IFLA WLIC 2019 Athens

Submitted on: 28.06.2019

Taking on the content discovery challenge: The NLB Case Study

Patrick Cher
Technology & Digital Services, National Library Board, Singapore, Singapore.
E-mail address: patrick_cher@nlb.gov.sg

Abstract:

This paper discusses four major projects undertaken by the National Library Board (NLB) to implement content discovery solutions that enhance discovery and accessibility to digitised archived content such as geo-referenced maps, photographs and audio-visual recordings. This paper falls under the "Libraries, archives and museums in dialogue: Improving access to complementary collections" topic.

In November 2012, NLB's role expanded beyond offering library collections when the National Archives of Singapore became an institution of NLB. Besides inheriting a treasure trove of primary research material about Singapore government's corporate memory, public and private records, its role expanded to include collection, preservation and management of those collections.

First, NLB had to consolidate the vast collections across its libraries and the archives to offer a unified research experience. A cross functional team consisting of content specialists and technologists was formed. The team's first priority was to harmonise content.

The Archives uses ISAD-G (General International Standard Archival Description) to describe archival records while NLB uses Dublin Core standard to describe its library collection. Harmonisation is needed to enable a common search interface that offers normalised search results and relevant facets to refine search results.

This paper will cover the harmonisation process as well as the user research studies to ensure that the labels speak the users' language, with words, phrases and concepts that are familiar to the users.

Second, having built a common foundation through the data harmonisation project, NLB developed new digital research services. One notable example is Spatial Discovery, a platform that allows researchers to search and discover maps from the libraries and archives. The maps are geo-referenced so that researchers can accurately trace the changing

landscapes. Beyond NLB, the harmonisation project's outputs have benefitted other local institutions in the cultural heritage business. The National Heritage Board and National Gallery of Singapore were able to leverage the common standards through knowledge and technology sharing via content partnership.

Third, the paper will discuss the NLB's recent experience in leveraging Linked Data and Machine Learning to improve content discovery and access. A common researcher's feedback is difficulties in drawing linkages between articles with other content. To address the problem, text analysis was used to surface complementary content relevant to the research topic. A linked data feature was also launched to help researchers see the relationships of entities mentioned in the piece of content. Researchers are now able to quickly find out more about the events, personalities and places mentioned.

Finally, beyond employing technologies for content discovery, NLB has also used Machine Learning to aid content description. Entity extraction technology was used to geo-tag one million digitised photographs so that they could be delivered on a map view. If done manually, it would be a massive undertaking. The NLB is currently in the midst of building its image recognition model to describe its image-based content. When completed, it would be able to assist staff in building metadata by detecting faces and places within images.

Keywords: subject analysis, discovery, accessibility, data linkage, integrated libraries and archives collections

---

1 INTRODUCTION

The National Library Board (NLB) of Singapore oversees the National Library, the Public Libraries and National Archives of Singapore (NAS). Its mission is to provide a trusted, accessible and globally connected library and information service.

A Salesforce research revealed that consumers expect companies to understand their needs and expectations.[1] Users expect a hyper convenient experience where seamless handoffs or contextualised engagement are important.

Through innovative use of technologies, NLB is able to provide users of libraries and archives with unified access to a myriad of information through the content discovery projects undertaken over the recent years.

2 DATA HARMONISATION PROJECT

The merger of the archives in November 2012 meant that NLB possesses a treasure trove of Singapore cultural and heritage related information from both public and private sources.

[1] "State of the Connected Customers," Salesforce Research, accessed May 18, 2019, https://www.salesforce.com/content/dam/web/en_us/www/documents/e-books/state-of-the-connected-customer-report-second-edition2018.pdf.

The combined collection size of the libraries and archives number over 30 million, spread over 50 digital touchpoints. This presented an overwhelming user experience, forcing users to work extra hard to fulfil their information needs.

The Data Harmonisation project aims to enhance the users' search and discovery experience of NLB's rich array of resources through a unified search. All records have to undergo the process of mapping and crosswalk to a new common standard.

2.1 Metadata Description Practices

Libraries organise its information at item level and group them into collections. On the other hand, the archives adopt a more granular approach. Records are described at multiple levels, allowing objects to be related in a hierarchy and linked to the collection.[2]

The libraries use Machine Readable Cataloguing (MARC21) for describing physical resources and Dublin Core Libraries Application Profile (DC-Library) for describing digital collections. Whereas, the archives use General International Standard Archival Description (ISAD-G) schema.[3]

Despite the differences in description standards, there are some key fields that are used by the libraries and archives. This finding meant that NLB can adopt a common metadata standard for a one-stop search engine. DC-library was chosen as the common schema.

2.2 Naming & Vocabulary Practices

Besides metadata description differences, naming conventions for people, places and organisations are dissimilar. Each collection is described and managed by different teams that do not share the same name headings or controlled vocabularies.[4] Consistent naming convention can distinguish similar records and facilitate effective and efficient information retrieval.

To address this issue, the team turned to Knowledge Organisation System (KOS). KOS contains authoritative names of entities in Singapore and Southeast Asia. It offers authoritative names in Singapore's 4 official languages – English, Malay, Chinese and Tamil, dialect and alternate names.[5]

[2] Shan Shan Chan and Haliza Jailani, "Data Harmonisation between National Library Board, National Archives and National Heritage Board of Singapore," Proceedings of International Conference on Dublin Core and Metadata Applications:241-243.
[3] Chan and Jailani, "Data Harmonisation," 241-243.
[4] Chan and Jailani, "Data Harmonisation," 241-243.
[5] Puay Eng Tang, Glenn Hong and Haliza Jailani, "Authoritative content to build trust in an age of information overload: The National Library Board of Singapore's experience," World Library International Conference 2018.

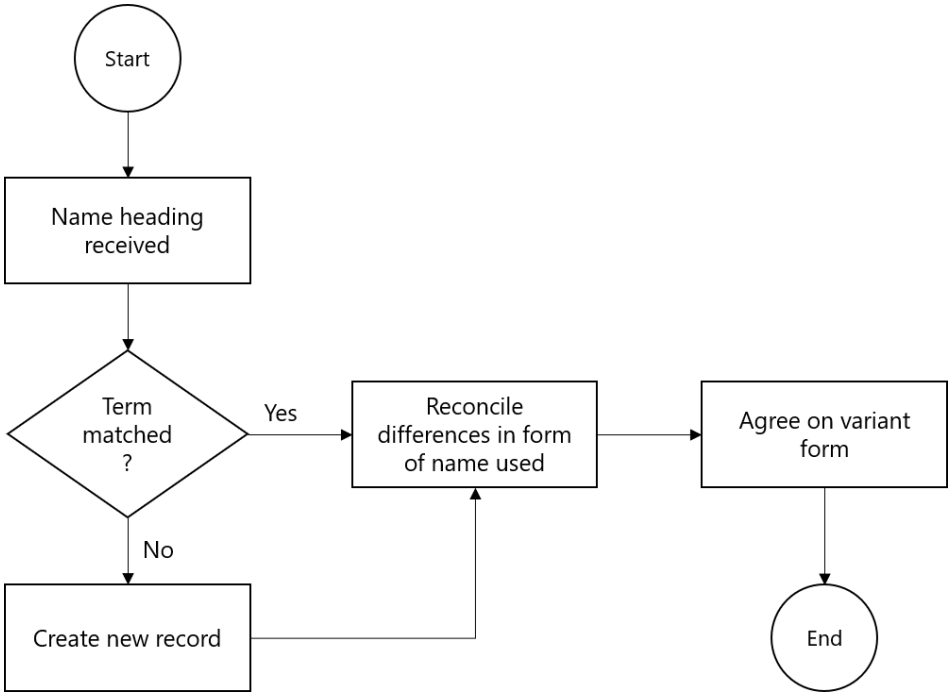


Fig. 1 – Mapping process of vocabularies used in the libraries and archives

Vocabularies and name heading used in the libraries and archives collections were mapped and consolidated into a controlled list. The list is managed using NLB's Taxonomy and Thesaurus Editor (TTE).[6] NLB is able to achieve standardisation in vocabularies through integration with its indexing systems.

3 ONESEARCH – UNIFIED SEARCH SERVICE

OneSearch (http://search.nlb.gov.sg), a unified search service launched by NLB in 2014, was designed with simplicity and ease of use in mind. Users can locate a range of books, magazines, audio-visual materials and other material.

3.1 Standardised data practices allowed IT team to build user-centric service

The standardised metadata and controlled naming and vocabularies practices allowed the IT team to direct its efforts on conducting user researches to gather qualitative and quantitative feedbacks.

[6] Chan and Jailani, "Data Harmonisation," 241-243.

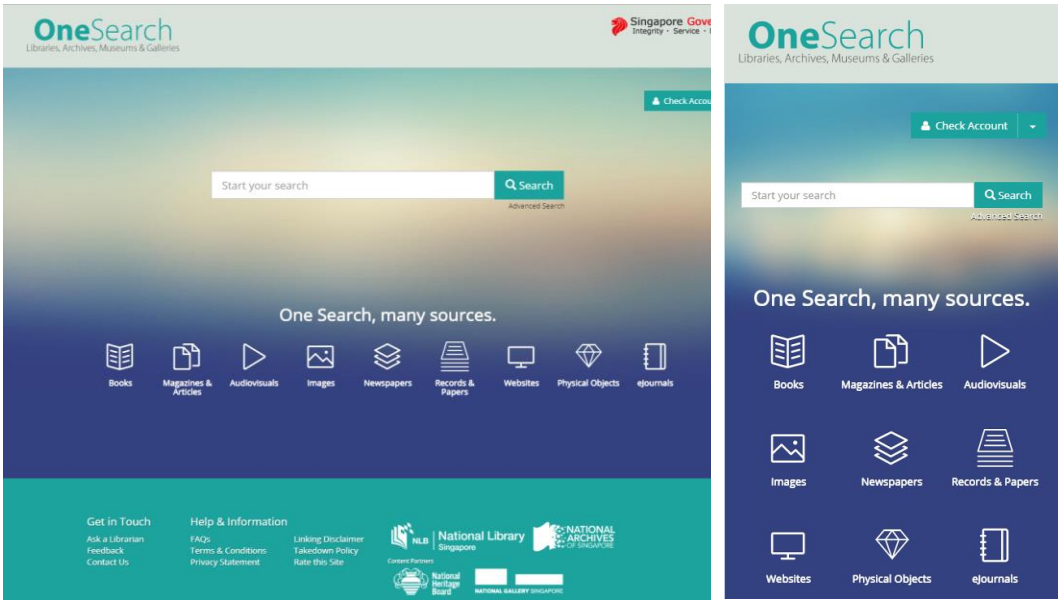


Fig. 2 – Screenshots of mobile-friendly interface. Left: Desktop, Right: Mobile

OneSearch adopted an agile application development approach. Throughout the development phase, incremental product releases were available so that stakeholder feedbacks could be sought. Furthermore, user research could be performed to gather data to make data-driven design decisions.

Traditional approach of displaying search results in an aggregated manner does not work well for NLB's collections. Text based records with more descriptive metadata like books and articles tend to rank higher than content with minimal metadata such as digitised photographs and maps. Advanced Web Ranking's research on Google Organic Click-through Rates History showed that click-through rates fall below 0.96 and 1.58 clicks on desktop and mobile respectively after the 20th search result.[7] Users are likely to drop out after 20 results. Through focus group discussions, the team adopted a "bento-style" presentation layer whereby each compartment holds content of a specific type. This gave equal opportunity for results of different types to be shown.

[7] Google Organic CTR History," Advanced Web Ranking, accessed May 19, 2019, https://www.advancedwebranking.com/ctrstudy/.

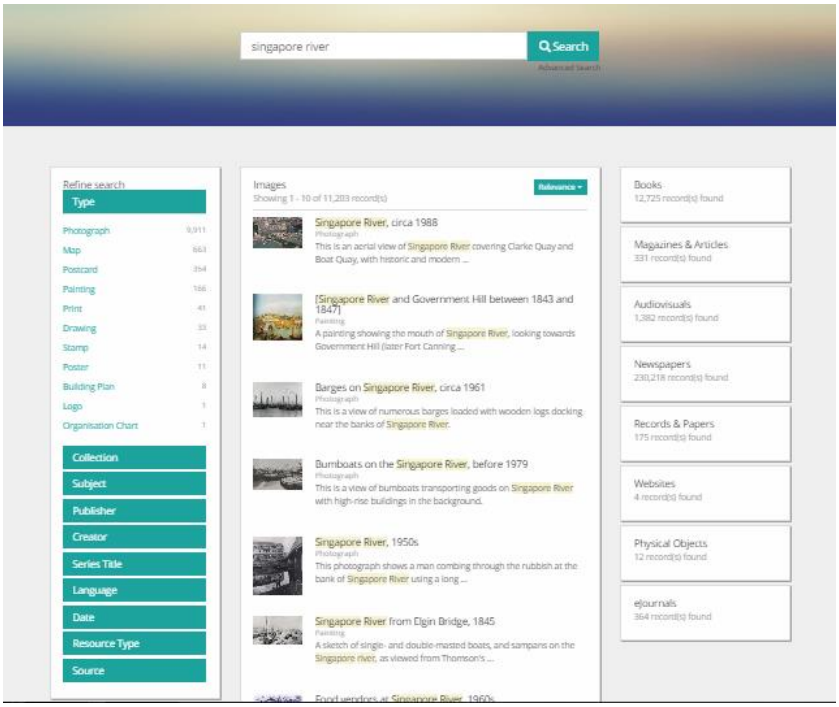


Fig. 3 – Screenshot of OneSearch's bento-style presentation layer

Next, to derive the number of compartments to offer, the team conducted online card sorting exercises. Participants provided feedbacks on the naming of each compartment, the number of compartments and sorted the sub-format types. The team was able to collect quantitative data to understand users' behaviour and make informed design decisions to better meet their needs and expectations.

Display Sequence	Content group	Content Group Description
1	Books	Books, eBooks & Digitised Rare Books
2	Magazines & Articles	Articles & Magazines
3	Audiovisuals	Audiovisual and Sound Recordings, Oral History Interviews
4	Images	Pictures, Posters, Maps & Building Plans
5	Newspapers	Digitised Newspapers
6	Records & Papers	Transcripts of Speeches & Press Releases, Government Records, Overseas Records & Private Papers
7	Websites	Archived Websites
8	E-Journals	Subscribed ejournals

Fig. 4 – An early version of grouping of content types

3.2 Standardised data practices enabled collaboration efficiency

The harmonisation project's outputs have benefitted other local institutions in the cultural heritage business. The National Heritage Board (NHB) and National Gallery of Singapore (NGS) were able to use the common standards built by NLB through knowledge and technology sharing.

By sharing the standard naming and vocabulary conventions, Singapore's museums and galleries were able to work in the same direction. Time and resources to create and maintain proprietary lists could be saved, allowing partners to share knowledge and focus on building their collections.

Seeing the benefits of the controlled list, NHB has since become an active contributor.[8] This win-win collaboration makes the authoritative list a valuable asset not only for institutions in the galleries, libraries, archives and museums industry but also the government.

Furthermore, adoption of the common conventions reduced the complexity and on-boarding time. Data specialists only had to crosswalk the names and vocabularies. The IT team did not have to worry about the impact of new content affecting the results' relevancy and facet filters. Instead, they can focus on harvesting partner's content and facilitate the acceptance tests. These practices allowed NLB to on-board new partners quicker with little changes to application design.

4 SPATIAL DISCOVERY

Spatial Discovery (http://search.nlb.gov.sg/spatialdiscovery) is a one-stop platform for users to explore, find and interact with high-resolution maps across the libraries and archives collections.

Fig. 5 – Landing page of Spatial Discovery (left) and interactive interface for exploring maps (right)

Cartographic materials contain important research information about events, places and people. Access to such material is controlled as frequent handling of the documents can add to the stress of environmental changes.

Although digitised maps are offered, the lack of interactive tools meant that the experience is not efficacious in replacing or matching the tactile experience of handling the physical copy. Another challenge is balancing the minimum file size for faster page load and acceptable quality for research.

The outputs of the harmonisation project allowed the team to focus on preparing the images for geo-referencing and build new technical capabilities in Geographical Information Systems (GIS).

[8] Chan and Jailani, "Data Harmonisation," 241-243.

4.1 Standardised data practices allowed NLB to build new research content and technical capabilities

The maps undergo a process called geo-referencing to prepare them to be overlaid on modern day map services such as Google Maps. Geo-referencing is the process of transforming a map image into a dataset with coordinate reference system. It involves selecting ground control points (GCPs) which will be used for the transformation, rotation and translation of the image.

4.1.1 Content Preparation for GIS

NLB's maps undergo a vigorous 4-stage process to prepare them for researchers' access on the Spatial Discovery website. The workflow employs some automation to speed up the process and for quality control.

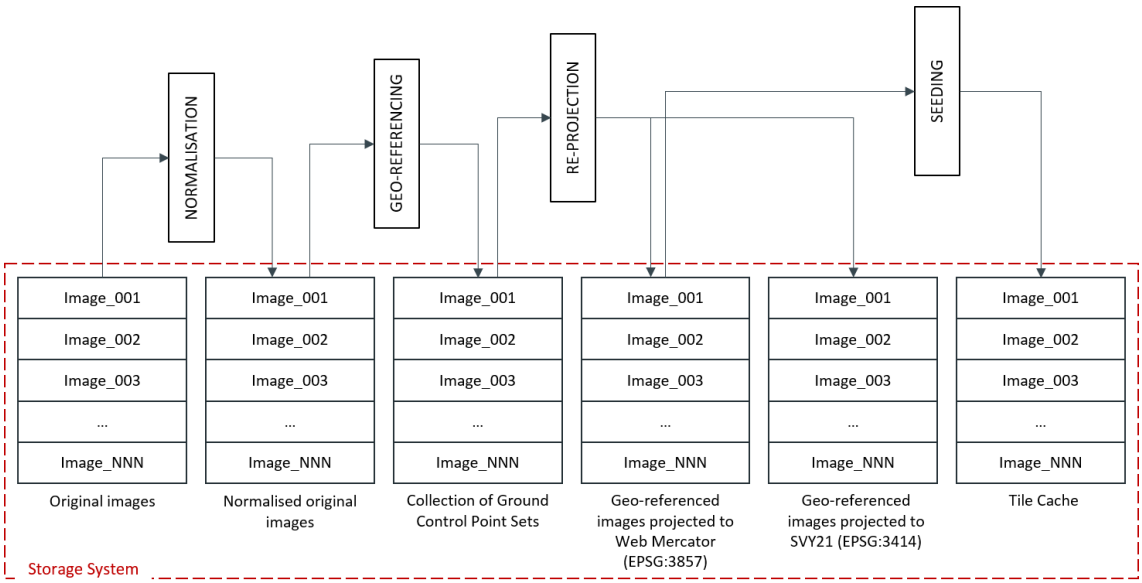


Fig. 6 – NLB's geo-referencing workflow

The 4-stage process are:

1. Normalisation
   Images are manually inspected and assessed if they could be geo-referenced. Torn or hand-drawn maps are removed at this stage. The remainder is converted to TIFF format and validated for possible errors.

2. Geo-referencing
   Geospatial specialists find pairs of corresponding points on the map images using customised open-source GIS software, Quantum GIS.

   Since 1960s, Singapore's land area expanded by almost 25% from 581 to 724.2 square kilometres.[9] This increase made the identification of corresponding points a challenge. To tackle this problem, the geospatial specialists and archivists picked 2

[9] Singapore Land Authority, "Total Land Area of Singapore," data.gov.sg, accessed May 20, 2019, https://data.gov.sg/dataset/total-land-area-of-singapore

reference maps per decade. This practice is based on Singapore's Urban Redevelopment Authority's practice in reviewing land use.

3. Re-projection
   A database containing landmarks in Singapore from 1950s to 2000s was setup to be used as Ground Control Points (GCPs) using the list of authoritative names in KOS as well as other government datasets.

   A script is executed to transform geo-referenced images into projected images using GCPs. NLB maintains geo-referenced maps in 2 coordinate projection systems, Web Mercator WGS84 (EPSG:3857) which is popular standard due to close association with Google and SVY21 (EPSG:3414) which is a standard used by the Singapore government.

4. Seeding
   Once re-projection is completed, smaller tiles of the processed images are generated for each zoom level. With tiling technology, smaller chunks of the entire map could be served on demand for faster loading speeds.

4.1.2 Interactive tools to enable meaningful research

To provide a meaningful research experience, interactive tools must be built to maximise the value of geo-referenced maps and GIS technology. Features on Spatial Discovery include:

- Layering – allows users to pick and view several maps as layers
- Transparency tool – adjust transparency of specific map to compare against the other layer(s)
- Show/hide cutline – show or hide image collars so that layers could be viewed as a seamless piece
- Peek through keyhole – look underneath selected map through a keyhole
- Slide to compare – compare 2 maps side-by-side
- Facet Filtering – filter results based on fields like date range
- Time scale – visually see the distribution of maps over years
- Permalink – share with others or bookmark their saved progress

4.2 Researchers can fulfil their requests anytime, anywhere

The implementation of Spatial Discovery has created value and positive impact for both users and NLB, leading to increased customer satisfaction.

Spatial Discovery improved accessibility to digitised maps, giving researchers the ability to self-service their requests anytime, anywhere. Spatial Discovery also reduced the need for

librarians and archivists to fulfil physical maps requests, allowing them to attend to higher value tasks such as attending to research questions.

5. ENABLING SERENDIPITOUS DISCOVERY WITH MACHINE LEARNING & LINKED DATA

Over the past decade, web technologies have been advancing at an increasingly rapid speed, outpacing the ability of traditional bibliographic formats like MARC in keeping up with the potential these technologies could offer.

As an organisation that strives for better service delivery, NLB could not pass on the potential of ideas behind Machine Learning and Linked Data to deliver content to users and facilitate serendipitously discovery.

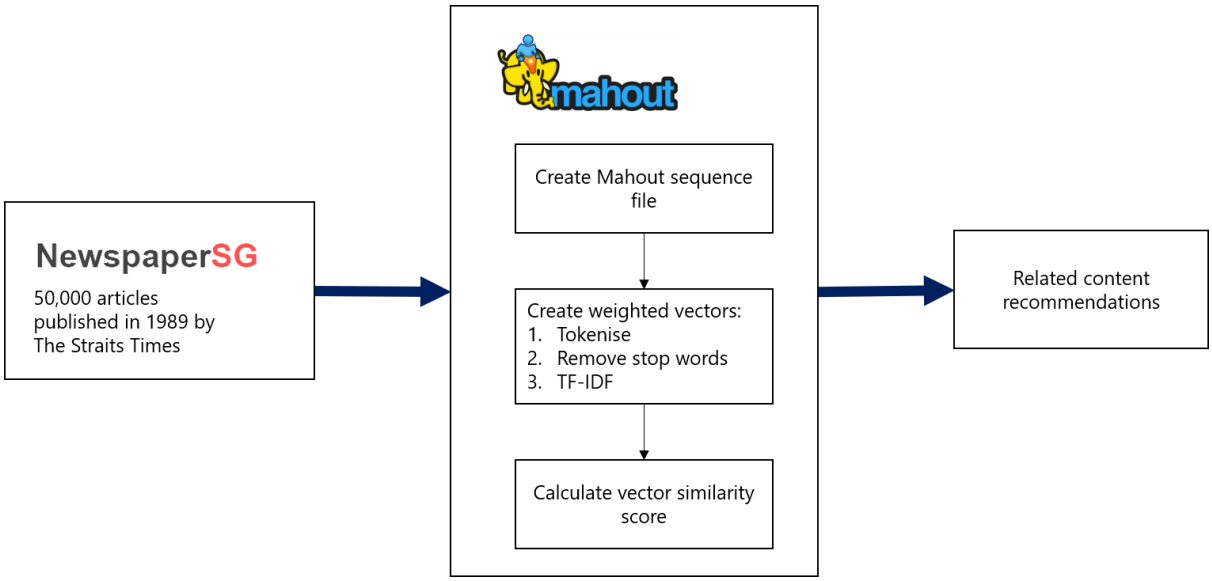
5.1 Harnessing Machine Learning to identify related content

Machine Learning (ML) is the science of getting computers to act without being explicitly programmed. It is provided access to data to learn for themselves. ML is increasingly pervasive that most users have used it without realising. For instance, facing a camera to identify themselves at security gantries or unlock their mobile devices.

5.1.1 Proof-of-Concept to ascertain feasibility

Textual Analysis is one of the ML techniques that could also be applied to aid content discovery. It is an automated process whereby information is extracted and classified from text available.

Apache Mahout was identified as the solution to perform this task.[10] A proof-of-concept (PoC) was conducted on NewspaperSG, an online archive of Singapore newspapers published since 1827. To manage the scope of the PoC, only newspaper articles published in 1989 by The Straits Times were used.



[10] Siang Hock Kia, Yi Chin Liau, Ian Ong, "Inter-connected Network of Knowledge – The NLB Journey," World Library International Conference (2014).

Fig. 7 – Process for generating related content recommendations using Mahout

Through Mahout, a list of similar newspaper articles for each of the 50,000 articles was generated. The similarity between articles was scored between 0 to 1, with 1 being an exact match.

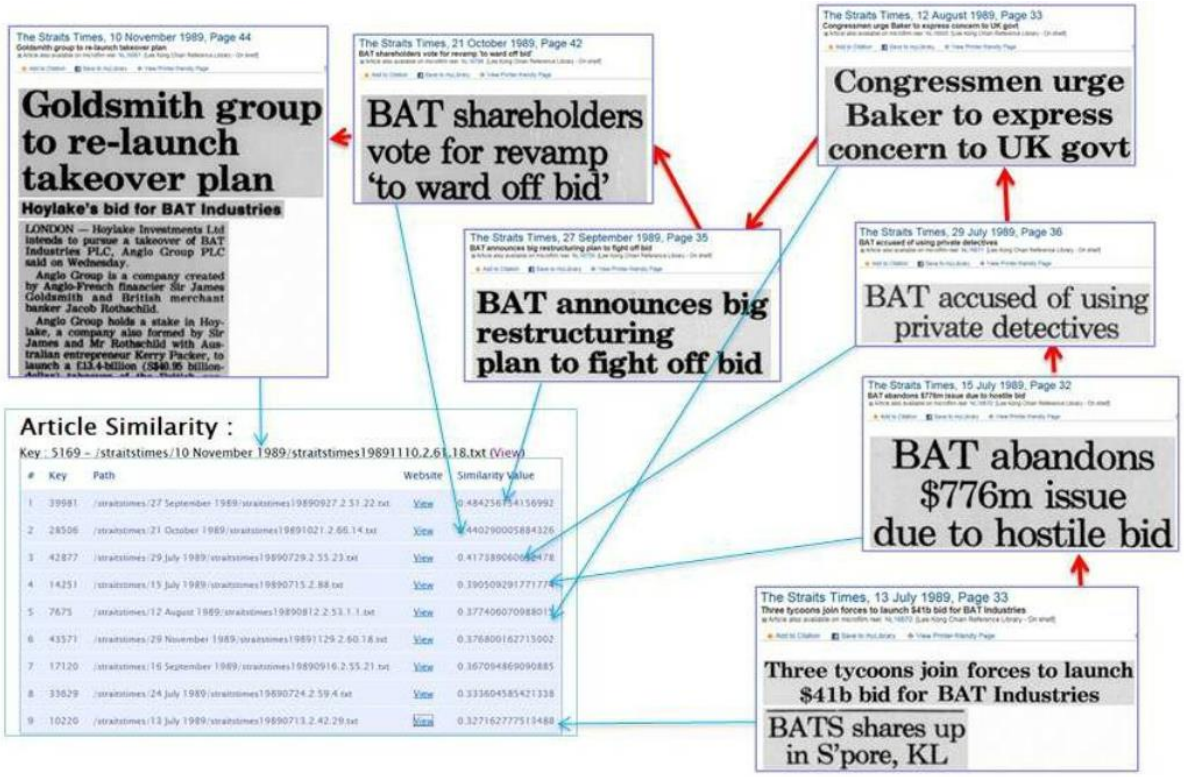


Fig. 8 – Sample output of the Mahout textual analysis

Manual verification performed by the content team confirmed the accuracy of the algorithm and readiness for production use. The team deployed the feature for other libraries and archives collections like PictureSG (http://eresources.nlb.gov.sg/pictures), Infopedia (http://eresources.nlb.gov.sg/infopedia/) and ArchivesOnline (http://nas.gov.sg/archivesonline).[11]

5.1.2 Benefits of Textual Analysis PoC

The Textual Analysis project has automated manual efforts of curating related content. Within 4 months of implementation, the project recorded 10% increased referral in page views due to the cross-selling opportunities presented.[12]

5.2 DISCOVERING RELATED PEOPLE, PLACES, ORGANISATIONS AND EVENTS USING LINKED DATA

Another strategic project is the Linked Data project, where the Web is used to connect related data that wasn't previously linked to increase accessibility and visibility to content.

[11] Kia, Liau and Ong, "Inter-connected Network of Knowledge".
[12] Kia, Liau and Ong, "Inter-connected Network of Knowledge".

Through the project, Linked Data Management System was setup to manage data. An NLB Data Model was created to govern its data structure. Existing bibliographic data was converted into Bibliographic Framework (BIBFRAME) format.[13]

5.2.1 Platform agnostic linked data widget allows integration with NLB's collections

Linked Data project was first piloted on HistorySG (http://eresources.nlb.gov.sg/history), a collection of 500 articles documenting Singapore's history since 1299.[14]

A platform agnostic widget was created to enable easy integration to existing digital properties such as Infopedia and HistorySG. The widget was created using HTML and Javascript. Mobile responsive design allowed it to function across screen sizes without affecting its functionalities.

The widget lists entities that were extracted from the respective articles which matched with entities in the Linked Data Management System. When a user clicks on an entity, he/she would be able to discover more information about it.

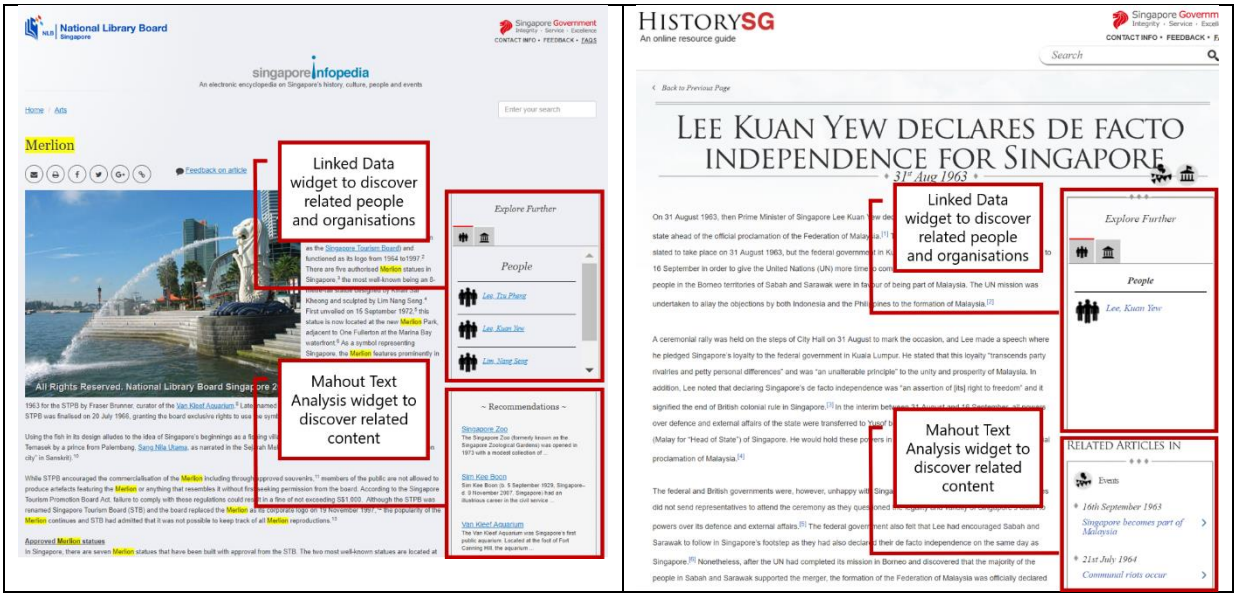


Fig. 9 – Various forms of recommendations on NLB's digital properties

NHB recognized the potential and benefits and integrated it with their collection of heritage material at RootSG (https://roots.sg) in 2017.

5.2.2 Overview of research topic about people and organisations through NLB's Knowledge Panel

One common issue with performing research is the inability to refine their search phrases due to the lack of knowledge on the research topic. Providing useful nuggets of information could help them reach a conclusion.

[13] Hanna Hussein, "Linked Data @ NLB," Singapore Journal of Library and Information Management (2015):20-34.
[14] Hussein, "Linked Data," 20-34.

Knowledge Panel was created with the intention of enhancing the search results with information gathered from other authoritative sources. The information is then presented to users in the form of an information box.

This way, users have an overview on research topics related to people or organisations. A search for “lee kuan yew” on OneSearch returns over 100,000 results. One can use the Knowledge Panel to get a synopsis of the life of Mr. Lee Kuan Yew, view his achievements and more.

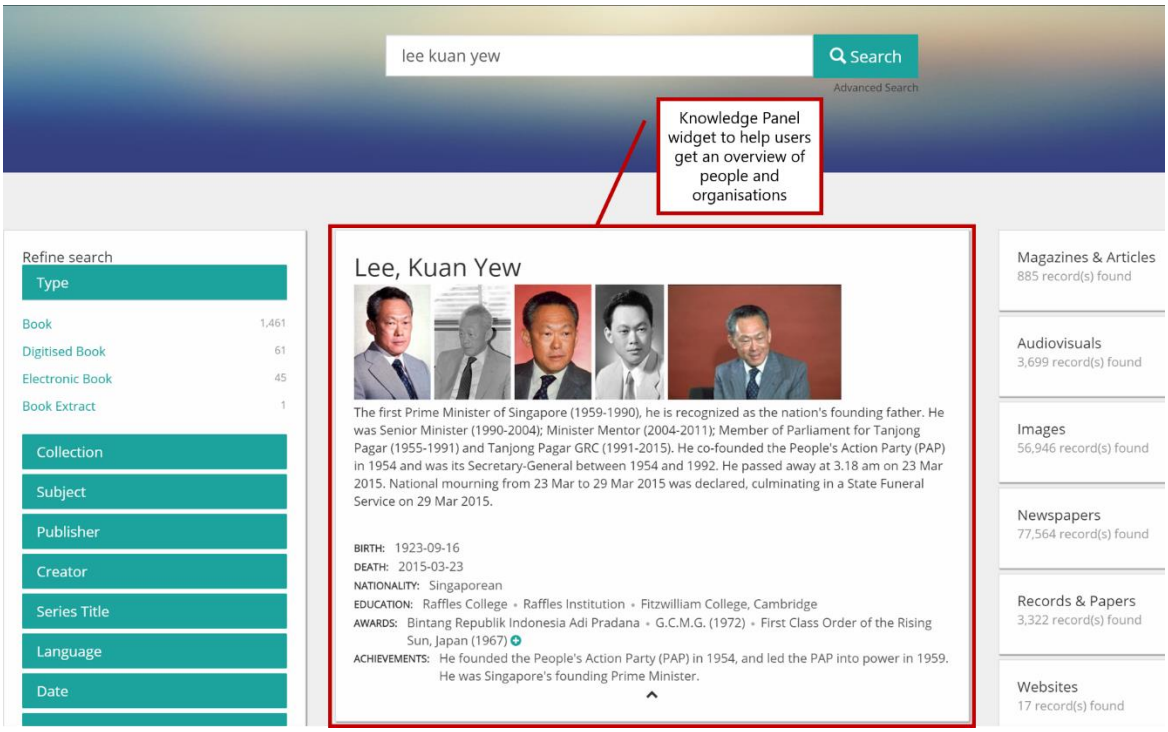


Fig. 10 – Knowledge panel helps users learn about topics related to people and organisations

6 MACHINE LEARNING PROJECTS FOR MACHINE-ASSISTED INDEXING

The application of ML could stretch beyond customer facing tasks. It could be applied for internal processes such as machine-assisted indexing.

6.1 Utilising Named Entity Recognition to extract information from documents

NLB possesses a collection of over 800,000 images, spread over 3 collections.

Substantial amount of research and verification has been done to describe each image. Without geo-tagging, it is difficult to visualise the image location with its accompanying metadata. GATE, an open source text analysis toolkit, was used to identify locations mentioned in each image.
Named Entity Recognition, also known as entity extraction, classifies entities that are present in a document into pre-defined categories such as people and places. This technique adds a wealth of semantic knowledge to a piece of content to help users understand their research topic.

KOS has an authoritative list of local historical places. However, the scope of the list does not cover present day landmarks. Data from GeoSpace, a government data sharing platform to build up its database was used to build up a comprehensive list of buildings, roads, monuments, historic sites and tourist attractions, complete with geographical coordinates in WGS84 and SVY21 coordinate projection systems.

Location features were identified by GATE using the list of landmarks. The identified features are then translated to the 2 coordinate systems and stored on NLB's Content Management System.

Fig. 11 – Workflow for extracting names of locations and plotting them on a map

With the images geo-tagged, the team designed a new service to present these images in a visual and interactive map view. Launched in Apr-2019 at the Archives Reading Room, PictureMap allows users to search for places, view photographs by clusters, filter images using time range and use historical maps as reference layers. Users are also able to toggle full screen mode for either map or grid view to utilise the screen space.

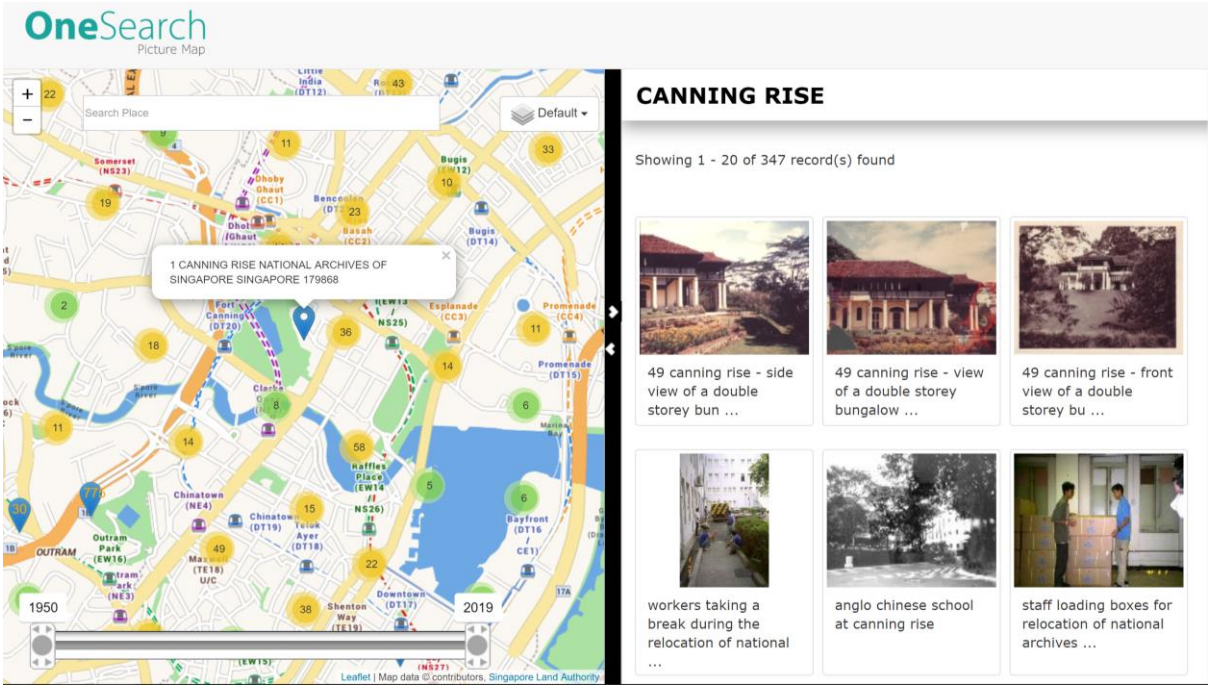


Fig. 12 – Geo-tagged images presented on PictureMap service

6.2 Using facial recognition to identify persons of interest

One challenge of describing image-based content is the implicit knowledge needed to perform the job accurately and efficiently. Images often contain clues like faces, places and objects but building up the knowledge requires time.

With advancements in artificial intelligence and innovations in deep learning and neural network, computer vision has been able to surpass humans in detecting and labelling objects. Computer vision could supplement description process for efficiency.

To keep the scope of the PoC manageable, only members of Singapore's 13th Parliament were identified. This is a starting point as the archives often receives photographs from government agencies with minimal metadata. The backlog of items pending description is lengthy before they could be available to users.

Like any ML project, training data is as important as the algorithm. The model gets increasingly accurate with more training. Despite the scope, the image quantity and quality still posed an obstacle. More than 100 quality images are needed for each member to ensure sufficient data for training and validating the model. In addition, the selected images have to meet these requirements:

1. 1 face to be present each image
2. Face must be within recommended range of angle e.g. <30 degree facing down and <45 degrees facing up
3. Face and eyes must be open and visible
4. Face must not be blocked by any item
5. No hand-drawn images

The team relied on the images in the libraries and archives and publicly accessible images on the Internet. In some cases, images were manually cropped so that there is 1 face present.

OpenFace, a free and open source facial recognition with deep neural networks, was used as the facial recognition engine.[15] Using the Docker image provided by the engine developers, the team focused its efforts on training and validating the model.

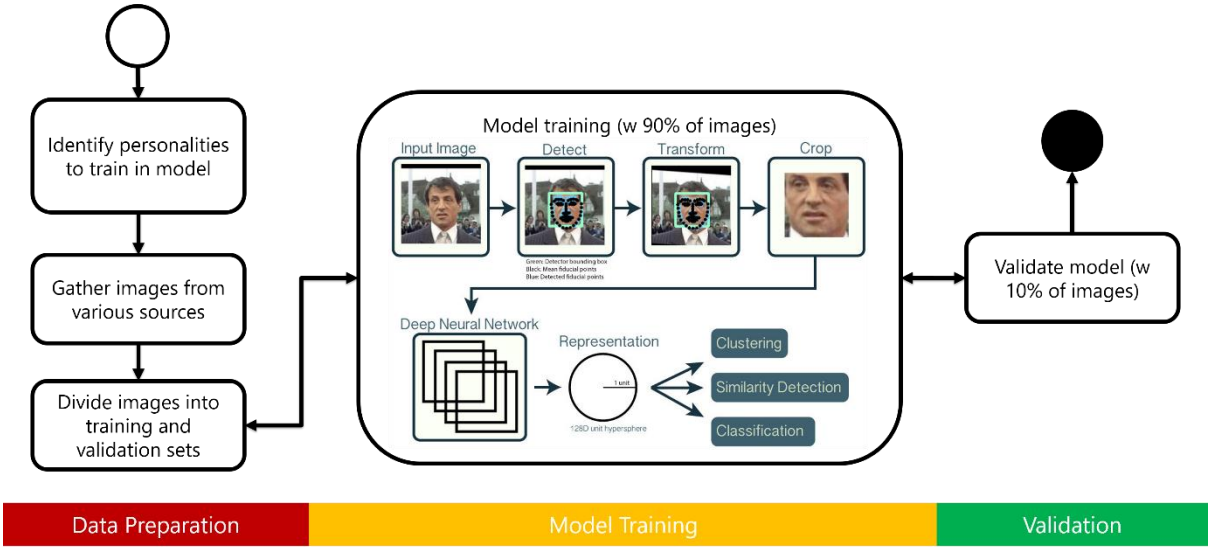


Fig. 13 – Workflow for training facial recognition model (adapted from OpenFace)

At the point of paper submission, the team is integrating the outputs of the model with the indexing systems. When completed, images of the same label would be clustered for ease of verification. The team would also progressively build up its database of faces in order to teach the model to recognise other local and regional personalities.

7 FUTURE PLANS

In this post-digital age, users expect digital technology to be a background utility, noticeable only in its absence. NLB will place more emphasis on harnessing ML to solve the content discovery gap.

There are plans for an intelligent data layer to provide a comprehensive semantic search service that enables users to draw linkages and understand relationships between different pieces of information so that a meaningful conclusion could be drawn. The layer will provide the foundation for machine assisted content curation and personalised recommendations to allow cross selling of content and programmes. This would provide targeted content to increase relevance of recommendations and trust in the NLB as an information source.

[15] Brandon Amos, Bartosz Ludwiczuk and Mahadev Satyanarayanan, "OpenFace", accessed May 19, 2019, https://cmusatyalab.github.io/openface/#openface.

REFERENCES

Amos, Brandon, Bartosz Ludwiczuk and Mahadev Satyanarayanan. *OpenFace*. 2016. Webpage. 19 May 2019. <https://cmusatyalab.github.io/openface/#openface>.

Chan, Shan Shan and Haliza Jailani. "Data Harmonisation between National Library Board, National Archives and National Heritage Board of Singapore." *Proceedings of International Conference on Dublin Core and Metadata Applications 2015*. 2015. 241-243.

Google Organic CTR History. 19 05 2019. <https://www.advancedwebranking.com/ctrstudy/>.

Hussein, Hanna. "Linked Data @ NLB." *Singapore Journal of Library and Information Management* (2015): 20-34.

Kia, Siang Hock, Yi Chin Liau and Ian Ong. "Inter-connected Network of Knowledge – The NLB Journey." *World Library International Conference*. Lyon, 2014. 19 May 2019. <http://library.ifla.org/876/1/208-kia-en.pdf>.

Salesforce Research. "State of the Connected Customer." 2018. *Salesforce Research: Customer Expectations Hit All-Time Highs.* Document. 17 May 2019. <https://www.salesforce.com/content/dam/web/en_us/www/documents/e-books/state-of-the-connected-customer-report-second-edition2018.pdf>.

Singapore Land Authority. *Total Land Area of Singapore*. 20 05 2019. <https://data.gov.sg/dataset/total-land-area-of-singapore>.

Tang, Puay Eng, Glenn Hong and Haliza Jailani. "Authoritative content to build trust in an age of information overload: The National Library Board of Singapore's experience." *World Library International Conference 2018*. Kuala Lumpur, 2018.