Virtual Archive as a prototype distributed data system for scientific knowledge base[♣]

© A.I.Osin E.P.Trushkina V.D.Kuznetsov

osin@izmiran.ru elena@izmiran.ru kvd@izmiran.ru

Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation

(IZMIRAN)

Abstract

This document outlines an attempt to develop guidelines for a low-barrier unified distributed data system ("Virtual Archive") based on modern standards. A prototype data system uses approach close to IVOA [1] but aims at a more general application area. Existing trends and standards in building distributed data systems are briefly discussed and Virtual Archive approach to specific issues laid out.

1 Introduction

Distributed archives and data systems are becoming routine nowadays. Although centralized archives and large data centers will remain playing important role especially in specific research areas, huge amount of data in many fields makes it unrealistic to maintain universal centralized data archives. Besides, more cross-disciplinary research projects require simultaneous access to data from various knowledge domains. With high bandwidth communication lines connecting the world and large number of various data sources available more feasible appears to be developing distributed interoperable archives tied together by a unified infrastructure which can be used for seamless data manipulation.

Recent years were marked by augmented activity in the field of data and data services 'virtualization'. Although significant progress has been made and a number of new technologies and systems such as Virtual Observatories matured, users still have to deal with the diverse and confusing collection of portals, services, query languages and data models. Dumb portals are slowly giving the way to semantically enhanced data systems, which use semantic information (vocabularies) to enhance data queries. Grid [2] community also grows and gains popularity. At the same time, many small research groups and centers are still using outdated technologies hampering productivity of their work and possibilities for international cooperation. A low-barrier system is what these small groups need to become fully integrated.

Astrophysical community has made significant progress recently due to IVOA activity and it is time for other science communities to embrace new information technology approaches. In particular, space and geosciences are characterized by large amounts of data collected at geographically distributed locations. More space projects are being targeted at the Earth and near Earth environment. It is a good opportunity for the international geophysical community to take advantage of the Electronic Geophysical Year, 2007-2008 (*e*GY) [3] and focus efforts on the modern approach to data processing: universal data discovery, open access to data, easy data sharing and effective data usage. The development of Virtual Observatories and Laboratories is the central feature of *e*GY. A number of such projects has started over the last few years and are still in the process of developing commonly accepted approaches and standards. Our project is targeted on the development of modern data system for the geophysical community in the first place.

2 Virtual observatories

The International Virtual Observatory Alliance (IVOA) has emerged in 2002 as an effort to unite international astronomical community, to develop and promote standards (tools and systems) for the utilization of the astronomical archives as an integrated and interoperable "virtual" observatory. IVOA has developed and adopted several key concepts and standards, building blocks of the future VO infrastructure, such as VOTable, UCD, VOResorce, VO Registry, VO Identifiers, etc. These concepts and standards are quite universal and can easily be adapted and successfully used in other knowledge areas as well.

Virtual Observatory Alliance is formed by member organizations from several countries including Russia. RVO [4] represents Russia's efforts in bringing astronomical community to international collaboration within the framework of the IVOA and filling the gap in the level of development and use of VO technologies in Russia. The core of RVO infrastructure is based on AstroGrid [5].

A number of VxOs and similar projects has recently been launched in space physics, solar-terrestrial physics and geophysics, such as VSO[6], EGSO[7], VSPO[8],

Proceedings of the 9th Russian Conference on Digital Libraries RCDL'2007, Pereslavl, Russia, 2007

VSTO[9], VHO[10], VMO[11], VGMO[12], VITMO[13], SPASE[14] and SPIDR[15]. Not all of them follow IVOA standards and many are still under development. The projects significantly differ in user interfaces and APIs and thus only partially solve the problem of the data access systems diversity.

Virtual observatories are often being built by first offering minimal, open and universal query interface and access to metadata of limited (discipline-specific) number of data archives and services. Simple browser based queries are routed to middleware which returns direct links to data providers based on a set of restrictions and specifications. At the same time, higher-level meta-observatories are expected to emerge like heliospheric-wide virtual observatory based on existing VO's in heliospheric, magnetospheric and ionospheric physics and VSO.

SPASE (Space Physics Archive Search and Extract) [14] is an international consortium of space physics data provider groups interested in making the data search and retrieval process easier for the space physics community. It is focused on developing space physics ontology for a model distributed data system for space physics and related research domains. SPASE data model and dictionary created for space physics is similar to that of IVOA for astronomy. SPASE ontology is used in VSPO (Virtual Space Physics Observatory) [8] and some other VxOs.

SPIDR (Space Physics Interactive Data Resource) is designed to allow a solar terrestrial physics customer to access and manage space physics data for integration with environment models and space weather forecasts. SPIDR is a distributed network of (currently 10) synchronous databases accessed via the World Wide Web. By enabling data mirroring and eliminating the network bottlenecks associated with transcontinental links, the distributed system architecture brings low latency in data visualization and fast data delivery.

The key concepts in the SPIDR architecture are the data basket (a collection of different space weather parameters selected from different databases for the same time interval) and space weather event. SPIDR databases include solar activity indexes, geomagnetic field variations and ionospheric parameters, sunspot numbers, solar wind parameters and cosmic ray intensity, environment parameters at the geostationary altitudes reported by GOES satellites, space images of the Earth in visible and infrared light obtained from US DMSP satellites and Solar images in various spectrum ranges.

VITMO (Virtual Ionosphere Thermosphere Mesosphere Observatory) is another virtual observatory in the field of solar-terrestrial physics which uses its own data model architecture composed of *regions, features, events and parameters*. VITMO builds on existing data centers and is based on the Scientific Resource Access System (SRAS) [16] and uses Java Server Pages (JSP), Java Servlet technology and relational database (Oracle). It plans to incorporate metadata representation using the eXtensible Markup Language (XML) and other technologies in the future.

3. Virtual Archive

3.1 In the beginning was the word

The Virtual Archive (VA) project initially emerged at IZMIRAN as an independent work aimed at providing common environment for the access to various local data archives. There is a number of research groups at IZMIRAN working in such fields as the Earth magnetism, ionospheric and magnetospheric research and solar physics. Considerable amount of data is accumulated from various facilities including ground based magnetic observatories, stratospheric and satellite measurements, network of topside ionospheric sounding stations, in-situ satellite measurements, cosmic ray monitors, complex space projects targeted at solar physics (CORONAS) and geophysical monitoring (COMPASS). The collected data though is poorly organized and lacks convenient unified access interface.

Virtual Archive [17] is a project aimed to develop a low-barrier unified distributed data system, which can be used to access data in various distributed data archives. The initial Virtual Archive architecture later slowly evolved in the direction of OAI and IVOA standards and recommendations but its primary goal remains the creation of a universal yet simple data access system. Minimal efforts are expected to be required from the data providers for the integration of their data into the VA.

Virtual Archive concentrates on data search and retrieval as well as certain basic preprocessing and visualization. As data processing modules can only be attached (using specific metadata) to certain known types of data, the two major tasks – data retrieval and processing should be implemented as two separate but interoperable subsystems. Virtual Archive is designed to be a pure data search and retrieval subsystem with possible prerequisites for subsequent attachment to the Virtual Laboratory. Virtual Laboratory should provide pointers to reusable software modules which can later be applied to process data discovered using Virtual Archive or create distributed workflow sequences.

3.2 UCD Semantics

UCDs (Unified Content Descriptors) [1], originally developed and used in ESO/CDS VizieR catalogue service (Ortiz et al. [18]) and later supported and enhanced by IVOA, can be used to describe many astronomical concepts. The main goal of using UCDs is to describe semantics of various quantities. UCD hierarchy is usually built by following “bottom-up” approach - describing what is found in existing datasets and comparing it with the existing ontology. Sharing common UCD among distributed nodes of a VO is extremely important for interoperability, but excessive level of detail restricts applicability of the UCD to specific research area (astronomy).

Authors and developers of UCDs acknowledge that creating and maintaining controlled vocabulary is an extremely difficult task. Indeed, introduction of universal and detailed UCD vocabulary seems to be extremely controversial and should probably not be attempted.

According to IVOA, UCD (UCD1+) is a string of "words" separated by semicolon. Words are composed of "atoms" separated by periods. Atoms are defined following the guideline: abbreviations should be kept to a minimum, and only if the result is not ambiguous. All UCD1+ words are grouped into 12 main categories: **arith, em, instr, meta, obs, phot, phys, pos, spect, src, stat, time**. It is quite obvious that the upper level atoms are extremely research domain specific and subjective (and are abbreviated).

UCDs can also be used in registries to locate relevant resources. But how can such a specific UCD list be used for registries embracing more general set of data archives ? This arguments led us to conclusion that a more general UCD vocabulary is needed in order to perform search query over a wider domain.

In our view, UCD taxonomy should only reflect specific research domain ontology as many common measurable quantities can be attributed to not just one research domain. Locally measured atmospheric pressure can be produced by a meteo station but can also be of interest for medical research (health effects) or ecology or cosmic ray physics. All physical quantities can be grouped under the first level atom **phys**. Another two fundamental atoms could be **geom** and **time**. The **geom** group should include all common geometrical quantities and parameters (including positional), such as sizes (width, length, height), geographical coordinates (latitude, longitude, altitude), distances, volumes and shapes (geom.volume, geom.radius).

On the other hand chemistry, biology, ecology, sociology etc. all have their own specific set of quantities other than physical or geometrical ones (their own ontology). Business databases include financial data such as currency rates or stock market quotes. These specific quantities can be collected under their own (chem, bio, eco, soc or fin) hierarchies (fin.currency.rur2usd, fin.quote.nasdaq.bestbid).

Selecting general versus discipline-specific vocabularies helps to build unified data systems but eliminates possibilities for detailed queries and may produce excessive hit results. This effectively takes us back to a simple WWW style keyword query. To overcome this controversy an approach is suggested which should combine general purpose hierarchy and discipline-specific vocabularies developed by various research communities. The unified structure is created by the introduction of corresponding name spaces (namespaces in UCDs are allowed but strongly discouraged by IVOA). All universal non-specific concepts can be collected under the Virtual Archive name space:

va:phys.temperature;va:bio.human

VxO-specific vocabularies then can save space by retaining only discipline-specific words.

3.3 VARegistry

IVOA defines a VO Registry as a special resource, a query service, for which the response is a structured description of resources. The resources described by a registry may be of any type. Registries are represented by two primary types: searchable and publishing. Searchable registries allow users and applications to search for resource records while publishing ones simply expose their resource descriptions to others for harvesting - collecting records in a centralised location (full registry). Local registries are usually publishing ones. Dublin Core Metadata Initiative (DCMI) [19] recommendations and standards are generally assumed as a good basis for resource description.

In order to assure low-barrier approach VA Registry can be implemented as an OAI Static Registry and thus support harvesting OAI-PMH style [20]. The only minimal requirement for data providers is to support local VA Registries as simple as a single XML file accessible via WWW. Just as IVOA extends DCMI resource metadata descriptions to address astrophysical parameters, VA allows for extended set of elements to describe certain specific keywords. We are also planning to extend static registry functionality to include simple search capability. Large data centers may wish to support full Registry interface implementation based on web services [21].

VA Registry and VATable (metadata) records may include "non-authoritative" records. By adding non-authoritative Resources Virtual Archive can take advantage of the open data sources which do not yet have ability or willingness to maintain their own VARegistry/VATable (Fig. 1).

As a basic data search service, Web-based VA interface can be used to discover data using simple keyword queries (Instrument, Facility, Time interval, etc.) as well as queries assisted by semantic entities (UCD).

3.4 VA metadata and VOTable format

VOTable is an XML-based format for tabular data endorsed by the International Virtual Observatory Alliance [1]. It is a central structure in the VO architecture that makes data interoperable and scalable. VA project intends to use different XML format for metadata representation. It's current working version ("VATable") was derived from VOTable.

Unlike VOTable, which was designed as a storage and exchange format for tabular as well as binary data (FITS) with particular emphasis on astronomical tables, VATable is intended to be simple format for metadata representation only. Keeping data and metadata separately allows to introduce a concept of "virtual data" when a process can perform certain preparatory actions without having to actually download the data. Using VATable format existing data archives can easily be described and integrated into the Virtual Archive. The data model of the current version of VATable can be represented as follows:

VATable	= Metadata + Resources
Resource	= Metadata + Tables
Table	= Metadata + Fields
Metadata	= Parameters + Descriptions

Metadata may also include specific instructions and pointers to the software modules. VATable structure is not fixed yet and may be changed and extended in the future.

3.5 Services

Distributed archives are usually built upon services. The appearance of the new standards for web based services, such as WSDL, SOAP and UDDI [22], opens clear way for data systems integration. Web services form the basic underlying technology for various VxOs. Registry queries and synchronization, data search queries, subsetting and format conversions prior to the data delivery, are all based on web services. A special new service is under development in the VA to support querying data availability. It is a common situation when time series data include gaps due to interrupted observation cycles or data loss. Attempts to access nonexistent data cause increased latency and delays. This seems to be a common problem but current systems often only include metadata in the form of start and end dates for a particular archive. VA plans to implement standard web services but also use REST approach when appropriate as well.

3.6 Current functionality

A user of the Virtual Archive starts session by looking for the relevant data by sending search query to the one of the VA Registries. Depending on the search domain he is then offered a search form specific to the selected domain. Another search query returns the list of relevant resources and allows to select some of them for saving. The selected resources are contacted and metadata (VATable) is then saved in the user work space. At this point the user have an opportunity to download actual data files and/or visualize them using built-in tools. A quick and straightforward access to the data is thus one of the main features of the VA.

Virtual Archive is seen as a data search and retrieval only service. Simple table browser and plotting facilities are the only tools that may be available at this level. For more elaborate data handling and processing data should be passed to a Virtual Laboratory. Workflow paradigm should be implemented to assist complex data manipulation. This is a future project, which should proceed as VA becomes mature enough.

4 Conclusions

A prototype distributed data system ("Virtual Archive") based on modern standards is described. VA architecture includes a number of provisions aimed at offering low-barrier approach to facilitate broad community of data providers from various knowledge domains. Among these special features are: simplified metadata (VATable) format, use of namespaces for extended UCD hierarchies, detailed metadata including data availability service and pointers to data processing modules, direct access and visualization of datasets and use of nonauthoritative resource records in order to widen the pool of available data sources.

References:

[1] IVOA: International Virtual Observatory Alliance http://www.ivoa.net

[2] Grid: Sharing distributed data http://www.globus.org

[3] EGY: Electronic Geophysical Year http://www.egy.org

[4] RVO: Russian Virtual Observatory http://www.inasan.rssi.ru/rus/rvo

[5] AstroGrid: http://www2.astrogrid.org

[6] VSO: Virtual Solar Observatory http://umbra.nascom.nasa.gov/vso

[7] EGSO: European Grid of Solar Observations http://www.egso.org

[8] VSPO: Virtual Space Physics Observatory http://vspo.gsfc.nasa.gov

[9] VSTO: Virtual Solar-Terrestrial Observatory http://vsto.hao.ucar.edu

[10] VHO: Virtual Heliospheric Observatory http://vho.nasa.gov

[11] VMO: Virtual Magnetospheric Observatory http://vmo.nasa.gov

[12] VGMO: Virtual Global Magnetic Observatory http://mist.engin.umich.edu/mist/vgmo/vgmo.html

[13] VITMO: Virtual Ionosphere Thermosphere Mesosphere Observatory http://vitmo.jhuapl.edu

[14] SPASE: Space Physics Archive Search and Extract http://www.spase-group.org

[15] SPIDR: Space Physics Interactive Data Resource http://spidr.ngdc.noaa.gov

[16] Immer, E., Daley R., Stock, J., Fortner, B., Jen, J. (2003), The Living With a Star Scientific Resource Access System: A Concept for Getting the Information to Do Science, *Eos Trans. AGU, 84*(46), *Fall Meet. Suppl., Abstract U22A-18*.

[17] VA: Virtual Archive http://va.izmiran.ru

[18] Ortiz, P. F., Ochsenbein, F., Wicenec, A., & Albrecht, M. 1999, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, eds. D. M. Mehringer, R. L. Plante, & D. A. Roberts (San Francisco: ASP), 379

[19] DCMI: Dublin Core Metadata Initiative http://dublincore.org

[20] OAI: Open Archives Initiative http://www.openarchives.org

[21] W3C: Web Services http://www.w3.org

♣This work has been partially supported by RFBR grant 05-07-90092

Appendix 1.

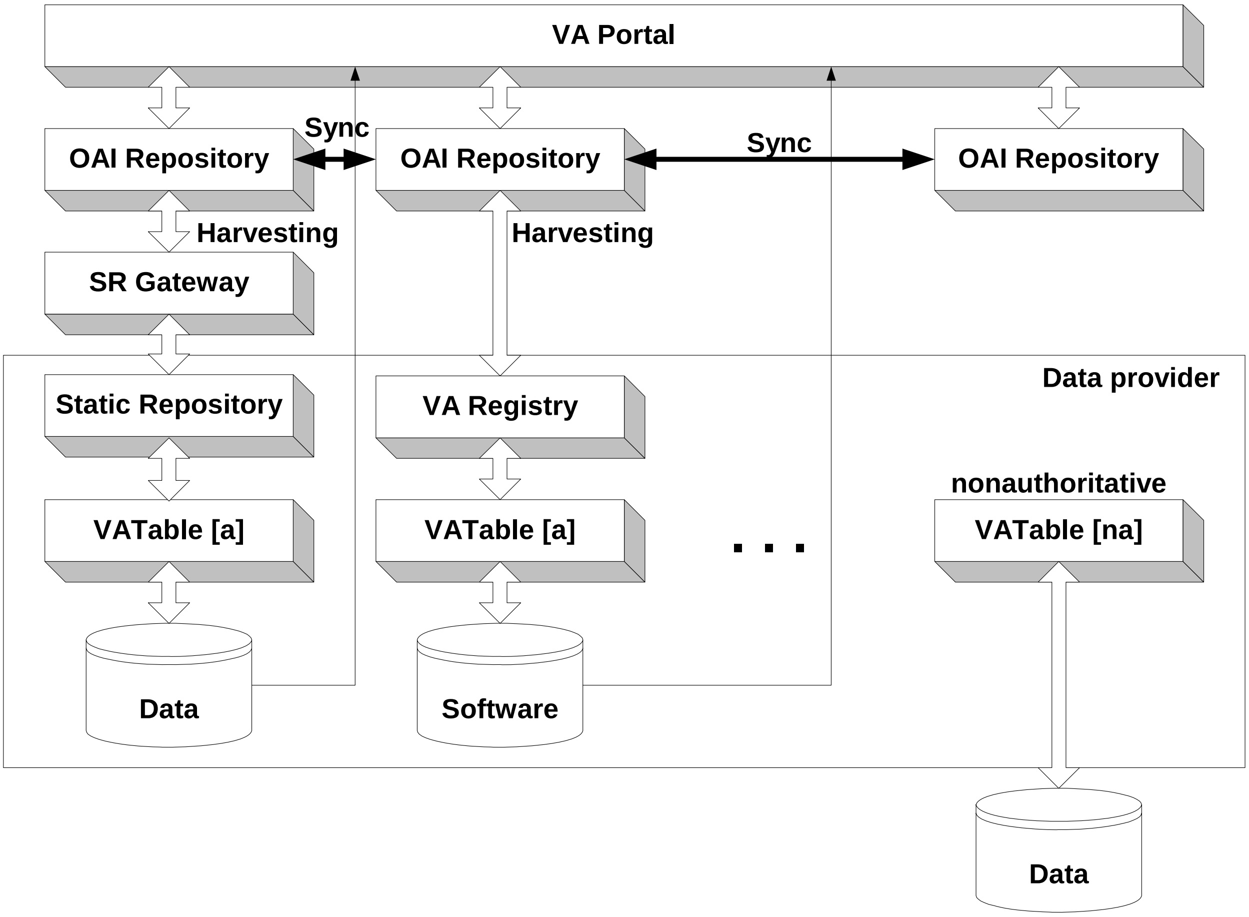

Fig. 1. Virtual Archive architecture