Data

With the set-up of the Linked Data platform, there is great potential for the development of new services that could be rolled out faster, and at a fraction of the cost and time required by current or traditional approaches.

The one-time effort in transforming data to RDF also allows the community, developers and other agencies to re-use and re-purpose the transformed data for new applications. NLB staff could also potentially use the data for curation of new content with minimal efforts.

Lastly, the implementation of Linked Data allows NLB to supplement its resources that have little or sparse data via linking to external datasets that have the required information. This also reduces duplicate and redundancies, for example, in updating of the NLB Authorities data since the data are already available from these other datasets.

Conclusion

The Linked Data project is a major milestone in NLB's move towards embracing the semantic web. There were many challenges and issues along the way that the team had to resolve to deliver the project.

The Web Feature was soft launched internally on 8 December 2015 and is under review till February 2016. During this review phase, the team will collate all feedback and prioritise them based on their impact to the Linked Data service, and propose how best to further improve the discovery and information-seeking experience for users.

The team is also looking forward to the next phase of the project where more data will be transformed. This includes getting cross-walked data from another NLB project which is a time-saving and cleaner approach than ingesting the raw records directly from the source systems.

The project team is also exploring potential external datasets for linking and developing new services that would utilise Linked Data, publishing of selected NLB Linked Data in the NLB

Open Data Initiative, and enabling Linked Data search via the NLB OneSearch platform. These activities would ensure that NLB Linked Data is utilised to its full potential for the benefits of NLB patrons.

References

Byrne, G., & Goddard, L. (November/December, 2010). The Strongest Link: Libraries and Linked Data. *16*(11/12). doi:10.1045/november2010-byrne

Gonzales, B. M. (2014). Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record. *Information Technology and Libraries, 33, No. 4*. doi:http://dx.doi.org/10.6017/ital.v33i4.5631

Hooland, S. V., & Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. London, UK: Facet Publishing.

International Council of Museums. (9 December, 2014). *The CIDOC Conceptual Reference Model*. Retrieved from CIDOC CRM Home page: http://cidoc-crm.org/

International Council on Archives. (2009). *ica.org*. Retrieved from International Council on Archives: http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

Library of Congress. (21 November, 2012). *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*. DC, Washington. Retrieved from Library of Congress: http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

RDF Working Group. (25 February, 2014). *RDF*. Retrieved from W3C Semantic Web: http://www.w3.org/RDF/

W3C SPARQL Working Group. (26 March, 2013). *SPARQL Query Language for RDF*. Retrieved from W3C: https://www.w3.org/TR/rdf-sparql-query/