Challenges of Digitizing Vernacular Newspapers & Preliminary Study of User Behaviour on NewspaperSG's Multilingual UI

Mazelan bin Anuar, Cally Law and Soh Wai Yee

Abstract

Singapore is a multicultural nation. About 76 per cent of Singapore's population is ethnically Chinese, 14 per cent Malay, nine per cent Indian and the rest comprising other ethnic groups. Borne out of a pragmatic need to operate in the global economy using the English language, Singapore adopted the bilingual education policy in 1966. As a result, the majority of the Singaporean citizens are bilingual today, able to communicate effectively in English and their mother tongues. Four official languages are used in Singapore, namely English, Mandarin, Malay and Tamil.

NewspaperSG is an online resource of Singapore newspapers published from 1831 to 2009. It was started in March 2009 with one major newspaper title in English, namely the Straits Times. Its coverage has now expanded to 29 titles, including Chinese, Malay and Tamil newspapers. With the introduction of these newspapers since August 2011, the multilingual interface was introduced in NewspaperSG to allow users to navigate the portal in English, Chinese or Malay.

The paper will highlight the process and challenges that the team faced during the digitization of the multilingual newspapers, especially newspapers in the Chinese language. It will also present the preliminary findings on the study of user behaviour in NewspaperSG's multilingual environment.

Introduction

Singapore was opened as a British trading post in 1819. Its first newspaper, Singapore Chronicle and Commercial Register, started in 1824. Since then more than 120 newspapers had been published in Singapore but many of them were short-lived and had ceased publication. However, these newspapers serve as important primary sources of information for researching into the nation's past and development. Historic and current newspapers are frequently consulted by students, researchers, government officers and working professionals at the Lee Kong Chian Reference Library primarily through microfilm. This collection currently stands at 22,000 reels and comprises newspapers in the four official languages of Singapore, namely English, Chinese, Malay and Tamil. There are newspapers in other languages as well such as Arabic, Malayalam and Punjabi.

Microfilming has been and still is the main means of preserving and providing access to the library's newspaper collection. All of our newspapers have been comprehensively microfilmed. The library continues to preserve all local papers through microfilming, though a commercial vendor now handles the actual microfilming process. The Legal Deposit Unit continues to store the preservation copy.

The library also subscribes to a number of news databases including Factiva, Proquest Newspapers Complete and Library Pressdisplay which allow users to perform keyword searches on local news content. Though the coverage of local news content may not be comprehensive, they have at least enabled our users to search for current news.

Newspaper Digitisation Project

In 2006, the National Library Board (NLB) started discussions with the newspaper publisher, Singapore Press Holdings (SPH) to explore the possibility of NLB digitizing all issues of The Straits Times. The Straits Times is the longest-surviving English language broadsheet in Singapore. It is also the most-read paper in Singapore and is a frequently consulted resource for both current and historical information. SPH agreed to let NLB digitize all back issues of the newspaper from its inception on 15 July 1845 right up to 2006. From 1 January 2007, SPH would digitally deposit every issue of the paper with NLB.

In March 2009, a trial service (soft launch) was released. Users can perform keyword searches on OCR (Optical Character Recognition) texts of newspaper content. In NLB's first agreement with SPH, the full newspaper content can be searched from the internet, but the full article may only be displayed at the computers in NLB libraries.

The trial service was a success and NewspaperSG was officially launched in January 2010. The content was increased from 1 to 17 newspaper titles (see Appendix 1 for the list of titles available in NewspaperSG now). And with the conclusion of new agreements with the newspaper publishers, a large extent of copyrighted content published prior to 1990 is now available for free to all Internet users. Free remote access was also extended to historical newspapers in the public domain. Full content from 1990 onwards continue to be accessible from NLB computers only.

When NewspaperSG was first started, it attracted an average of 18,000 visits and 70,000 page views every month. Now the average monthly visits and page views have increased to 100,000 and 450,000 respectively. To date, there are close to 18 million articles available in NewspaperSG.

Challenges of Digitizing Vernacular Newspapers

At present, 29 titles are available in NewspaperSG and these include eight non-English newspapers since August 2011 of which three are Chinese newspapers, three Malay newspapers and the other two Tamil newspapers. Only Lianhe Zaobao, Sin Chew Jit Poh (both Chinese) and Berita Harian (Malay) are available for keyword search (full service). The others can only be viewed through browsing via the Preview section with no keyword search function capability. In the case of Sin Chew Jit Poh, only certain issues have been made available for keyword search at this point. Please see Table 1 below.

Non-English Newspapers in NewspaperSG				
Full service		**Preview**		
Searchable		**Non-searchable (browse only)**		
Chinese	**Malay**	**Chinese**	**Malay**	**Tamil**
Lianhe Zaobao (1983-2008)	Berita Harian (1970-2008)	Nanyang Siang Pau (1923-1983)	Warta Malaya (1933-1941)	Singai Nesan Tamil Journal (1887-1890)
Sin Chew Jit Poh (1979-1983)		Sin Chew Jit Poh (1951-1983)	Warta Perang (1941)	Tamil Murasu (1936-2008)

Table 1

Lianhe Zaobao and Berita Harian were the first non-English newspapers to be made available in NewspaperSG. After their inclusion, there were requests for the Tamil newspapers and other Chinese and Malay newspapers to be made accessible via NewspaperSG as well. Due to constraints such as the complexity of the layout of the Chinese newspapers and the unavailability of suitable Tamil OCR software, NLB decided to introduce the Preview feature as an interim solution.

NLB outsources all digitization. Vendors were invited to propose a solution for digitizing the newspapers from microfilm, and provide a content delivery system for the digitized newspapers. Among the criteria was that the solution utilised open standards as opposed to proprietary formats, and that where articles continued over more than one page, they had to be linked. Several proposals were received, ranging from very simple image-only solutions to custom developed solutions. The chosen proposal was one that followed closely to the Library of Congress' NDNP specifications extended to article-level, coupled with a customized Greenstone solution for content delivery, very much like the National Library of New Zealand's Papers Past.

A small sample of a few issues were processed and checked each time. The checking process was a combination of checking every article on every page and examining the accompanying XML files, as well as making sure that continuations were correctly linked, and that the article types were correctly assigned before full-scale production begin.

The process for digitizing the Malay newspaper Berita Harian is the same as the English newspapers. This is because Berita Harian employs the roman script and apart from the language difference that slows down the process of matching articles to illustrations or matching same articles that ran over more than one page, there was no major difficulty in ensuring OCR accuracy of at least 70%. It helps that Berita Harian is a comparatively recent newspaper which dates back to 1957. However, only issues from 1970 to 2008 of Berita Harian have been digitised as the issues prior to 1970 were not covered under the agreement with its publisher (SPH) as there were some uncertainties regarding the copyright ownership. Historically, Singapore was part of Malaya (Malaysia) and Berita Harian was distributed both in Malaya and Singapore before the original publishing company (Straits Times Press) split and the two publishers operated independently in Malaysia and Singapore after 1972.

The process for digitization the Chinese newspapers is similar to the English newspapers as well. However, there are added steps to ensure the quality of the OCR accuracy. Please see Diagram 1 that illustrates the work carried out in Singapore. Diagram 2 illustrates the tasks performed in China which involved significant human input to analyse the digital content before and after OCR to ensure a quality product.

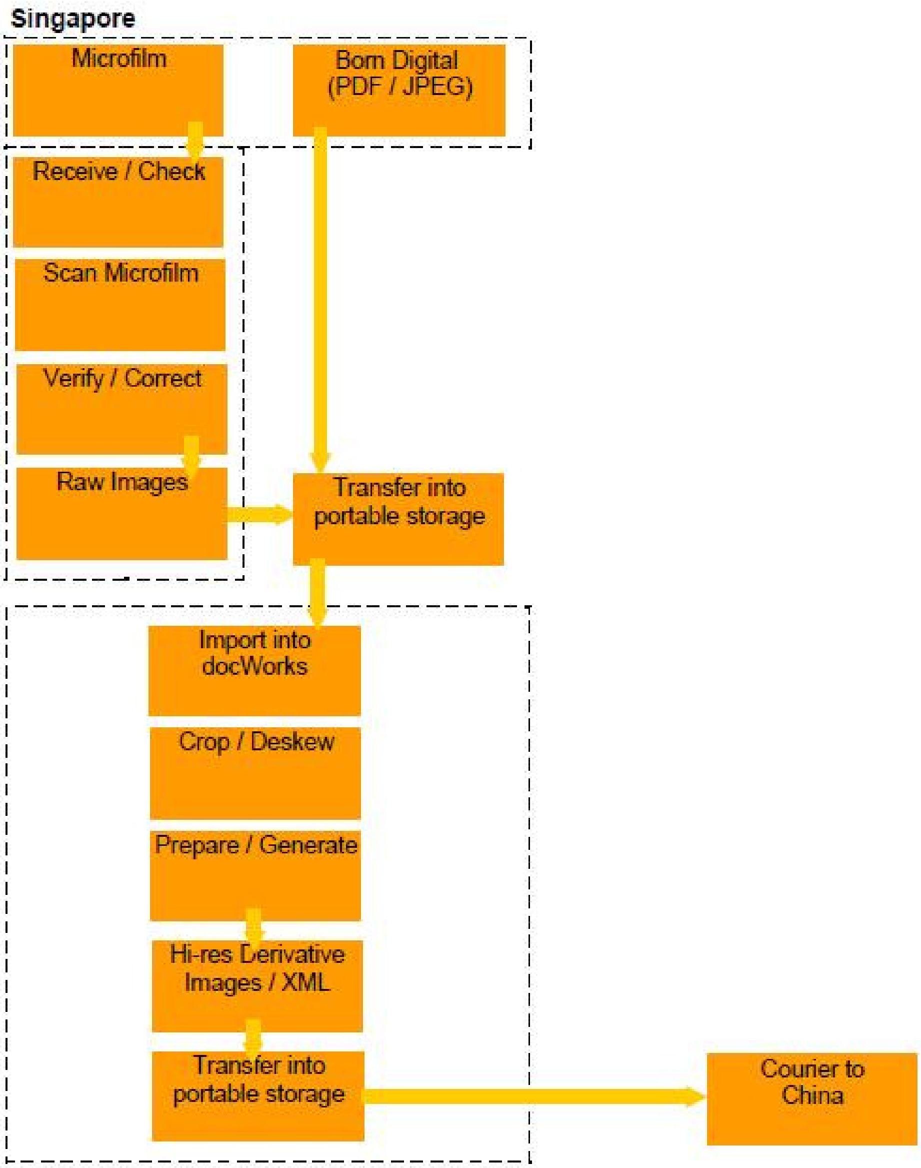
Singapore
Microfilm
Born Digital (PDF / JPEG)
Receive / Check
Scan Microfilm
Verify / Correct
Raw Images
Transfer into portable storage
Import into docWorks
Crop / Deskew
Prepare / Generate
Hi-res Derivative Images / XML
Transfer into portable storage
Courier to China

Diagram 1

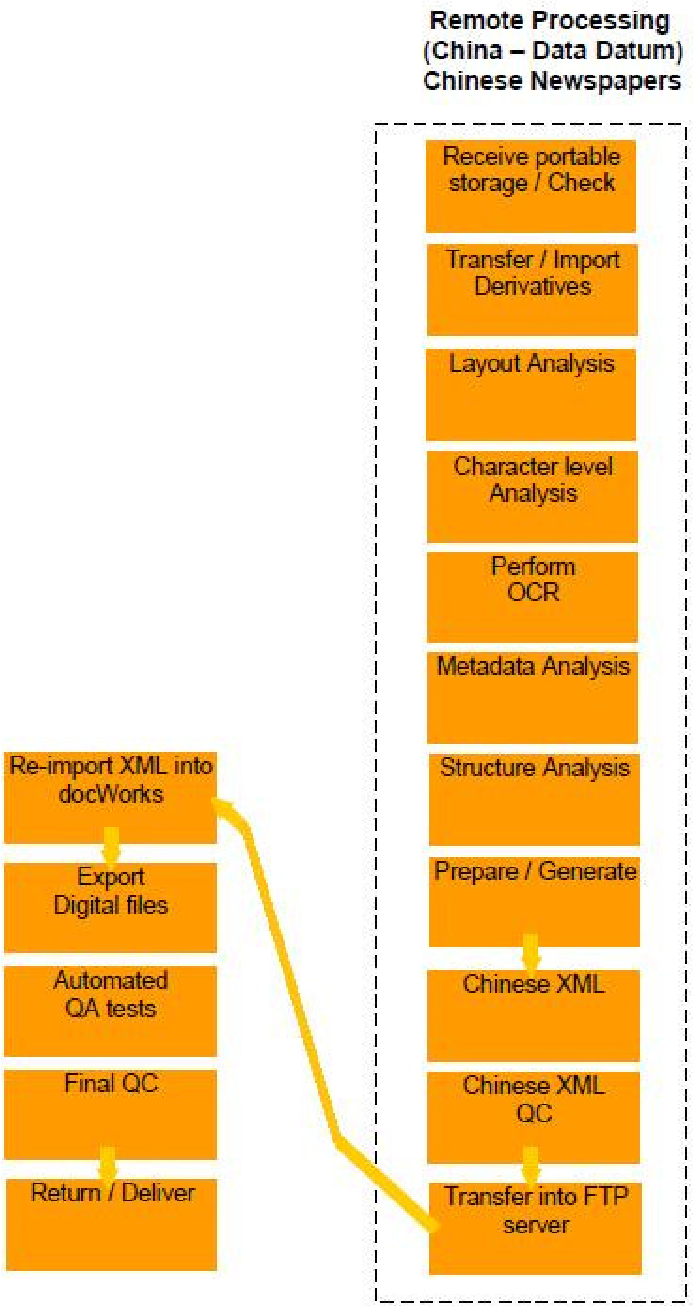

Diagram 2

The difficulties of digitizing Chinese newspapers had already been discussed by Chunming Li Wei Zhang in his paper "Microfilming and digitizing of newspapers in China". NLB face the same technical problems in our effort to digitize the Chinese newspapers in our microfilm collection. The layout and font of the Chinese newspapers, especially the older ones were inconsistent and complicated. Both vertical and horizontal typesetting were employed, even within the same page (see Illustration 1). Older Chinese newspapers prior to 1970 employed traditional Chinese font and typically crammed as many articles possible in a page (up to 100 articles) compared to 40 articles in a page for Chinese newspapers employing the simplified Chinese script (see Illustration 2). The gap between columns could be so narrow in the older Chinese newspapers that it made it hard to differentiate the different articles or to even determine the sequence of an article as the layout did not always follow a logical arrangement (see Illustration 3). There were also occasions where bigger size font was used for the article text which made the text appeared like the headline instead (see Illustration 4). To make matters worse, the quality of the digital images is affected as a result of the poor quality of the microfilm images especially when the original newspaper was microfilmed after its physical condition deteriorated. There is also the issue of curvature caused by the binding of the physical newspapers rendering the microfilm images captured to be blurry (see Illustration 5). All these impact directly on the productivity of the vendor's operators as their workload increase to almost 300% per page. The operators would only need to identify about 30 zones (groups) per page for the English newspapers but could be required to identify up to 200 zones per page for a Chinese newspaper.

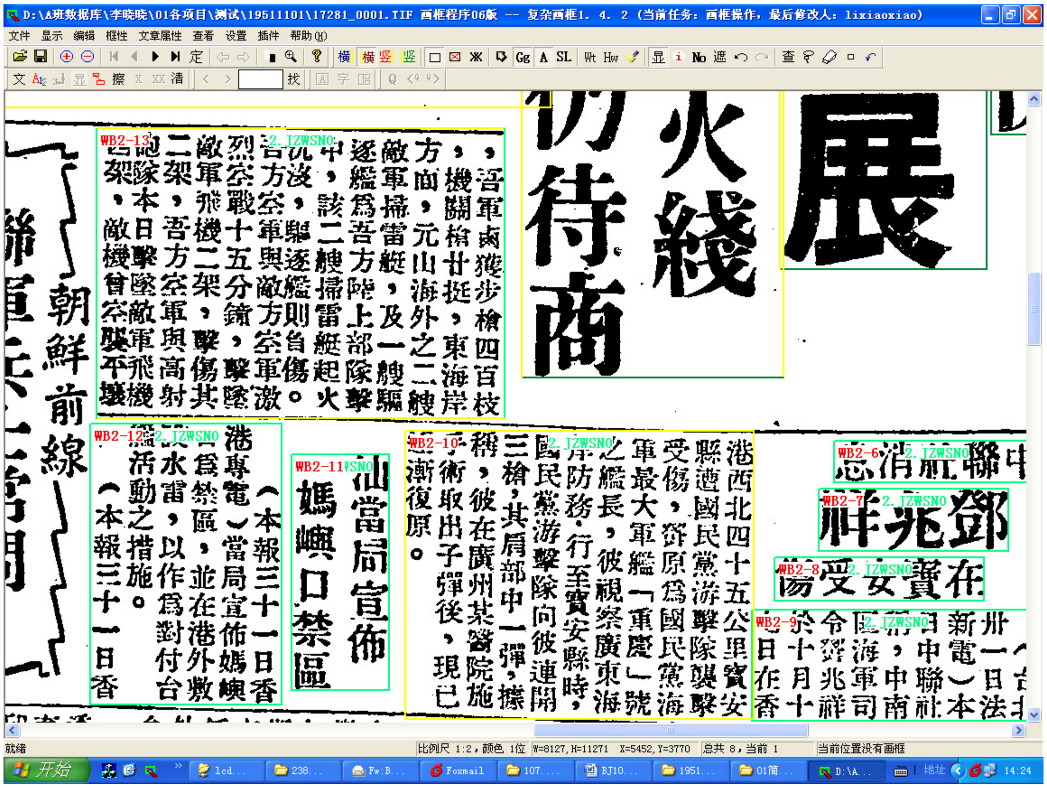
D:\A班数据库\李晓晓\01各项目\测试\19511101\17281_0001.TIF 画框程序06版 -- 复杂画框1. 4. 2 (当前任务：画框操作，最后修改人：lixiaoxiao)

Illustration 1

vertical and horizontal typesetting in the same page

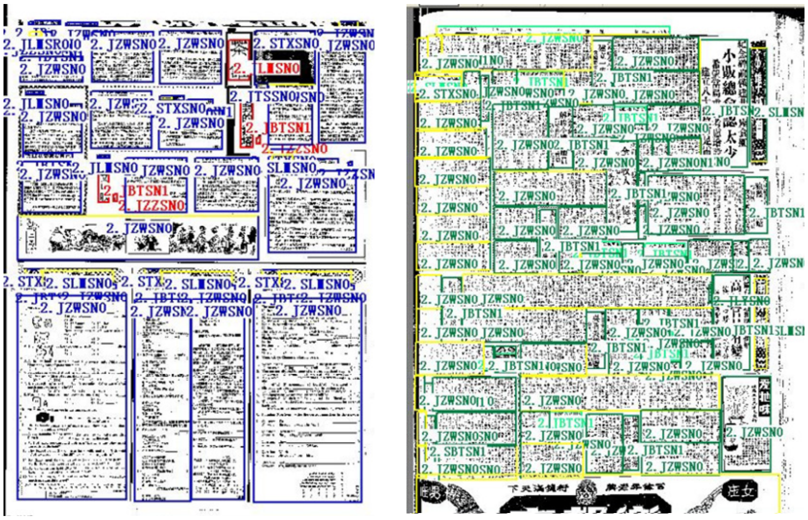

Illustration 2

Comparing the layout of more recent (left) and older Chinese newspapers (right)

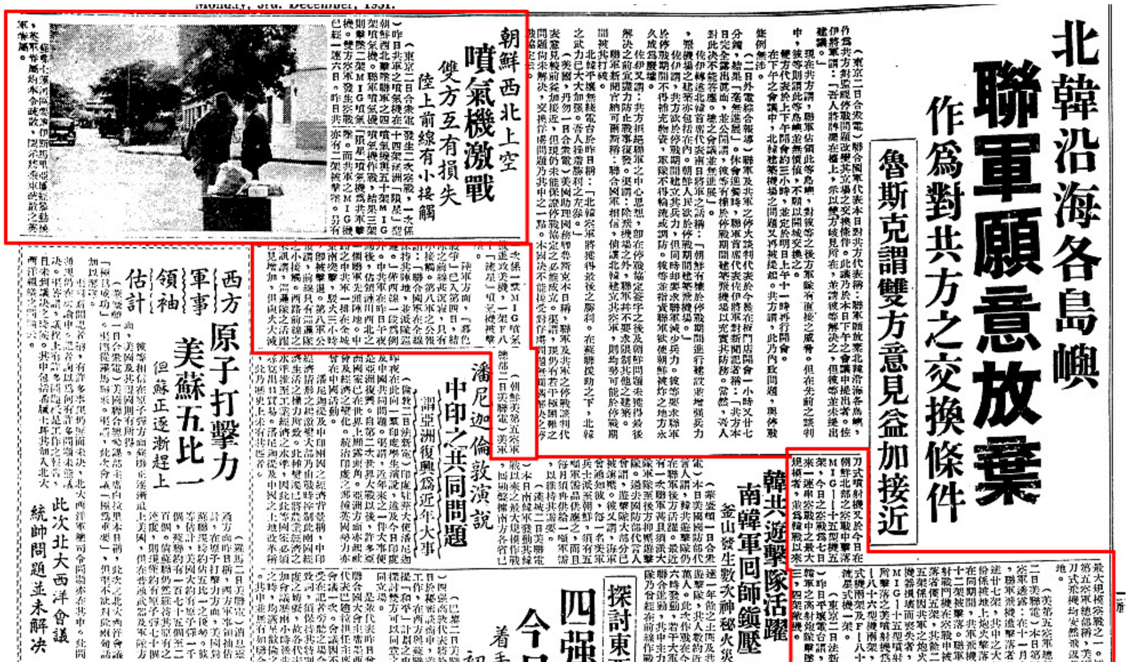

Illustration 3

Illogical sequencing of article

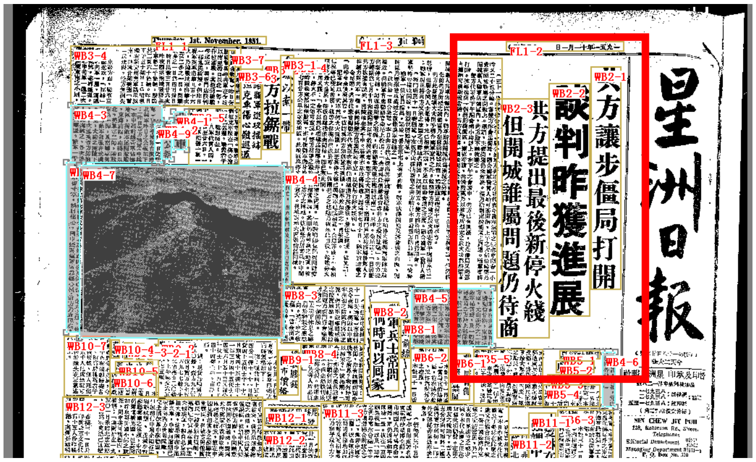

Illustration 4

Inconsistent font for text (appeared like a headline)

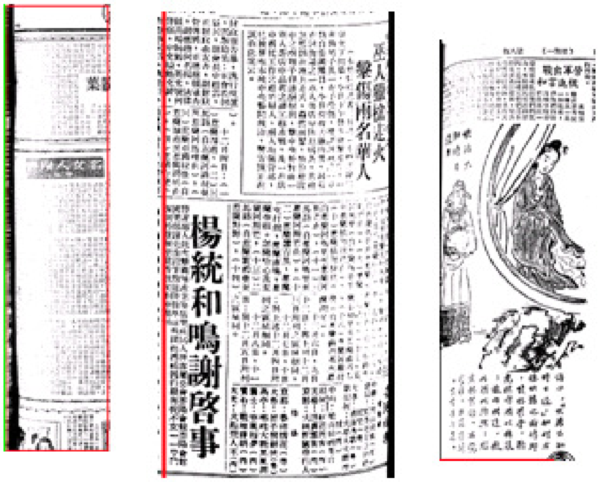

Illustration 5

Problem of curvature rendering blurred text

Taking the above factors into consideration, NLB agreed with our vendor to work on the more recent Chinese newspapers that use the simplified Chinese script with simpler layout. More resources will be devoted for the digitization of the older Chinese newspapers and they are expected to be made available in NewspaperSG fully searchable by mid-2014. Thus at present only Lianhe Zaobao (1983-2008) and Sin Chew Jit Poh (1979-1983) are made fully searchable in NewspaperSG. Older issues of Sin Chew Jit Poh and Nanyang Siang Pau in the Preview section (browse only) will eventually be made fully searchable as well. However, the Malay newspapers (Warta Malaya and Warta Perang) in jawi (Arabic script) and the Tamil newspapers (Tamil Murasu and Singai Nesan Tamil Journal) in the Preview section will remain for browsing only until suitable jawi and Tamil OCR softwares that would be compatible with NewspaperSG's systems are located. Finding the solution to this problem represents NLB's immediate challenge in our efforts to digitize Singapore newspapers.

Multilingual User Interface in NewspaperSG

Before August 2011, only English newspapers were available in NewpaperSG. When the newspapers Lianhe Zaobao (Chinese) and Berita Harian (Malay) were released in NewspaperSG, NLB took the opportunity to introduce the multilingual user interface as a new service feature to allow users to navigate the microsite either in English, Chinese or Malay. Singapore's population is made up about 76 per cent Chinese, 14 per cent Malays and nine per cent Indians with the remaining one per cent made up of other ethnic groups. The population census conducted in 2000 estimated that 70% of Singaporeans are literate in English. This is because English is the working language in Singapore and since 1966, with the introduction of the bilingual education policy, more Singaporeans, especially the younger generations, preferred using English than their mother tongues.

To find out if the multilingual UI serves the information needs of NewspaperSG's users, the usage statistics were analyzed to discern patterns of users' behaviour. We obtained the number of times the landing pages of the Chinese and Malay interfaces were accessed on a monthly basis (see table 2). The figures are low given that NewspaperSG receives an average of 45,000 unique visitors every month. The number of times in a month users switched to the Malay interface (English is the default interface) has been fairly consistent (an average of 350 hits a month). However, the Chinese interface posted a significant increase in number of usage in June 2012 (double) after averaging around 600 hits a month since November 2011.

Month	Chinese interface	Malay interface
Sep 11	809	469
Oct 11	881	463
Nov 11	613	366
Dec 11	572	346
Jan 12	546	333
Feb 12	689	427
Mar 12	704	416
Apr 12	555	392
May 12	630	423
Jun 12	1276	375

Table 2

When we obtained the access statistics to the non-English newspapers, we discovered that the usage of Lianhe Zaobao posted the highest number (so far) in June 2012, corresponding to the increase in the switch to Chinese language interface in June 2012. Table 3 below indicates the number of usage for the different non-English newspapers.

	Chinese			Malay			Tamil	
	Full service	Preview only		Full service	Preview only		Preview only	
Months	Lianhe Zaobao	Sin Chew Jit Poh	Nanyang Siang Pau	Berita Harian	Warta Malaya	Warta Perang	Tamil Murasu	Singai Nesan Tamil Journal
Aug 11	229	-	-	1466	-	-	-	-
Sep 11	979	-	-	8014	-	-	-	-
Oct 11	818	1692	-	2901	-	-	-	-
Nov 11	637	2612	-	3394	-	-	-	-
Dec 11	1402	n.a.	-	1594	-	-	-	-
Jan 12	912	3194	-	2174	-	-	-	-
Feb 12	1126	1915	-	1882	-	-	23	-
Mar 12	2695	18953	-	3977	-	-	298	-
Apr 12	4994	8808	8515	5081	-	-	840	-
May 12	4374	9062	17448	3995	10	12	1100	12
Jun 12	20107	7794 (full service)	8077	11950	34	7	204	49

Table 3

The increase in the number of usage for the full service Chinese and Malay newspapers (Lianhe Zaobao and Berita Harian respectively) in June 2012 could be linked to the re-submission of the newspapers' content for Google to index performed in June 2012. This has allowed more users to discover NewspaperSG's content via Google.

The usage for the non-English newspapers had been low possibly because the content available for these non-English newspapers had been low as well. There are almost 18 million articles available in NewspaperSG now. However, the number of non-English articles only surpassed the 4 million mark in June 2012 (see Table 4). This is because content is released in batches in NewspaperSG, as and when a sizeable amount of content is available after the digitization process is completed.

	All titles in full service	Berita Harian	Lianhe Zaobao	Sin Chew Jit Poh
Jan 12	12,013,812	32,2565	44,9329	
Feb 12	15,146,012	49,4889	1,473,173	
Mar 12	16,688,383	1,493,371	1,995,318	
Apr 12	16,688,383	1,493,371	1,995,318	
May 12	16,688,383	1,493,371	1,995,318	
Jun 12	17,916,064	1,493,371	2,865,351	45,9614

Table 4

Collection size (number of articles)

An online survey was also conducted in June 2012 to gather feedback from our users on the multilingual UI feature. The survey was conducted in three languages (English, Chinese and Malay). The survey in each language was placed on the main landing page of that language interface. A total of eight questions were asked (please see Appendix 2) and 67 complete responses were received.

Out of the 67 responses, only two respondents did the survey in Chinese, none in Malay and 65 in English. Similarly, only three respondents would normally use the Chinese interface, none would normally use the Malay interface and 64 would normally use the English interface. Only 22 respondents (32.8%) would switch to Chinese or Malay interface when searching for Chinese or Malay articles but 25 respondents (37.3%) found the Chinese/Malay interface useful. 29 respondents (43.3%) were not aware of the availability of the Chinese/Malay interface and 33 respondents (49.3%) would use or continue to use the Chinese/Malay interface.

The statistical analysis and survey findings have provided some insights but as yet not sufficient to make any definite conclusion. We will continue to monitor the usage of the non-English newspapers to see its correlation between usage of non-English interface and number of non-English articles available in NewspaperSG. However, we expect the usage for the non-English newspapers and the non-English interface to increase as we release more content in NewspaperSG and more of our users become aware of the availability of the multilingual UI feature.

At present it is likely NLB will continue to provide the multilingual UI as significant percentage of users have indicated that they found the feature useful and will use/continue to use it. NLB will also have to address the issue of lack of awareness among our users on the availability of the multilingual UI. We believe that many of our users are used to the English interface and therefore do not switch to the non-English interface to find information from the non-English newspapers as the search function in any language interface allows input in any language (even Chinese characters). We intend to conduct a focus group interview to understand how non-English newspapers are searched and used. It is also likely that NLB will introduce the Tamil interface when we are able to locate for a suitable Tamil OCR software that is compatible with our systems. We have received feedback and request for a Tamil interface.

Appendix 1

Titles available in NewspaperSG

Titles	Issues Digitised
Berita Harian/Minggu	1970 - 2008
The Business Times	2007–2009
Daily Advertiser	1890 - 1894
Eastern Daily Mail and Straits Morning Advertiser	1905 - 1907
联合早报(Lianhe Zaobao)	1983 - 2008
Malayan Saturday Post	1924 - 1933
Mid - day Herald	1895 - 1896
Singapore Chronicle and Commercial Register	1831 - 1836
星洲日報(Sin Chew Jit Poh)	1979 - 1983
Singapore Daily News	1932 - 1933
The Singapore Free Press	1925 - 1962
The Singapore Free Press and Mercantile Advertiser (1835–1869)	1835 - 1851
The Singapore Free Press and Mercantile Advertiser (1884–1942)	1884 - 1942
Singapore Weekly Herald	1888 - 1889
Straits Advocate	1889
Straits Chinese Herald	1894
Straits Eurasian Advocate	1888
Straits Mail	1894 - 1895
Straits Observer (Singapore)	1874 - 1897
Straits Telegraph and Daily Advertiser	1899
The Straits Times	1845 - 2009
Straits Times Overland Journal	1869 - 1881
Straits Times Weekly Issue	1883 - 1893
TODAY / Weekend TODAY	2000 - 2009
Weekly Sun	1910 - 1913

Titles in Preview section

Newspaper	Issues Digitised
星洲日報 (Sin Chew Jit Poh)	1951 - 1983
南洋商报 (Nanyang Siang Pau)	1923 - 1983
Tamil Murasu (淡米尔之声)	1936 - 2008
Singai Nesan Tamil Journal	1887 - 1890
Warta Malaya	1933 - 1941
Warta Perang	1941

Appendix 2

Please help us answer the following questions regarding the multilingual interface feature in NewspaperSG. Your feedback is useful to improve the service of NewspaperSG.

1. Which language interface do you normally use to find information in NewspaperSG?

   English Chinese Malay

2. Do you switch to the Chinese/Malay interface to find Chinese/Malay articles?

   Yes No

3. Is the Chinese/Malay interface is useful to you?

   Yes No

4. If no, why?

   ____

   ____

5. Were you aware of the availability of the Chinese/Malay interface?

   Yes No

6. Would you use/continue to use the Chinese/Malay interface?

   Yes No

7. If no, why?

   ____

   ____

8. If you are given an opportunity to improve the multilingual user interface, what would you suggest?

   ____

   ____

Thank you.