A platform for assessing the current state of data repositories in Korea, Japan and China

Open Repositories 2014
June 9-13, Helsinki, Finland

Suntae Kim
stkim@kisti.re.kr
To analyze the data & identify 6 kinds of features
- China, Japan and Republic of Korea operate 288 repositories (42.2% of Asia, 8% of global)
- The repositories that provide the content in Korean was 0.72%.
- The 2,519 repositories are registered in OpenDOAR and ROAR registry services.
- Korea has incomplete repository operation in subject area and paper.
- Korea and Japan, the repositories in 'Health and Medicine' area are in the top.

To develop the OpenDOAR & ROAR Crawlers
To collect the repository information
To identify the collected data
Work flow of the Crawlers

National Code
ISO 3166-1 Search

To Build the OpenDOAR query

To ingest the OpenDOAR XML

To parse & analyze the XML DOM

To store the processed data in DB

Repository Type Search

To refer to the OpenDOAR

ROAR XSD Analyze

To request the ROAR XML

To ingest the ROAR XML Data

To parse & analyze the XML DOM

To store the processed data in DB

Work flow of the OpenDOAR Crawler

Work flow of the ROAR Crawler
Development Environment

- **Programming language**: Java
- **Java virtual machine**: jdk1.7.0_45
- **DBMS**: MySQL 5.6.15
- **Execution environment**: OS - windows 7
- **Editing environment**: eclipse (Version: Indigo Service Release 2 / Build id: 20120216-1857)

<table>
<thead>
<tr>
<th>G_itemCnt</th>
<th>K_itemCnt</th>
<th>K_percent</th>
<th>C_itemCnt</th>
<th>C_percent</th>
<th>J_itemCnt</th>
<th>J_percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>164400861</td>
<td>1332674</td>
<td>0.81</td>
<td>472803</td>
<td>0.29</td>
<td>1416966</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>language</th>
<th>Repository Cnt</th>
<th>Nation</th>
<th>Repository Cnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>106</td>
<td>China</td>
<td>37</td>
</tr>
<tr>
<td>Japanese</td>
<td>144</td>
<td>Japan</td>
<td>143</td>
</tr>
<tr>
<td>Korean</td>
<td>17</td>
<td>Korea, Republic of</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>global_Rcnt</th>
<th>asia_Rcnt</th>
<th>asia_percent</th>
<th>cjk_Rcnt</th>
<th>cjk_percent_inGlobal</th>
<th>cjk_percent_inAsia</th>
</tr>
</thead>
<tbody>
<tr>
<td>2514</td>
<td>446</td>
<td>17.74</td>
<td>194</td>
<td>7.72</td>
<td>43.50</td>
</tr>
</tbody>
</table>
The status of data repositories in each country can be analysed in real time and topics that require more support from data repositories can be identified using the established platform.

Using **Rest-full API from OpenDOAR**, structured information on global repositories is collected and **it is utilized to analyse the status of Korean, Chinese, and Japanese repositories** in comparison with global and Asian ones.

For future study, a module for collecting repository information from other than OpenDOAR needs to be developed.

In addition, development of a Query Builder module and user-interface that can enable multi-dimensional analysis should be followed.