VO service in Japan : Registry service based on Apache Solr and SIA v2 service for Japanese Facilities

Yuji Shirasaki, Christopher Zapart, Masatoshi Ohishi, Yoshihiko Mizumoto (NAOJ)

National Astronomical Observatory of Japan

Currently more than 20 thousands of VO services are registered in the VO registry database. Keyword Search is the most popular way to find a resource. There can be a lot of ways to implement this capability, and the performance depends on how to index the document describing the resource metadata.

Apache Solr is an open source search platform and uses the Lucene Java search library for full-text indexing. The document is indexed after the process that removes stop words (not adequate for a keyword e.g. a, the, and, is, are ...) and stems a word to a root word (e.g. clustering to cluster). Thus for the query of a keyword cluster, the Solr search system will return documents that contain a word cluster, clustering, or clustered. Solr also can handle single- or multi-token synonyms and abbreviation. For an example, SN, supernovae, and SNe are translated to supernova in the process of indexing. So it increases the probability for a given keyword to hit a desired resource metadata. In order to incorporate this feature, we upgraded the registry service behind the JVO portal to the one based on Solr.

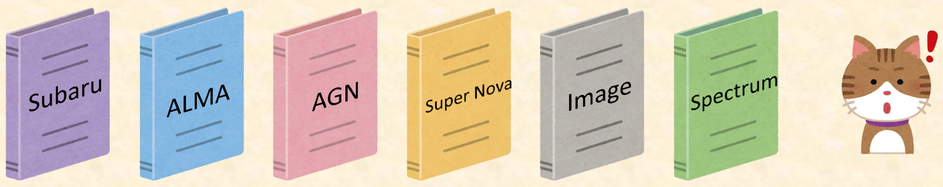
The data from Nobeyama Legacy project were released from the JVO portal on 1st June in 2018. The data are now distributed also through the most recent VO standard interface called SIA-v2. The data of ALMA and Subaru telescope are also accessible through the SIA-v2. We are now working on distributing the data of Hyper Suprime-Cam SSP/DR1 and also the data of Hitomi satellite under the collaboration with C-SODA/JAXA.

We present those our recent development on Japanese Virtual Observatory system. http://jvo.nao.ac.jp/portal/

1. Classification of VO resources

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

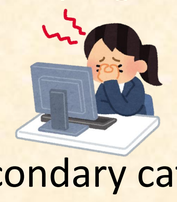
auto-classification based on the full text in the RM.



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.).

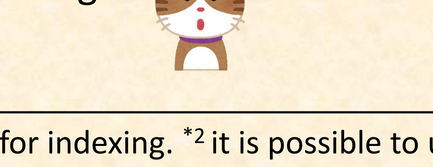
The list of the categories needs to be updated regularly, but the manual intervention is a real bottleneck ... We need AI !

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.

tokenizer	many	Galaxies	clustered	around	AGN	
	↓	↓	↓	↓	↓	
stop filter	X	Galaxies	clustered	X	AGN	
		↓	↓	↓	↓	↓
synonym filter		Galaxies	clustered	active	galactic	nucleus
		↓	↓	↓	↓	↓
stem filter		galaxy	cluster	active	galactic	nucleus

These words are used as indexes for the given text.

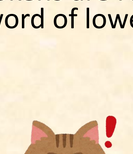
input text is split into tokens.

stop words[*1] are removed.

synonyms are replaced with a representative single- or multi-token.

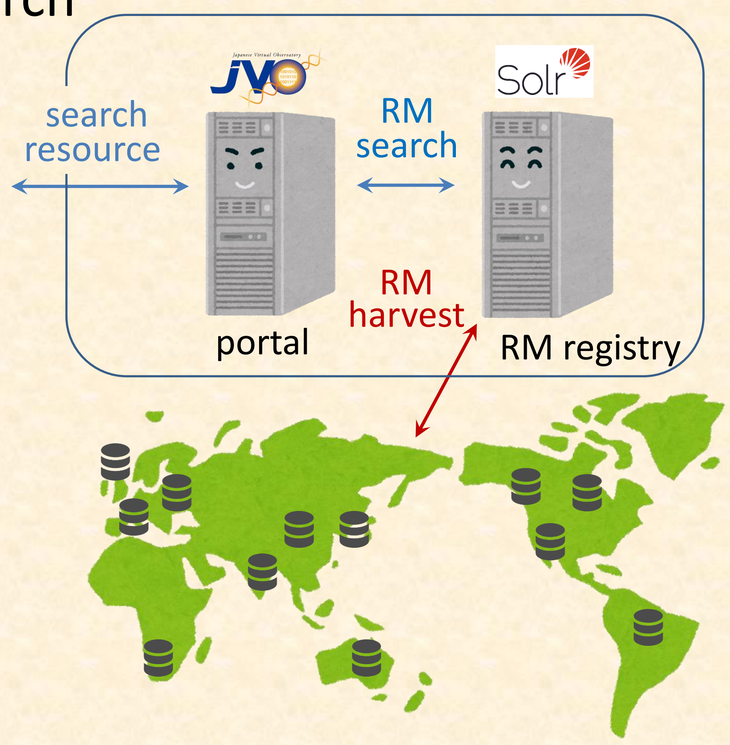
tokens are replaced with a stem word of lowercase.

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

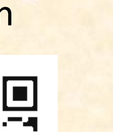


4. Architecture of RM search

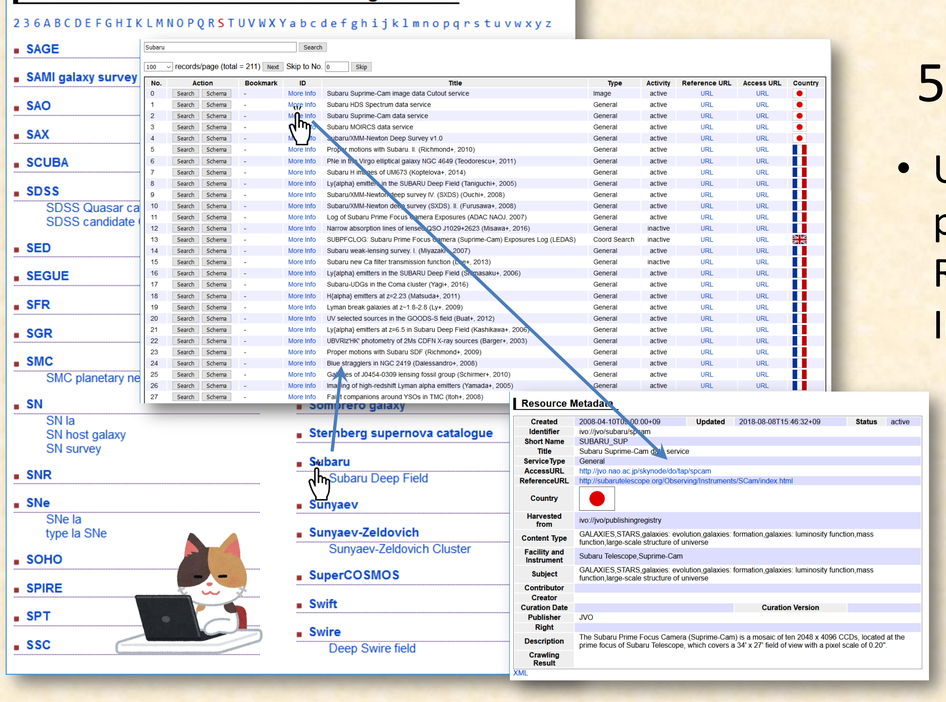
- Solr-based RM registry is located behind the JVO portal.
- JVO portal provides a GUI for searching a resource.
- RM registry harvests RM from the publishing registries.



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.

Title	SIA v1	SIA v2	SSA 1.1	CS	TAP	Obs Core
ALMA VO Service		○			○	1.1
Subaru Suprime-Cam data service	○	○			○	1.1
Subaru MOIRCS data service	○				○	
Subaru HDS Spectrum data service			○		○	
Nobeyama Radio Telescope FITS archive		○			○	1.1
AKARI Far-infrared All-Sky Survey Maps	○				○	
AKARI Point Source Catalog Public Release 1					○	

References
[1] https://github.com/aneesha/RAKE
[2] Shirasaki, Y. et al. 2012, ASP Conf Vol 461, p451, ADASS XXI

*1 stop words are words which are the most common words and not useful for indexing. *2 it is possible to use different analysis patterns for indexing of text and query.