1. Classification of VO resources

~ 20,000 resource metadata (RM)
how to classify (grouping) ?
content metadata ?
some of them are poorly fulfilled.

auto-classification based on the full text in the RM.

2. How do we classify ?

define categories  key words / phrases
extract key words and phrases from the RMs
We used a Python implementation of the Rapid Automatic Keyword Extraction (RAKE) algorithm [1].
after long-long manual inspection ...

776 primary categories and 1,013 secondary categories are defined (as of 2018 Nov.).
The list of the categories needs to be updated regularly, but the manual intervention is a real bottleneck .... We need AI!

3. Database

RMs belonging to a given category are searched for with an backend database upon a request.

We selected Apache Solr for the database.

Apache Solr is an open source search platform.
Customizable text analyzer build from a sequence of tokenizers and filters.

An example of how the text “many Galaxies clustered around AGN” is indexed.

<table>
<thead>
<tr>
<th>tokenizer</th>
<th>many</th>
<th>Galaxies clustered</th>
<th>around</th>
<th>AGN</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop filter</td>
<td>X</td>
<td>Galaxies clustered</td>
<td>X</td>
<td>AGN</td>
</tr>
<tr>
<td>synonym filter</td>
<td>Galaxies clustered</td>
<td>active</td>
<td>galactic</td>
<td>nucleus</td>
</tr>
<tr>
<td>stem filter</td>
<td>galaxy</td>
<td>cluster</td>
<td>active</td>
<td>galactic</td>
</tr>
</tbody>
</table>

input text is split into tokens.
stop words[1] are removed.
synonyms are replaced with a representative single- or multi-token.
tokens are replaced with a stem word of lowercase.

These words are used as indexes for the given text.

Query input is analyzed in the same way[2], so “galaxy clustering active galactic nuclei” will match with the text.

4. Architecture of RM search

• Solr-based RM registry is located behind the JVO portal.
• JVO portal provides a GUI for searching a resource.
• RM registry harvests RM from the publishing registries.

5. JVO Index

• Using the key words & phrases extracted from the RMs, indexing service “JVO Index” was developed.
• JVO Index service provides an overview of the resources which is available in the VO world.

6. VO Services of Japanese Facilities

Major data resources published from the JVO and the adapted standards are summarized in the table below. JVO VO service toolkit [2] was used to build these services.

<table>
<thead>
<tr>
<th>Title</th>
<th>SIA v1</th>
<th>SIA v2</th>
<th>SSA 1.2</th>
<th>CS</th>
<th>TAP</th>
<th>Obs Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMA VO Service</td>
<td>O</td>
<td>O</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subaru Suprime-Cam data service</td>
<td>O</td>
<td>O</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subaru MOIRCS data service</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subaru HDS Spectrum data service</td>
<td>O</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nobeyama Radio Telescope FITS archive</td>
<td>O</td>
<td>O</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AKARI Far-infrared All-Sky Survey Maps</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AKARI Point Source Catalog Public Release 1</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References
[1] https://github.com/aneeshu/RAKE