Astronomical Data Analysis Software and Systems XIX
ASP Conference Series, Vol. 434, © *2010*
Y. Mizumoto, K.-I. Morita, and M. Ohishi, eds.

Lessons Learned during the Development and Operation of Virtual Observatory

Masatoshi Ohishi, Yuji Shirasaki, Yutaka Komiya, and Yoshihiko Mizumoto

National Astronomical Observatory of Japan, 2-21-1, Osawa, Mitaka, Tokyo, Japan

Naoki Yasuda

Institute for the Physics and Mathematics of the Universe, The University of Tokyo, Kashiwa, Chiba, Japan

Masahiro Tanaka

University of Tsukuba, Tsukuba, Ibaraki, Japan

Abstract. In the last a few years several Virtual Observatory (VO) projects have entered from the research and development phase to the operations phase. The VO projects include AstroGrid (UK), Virtual Astronomical Observatory (former National Virtual Observatory, USA), EURO-VO (EU), Japanese Virtual Observatory (Japan), and so on. This successful transition from the development phase to the operations phase owes primarily to the concerted action to develop standard interfaces among the VO projects in the world, that has been conducted in the International Virtual Observatory Alliance. The registry interface has been one of the most important key to share among the VO projects and data centers (data providers) with the observed data and the catalog data. Data access protocols and/or language (SIAP, SSAP, ADQL) and the common data format (VOTable) are other keys. Consequently we are able to find scientific papers so far published.

However, we had faced some experience during the implementation process as follows:

- At the initial stage of the registry implementation, some fraction of the registry meta data were not correctly set, or some meta data were missing. IVOA members found that it would be needed to have validation tools to check the compliance before making the interface public;
- It seemed that some data centers and/or data providers might find some difficulties to implement various standardized interfaces (protocols) in order to publish their data through the VO interfaces. If there were some kind of VO interface toolkits, it would be much easier for the data centers to implement the VO interfaces;
- At the current VO standardization, it has not been discussed in depth on the quality assurance on the published data, or how we could provide indexes on the data quality. Such measures would be quite helpful for the data users in order to judge the data quality. It would be needed to discuss this issue not only within IVOA but with observatories and data providers;
- Past and current development in the VO projects have been driven from the technology side. However, since the ultimate purpose of the VOs is to accelerate getting astronomical insights from, e.g., huge amount of data or multi-wavelength data, science driven advertisement (including schools to train astronomers) would be needed;

- Some data centers and data providers mentioned that they need to be credited. In the Data-Centric science era it would be crucial to explicitly respect the observatories, data centers and data providers;

Some suggestions to these issues are described.

1 Introduction

New era for astronomy has started; new telescopes that are being built or planned, such as ALMA, LSST, SKA and others, have very high data production rates. For example, ALMA will produce observed data a few Peta bytes per year, and LSST will produce 200 Pete bytes for expected six years of operation. Thus it would be possible to conduct "Data-Intensive Astronomy" by using survey data toward very large angular coverage and/or by using multi-wavelength data covering the radio to the gamma-ray. However, the traditional data query and data analysis systems usually do not correspond to such amount of data, and it is foreseen that astronomers may suffer from handling such amount of flooded data.

Astronomical Virtual Observatories (hereafter VOs) aim to support and accelerate data-intensive astronomical research by incorporating latest information technologies (e.g., database management, data federation through the web technologies, federation of remote storages and data analysis engines). VOs have been regarded as a new research infrastructure in the 21st century, and a definition of VOs is given "A virtual observatory (VO) is a collection of interoperating data archives and software tools which utilize the internet to form a scientific research environment in which astronomical research programs can be conducted. In much the same way as a real observatory consists of telescopes, each with a collection of unique astronomical instruments, the VO consists of a collection of data centers each with unique collections of astronomical data, software systems and processing capabilities."[1]

In 2002, the International Virtual Observatory Alliance (IVOA)[2] was established, aiming to standardize several protocols (such as meta-data description and access, data queries, data models, data format, and others) that are essential to build a data-intensive astronomical research environment. There are 17 members in the IVOA, and IVOA has already approved several essential standards toward building interoperable VO environment. As the result some national and regional VO projects have transited from the research and development phase to operation phase. However, not many astronomers have not used the VO portals and/or VO tools, consequently not so many academic papers that were produced through the VO environments have published so far.

---

[1] `http://en.wikipedia.org/wiki/Virtual_Observatory`

[2] `http://www.ivoa.net/`

In this paper I would describe the current status of the VO environment in the world, what we could learn from the past VO developments, and what we should do toward much more scientific outputs by utilizing the VO environment.

2 Latest Status of Virtual Observatories

2.1 Standardization in IVOA and VOs in Operations

Since 2002 there have been extensive efforts towards IVOA Recommendations[3] that enable all the VO projects and data centers to interoperate among them, i.e., data discovery by using meta-data standards (registry), data access (images, spectra, time-series data, and catalogs), output data format (VOTable), federation among data analysis tools, and others. The standardization process was so painful, because the IVOA members consist of astronomers, software engineers and others where different primary interests were mixed up. There were different philosophies, ideas, aims, intentions, views towards an IVOA Recommendation. In many occasions people need to be very patient and have to compromise. Some people were so reluctant to compromise. However, the IVOA members were able to establish good relationship even during coffee breaks and lunch/dinner times, and were able to trust to each other, which finally lead to approve many standards.

The established standards have been found to be very effective, after people implemented individual VO services systems. Even in 2004, some VO services were interoperable by adopting the IVOA standards ! After that some VO projects (AstroGrid in UK, NVO in the USA (its new name is VAO – Virtual Astronomical Observatory), Euro-VO, JVO, and others) transited from the research & development phase to the operations phase. As the result, more than 6000 resources are shared by the astronomy communities in the world, and other science disciplines have tried to construct similar data sharing environment – “Virtual Observatories”.

2.2 A Case for Japanese Virtual Observatory

As an example of the VO system in operation, we will describe the portal system of the Japanese Virtual Observatory[4]. JVO portal is designed to seamlessly access to multi-wavelength, federated databases and data analyses systems for astronomers through high speed network facility. We have kept the same basic concept as was described in Ohishi et al. (2004). Figure 1 shows the top page of the JVO portal system.

The top menu is categorized into several parts. “Data Search” corresponds to retrieval of images, spectra and catalogs from registered data services in the world. “Quick Search” is the easiest one to find any data service that is satisfied by the given key words. If a user already knows which data service(s) to access, the user can use the “Single VO Service” or “Multiple VO Services”. “JVO Sky” displays observed data on the celestial plane by using the Google Sky API,

---

[3] `http://www.ivoa.net/Documents/`

[4] `http://jvo.nao.ac.jp/portal/`

and a user can easily find and download necessary data. The “Service Search” corresponds to the registry service, where each menu is linked to the “Data Search” service. Therefore a user can search for interested data service(s) by giving some key words or from categorized data services.

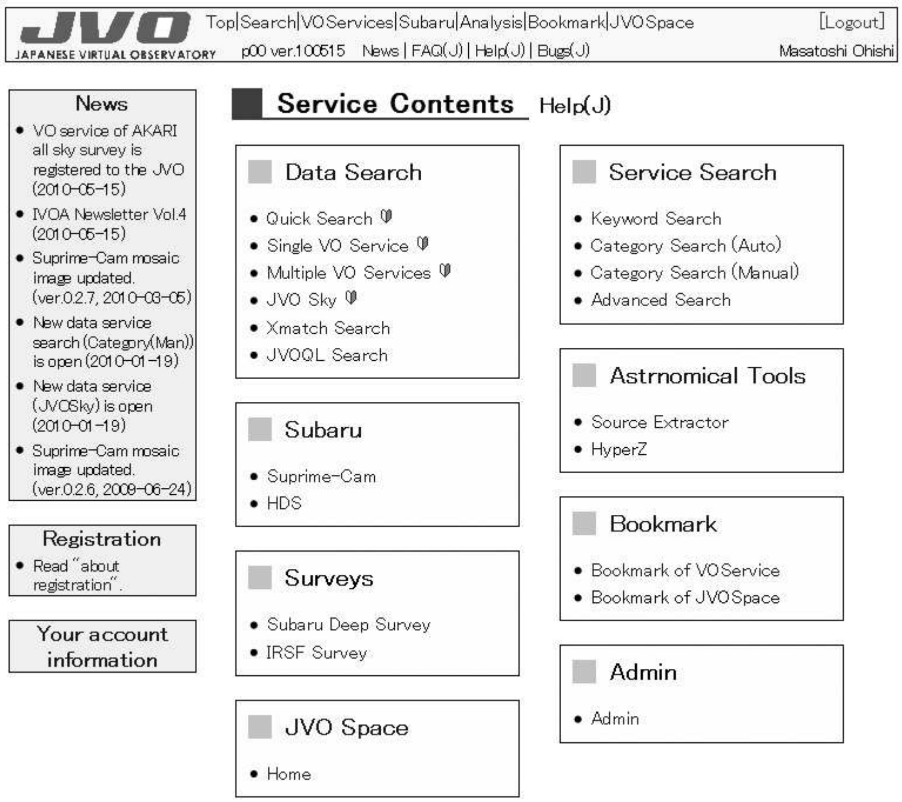

Figure 1. Top menu panel of the JVO portal system. Some menus with a leaf-like image are designed for novice users.

The “Subaru“ and the “Surveys” are the unique data services in the JVO portal system. The Suprime-Cam is a wide field camera equipped with the Subaru telescope, and produces the largest amount of image data from Subaru. The JVO portal system provides not only the raw data but calibrated and mosaicked image data that users can easily feed them into their data analysis softwares (Shirasaki et al, 2009). HDS provides with the high dispersion spectra in the optical region, and the “IRSF Survey” gives the observed data by the

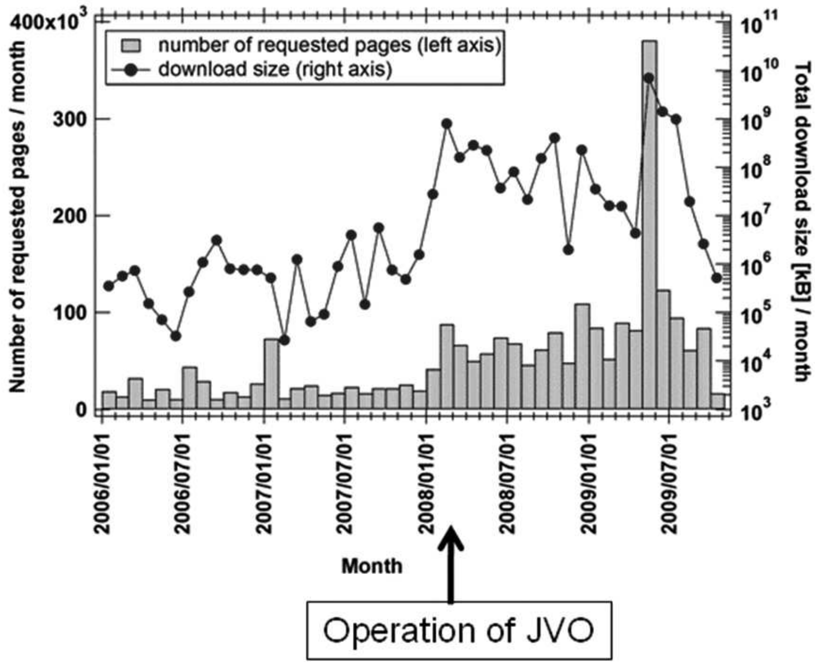

Figure 2. A statistical access log to the JVO portal system. The downloaded amount of data is between a few hundreds Giga bytes to about 1 Tera bytes per month after entering the operation phase.

InfraRed Survey Facility[5] that is located in South Africa and is operated by the Nagoya University, Japan.

The JVO portal system has been in its operation phase since March 2008. Figure 2 shows the access log data to the JVO portal system since January 2006. At the beginning of this access log data, the JVO portal system was under experimental operation and was not well advertised. Thus the the downloaded amount of data was around 1 Giga bytes per month. But after the official announcement in March 2008 that the JVO portal system would enter the operation phase, the downloaded amount of data jumped up by about 2 to 3 orders of magnitude. We were able to analyze the source sites of the access, and it was found that the access come from all over the world. The mostly downloaded data was those from the Suprime-Cam of the Subaru telescope, thus it was understood that many astronomers really hoped to retrieve the Subaru data for their scientific use. It should be noted that some users of the JVO portal include amateur astronomers. This would mean that the VO systems would provide good means for public outreach and educational use.

---

[5] `http://www.z.phys.nagoya-u.ac.jp/~irsf/index_e.html`

2.3 Scientific Outputs through VOs

IVOA publishes newsletters biannually[6]. The primary purpose of the IVOA newsletters is to highlight VO applications that were developed by the IVOA members. At the same time, each newsletter contains links to find some recent papers about VO-enabled science and other VO-related papers. However it is found, as of October 2009, there are about 110 refereed science papers so far by using the VO facilities in the world. It was striking fact for the IVOA members, and it is needed to analyze why there are such small numbers of refereed science papers whereas in ADS there are more than 1400 non-refereed papers that contain "virtual observatory" in its abstract.

3 Towards More Science Outputs through VOs

It is found that most of the 1400 non-refereed papers describe technical developments. This is understandable since these papers were published in the technology development phase. However the ultimate purpose of the VOs would be to accelerate data-intensive astronomical research, and we need to be well prepared before the data avalanche actually occurs within a few years. There would be several issues for VO developers to consider as shown below.

3.1 More Science-Driven Development

Since the VOs aim to accelerate data-intensive astronomical research, it would be very useful to survey science use cases from astronomers. Such use cases could be used to determine the priorities of future developments and to examine if the developed VO functionalities fulfill such science requirements. Further data manipulations based on such science use cases would lead to realistic demonstrations for astronomers what VOs really can achieve. Such science-driven development would be crucial towards the realization of the "VO era".

It should be noted that IVOA has set up "Science Priority Committee" to collect a few science use cases from astronomy communities. The task of this committee is not to select high priority areas in astronomy, but to find priorities in future developments of the VO technologies and VO tools. Thus the idea of the "Science Priority Committee" is quite similar to the paragraph above.

3.2 Easy Toolkit in Taking Up VO interfaces

It has been the policy in IVOA that IVOA will only provide interoperable standards among individual VO projects, data providers, VO tool providers. Thus those who wish to introduce VO environment may need to implement its own system unless they decide to duplicate an existing VO system. It is easy to install VO tools (e.g., Aladin, TOPCAT, VOSpec) to a local machine, however, it would need much more effort to implement a new VO system from a scratch and such implementation may introduce unwanted bugs.

Thus it would be desirable to provide an easy-use toolkit with new comers to the VO world. As an example, JVO developed a toolkit to build a(n) (obsolete)

---

[6] `http://www.ivoa.net/newsletter/`

SkyNode portal some years ago. It was possible for a data center in Japan to construct a basic VO environment by using this toolkit. Another way would be to provide a validator for a new developer, which could be used to examine if the implemented software has been in accordance with the IVOA Recommendations. For example, IVOA has developed a validator for the registry service.[7] Similar validators for other IVOA standards would assist new developers to examine if the implemented VO system comply with the interoperable standards.

3.3 Training Courses and Tutorials

Even if the VO systems are science-driven, astronomers may not know how to utilize the VO systems. It should be recognized that most of astronomers have no knowledge on the SQL and other technical terms. Thus it would be crucial to the VO projects to prepare manuals (both the novice and advanced versions), sample science use cases than can be traced by users, and to hold training courses and/or tutorials regularly. The manuals could be prepared in English, but it would be desirable that the manuals are written in the local language.

The primary target of the tutorials should be younger people (e.g., undergraduate and graduate students, and postdoctoral fellows), since younger people usually show strong interests on "new" systems and it is possible for them to learn quickly how to use the VO systems. Further younger students may have sufficient time to practice and to write papers. Once younger people get used to use the VO systems, they may talk to their supervisors that the VO systems are so good. Similar thing was observed at the beginning of the Web use; young students enjoyed to use the Mosaic (the first web browser) and elder people followed these young people.

3.4 Quality Assurance or Quality Index

Current VO systems can provide users with multi-wavelength observed data by various telescopes. There are more than 6000 resources than can be accessed. However, it is usually experienced that too many data sources are hit and it is sometimes difficult to find which data should be actually usable for his/her research purpose. This is because most of the observed data is not associated with information on its data quality and it would be difficult to query by means of the data quality information. Of course there are exceptions; survey data taken by dedicated telescopes are usually well controlled in the data quality.

The problem would be that there has not been an established quality index. For an image quality information may vary from pixel to pixel. For a spectral data quality information may vary as a function of wavelength or frequency. It would be an urgent issue for the IVOA to develop a standard to express such quality indexes. And it is ultimately needed for the VO projects and (real) observatories to establish how to assure the quality of observed data. Such quality assurance would be crucial for the data-intensive science era. No one would not be able to examine data quality of all the data, and the data analysis would be made through pipeline systems or workflows. There would be no room for an astronomer to interact with such a huge system.

---

[7] `http://rofr.ivoa.net/regvalidate/regvalidate.html`

3.5 Credit to Data Centers and Data Providers

Finally we point out that some people in data centers or data providers expressed concerns if their efforts to make all the data available to the VO world could not be acknowledged. If nobody acknowledges to data centers and data providers, there would be no incentive for data centers and data providers to publish the data to the VO world. This would be a serious problem not only for data centers and data providers but also to astronomers, i.e., users of the VO systems. Therefore it is needed for the IVOA to develop a standard mechanism to collect and provide data curation information. Such a mechanism may lead to acknowledge data centers and data providers, and then data centers and data providers could apply new budget to maintain and upgrade their activities.

4 Summary

Astronomical Virtual Observatories have been regarded from other science disciplines as a "successful model" in sharing distributed data. However we have experienced several difficulties in the development process of the VO standards and systems, and have found some important issues to be tackled toward the successful data-intensive science. Our real success would come when so many scientific insights have been produced through the VO systems. It is expected that the real success will be realized soon.

Acknowledgments. This work was supported by the Japan Society for Promotion of Sciences (JSPS) Core-to-Core Program and Grant-in-aid for Information Science (15017289, 16016292, 18049074 and 19024070) carried out by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References

Ohishi, M. *et al.* 2004, in ASP Conf. Ser. 314, ADASS XIII, ed. F. Ochsenbein, M. Allen, & D. Egret (San Francisco: ASP), ASP Conference Proceedings, Vol. 314, 296 (2004)

Shirasaki, Y. *et al.* 2009, in ASP Conf. Ser. 394, ADASS XVII, ed. Robert W. Argyle, Peter S. Bunclark, & James R. Lewis (San Francisco: ASP), ASP Conference Proceedings, Vol. 411, 396 (2009)