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

IFLA 2012 Preconference
Mikkeli, Finland
7-9 August 2012
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
Newspaper SG

Trial service in 2009 (Straits Times)
NewspaperSG

NewspaperSG is an online resource of current and historic Singapore and Malaya newspapers. You can search our digital archive of newspapers published between 1831-2006 or find information on over 200 newspaper titles in the National Library's microfilm collection.

Officially launched in Jan 2010 (17 titles)
### Non-English newspapers

<table>
<thead>
<tr>
<th>Full service</th>
<th>Preview</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Searchable</strong></td>
<td><strong>Non-searchable (browse only)</strong></td>
</tr>
<tr>
<td>Chinese</td>
<td>Malay</td>
</tr>
</tbody>
</table>
Digitisation Process

Singapore

- Microfilm
- 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

Remote Processing (China – Data Datum)

- Chinese Newspapers
  - Receive portable storage / Check
  - Transfer / Import Derivatives
  - Layout Analysis
  - Character level Analysis
  - Perform OCR
  - Metadata Analysis
  - Structure Analysis
  - Prepare / Generate
  - Chinese XML
  - Chinese XML QC
  - Transfer into FTP server

- Re-import XML into docWorks
  - Export Digital files
  - Automated QA tests
  - Final QC
  - Return / Deliver
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
layout of recent Chinese newspapers
4 headlines in 1 page

新明日报
父女乱伦案
女郎认罪

港周刊报道
华仔老婆
怀孕

遭遇20人追打
男子逃命

名媛
月花
30万
离婚每月索45万
curvature = blurred text
乾 隆

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

干 隆

干 means dry, not auspicious at all
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

1995年经济增长8.9%，今年料介于7至8%

世界银行报告 发展中国家平均收入增速再远远超过世界水平
经济发达但还不是优雅社会
Digitising Tamil newspapers

No suitable OCR software
Future challenges

Other language newspapers

- Arabic
- Malayalam
- Punjabi
Multilingual UI

allow users to navigate the portal in English, Chinese or Malay
## Chinese & Malay landing page stats

<table>
<thead>
<tr>
<th>Month</th>
<th>Chinese interface</th>
<th>Malay interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep 11</td>
<td>809</td>
<td>469</td>
</tr>
<tr>
<td>Oct 11</td>
<td>881</td>
<td>463</td>
</tr>
<tr>
<td>Nov 11</td>
<td>613</td>
<td>366</td>
</tr>
<tr>
<td>Dec 11</td>
<td>572</td>
<td>346</td>
</tr>
<tr>
<td>Jan 12</td>
<td>546</td>
<td>333</td>
</tr>
<tr>
<td>Feb 12</td>
<td>689</td>
<td>427</td>
</tr>
<tr>
<td>Mar 12</td>
<td>704</td>
<td>416</td>
</tr>
<tr>
<td>Apr 12</td>
<td>555</td>
<td>392</td>
</tr>
<tr>
<td>May 12</td>
<td>630</td>
<td>423</td>
</tr>
<tr>
<td>Jun 12</td>
<td>1276</td>
<td>375</td>
</tr>
</tbody>
</table>
## Usage stats

<table>
<thead>
<tr>
<th>Months</th>
<th>Lianhe Zaobao</th>
<th>Sin Chew Jit Poh</th>
<th>Nanyang Siang Pau</th>
<th>Berita Harian</th>
<th>Warta Malaya</th>
<th>Warta Perang</th>
<th>Tamil Murasu</th>
<th>Singai Tamil Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 11</td>
<td>229</td>
<td>-</td>
<td>-</td>
<td>1466</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sep 11</td>
<td>979</td>
<td>-</td>
<td>-</td>
<td>8014</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Oct 11</td>
<td>818</td>
<td>1692</td>
<td>-</td>
<td>2901</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Nov 11</td>
<td>637</td>
<td>2612</td>
<td>-</td>
<td>3394</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dec 11</td>
<td>1402</td>
<td>n.a.</td>
<td>-</td>
<td>1594</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jan 12</td>
<td>912</td>
<td>3194</td>
<td>-</td>
<td>2174</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Feb 12</td>
<td>1126</td>
<td>1915</td>
<td>-</td>
<td>1882</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>-</td>
</tr>
<tr>
<td>Mar 12</td>
<td>2695</td>
<td>18953</td>
<td>-</td>
<td>3977</td>
<td>-</td>
<td>-</td>
<td>298</td>
<td>-</td>
</tr>
<tr>
<td>Apr 12</td>
<td>4994</td>
<td>8808</td>
<td>8515</td>
<td>5081</td>
<td>-</td>
<td>-</td>
<td>840</td>
<td>-</td>
</tr>
<tr>
<td>May 12</td>
<td>4374</td>
<td>9062</td>
<td>17448</td>
<td>3995</td>
<td>10</td>
<td>12</td>
<td>1100</td>
<td>12</td>
</tr>
<tr>
<td>Jun 12</td>
<td>20107</td>
<td>7794</td>
<td>8077</td>
<td>11950</td>
<td>34</td>
<td>7</td>
<td>204</td>
<td>49</td>
</tr>
<tr>
<td>Month</td>
<td>All titles in full service</td>
<td>Berita Harian</td>
<td>Lianhe Zaobao</td>
<td>Sin Chew Jit Poh</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------</td>
<td>---------------</td>
<td>---------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan 12</td>
<td>12,013,812</td>
<td>32,2565</td>
<td>44,9329</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feb 12</td>
<td>15,146,012</td>
<td>49,4889</td>
<td>1,473,173</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mar 12</td>
<td>16,688,383</td>
<td>1,493,371</td>
<td>1,995,318</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr 12</td>
<td>16,688,383</td>
<td>1,493,371</td>
<td>1,995,318</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 12</td>
<td>16,688,383</td>
<td>1,493,371</td>
<td>1,995,318</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jun 12</td>
<td>17,916,064</td>
<td>1,493,371</td>
<td>2,865,351</td>
<td>45,9614</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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