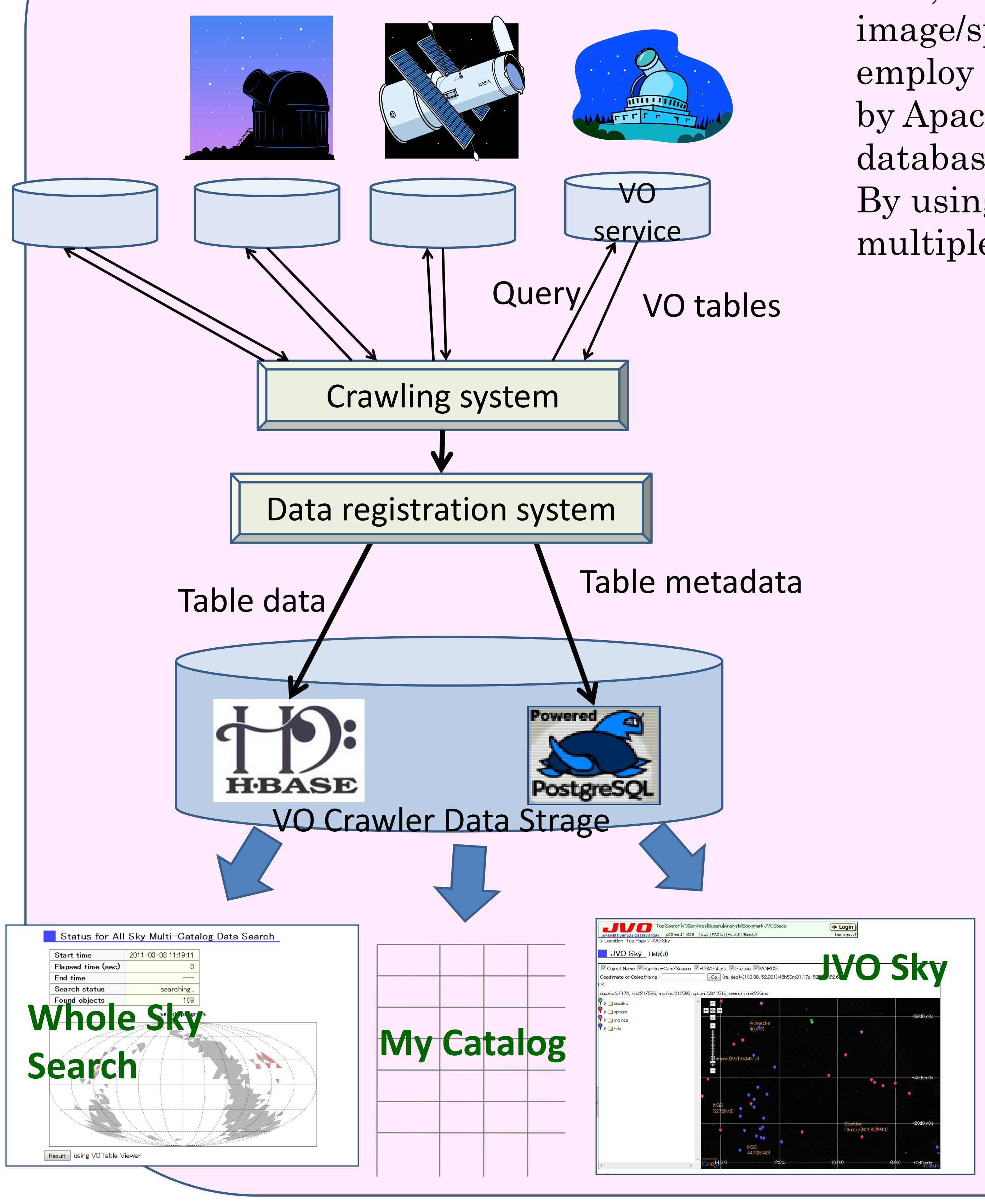
VO Crawler :
a crawling system for Virtual Observatory services

We report on the development of VO Crawler under the Japanese Virtual Observatory (JVO) project. VO Crawler accesses all the Virtual Observatory (VO) Services around the world, and cache data over the whole sky. As all the data and metadata are managed in a single system, it enables quick access of and searches on huge data, and to find location of VO data on a sky map.
Our goal is that users can find data about objects with user given characteristics from whole sky without bothering a huge variety of data sources and huge data size.

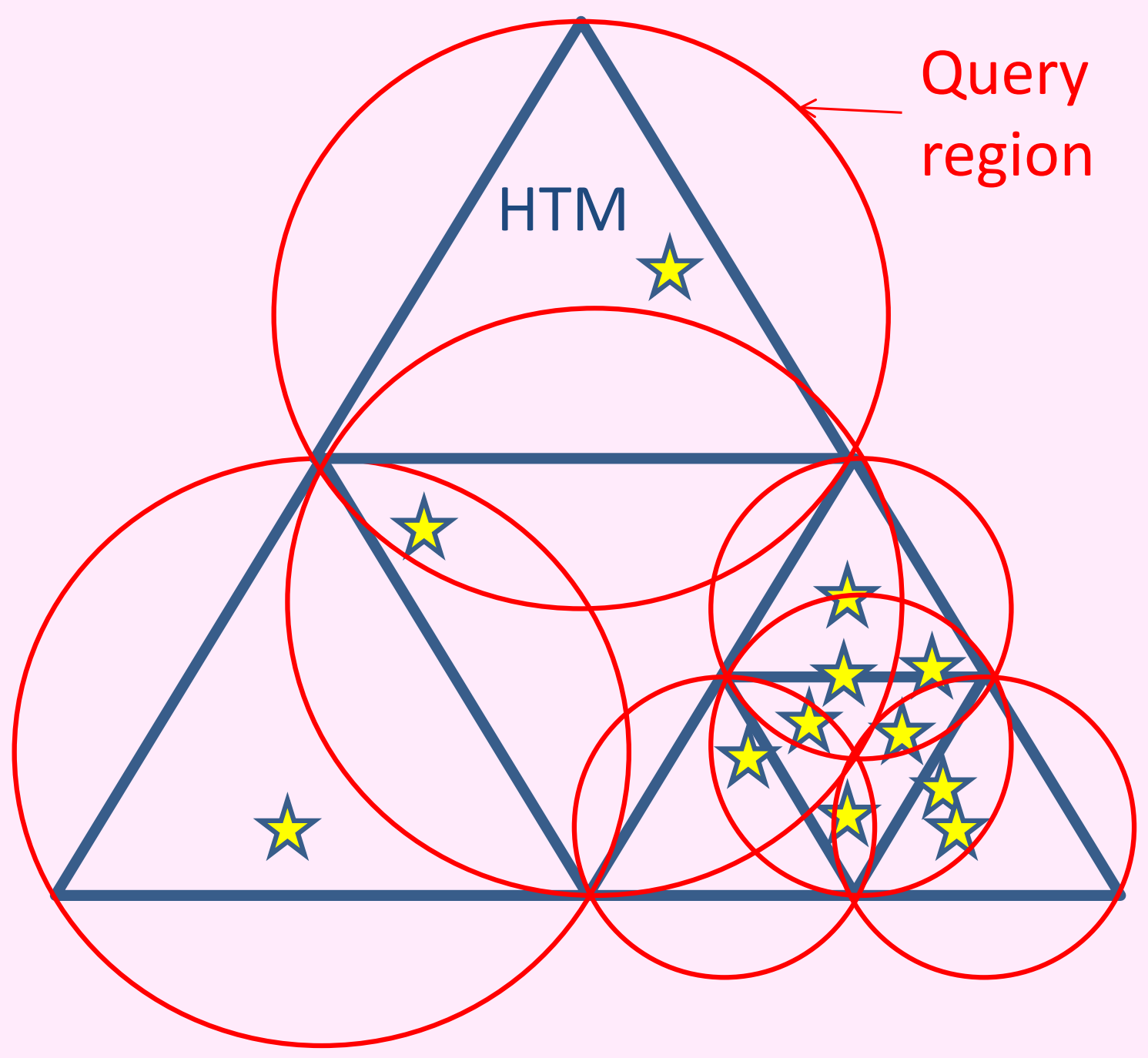
Yutaka Komiya (NAOJ), Y. Shirasaki (NAOJ),
M. Ohishi (NAOJ), S. Eguchi (NAOJ),
Y. Mizumoto (NAOJ), Y. Ishihara (Fujitsu),
J. Tsutsumi (Fujitsu), T. Hiyama (Fujitsu),
H. Nakamoto (SEC), & M. Sakamoto (SEC)

VO Crawler



VO Crawler compiles the data from all active VO services. All catalog data, and image & spectrum metadata with their access URI of the image/spectrum data are stored in VO Crawler Data Storage. We employ Hadoop; a software for distributed processing system developed by Apache project, and the retrieved data are managed with HBase; a database for Hadoop. Metadata of tables are stored in PostgreSQL.
By using stored data, high-speed access to huge data provided by multiple instruments has been achieved.

The VO Crawler repeats radial search until total search area covers the whole sky.
A search region for a single query is a circle which covers a triangular area represented by one HTM ID. We start from the HTM level 2. When too much data (>10,000 rows) is contained in a region, we subdivide the search region into smaller areas represented by the higher level of HTM, and query again.



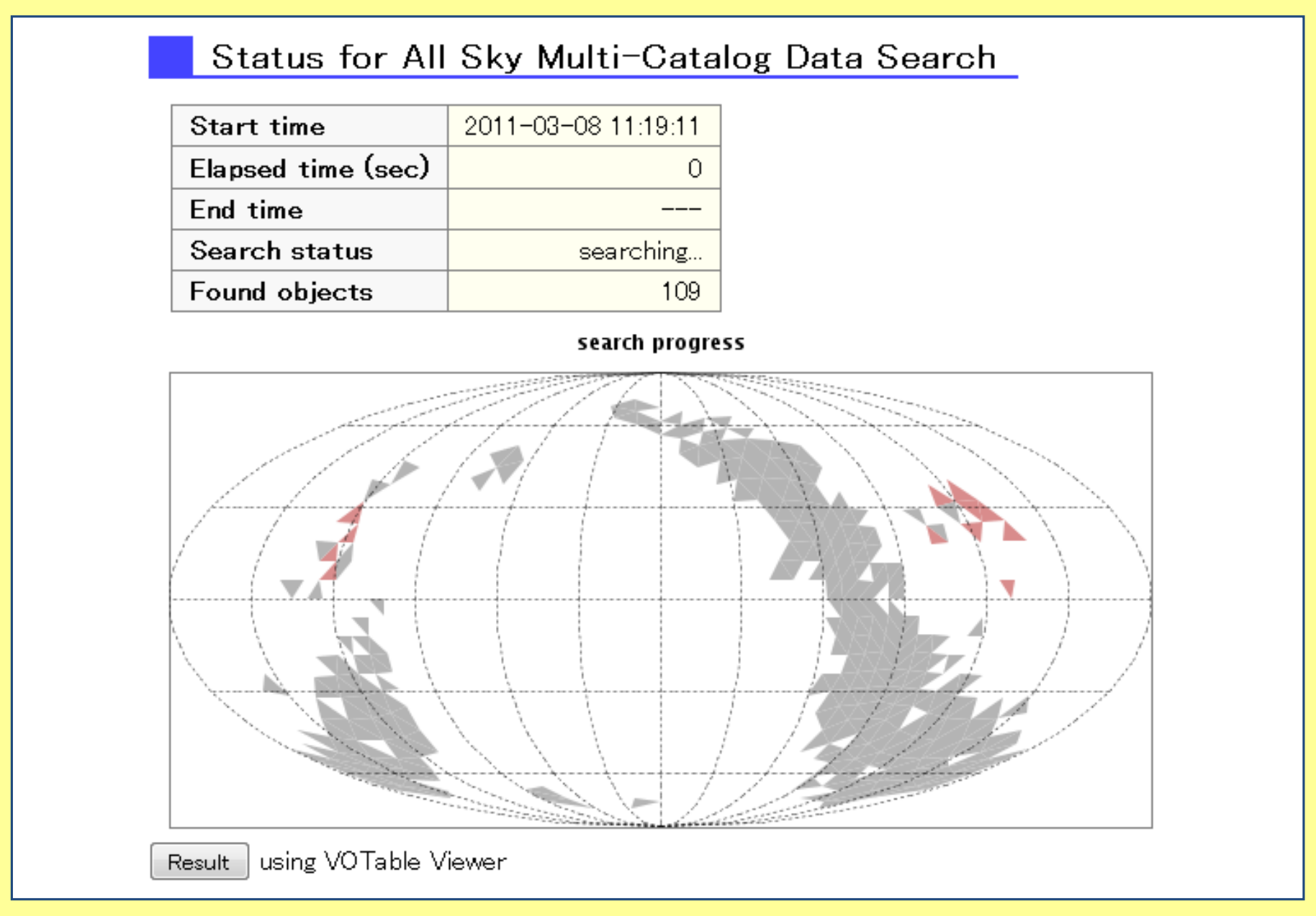
Interim results of a trial run of VO Crawler

Number of records:	~230 million
Total amount of data:	5.3TB
Execution time:	32 days

10 parallel queries are performed.
Among 11,632 target services, crawling for 2551 services finished successfully. For 914 services, crawling failed. Most of the remaining 7963 services are tables in VizieR ,and crawling for most of other services finished.

Whole Sky Search

We are developing a Whole Sky Search system, which searches objects with specified characteristics, such as colors and magnitude, from the data retrieved from multiple VO services. It searches objects from the whole sky without requiring to specify a coordinate to be searched.
In advance, flux(magnitude) data of all objects in the all catalogs in the VO Crawler Data Storage is extracted, and converted into tables with the common separated format. The Whole Sky Search System search objects from the prepared table using Hadoop.

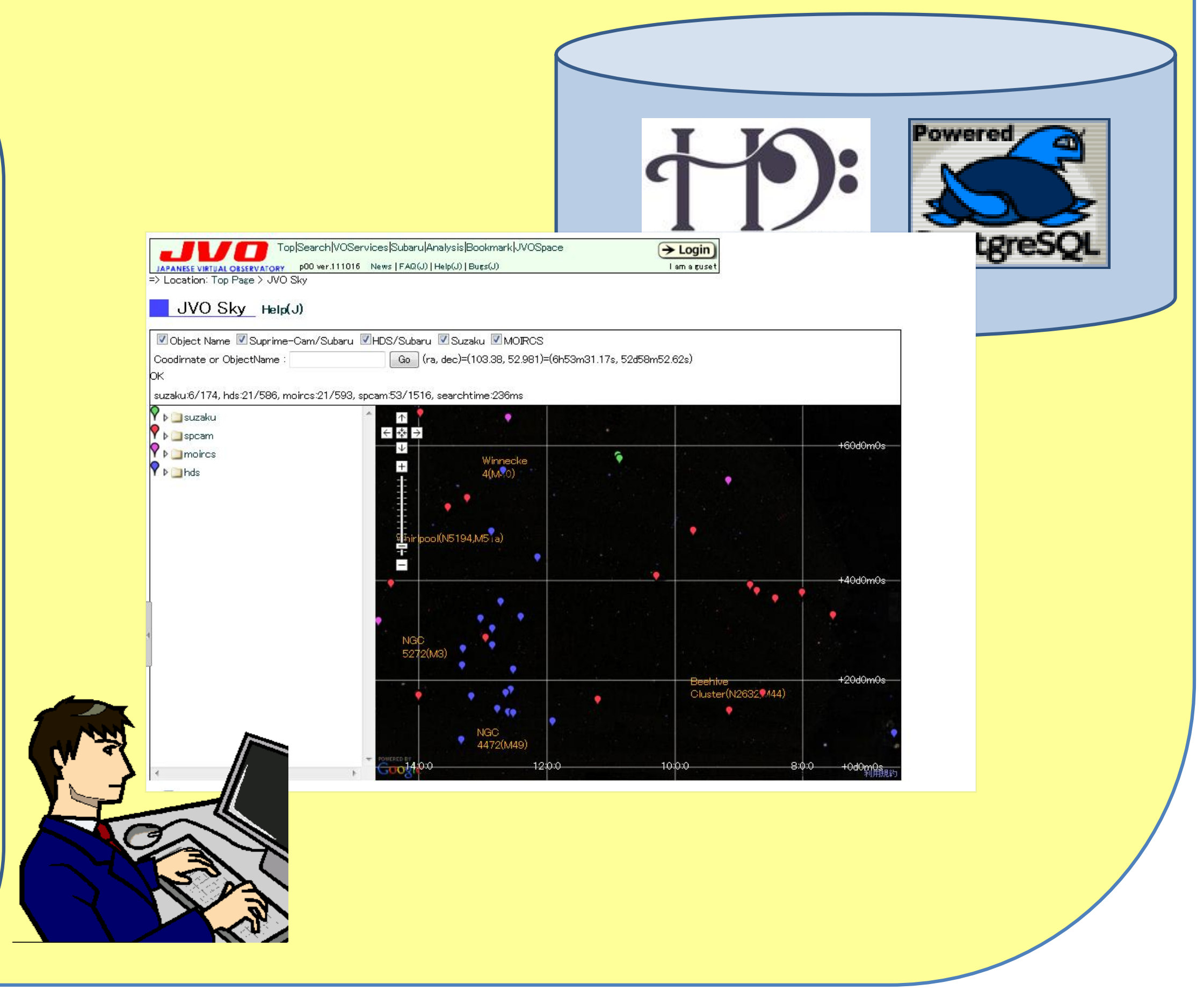

Status for All Sky Multi-Catalog Data Search

Start time	2011-03-08 11:19:11
Elapsed time (sec)	0
End time	---
Search status	searching...
Found objects	109

A snapshot of the Whole Sky Search system using SDSS and 2MASS catalog (they are not data stored by VO Crawler). Grey region shows searched area and red region shows area with objects that meet the query conditions.

JVO Sky

We have been developing JVO Sky, which is a graphical user interface (GUI) to display observed area on a sky map by using the Google Sky API, and until now data obtained by the Subaru telescope and the Suzaku satellite are registered in the JVO Sky.
We are planning to visualize distribution of the data stored by the VO Crawler on the JVO Sky. Number density of stored data in each HTM ID, and observed area of image data will be displayed on the sky map. Users may select data by band-name, instrument names, and/or region on the sky map.
Currently, we are developing a classifying and totaling system for stored data.



My Catalog Builder

We are developing a system to build custom-made catalogs from the stored data.
Users may assign column names and/or keywords for a “My Catalog”. In response, My Catalog Builder searches metadata and column description of catalogs in VO Crawler Data Storage, and find candidates of data for the My Catalog. Among the candidates, users select data to be registered on the My Catalog.
Users can make catalogs with all the data of the specified physical quantities of objects in the specified class from data in all the VO services.