IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 19.07.2018

Using Named Entity Recognition for Automatic Indexing

Rachael Goh
Resource Discovery and Management, National Library Board, Singapore.
E-mail address: Rachael_goh@nlb.gov.sg

Abstract:

Automatic indexing has been used by libraries to index their collections for many years. Recent technological advances have allowed for a refinement of automatic indexing, providing quicker and more accurate results.

In 2016, the National Library Board (NLB) embarked on a Named Entity Recognition (NER) project which leveraged on natural language processing techniques to extract names of entities such as people, organisations and places that are found in documents such as articles. Articles from NLB's Infopedia (eresources.nlb.gov.sg/infopedia) and HistorySG (eresources.nlb.gov.sg/history) were identified for extraction. This has since enabled users to search for related resources in NLB's vast collections.

In late 2017, the NER project was extended to include extraction of topics mentioned in articles and metadata records. This automated indexing was done using NLB's controlled vocabularies such as Events, Historical Events, Programmes, Legal Acts, Awards, Time, and SingHeritage. Data from other agencies, National Archives of Singapore (NAS) and National Heritage Board (NHB) collections were also included. A smaller scope of the project includes running the documents against external vocabularies such as GeoNames and Wikidata.

This paper will discuss the process, the method used to evaluate the accuracy of the named entities extracted, the NER issues discovered across the collections, and the challenges faced in ensuring that the extraction process is improved after verification. Lastly, the paper will also discuss how the NER results are used to support NLB's digital cataloguing, for example, by adding the extracted entities to the subject field in the metadata records without manual cataloguing for collections with minimal or no subject headings. This allows the collections to be searched by subjects in NLB's OneSearch platform (search.nlb.gov.sg), where metadata from 3 cultural institutions (library, archive and heritage) are aggregated via an integrated interface to enable a single search. The NER results are also used to support NLB's Linked Data services such as the Linked Data widget that identifies entities from the website articles and links them to relevant resources from other collections. In this way, we supplement manual indexing by creating additional access points to the articles, while at the same time creating links to our catalogue records. Implemented in NLB's Infopedia and HistorySG websites, this service is being explored for future implementation in other websites such as NAS' Archives Online (nas.gov.sg) and NHB's RootsSG website (roots.sg).

Keywords: named entity recognition, automatic indexing, linked data, information retrieval.

I. INTRODUCTION

This paper addresses the use of Named Entity Recognition (NER) in the National Library Board (NLB)'s linked data project, which aims to "explore new ways in improving the information-seeking experience of its users" (Hussein, 2015). NER presented an avenue for NLB to use the unstructured data in our digital collections to connect users to more and varied collections within NLB, which they might not know exist. NER also allows us to extract named entities contextually, through the automatic indexing process, thus effectively utilising the unstructured data to create links to other resources within the larger NLB collection. Another objective of using NER was to explore how well automatic indexing on named entities could be applied to a collection of metadata that did not have any subject description.

Within the National Library Board (NLB)'s collections, there are collections hosted on our microsites. These collections feature curated information on Singapore. For example, Infopedia (http://eresources.nlb.gov.sg/infopedia/) contains articles related to historical events, arts, culture, economy, government and key personalities, while HistorySG (http://eresources.nlb.gov.sg/history/) contains snippets and short articles about Singapore's history. The resources on these microsites are also connected with resources from the National Archives and the National Heritage Board via NLB's OneSearch (http://search.nlb.gov.sg). This consolidated search platform offers keyword and faceted searching to enable users to accurately find the resources they might need.

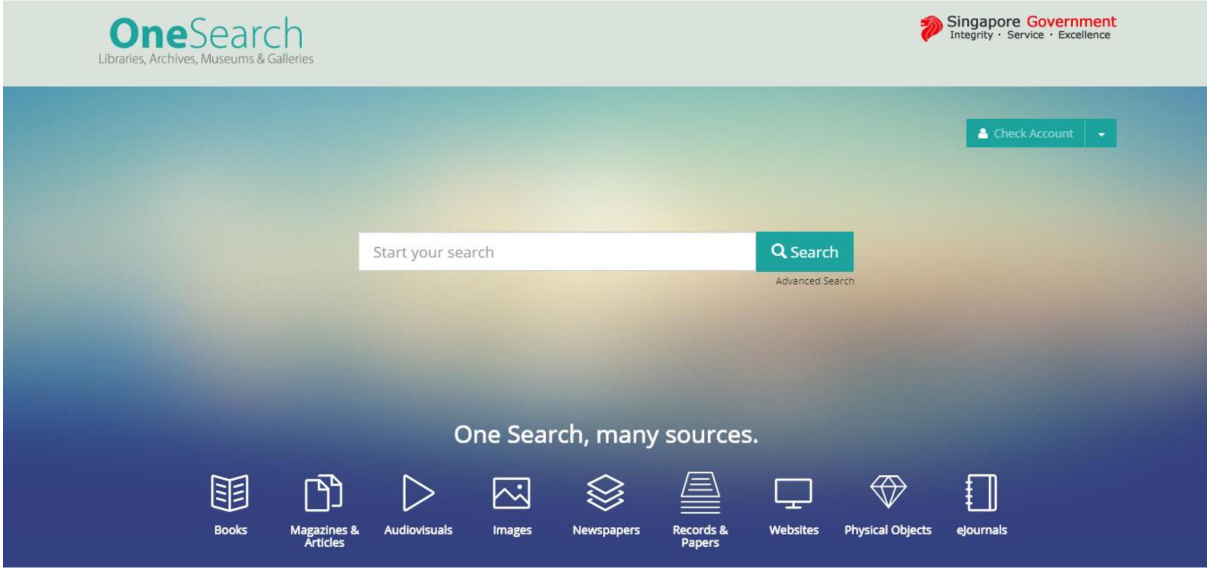


Figure 1: OneSearch Landing Page

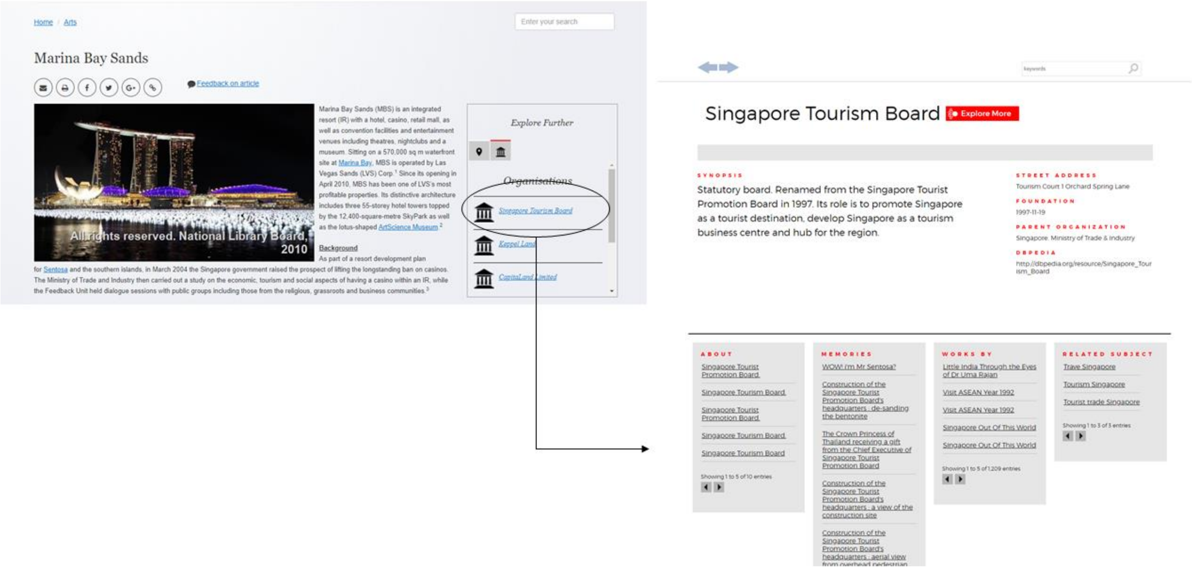
II. LITERATURE REVIEW

Automatic indexing is not new – it has been used to assign keywords to documents using a controlled vocabulary for many years (Jones, 1974). Traditionally, indexing is the process of "assigning a limited number of keywords to a document, keywords which indicate concepts that are sufficiently representative of the document" (Vallez, Pedraza-Jimenez, Codina, Blanco, & Rovira, 2015). With the advent of search engines and quick information retrieval, it has seemed to expand its scope in some scenarios to include keywords mentioned in the resource, especially named entities (United States Patent No. US9135238B2, 2015). Automatic indexing is the same concept, but without human intervention. Since its introduction, there has been a continuous debate over the accuracy of automatic indexing. One side argues that current Artificial Intelligence (AI) technology is not robust enough to replace manual indexers. Furthermore, the process of training the AI itself takes up a fair amount of resources, and the training sets used currently still require the need of people with specialised knowledge. Automatic indexing generally consists of two ways to interact with keywords – keyword extraction (Wu, Li, Bot, & Chen, 2005; Beliga, 2014), which focuses on frequency of words in the content and the whole collection and use it to determine extraction, and keyword assignment (Yang, Zhang, Li, Yu, & Hao, 2014), which focuses on matching terms in a controlled vocabulary to words found in the content being analysed.

In the case of NLB's usage of NER, we are utilising a combination of both methods. Firstly, the relevant documents are analysed using an NER process, where entities are extracted based on the Part-of-Speech (POS) in which they appear in the document. However, to modify the results to suit our needs of localised entities, the extracted keywords have to match a given controlled vocabulary of both named entities (Person, Places, Organisations), as well as more topical words belonging in the categories of Events, Historical Events, Programmes, Legal Acts, Awards, Time, and SingHeritage (terms which represent the heritage of Singapore). Furthermore, the controlled vocabulary included variants – terms which represent the same entity were consolidated under the preferred term for that entity. This process will be expanded upon later in the paper.

Before delving into the work NLB has done, it is essential to briefly discuss NER, and its uses for automatic indexing. NER is generally seen as a portion of Natural Language Processing (NLP), and a method for information extraction (Roy, 2017).It is used to identify and classify entities found in a document. While it might be experimental in nature, NER was chosen as our Knowledge Organisation System (KOS) was already set up for cataloguing purposes. Thus, we had the necessary ontology and controlled vocabulary for the NER to be able to classify entities and reduce problems in segmentation (for example, "National Library Board of Singapore" is to be understood as one entity despite it consisting of possibly two separate entities – "National Library Board" and "Singapore"). We hoped to tap into our existing resources to lessen the manual work needed for automatic indexing. NLB's decision to use NER to extract more topical keywords is a risky one – such introduction might introduce noise to the NER process as it does not only need to have patterns set up to account for named entities, but also more abstract words. However, as our controlled vocabulary is strictly moderated, we hoped that that would help prevent some of the noise being introduced.

All in all, the aim of embarking on these projects was to increase the discoverability of our resources and collections. By indexing our microsites and linking the indices through a Linked Data widget, we introduced connectivity between our resources. Through our first foray into NER back in 2016, we have introduced a Linked Data Widget onto two of our microsites. This widget contains entity-level indices i.e. names of entities and uses them as an access point to our entity pages.



Marina Bay Sands

Marina Bay Sands (MBS) is an integrated resort (IR) with a hotel, casino, retail mall, as well as convention facilities and entertainment venues including theatres, nightclubs and a museum. Sitting on a 570,000 sq m waterfront site at Marina Bay, MBS is operated by Las Vegas Sands (LVS) Corp.[1] Since its opening in April 2010, MBS has been one of LVS's most profitable properties. Its distinctive architecture includes three 55-storey hotel towers topped by the 12,400-square-metre SkyPark as well as the lotus-shaped ArtScience Museum.[2]

Background

As part of a resort development plan for Sentosa and the southern islands, in March 2004 the Singapore government raised the prospect of lifting the longstanding ban on casinos. The Ministry of Trade and Industry then carried out a study on the economic, tourism and social aspects of having a casino within an IR, while the Feedback Unit held dialogue sessions with public groups including those from the religious, grassroots and business communities.[3]

Singapore Tourism Board

SYNOPSIS

Statutory board. Renamed from the Singapore Tourist Promotion Board in 1997. Its role is to promote Singapore as a tourist destination, develop Singapore as a tourism business centre and hub for the region.

STREET ADDRESS

Tourism Court 1 Orchard Spring Lane

FOUNDATION

1997-11-19

PARENT ORGANIZATION

Singapore. Ministry of Trade & Industry

DBPEDIA

http://dbpedia.org/resource/Singapore_Tourism_Board

ABOUT

- Singapore Tourist Promotion Board
- Singapore Tourism Board
- Singapore Tourist Promotion Board
- Singapore Tourism Board
- Singapore Tourism Board

Showing 1 to 5 of 10 entries

MEMORIES

- WOW! I'm Mr Sentosa?
- Construction of the Singapore Tourist Promotion Board's headquarters : de-sanding the bentonite
- The Crown Princess of Thailand receiving a gift from the Chief Executive of Singapore Tourist Promotion Board
- Construction of the Singapore Tourist Promotion Board's headquarters : a view of the construction site
- Construction of the Singapore Tourist Promotion Board's headquarters : aerial view from overhead pedestrian

WORKS BY

- Little India Through the Eyes of Dr Uma Rajan
- Visit ASEAN Year 1992
- Visit ASEAN Year 1992
- Singapore Out Of This World
- Singapore Out Of This World

Showing 1 to 5 of 1,209 entries

RELATED SUBJECT

- Trave Singapore
- Tourism Singapore
- Tourist trade Singapore

Showing 1 to 3 of 3 entries

Figure 2: Linked Data Widget to Entity Page

As seen in Figure 2, once a user selects an entity on the Linked Data widget, they are brought to the entity page corresponding to the entity. On this page, they are given more information about the entity, as well as access points to other resources related to the entity, both physical and digital. Through automatic indexing, we can interlink our resources, creating an ecosystem in which different collections can interact with each other, and are discoverable through each other.

III. METHODOLOGY AND FINDINGS

In 2017, NLB embarked on an extension of the NER project and scope. We introduced more databases to be analysed, including both unstructured data and metadata as input. All in all, there were about fifty thousand resources to be analysed, with about ten thousand of these being metadata records. Furthermore, we explored the possibility of using the topics (SingHeritage) extracted to populate the metadata of records from the National Archives of Singapore (NAS). NAS has millions of records on their digital catalogue (http://www.nas.gov.sg/archivesonline/) which do not allow for subject filtering on NLB's consolidated search platform due to their lack of keywords in their subject field. However, the sheer size of the collection meant that it would require an enormous amount of effort to manually enter indices for these documents to make them searchable. For reference, the SingHeritage controlled vocabulary includes terms such as:

Table 1: Examples of terms in SingHeritage Vocabulary

Terms	Vocabulary Class
Arts Organisations	SingHeritage
Badminton	SingHeritage
Heritage and Culture	SingHeritage
National Monuments	SingHeritage

These terms were crafted to fit the scope of the domain which our collections covered. As mentioned above, the consolidated search platform consists mostly of resources from cultural organisations, and the terms in our SingHeritage vocabulary class reflect as such.

Before embarking on the NER process, preparations had to be made to prepare the datasets which were going to be analysed. For this extended scope, we used nine datasets, consisting of four metadata datasets, and five unstructured data datasets:

Table 2: Datasets used for NER

Name of dataset	Source	Type of resource
Infopedia	NLB	Unstructured Data
HistorySG	NLB	Unstructured Data
Roots.SG	NHB	Unstructured Data
Speeches and Press Releases	NAS	Unstructured Data
Oral History Transcripts	NAS	Unstructured Data
Photographs	NAS	Metadata
Audio-visual Materials	NAS	Metadata
Maps and Building Plans	NAS	Metadata
Posters	NAS	Metadata

Metadata was chosen as datasets due to the high volume of resources in NAS catalogue which were non-textual in nature. Furthermore, the current metadata of NAS records did not contain any keywords in their subject, making them impossible to filter for on NLB's consolidated search platform using subject headings. By extracting the named entities from these records, it would then be possible for the records to be integrated into the current Linked Data ecosystem through the widget.

As we had decided to run NER only with content in English, we had to filter out non-English contents. While this was not a problem for the metadata datasets, or the curated articles found in NLB and NHB's sites, it was prevalent in the Oral History dataset. The Oral History and Speeches and Press Releases datasets also came mostly in PDFs, and not all of them had been subjected to Optical Character Recognition (OCR) processes. Thus, those had to be removed as the NER process would require text to be effective.

Once the datasets were deemed to be suitable for NER, they were sent to a vendor who did the process for NLB. We were to have three rounds of analysing, with a check for accuracy and possible issues between each round. We also sent them our controlled vocabulary in two formats – one with the variants for each key term, and one with the bibliographic information regarding the terms.

Before the first round of extraction and assignment was done, we picked out 90 samples, proportionately from each dataset, to do manual indexing on. This was done to have a basis of comparison and accuracy checking when the first round of NER was finished. Accuracy was calculated using the Information Retrieval Evaluation measures, with equal weightage on both Recall and Precision. The equation which was used is as followed:

$$F_{\beta=1} = \frac{2PR}{P+R}$$

where P = Precision and R = Recall

In the case of the automatic indexing done through NER for our resources, precision refers to the fraction of correctly (i.e. are in the resource) extracted entities among all the entities extracted by the NER process. On the other hand, recall refers to the fraction of correctly extracted entities among the entities found in the resource which are in the controlled vocabulary. An equal weightage of both was selected as our end goal (automatic indexing) did not seem to favour one or the other metric in terms of accuracy and usability. For contractual acceptance, we were expecting an accurate retrieval percentage of 80% on average across the nine datasets used.

Verification: Round 1

From the very first round of accuracy verification, it was clear that there were major issues which had to be fixed for NER to work as an automatic indexer for NLB and related agencies' resources. One major issue was the wrongful extraction of entities due to their possible variations matching commonly used words. For example, if a sentence in the resource contained the word "most", the NER system might extract "Museum of Shanghai Toys" as an index. This was a result of "MoST" being an acceptable variant for the above entity. This issue was further exacerbated by variant forms which were also nouns, which made using Part-of-speech differentiation for these variants more difficult. One example of such an entity is "Singapore Centre for Social Enterprise" with the variant "raiSE". The table below lists a few more examples across the vocabulary classes to show the prevalence of this issue:

Table 3: Examples of common word variants

Entity Name (Preferred)	Variant Name	Vocabulary Class
Singapore Institute of Technology	SIT	Organisation
East Asia School of Theology	EAST	GeoBuilding
Yuan, Dian	Men	Person

There were also instances of entities not being extracted due to the type of punctuation found in its' name. This was especially prevalent in entities which contain an apostrophe, such as "St John's Island". When this was brought up to the vendor, they mentioned that it was due to the "regular expression" they had used to filter the entities in their NER processing. For example, they might have chosen to not extract entities which are not fully made up of alphanumerical elements, thus causing these entities with punctuations to not be extracted. This was rectified quickly and was not an issue in subsequent runs of the NER process. This instance (along with other instances of issues) really helped to emphasise the necessity of multiple runs to improve upon the accuracy, and also to assess the structure of our controlled vocabulary, and how it can be altered for the purpose of NER to enable more accurate and precise indexing.

Another issue we observed was a problem that occurs in other uses of NER as well – the partial match of entities. However, the occurrences of partial matches in our NER project seemed to be due to the terms found in our controlled vocabulary. For example, "York Hill Home" was not extracted as an entity as it did appear in our controlled vocabulary as an entity, but the entity "York Hill" was extracted as it was a part of our vocabulary.

The extraction of entities also allowed us to see the shortcomings of some of the databases, and our controlled vocabulary, used in the NER project.

1. Oral History: While the PDFs used for the NER process were checked to be OCR-ed, the quality of the text on the PDFs left much to be desired. Due to the inaccurate text, some entities were left out, while others were wrongly extracted due to the same error.
2. Speeches and Press Releases: As the documents used for this process were either PDFs or HTMLs of official press releases or speeches, they contained a header and/or footer. However, these portions were also text used in extraction, and the multitude of variations made it difficult for the vendor to differentiate which part of the text was a header and/or

footer, or actual content. We did not want entities which only appeared in the header and/or footer to be on the widget as it would not accurately encompass the context of the resource.

3. Controlled Vocabulary: The dictionaries used in the NER project came from our KOS system, which primary use is for cataloguing digital materials. As such, the entities from it followed the format of the Library of Congress Name Authority, which had idiosyncrasies such as separating the last name and first name with commas.

As can be seen from the above, automatic indexing using NER is not perfect. There were numerous issues being discovered, coming from both the NER process as well as the data used in the process. For automatic indexing to work for NLB, it was necessary to fine-tune the process, as well as ensure that the documents used in the process were suitable for automatic indexing.

At the end of the first round, metadata content was seen to have met the retrieval requirement. However, some of the unstructured data resources had not. For accuracy verification, mistakes made due to the problems with our OCR files were not included, as the accuracy verification was designed to only evaluate the NER system itself.

Table 4: Precision, Recall, and Retrieval by Content Type (Round 1)

Type of resource	Precision	Recall	Retrieval (%)
Metadata	92%	97%	94%
Unstructured Data	68%	95%	79%

As can be seen by the aggregated results, Recall is significant higher than Precision in the content resources. What this seems to signify is that the NER process was able to accurately extract the entities found in the resource, but at the same, returned a fair amount of noise.

Verification: Round 2

We quickly realised that the second round of verification was necessary for us to fine-tune the NER process. Due to the changes we requested for the second run of the NER process to rectify the issues identified in the first round of verification, it resulted in errors which did not occur in the first run. This resulted in a few of the dataset's retrieval percentage dropping drastically. It should be noted however; it did resolve some of the issues we presented to the vendor after the first round. While it was not possible for the vendor to completely isolate and ignore the headers and/or footers of a resource, they were able to solve the issue of common words by introducing stricter NER rules and parameters.

Table 5: Precision, Recall, and Retrieval by Content Type (Round 2)

Type of resource	Precision	Recall	Retrieval (%)
Metadata	92%	72%	78%
Unstructured Data	82%	85%	82%

For the second round, we also took a closer look at the Time vocabulary and how it was being extracted. NLB's Time vocabulary consist of each individual year, as well as periods of time. For example, "1920" was an entity in the dictionary, as well as "1920s". The purpose of including the Time vocabulary was to possibly group resources together by year, allowing users to learn about the events and happenings in Singapore in a particular year or time frame. However, through analysing the results, it was clear that NER would just take every year mentioned in the resource and extract it as an entity. While this might work for the metadata resources, it did not work for our unstructured resources. For example, here is an abstract from our Oral History dataset:

> *OW: …. A year later I was promoted to Standard III in the same school. This was a mixed class and it was in this class that I began mixing with non-Malay students. That would be about **1934**, I think. I completed my primary education – Standard V, in **1936**. Then went to Raffles Institution – Standard VI in 1937. In **1941** I was in the Senior Cambridge…"*

As can be seen from the abstract, there are three different years in just one paragraph alone, and none of them can serve as an index as they do not encompass the entire idea of the resource.

Verification: Round 3

Having three rounds enabled us to achieve an optimal state of the NER process which did not compromise retrieval accuracy for a few possibly unavoidable inaccurate extractions. It also allowed us to focus on deciding which datasets to deploy the Linked Data widget on during the last round of verification. Aside from the accuracy of the entities being extracted, the important parameter was how useful the entities extracted were for indexing the resources to create paths between them. To evaluate this, we looked at the entities extracted in relation to the resource in which they came from i.e. to determine if the entities extracted accurately index the resource, rather than just being extracted because it was briefly mentioned in the resource. When we did so, we realised that the "Title" field we had been using for entity extraction in the photographs dataset (metadata) referred to the collection title rather than an individual title. Therefore, there were entities extracted which were not pictured or mentioned in the photograph.

Another dataset which we decided not to implement the widget on was the Oral History dataset. Due to the unsatisfactory quality of the OCR text on the PDFs, it would not be advisable to use the text to derive indices through NER (which is inherently text-based).

The last round of verification also showed us that our SingHeritage (topic) vocabulary was currently not suitable for use of indexing. Firstly, NER was designed to be used to extract named entities i.e. proper nouns. Our topical vocabulary mostly contains words which encapsulate a concept i.e. common noun. The NER process we utilised did not consider more traditional automatic indexing methods such as word frequency in a collection to determine if a topic was accurately representing the resource. Secondly, the SingeHeritage vocabulary was not created with a domain in mind, as it was meant to be used for multiple domains. Furthermore, some of the databases used were not restricted to a specific domain as well. For example, the Posters databases included posters of any topic, be it about a movie screening or a campaign for eco-consciousness. Thus, the decision was made to not go forward with the population of the subject field for NAS resources. Instead, NLB will continue to explore other options to further expose these resources beyond the Linked Data widget.

IV. IMPLICATIONS

Through the two NER projects that NLB has embarked on, much has been learnt about the feasibility of using NER to conduct automatic indexing instead of using tools designed specifically for automatic indexing. One major takeaway is that since NER is being used, it is more viable to use it for extracting indices which are named entities, rather than conceptual words. This will still allow us to connect different resource collections with each other through the entities found in the content or metadata of the resource. If it is necessary to index using conceptual words, it might be more prudent to use NER alongside a different tool to compensate for the broadness that conceptual words can introduce to the indices.

Another major takeaway is that automatic indexing is not fully automatic. What this means is that while the process itself can be automatic, there is a significant amount of preparation work to be done on the content being processed, and there is a need for a manual indexer to check the output given by automatic indexing to determine if the process has worked effectively. Furthermore, if no vocabulary or ontology exist before the start of automatic indexing, it has to be created and curated by people who have knowledge of the specific domain(s).

Lastly, automatic indexing, using NER or otherwise, does not guarantee 100% accuracy. Instead, what we can do is to implement measures and steps to mitigate as much inaccuracies as possible, to make automatic indexing useful to our work through an improvement in quality, and a reduction in time spent.

V. NEXT STEPS

Moving forward, NLB plans to continue exploring NER and other automatic indexing processes and software to ensure that our growing collection of resources is linked in the Linked Data ecosystem. This will be a challenge as in its current implementation, NER is done on existing batches of collections. It will be necessary to consider the workflow of keeping up with new records and ensuring that our records remain discoverability amongst each other.

NLB also plans to evaluate the usefulness of the SingHeritage and Time vocabulary and consider what can be done to make these vocabularies more viable for our automatic indexing needs.

VI. CONCLUSION

Through the linked data widget, NLB has successfully created more linkages between its resources using the named entities found in both the unstructured data and metadata. Following this process, we will be launching the widget on the chosen NAS databases as mentioned, as well as a refresh of the widget on the sites it currently already populates.

Our team will also be looking into other services which can utilise the entities extracted from the two NER projects.

References

Beliga, S. (2014). Keyword extraction: a review of methods and approaches.

Bunescu, R. C., & Pasca, A. M. (2015). *United States Patent No. US9135238B2.*

Hussein, H. (2015). Linked Data @NLB. *Singapore Journal of Library and Information Management, 44,* 20-34.

Jones, K. S. (1974). Automatic Indexing. *Journal of Documentation, 30*(4), 393-432.

Roy, S. (2017). Named Entity Recognition. *AKGEC International Journal of Technology, 8*(2), 38-41.

Vallez, M., Pedraza-Jimenez, R., Codina, L., Blanco, S., & Rovira, C. (2015). A semi-automatic indexing system based on embedded information in HTML documents. *Library Hi Tech, 33*(2), 195-210.

Wu, Y.-F., Li, Q., Bot, S. B., & Chen, X. (2005). Domain-specific keyphrase extraction. *Proceedings of the 2005 ACM CIKIM International Conference on Information and Knowledge Management*, (pp. 283-284). Bremen.

Yang, S., Zhang, B., Li, S., Yu, C., & Hao, Q. (2014). Keyword extraction using multiple novel features. (C. Zhai, Ed.) *Journal of Computational Information Systems, 10*(7), 2795-2802.