

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

IFLA 2012 Preconference

Mikkeli, Finland

7-9 August 2012

Singapore



1819 – start of modern Singapore

Population : 5.2 m

Land area : 714.3 sq km

Official languages : English, Chinese, Malay & Tamil

Newspaper Collection

- More than 120 titles since 1824
- 22,000 microfilm reels; low usage
- Factiva, Proquest Newspapers Complete and Library Pressdisplay allow keyword searches on local news content; incomplete content



NewspaperSG

[FORUM/](#)
[FAQ/](#)
[SEE EVERYTHING/](#)

SINGAPORE PAGES / NEWSPAPERS

SEARCH NEWSPAPERS

ADVANCED SEARCH / HELP

ILB NEWSPAPERS

- ✘ Introduction

BROWSE NEWSPAPERS

- ✘ By Title
- ✘ By Language
- ✘ By Date

[Terms of Use](#)



SINGAPORE Newspapers

A WINDOW TO THE PAST

Read entire issues of Singapore newspapers from as far back as the 1800s, and take a virtual step back to when significant historical events occurred.

BROWSE NEWSPAPERS

BY TITLE



BY LANGUAGE



BY DATE



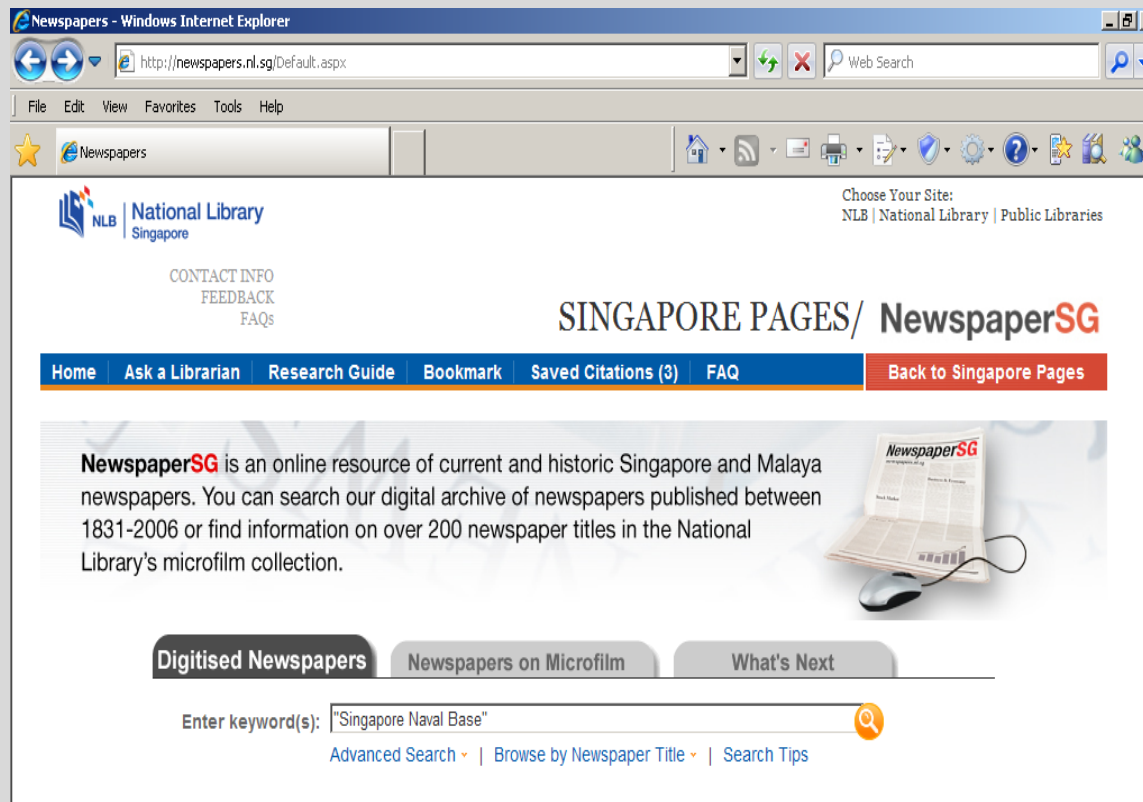
FEATURED

THE STRAITS TIMES
5 October 1977



Trial service in 2009 (Straits Times)

NewspaperSG



The screenshot shows the NewspaperSG website in a Windows Internet Explorer browser window. The address bar displays <http://newspapers.nl.sg/Default.aspx>. The website header includes the NLB Singapore logo and navigation links for CONTACT INFO, FEEDBACK, and FAQs. A "Choose Your Site:" dropdown menu is set to "NLB | National Library | Public Libraries". The main navigation bar contains links for Home, Ask a Librarian, Research Guide, Bookmark, Saved Citations (3), FAQ, and a prominent "Back to Singapore Pages" button. The main content area features a descriptive paragraph about NewspaperSG as an online resource of current and historic Singapore and Malaya newspapers, published between 1831-2006. An image of a laptop displaying the NewspaperSG interface is shown next to the text. Below the text are three tabs: "Digitised Newspapers" (selected), "Newspapers on Microfilm", and "What's Next". A search bar contains the keyword "Singapore Naval Base" and includes a search button. Below the search bar are links for "Advanced Search", "Browse by Newspaper Title", and "Search Tips".

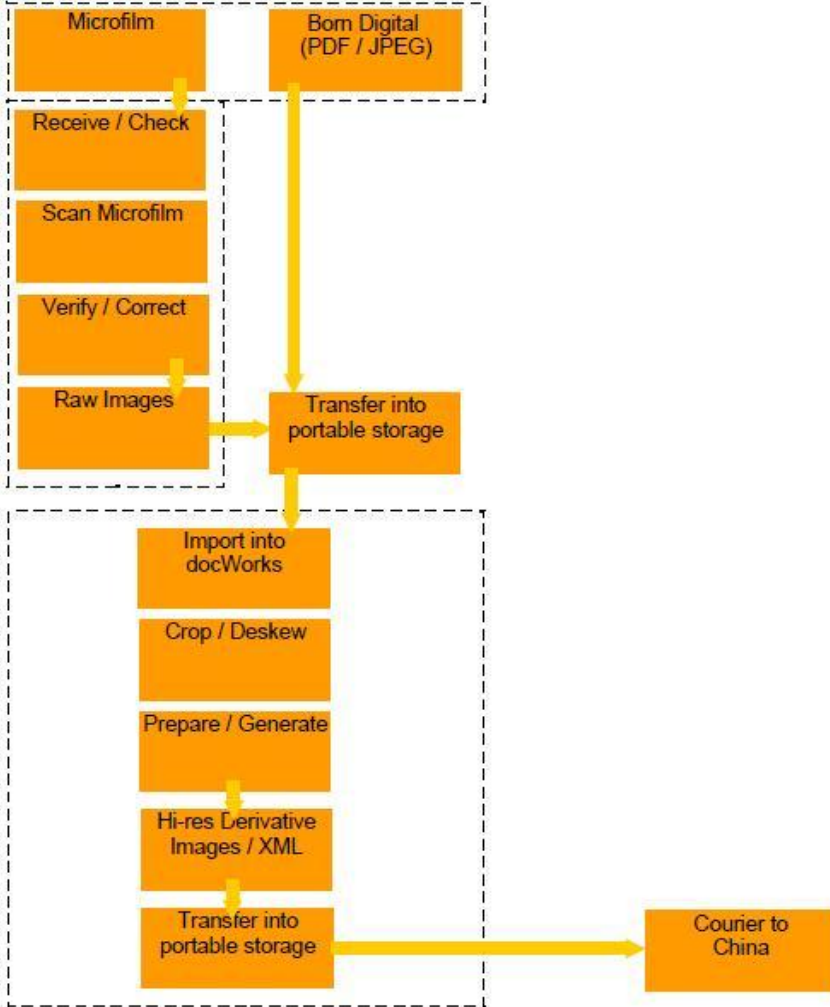
Officially launched in Jan 2010 (17 titles)

Non-English newspapers

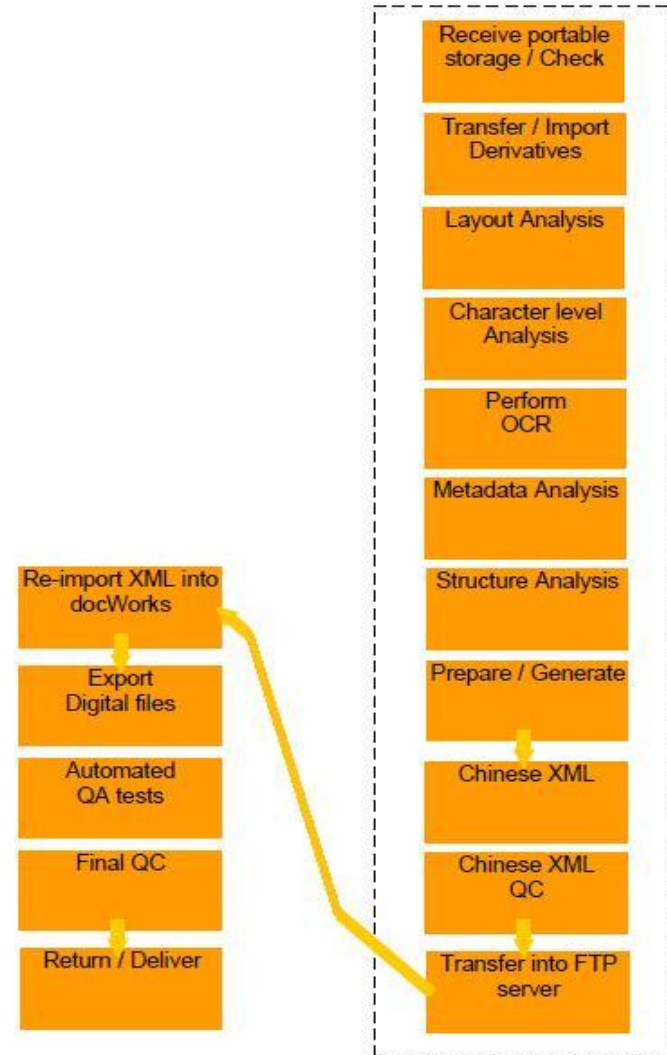
Full service		Preview		
Searchable		Non-searchable (browse only)		
Chinese	Malay	Chinese	Malay	Tamil
Lianhe Zaobao (1983-2008)	Berita Harian (1970-2008)	Nanyang Siang Pau (1923- 1983)	Warta Malaya (1933-1941)	Singai Nesan Tamil Journal (1887-1890)
SinChew Jit Poh (1979-1983)		Sin Chew Jit Poh (1951-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

Digitisation Process

Singapore



Remote Processing (China – Data Datum) Chinese Newspapers



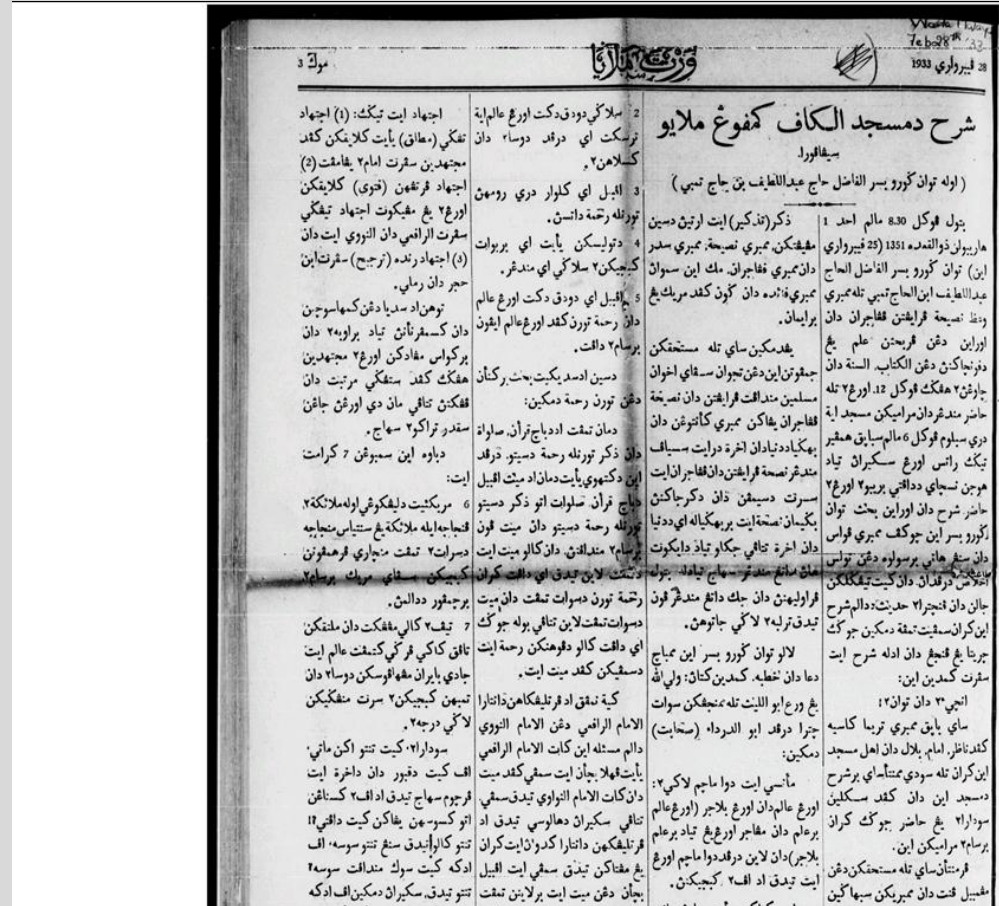
Digitising Berita Harian

- Malay newspaper
- Roman script
- Language barrier – slows down matching process of articles to illustrations or articles that ran over more than one page



Warta Malaya & Warta Perang

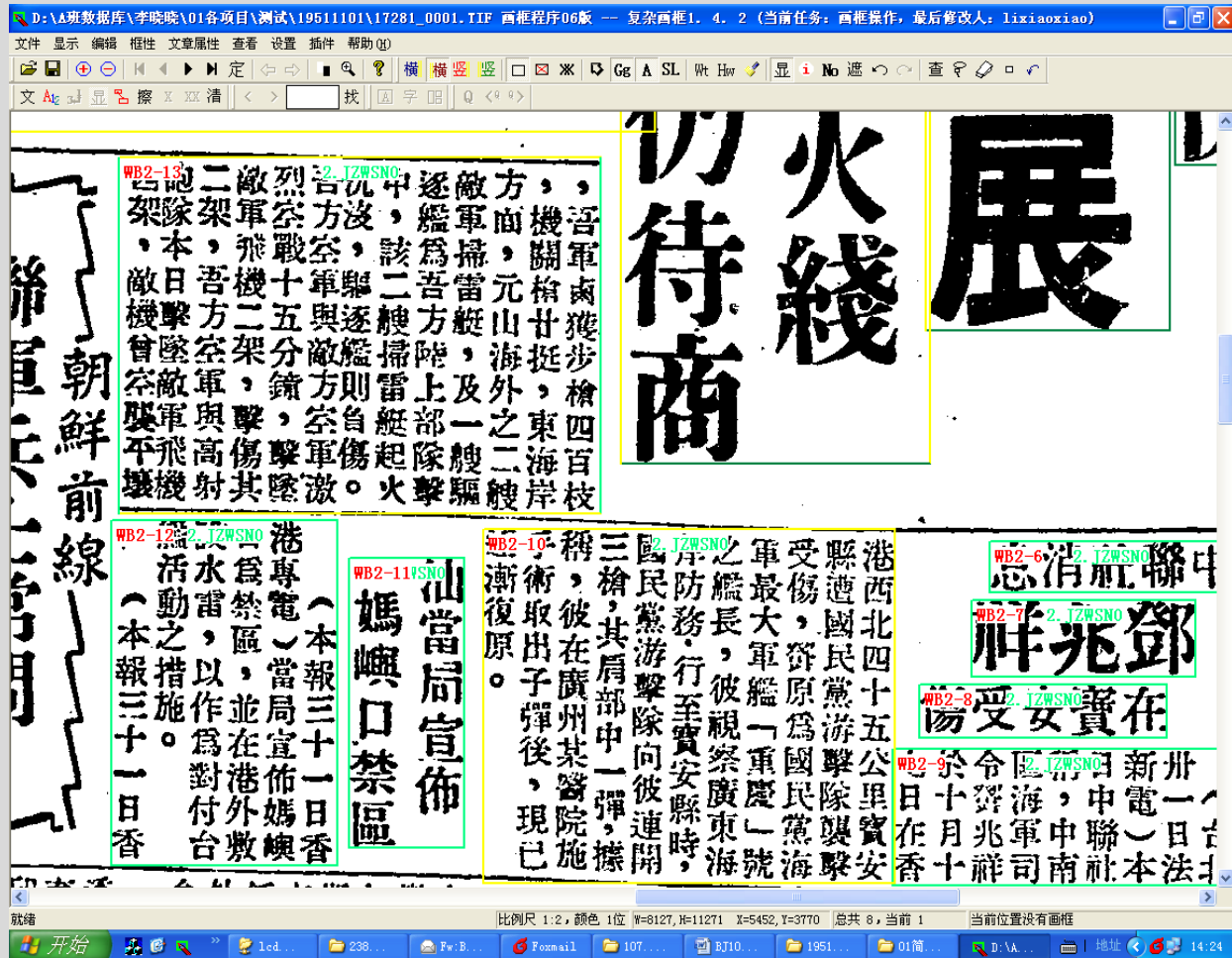
- Malay newspapers
- Jawi script (Arabic)
- No suitable OCR software



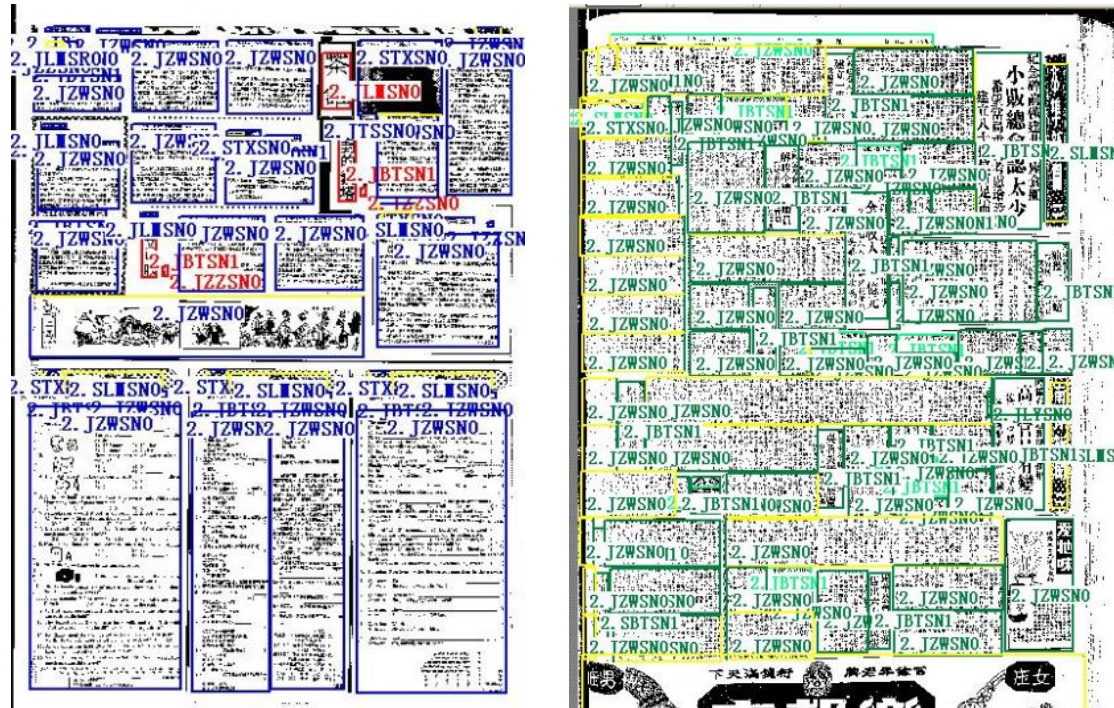
Digitising Chinese newspapers

- Inconsistent layout
- Traditional and simplified fonts

vertical & horizontal typesetting



layout of Chinese newspapers



illogical sequencing of articles

北韓沿海各島嶼

聯軍願意放棄

作為對共方之交換條件

魯斯克謂雙方意見益加接近

【東京二日合衆電】聯合國軍代表本日對共方代表稱：聯軍願將北朝鮮沿海各島嶼，作為對共方對聯軍停戰問題改善其立場之交換條件。此議乃於本日下午之會議中提出者。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

現在共方謂：聯軍佔領此等島嶼，對彼等之後方軍區有直接之威脅。但在先期之談判中，彼等則謂此等島嶼並無價值，不願以開城交換之。此等島嶼之交換，乃在十一月廿七日之會議中，北朝鮮代表表示願將此等島嶼，作為對共方對聯軍停戰問題改善其立場之交換條件。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

【東京二日合衆電】聯軍及共軍之停戰談判代表於今日在板門店開會一小時又廿七分，結果「毫無進展」。聯軍代表表示，對於停戰談判，雙方意見益加接近。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

【東京二日合衆電】聯軍及共軍之停戰談判代表於今日在板門店開會一小時又廿七分，結果「毫無進展」。聯軍代表表示，對於停戰談判，雙方意見益加接近。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

朝鮮西北上空 噴氣機激戰

雙方互有損失

【東京二日合衆電】發生二次激戰，一次係在朝鮮西北上空，噴氣機激戰，雙方互有損失。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」



【東京二日合衆電】發生二次激戰，一次係在朝鮮西北上空，噴氣機激戰，雙方互有損失。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

韓共遊擊隊活躍

南韓軍回師鎮壓

【華盛頓一日合衆電】南韓軍回師鎮壓韓共遊擊隊活躍。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

潘尼迦倫敦演說

中印之共同問題

【倫敦一日合衆電】潘尼迦在倫敦演說，論及中印之共同問題。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

西方原子打擊力 美蘇五比一

但蘇正逐漸趕上

【華盛頓一日合衆電】美國與蘇聯之核武力量，美蘇五比一，但蘇正逐漸趕上。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

四強

今日

【華盛頓一日合衆電】四強今日在板門店開會。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

統帥問題並未解決

【華盛頓一日合衆電】統帥問題並未解決。魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

魯斯克謂：「吾人將詳加考慮，亦以雙方意見所在，並請彼等解決之，但彼等並未提出建議。」

layout of recent Chinese newspapers



4 headlines in 1 page



curvature = blurred text



Traditional Fonts OCR

This image displays a complex grid of traditional Chinese characters and symbols, likely from a historical document or a specialized font set. The characters are arranged in a dense, somewhat chaotic pattern. A prominent feature is a large red circle highlighting a specific character, '韃' (Tā), which is part of the word '韃靼' (Tatars). Other visible characters include '印尼椰' (Indonesian), '復敬言' (Respectful speech), and '暫停' (Temporary stop). The overall layout is highly detailed and visually busy, illustrating the challenge of OCR for traditional fonts.

Traditional Fonts OCR



乾隆



干隆

Qianlong (乾隆), 乾 means heaven, 隆 means eminence, which means "Lasting Eminence".

干 means dry, not auspicious at all

Characters are Words

南洋·星洲 1st Copy

联合早报

Lianhe Zaobao

2 JAN 1996

经济发达但还不是优雅社会

吴总理：国人须改善社会行为和态度以配合富裕生活

吴作栋总理说，经济合作与发展组织把我国列为发达国家，不过，我们要成长为一个优雅的社会，还有一段很长的路要走。

吴总理认为，我们要拥有优雅的生活环境，以配合富裕的物质生活，国人必须改善社会行为和生活态度，破除旧习惯，并且树立新的行为标准。

他说，他已经告诉健康发展局，在选择翻新屋宇时，应当把环保的评估程度列为考虑因素。

“我们必须使国人不只为自己的家着想，同时也为他人着想，以尊重的态度对待公物和公共设施，正如他们照顾自己家里的财物一样。”

吴作栋总理是在1995年的新年献词中，作出上述讲话。他表示，我国经济快速增长，但我们的社会行为以及一些人的言行举止，却未能赶上这种速度。

“我们在公共场合乱扔垃圾、胡乱停放车辆、破坏环境等行为的普遍存在，在公众场合因小事而争风吃醋、侮辱或破坏环境而出现的破坏花草等。”

吴总理说，这些人破坏了别人愉快的生活环境，也给游客留下不良的印象，以为新加坡人都是野蛮和没有教养的。

我国人均国内生产总值已超过西班牙，达到约2万4000美元，但吴总理认为，我们的社会行为水平，却远逊于西班牙。他吁请国人加倍努力，以便迎头赶上。

他表示，除了通过教育以及由家长、教师和社会各界领袖以身作则发挥榜样作用，罚款、惩罚性措施以及反面宣传也是必要的，以促使国人铭记反社会行为是不对的，不能被接受的，并且必须纠正过来。

——吴总理

吴作栋总理说，我国经济在1995年取得8.9%的增长。根据劳工部预测，我国今年的经济增长将介于7%到8%之间。

吴总理指出，尽管面对激烈的全球性竞争，我国去年吸引956亿美元的制造业投资承诺，创下最高记录。这些投资承诺，主要是高增值的化工和电子工程项目。

今年的全球环境可能仍然对我们有利。不过，吴总理指出，我们在美国和亚洲的主要市场，增长可能进一步放慢下来。

“虽然日本正在复苏，不过，它的结构性改革和金融问题还是令人感到担忧。”

吴总理说，作为一个全面的商业中心，我国所有主要的经济领域都取得良好的增长。总工资增长了7%，通货膨胀率却只有1.8%那么低，商业服务投资承诺则高达11亿美元。

吴总理表示，我国去年取得的8.9%经济增长，是比较能长期维持的增长率。令人欣慰的是，这正好缓和了成本的压力。

“世界银行正在复苏，不过，它的人都是看错、没有教养的。刚过路的人，肯定对当地人的热情、有礼以及乐于助人精神留下深刻的印象。我们的人均国内生产总值虽然已超越西班牙，但我们的社会行为水准却在西班牙之下。我们还需要加倍努力才能赶上。”

“我们要怎样努力，才能提升我们的社会行为呢？”

选择翻新屋宇 清洁程度是考虑因素

今日继续跃进，以定崭新的年轮，新的国家定位。

世界银行报告
发展中国家人均收入增速再远远超过世界水平

《华盛顿邮报》根据世界银行昨天发表的一份报告，1994年世界人均收入增长了1%，结束了连续三年的下降趋势。

《世界银行1996年年报》说，在低收入国家里，中国和印度已成为世界上主要的工业品生产国和出口国。

世界银行说：“发展中国家（前苏联除外）人均收入的增长速度再次远远超过世界人均收入的增长速度。”

世界银行年度报告汇集了国际经济方面的数据。报告说，除前苏联共和国之外，低收入和中等收入国家的人均收入增长了2.5%，其中亚洲国家，特别是中国8.2%，其次是韩国，增长率为7.8%。

人均生产总值增长
新加坡中国并列第三

中国和新加坡并列第三，人均国内生产总值的增长率都是6.9%；挪威名列第五，为6.6%；马来西亚名列第六，为6.5%；智利名列第七，增长率为6.2%。

1994年，卢森堡的人均收入最高，为5万9850美元；其次是瑞士，人均收入为3万7180美元；日本的人均收入是2万4630美元，名列第三。

人均收入排名第四至第七名的国家依次是：丹麦，2万8110美元；挪威，2万

Characters are Words

电脑洗眉

电脑操作, 20分钟一次完成 100% 安全, 绝无伤害, 无痛无感, 100% 成功改变眉型, 失败的纹眉, 纹眼线, 高低不平, 太宽太窄, 太黑, 太深, 太暗, 纹眉型, 纹眼线, 再补纹更美观性的眉型。

地址: 2414280
电话: 9128072647
手机号码: 2583720
网路地址: 2522917

美丽园

南洋·星洲 1st Copy

联合早报

Lianhe Zaobao

2 JAN 2008

新加坡报业控股出版

KING'S

经济发达但还不是优雅社会

吴总理：国人须改善社会行为和态度以配合富裕生活



吴作德总理说，经济合作与发展组织把我国列为发达国家，不过，我们要成长为一个优雅的社会。

吴总理说，以配合富裕社会行为标准的

的行为标准。

他说，我们必须在保持良好

的声誉。

“我们将继续

改善环境，也绝对

加派人手来保护

我们的国家。

他表示，新加坡

```

- <TextBlock ID="P1_TB00003" HPOS="301" VPOS="1045" WIDTH="2308" HEIGHT="207" STYLES="TX_0 PAR_LEFT">
- <TextLine ID="P1_TL00003" HPOS="301" VPOS="1046" WIDTH="2292" HEIGHT="194">
  <String ID="P1_ST00009" HPOS="301" VPOS="1046" WIDTH="168" HEIGHT="177" CONTENT="经" WC="0.89" CC="1" />
  <String ID="P1_ST00010" HPOS="476" VPOS="1046" WIDTH="187" HEIGHT="190" CONTENT="济" WC="0.78" CC="2" />
  <String ID="P1_ST00011" HPOS="671" VPOS="1046" WIDTH="179" HEIGHT="187" CONTENT="发" WC="0.89" CC="1" />
  <String ID="P1_ST00012" HPOS="863" VPOS="1046" WIDTH="183" HEIGHT="194" CONTENT="达" WC="0.78" CC="2" />
  <String ID="P1_ST00013" HPOS="1055" VPOS="1046" WIDTH="187" HEIGHT="188" CONTENT="但" WC="0.78" CC="2" />
  <String ID="P1_ST00014" HPOS="1250" VPOS="1046" WIDTH="183" HEIGHT="183" CONTENT="还" WC="0.78" CC="2" />
  <String ID="P1_ST00015" HPOS="1448" VPOS="1046" WIDTH="179" HEIGHT="179" CONTENT="不" WC="0.89" CC="1" />
  <String ID="P1_ST00016" HPOS="1638" VPOS="1046" WIDTH="183" HEIGHT="185" CONTENT="是" WC="0.78" CC="2" />
  <String ID="P1_ST00017" HPOS="1827" VPOS="1046" WIDTH="187" HEIGHT="183" CONTENT="优" WC="0.89" CC="1" />
  <String ID="P1_ST00018" HPOS="2019" VPOS="1046" WIDTH="187" HEIGHT="185" CONTENT="雅" WC="0.78" CC="2" />
  <String ID="P1_ST00019" HPOS="2217" VPOS="1046" WIDTH="183" HEIGHT="188" CONTENT="社" WC="0.78" CC="2" />
  <String ID="P1_ST00020" HPOS="2414" VPOS="1046" WIDTH="179" HEIGHT="185" CONTENT="会" WC="0.89" CC="1" />
    
```

渣打程度考量因素

其中亚洲国家，特别是中国 2万8110美元；挪威，2万

Digitising Tamil newspapers



No suitable OCR software

Future challenges

Other language newspapers

- Arabic
- Malayalam
- Punjabi

Multilingual UI




National Library
Singapore

CONTACT INFO
 FEEDBACK
 FAQs

Choose Your Site:
 NLB | National Library | Public Libraries

SINGAPORE PAGES/ NewspaperSG

[主页](#) [参考服务](#) [馆藏专题指南](#) [书签](#) [保存题录信息](#) [常见问题解答](#)

[Back to Singapore Pages](#)

选择语言: [English](#) | [中文\(简体\)](#) | [Bahasa Melayu](#)
 请抽空参与有关本站的多语言用户界面的[调查](#)。

NewspaperSG 收录了曾在新加坡与马来亚出版的旧报纸资料。通过这个网上数字报纸资源平台，图书馆读者可以浏览与检索1831年至2009年出版的报纸。读者也可以亲自到图书馆使用缩微胶卷，查询200多种报章资料。




 读者目前已可通过NewspaperSG查找《[联合早报](#)》和《[每日新闻](#)》的文章。同时，NewspaperSG也开始采用多语言用户界面，让读者选择使用英文、中文或马来文的界面浏览本网站。除此之外，读者目前也可通过预览服务的栏目抢先浏览《[星洲日报](#)》、《[南洋商报](#)》和《[淡米尔之声](#)》。

[数字报纸](#) [馆藏报章缩微胶卷索引](#) [预览](#)

请输入报章名称: 

[高级检索](#) | [帮助](#)

选择日期:
 开始 (dd/mm/yyyy)  结束 (dd/mm/yyyy) 

出版地:

选择语文:

★ 热门报章

- Berita Harian
- Lian He Zao Bao (联合早报)
- Tamil Murasu (தமிழ் முரசு)
- The Straits Times
- TODAY

allow users to navigate the portal in English, Chinese or Malay

Chinese & Malay landing page stats

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

Usage stats

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	8077	11950	34	7	204	49

Collection size (number of articles)

	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

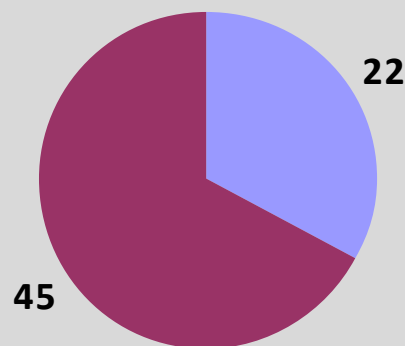
responses to online survey

- 67 completed responses.
- Only 2 respondents did the survey in Chinese; the rest in English.
- Only 3 respondents normally use the Chinese interface ; the rest normally use the English interface.

- **Do you switch to the Chinese/Malay interface to find Chinese/Malay articles?**

Yes – 22 (32.8%)

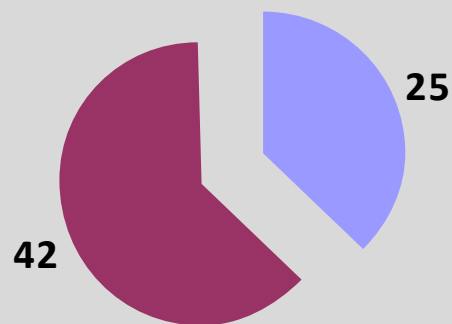
No – 45 (67.2%)



- **Is the Chinese/Malay interface useful to you?**

Yes – 25 (37.3%)

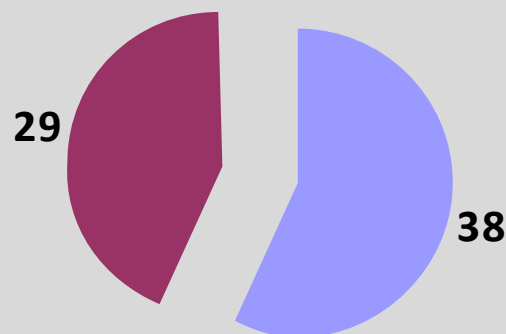
No – 42 (62.7%)



- **Were you aware of the availability of the Chinese/Malay interface?**

Yes – 38 (56.7%)

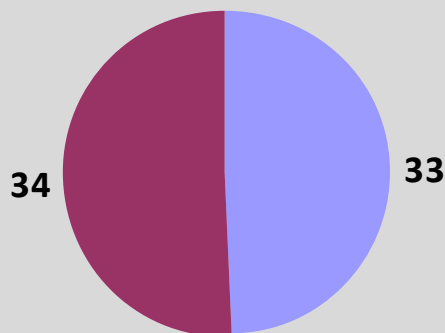
No – 29 (43.3%)



- **Would you use/continue to use the Chinese/Malay interface?**

Yes – 33 (49.3%)

No – 34 (50.7%)



Conclusion

- Low usage for non-English interface
- Low usage for non-English newspapers
- Currently, low number of articles for non-English newspapers
- More efforts to spread awareness; need to understand how language newspapers are searched and used
- Tamil interface

Thank you