Digital Humanities from the Ground Up: The Tamil Digital Heritage Project at the National Library, Singapore

Sharmini Chellapandi, National Library Board, Singapore

The Asian Conference on Literature, Librarianship & Archival Science 2016
Official Conference Proceedings

Abstract
To commemorate Singapore's 50th year of independence in 2015, a community-led group sought to present a gift to the nation in the form of a digital collection of 50 years of Tamil literary writing in Singapore. The National Library, Singapore as the national repository of Singapore's published heritage was approached to be a key partner to create this digital archive. This saw two unique features taking place - digital humanities from the ground-up and secondly, a ground-breaking initiative in which Tamil content was digitised and made searchable online through optical character recognition (OCR). This project was challenging as Tamil OCR is still a developing technology and the objective to digitise and OCR Tamil works is an endeavor that has not been attempted on this scale before.
This presentation will highlight how writers, teachers and the National Library collaborated to bring justice to books that were not easily available and create a ready-resource of local literary works by transforming physical books into a digital resource. Extensive community resources were mobilized to annotate and proofread these books. As a result, not only has the historical record of Tamil creative writing been preserved but it has become the most comprehensive resource of Singapore Tamil literature available and this has opened up many possibilities in teaching and learning as well as in raising awareness to a wider audience.

Keywords: digital humanities, Tamil, literary arts, OCR, digitization, archive

Introduction

Singapore celebrated its 50th year of independence in 2015, marking a transformation from a British colony to a modern city-state and financial hub.

An island nation in the heart of Southeast Asia, Singapore has a multi-cultural and multi-ethnic population. The three major ethnic groups are the Chinese who form the majority at 74.2%, the Malays (13.3%) and Indians (9.1%) and a substantial number of Eurasians, Europeans and other Asians. English is the lingua franca and one of the four official languages along with Chinese, Malay and Tamil (spoken by a majority of Singaporean Indians), which are aligned to the country's major ethnic groups.

This paper will highlight how writers, teachers and the National Library Board (NLB) collaborated and implemented a ground-up initiative to bring justice to books that were not easily available. It will also share how through this initiative, a ready-resource of local literary works was created by transforming physical books into a digital resource that were not only annotated but searchable online.

Tamil Literature in Singapore

NLB manages the National Library (NL), 26 Public Libraries and the National Archives. NL's statutory mandate is to collect, organise, preserve and make accessible Singapore's published heritage. NL also seeks to promote and inculcate knowledge of Singapore's history and multicultural heritage, by collecting early documentation and unique materials.

Legal Deposit is one of the statutory functions of NL and through this, all works by Singapore writers, producers and publishers are collected and preserved to become a part of Singapore's heritage. Given its statutory role, there is an expectation for NL to safeguard Singapore's literary heritage and play a more active role in promoting the appreciation and creation of local literary works. Thus, the NL aims to build a comprehensive collection on the development of literary arts in Singapore in all four languages. The table below provides a brief overview of the coverage and scope of the literary arts collections at NL.

Colonial Era 1826 - 1941	**Post-War Period 1945 - 1965**	**Modern Period 1965 -**
• Literary publications and manuscripts	• Comprehensive collection of published works by Singapore writers and literary publications • Primary sources (e.g. manuscripts, personal papers) from Malayan-Singapore and Singapore writers	

Table1: Overview of the literary arts collections

Tamil Literary Arts at the National Library

The literary arts collection at NL comprises published and unpublished materials in English, Chinese, Malay and Tamil. There is a comprehensive Tamil literary arts collection of major published works from 1945 received through donations and Legal Deposit. The collection also has works by local authors published outside Singapore.

In the 1950s to 1970s, Tamil literary works were largely published in newspapers and literary periodicals, radio plays and drama scripts as these were the more popular and accessible platforms for the community in general.

Major themes of Singapore Tamil literature mirrored Singapore's development through the years. Many books published in the 1960s to 1980s dealt with the recognition of Tamil language and society in independent Singapore and Malaysia. Politics, socio-economic issues, Indian traditions in modern society as well as religion and atheism were some of the common strands in literary works published from the 1990s. A younger generation of writers including a large number of expatriates, who had settled in Singapore, were writing prolifically and also becoming more active across digital platforms.

The Tamil Digital Heritage (TDH) Project

To celebrate Singapore's 50th year of independence, a group of community leaders passionate about the Tamil language, wanted to present a gift to the nation. They wanted to give voice and visibility to 50 years of Singapore Tamil literature and approached the NL to be a key partner. The project received widespread interest and support from the community and was officially launched in October 2013 with a target to present this gift to the nation in August 2015.

The objective of the TDH project was to digitise and perform optical character recognition on Tamil works. The digitised collection, consisting of novels, short stories, poems, plays and essays published between 1965 and 2015, would be housed in an online platform BookSG, hosted by NLB.

Although there was a small and strong community of writers, many early publications were either out of print or hard to find. A number of older authors did not own some of their early works and relied on the library's collections. At the same time, the lack of Singapore Tamil literary works at local bookstores, their limited print runs and lack of widespread access had an impact on the teaching and learning community. Many teachers bemoaned the lack of local books with content that could inspire the younger generation and for them to have a connection with.

Fortunately, NL as the national repository and supported by Legal Deposit, had a comprehensive collection. These books were in safe custody but there was minimal access and usage.

The TDH Project and Digital Humanities

The community-led group wanted to create a digital archive so that these literary works could be digitised and made accessible in a new and unprecedented manner. They had strong networks within the community and were able to reach out to Tamil language teachers and writers; two important groups that had direct relevance and impact on this endeavour. The only way to bring justice to Singapore Tamil literature was to find a novel way to bring out this hidden collection. However, how could justice be done to these books?

Digitisation was the obvious answer but this was not a straightforward or simple solution. There are two parts to digitising text-based print materials: scanning and Optical Character Recognition (OCR). OCR software recognizes text in a document image by analyzing images through image processing and pattern recognition. When analogue texts are converted into digital formats, they are scanned to create a photo image and have OCR technology applied to enable the scanned texts to be searchable.

OCR consists of a number of preprocessing steps followed by the actual character recognition. These include gray scale conversion, skew detection and correction followed by character segmentation. The types of preprocessing algorithms and tasks used on a scanned image would depend on many factors such as the quality of the paper, resolution of the scanned image, the amount of skew in the image (which refers to orientation of the document), the format and layout of the images and text, the script used and whether the characters are typed or handwritten.

OCR for Tamil content is still a developing technology. Tamil books could be digitised in terms of scanning but they were not OCR-ed. Thus, Tamil e-books and other digitised texts could only be read like a photo image and were not searchable. Tamil OCR had been tested out within academia for research and development and there were many research papers on this topic. However, when the TDH project started in late 2013, this software was not readily available nor was it tested on a larger-scale. Many digitisation vendors were not fully equipped to handle Tamil OCR competently. Furthermore, the paper quality of the Tamil books was poor and thin which meant that several layers of corrections had to be made get the texts OCR-ready.

There were, thus, many challenges that needed to be overcome for this project. This included engaging a vendor with the capabilities to fine-tune and further develop the technology for Tamil OCR while concurrently working on a steady pipeline of digitised texts sent by NL for OCR conversion. This challenge was further compounded by the deadline that was tied to Singapore's national day celebrations.

Despite the challenges, OCR for digitised Tamil text was a ground-breaking initiative. To make the digital archive a reality and to do justice to this project, the collaboration and involvement stakeholders were critical.

Collaborating with the Community

The TDH project was an example of digital humanities from the ground-up. The project was mooted by the community and gained widespread support within the larger community. It was spearheaded by the TDH Working Group made up of community leaders and representatives from the media and education sectors, who provided the strategic direction. They also galvanised community support and involvement. Tamil language teachers from Singapore schools as well as a community of Tamil writers, many of whom were the very authors of the literary works to be digitised, were the other important blocks who believed in the objectives of this project and helped to make it a reality despite the many challenges.

As a key partner, NLB was responsible for the technical and professional aspects of the project. It took charge of vendor management, digitisation and meta- tagging as well as the ingestion and upload of digitised content. NLB and the community volunteer groups worked together to secure copyright permissions from authors and publishers to enable the digitisation of the physical books, in media relations and publicity as well as in the annotation and proofreading of the books. The community, in turn, oversaw volunteer acquisition and management.

For the purposes of this paper, which focuses on the ground-up effort of the community, two important tasks - annotation and proofreading - which depended heavily on their support will be discussed.

Annotations

The annotation of books was the first major task that involved a large group of volunteers. The community had shared that annotations would be a useful addition to the bibliographic details as these would facilitate online searching based on keywords and a quick summary of the book without having to read it in full or download the full text.

The annotations took place from November 2013 to April 2014 soon after the project was launched in October 2013. Several rounds of discussions were held with the TDH Working Group to ensure that the objectives for this exercise were clear as well as to draw up a list of guidelines on how to annotate with sample annotations for the different genres such as novels, short stories and poetry anthologies. Following this, several briefing sessions were conducted in November and December 2013 to a large group of volunteers.

This group was largely made up of writers. This garnered popular support as the writers felt a sense of ownership towards a project that they felt was meaningful. An annotation coordinator, a senior teacher who was also a part of the TDH Working Group, had oversight over this exercise. Each group was led by a group leader who was tasked with advising group members and reviewing their annotations.

Prior to start of the annotation exercise, a special session was organised for Singapore writers to share more details about the project and to obtain their copyright permission to enable NLB to proceed with digitisation and upload of the digitised works onto the BookSG portal. At the same time, writers who had volunteered to assist in the

annotations were able to borrow the library books. Their loan duration was extended to ensure there was sufficient time to read and complete their annotations.

Annotation Steps and Processes

The NLB had prepared a listing of the basic bibliographical details such as name of author, title and publisher name. The annotator's task was to enter the printer details as this information was usually not captured by cataloguers. Tamil book publishing was unique as authors typically double up as publishers in terms of copy-editing and proof-reading. The final manuscript or word document would then be handed over to the printer for printing. Entering details on the printer was unique to this collection and to the Tamil collection as a whole. In terms of the annotation, the volunteers had to succinctly provide a summary about the book assigned containing details about the genre as well as main points that would help the user know more about the content of the book. At the end of this phase, over 450 books were annotated.

Figure 1 below is an example of an annotation template while Figure 2 shows a sample annotation for novels.

S/N	Annotator's Name	Author	Title	Printer/ Publisher	Annotation
எண்	குறிப்புரையாளரின் பெயர்	நூலாசிரியரின் பெயர்	நூலின் தலைப்பு	அச்சகம்	குறிப்புரை

Figure 1: Annotation Template.

தலைப்பு	அச்சகம்	குறிப்புரை
நினைவுகளின் கோலங்கள்	பாமா பிரசுரம், சென்னை 600002	எழிலரசி என்பவள் சிங்கப்பூரில் பிறந்து வளர்ந்தவள். அவள் பெற்றோர் தமிழ்நாட்டில் உள்ள ஒருவனுக்கு அவளின் சம்மதமின்றி அவளுக்குத் திருமணம் செய்துவைக்கின்றனர். அங்கு எழிலரசி படும் அவதிகளைச் சித்தரிப்பதாக இந்நாவல் அமைந்துள்ளது.

Figure 2: Sample Annotation for novels.

Proofreading

The second and more complicated and intensive task that tapped on a larger pool of volunteers was the proofreading exercise. This took place from April 2015 to August 2015. There was a delay of over 5 months due to the difficulties faced by the vendor in completing the preprocessing tasks and creating a searchable PDF of a reasonable quality that could be replicated for all the text files.

The purpose of the proofreading exercise was to conduct quality-checks on OCR-ed texts to ensure that the content was identical to the original. As opposed to

proofreading a text before it goes into publication, which was the conventional understanding of the term, proofreading in this instance referred to pattern-recognition to ensure that the characters in the OCR-ed texts were identical to the original, including errors that were inherent in the original book.

In addition, as mentioned in earlier sections, the technology for Tamil OCR was still developing, there were many errors in the OCR-ed texts that had to be highlighted. Volunteers were advised to look out for spelling errors, wrong sentence sequences, reverse order of characters or missing punctuation marks. All errors were to be highlighted for the vendor to correct and to create an updated version. The general guidelines were for proofreaders to continue checking line by line if the overall accuracy rate was approximately 80%. If texts consistently had many errors (ie more than 20% errors per page), these would be flagged for the vendor to re-do the OCR for that particular title. Figure 3 below is an example of checking the OCR-ed text (right panel) against the original (left panel). The highlighted words indicate errors.

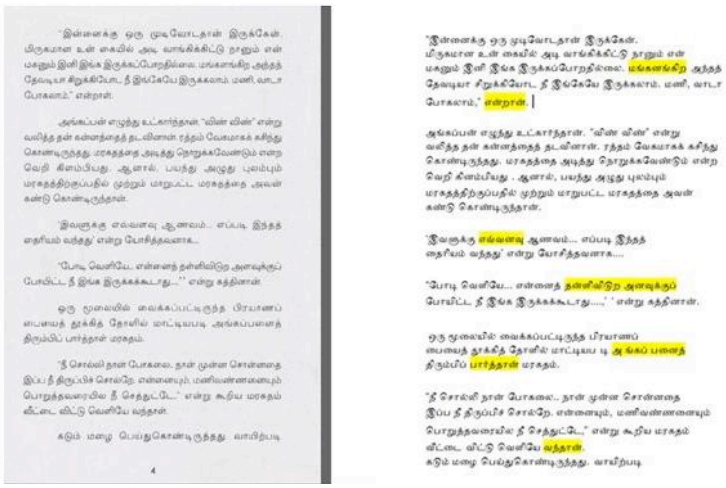

Figure 3: Sample proofreading of OCR-ed text (right) with the original (left)

Figure 4 below is an example of a page with many errors. Books that fall in this category were flagged for the vendor's follow-up action.

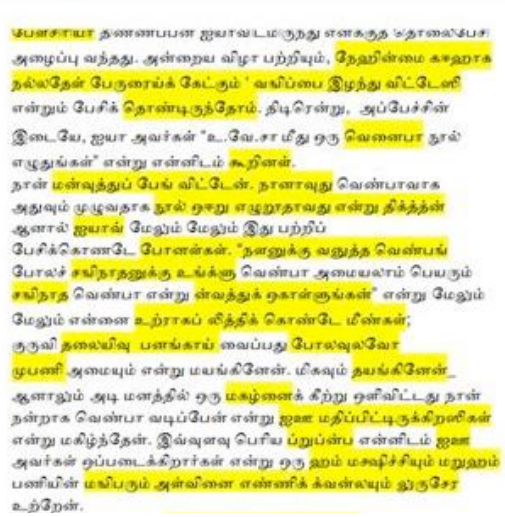

Figure 4: Example of an OCR-ed text with many errors

Proofreading steps and processes

Following the same approach as the annotation exercise, a coordinator from the TDH Working Group who was also a senior Tamil teacher, was appointed to oversee the proofreading exercise. The coordinator worked with group leaders who were each assigned 10 to 12 members. Each member was allocated between 1 to 3 books, depending on their schedule, to proofread. The group leader helped to guide their members and conducted quality-checks of the proofread files. The coordinator conducted random quality checks on these files and handed these over to the NLB representative at regular intervals. NLB conducted a final review before the files were sent to the vendor.

Given the large file sizes, Google Docs was used as the platform to allow the upload of new and completed files. Two separate accounts were created; one for the volunteers and the second for the vendor. NLB oversaw both accounts and ensured that the process flow was adhered to given the numerous files and versions of files that were either new, corrected or in need of re-work and uploading them in the correct folders.

The unique aspect of the collaboration with the community was their dedication and conscientiousness. Firstly, the tasks assigned had clear instructions and guidelines and these were effectively communicated to all volunteers. Secondly, many volunteers who were school teachers, worked after hours and during their rest days to complete their assigned tasks. Both the annotation and proofreading exercises took up considerable effort and time with over 200 volunteers from Tamil language teachers, writers, community groups and staff from NLB coming together to work over a compressed period of time. It was a combined community and NLB effort. At the end and in time for the launch evet in August 2015, over 450 books were annotated and 55,000 pages (350 books) proofread.

The TDH collection in BookSG

BookSG is an online portal of the NLB containing digitised books and other printed material. The digitised TDH collection was uploaded and made accessible in this portal (see Figure 5).

Figure 5: Home page of the TDH Collection in BookSG

As the copyright owners had granted permission for these works to be made accessible online as well as downloaded for research and educational purposes, with the TDH collection ingested and uploaded in BookSG, over 350 Singapore Tamil literary works finally saw the light of day. Users could either read the digitised book online or download and read it offline.

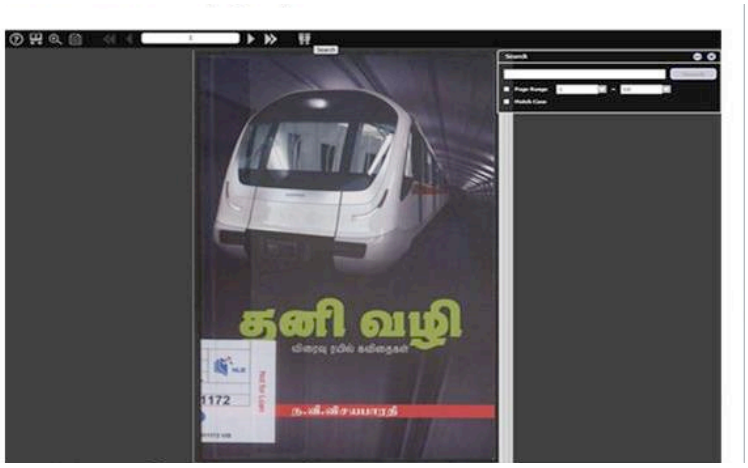

Figure 6: Reading the digitised e-book

When reading the book online, a 'flipviewing' page opens and clicking on the binoculars icon opens up a search box on the top right corner of the screen.

Users are able to search for keywords and phrases by entering them into the searchbox. The search word(s) and the corresponding pages where these words or phrases can be found will be highlighted. This is a feature that was not possible in digitised Tamil texts before due to the non-usage of OCR technology in this manner.

Conclusion

This paper aimed to highlight how writers, teachers and the National Library collaborated to bring justice to books that were not easily available and create a ready-resource of local literary works by transforming physical books into a digital resource. By creating this digitised collection online and working together, justice was served on several fronts.

Firstly, this was a ground-up effort by the Indian community and the mobilisation of community resources which propelled the project forward and gave rise to media interest and publicity. Secondly, through the community's dedication, Tamil OCR became a reality and this was 'ground-breaking' as this was an endeavour that had not been attempted on this scale before. Thirdly, the project had opened up many possibilities especially to the Tamil teaching and learning community and the Tamil diaspora. The TDH Collection has become the most comprehensive resource of Singapore Tamil literature available. Finally, for Singapore Tamil writers, this collection has become the most reliable method to preserve their works at minimal cost to them and has provided a platform that was visible and easily accessible.

In conclusion, this project, while attemptng to implement a unique and unprecedented technology, is also a study of how greater collaboration between various stakeholders with a common objective could be implemented

References

Contact Singapore. (2016). Facts & Figures. Retrieved from https://www.contactsingapore.sg/en/professionals/why-singapore/about-singapore/facts-and-figures

National Library Board, Singapore (2016). Statutory Functions of National Library. Retrieved from http://www.nlb.gov.sg/About/StatutoryFunctionsofNationalLibrary

National Library Board, Singapore (2016).Tamil Digital Heritage Collection Retrieved fromhttp://eresources.nlb.gov.sg/printheritage/browse/Tamil_Digital_ Heritage_ Collection.aspx

Kannan, J., & Prabhakar, R. (2009). A Comparative Study of Optical Character Recognition for Tamil Script. European Journal of Scientific Research, 35 (4). Retrieved from http://www.eurojournals.com/ejsr.htm

Seethalakshmi, R., Sreeranjani, T.R., Balachandar, T., Singh, A., Singh, M., , R., Ritwaj (2005). Optical Character Recognition for printed Tamil text using Unicode. Journal of Zhejiang University Science, 6A, pp. 1297-1305. doi: http://dx.doi.org/10.1631/jzus.2005.A1297

Contact email: Sharmini_Chellapandi@nlb.gov.sg