

FROM DATA TO KNOWLEDGE IN EARTH SCIENCE, PLANETARY SCIENCE, AND ASTRONOMY

Elaine R. Dobinson, Joseph C. Jacob, and Thomas P. Yunck

*Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, California 91109-8099
USA*

Email: {Elaine.Dobinson,Joseph.Jacob,Thomas.Yunck}@jpl.nasa.gov

ABSTRACT

This paper examines three NASA science data archive systems from the Earth, planetary, and astronomy domains, and discusses the various efforts underway to provide their science communities with not only better access to their holdings, but also with the services they need to interpret the data and understand their physical meaning. The paper identifies problems common to all three domains and suggests ways that common standards, technologies, and even implementations be leveraged to benefit each other.

INTRODUCTION

The past several decades of scientific exploration of the Earth, the sky, and the entire Solar System from space has produced an abundance of data. Remote sensing and in situ instruments are measuring numerous physical quantities and producing a wide range of numerical and imaging data sets. The curation of these data is, in most cases, the responsibility of the national archives for the appropriate scientific disciplines and subject area, while the analysis of the data is left to the individual science investigator. Thus, within the United States alone, there are the Earth, planetary, and astronomy data archives, each housing terabytes of data collected by satellites flown since the early seventies. The data stored in each of these archives are the original "raw" data, plus various higher-level processed data, usually converted to meaningful physical parameters for scientific study. Access to these data is generally very good, as most digital data are electronically available, and the data archive systems have done a thorough job of "cataloging" their holdings. Yet, the scientific process of converting these data into a real understanding of the universe and its physical phenomena is arduous at best. The sheer volume and diversity of data pose formidable problems to the science investigator, and the data preparation phase of an investigation typically accounts for more than 80% of the task. Furthermore, the same preparation activities – data acquisition, format transformation, re-gridding, re-projection, visualization, etc. – are done repeatedly from investigator to investigator, regardless of the domain, and from one study to the next. However, recent development activities, by information technology (IT) professionals, working together with their science colleagues, have focused their attention to alleviating these problems. In the remaining sections of this paper we examine some of these separate efforts underway in the various science domains listed above, with the goal of recommending common IT solutions across these different domains, and leveraging on emerging IT standards and developments for the benefit of all.

EARTH SCIENCE DATA SYSTEMS

The Earth science data systems have, by far, the largest and most diverse collection of data, having benefited from a wealth of attention on studying our home planet from space. The data resulting from these studies span long time periods, and consist of many remote and in situ measurements, collected by government agencies around the world. While this paper focuses its discussion to only those data collected by NASA and stored in the Earth Observing System Data and Information System (EOSDIS), the problems and solutions discussed have broad application across agencies and across the globe.

Observing the Earth from space entered its prime during the 1980s when the technology of conducting observations from Earth orbiting satellites became mature. The study of global climate change became a world-wide focus and NASA joined other agencies to measure key climate parameters. Archiving these measurements was the responsibility of the individual agencies, and data systems sprung into being at NASA, NOAA, USGS, as well as in Canadian, European, and Japanese space agencies. International committees formed to work through various standards issues so that the data could be shared among all, and these committees are still at work today.

In the 1990's a new era of the Earth Observing System began in the US, and NASA designed and launched multi-sensor platforms that would repeatedly orbit the Earth and measure land, ocean, and atmospheric parameters continuously over a period of many years. Designed as its flagship series of satellites, Terra, Aqua, and the recently launched Aura have sent back hundreds of terabytes of data to help Earth scientists determine the effects of mankind on the Earth system. To house and distribute these data, a new distributed system of data archives was devised, spanning the land, oceans, atmosphere, and other Earth science data sets, and organized in such a way as to hide their physical distribution. Each of the new Distributed Active Archive Centers (or DAACs) was to be part of a larger whole, but was responsive to its subset of the Earth science community. From the beginning, these DAACs were built to a common design and interface, and data format standards were developed to cover the majority of the data. The emerging standard format for these new data (HDF-EOS), based on an existing format, but with geo-location added in, would apply to many of the data collected by the EOS spacecraft, and a series of tools were developed to read and use data stored in this standard format. Thus all data in compliance would be interoperable. This was all well and good, except that not every data producer would use the standard the same way, not all data would be sampled in the same way, not all data would be at the same resolution, and not all data would be collected in the same time period. Thus, additional transformations are still required in order to compare different datasets with each other, not to mention all the legacy data to fit into this "system", leaving the individual scientist with an even more perplexing set of problems to solve. Hence, because these tasks are overwhelming to the individual scientist, each instrument team tends to focus on its own data issues for that instrument – calibration, validation, re-sampling, re-gridding, aggregating – and users of these data tend to work within this team's purview. Interdisciplinary investigations, that is, investigations that integrate data from more than one instrument and more than one topic area, remain extremely difficult. One notable exception to this is the emergence of widely accepted standards for geospatial data, promulgated by the OpenGIS Consortium, which has made tremendous inroads into this problem for all geo-spatial data. However, generic solutions for other data types and other domains still are lacking.

Several recent efforts are currently underway at NASA to aid the interdisciplinary investigator in this dilemma, and to promote the study of Earth System Science and make it feasible. One such effort begun this year is the General Earth Science Investigation Suite (GENESIS) project at the Jet Propulsion Laboratory (JPL) [Yun04]. GENESIS is a NASA-sponsored partnership between the Jet Propulsion Laboratory, academia, and three NASA data centers to develop a new suite of web services tools to facilitate multi-sensor investigations in Earth System Science. These tools will offer versatile operators for data access, subsetting, registration, fusion, compression, and advanced statistical analysis. They will first be deployed in a model server at JPL, and later released as an open-source toolkit to encourage enhancement by independent developers. While the initial work will focus on four premier atmospheric sensors – AIRS, MODIS, MISR, and GPS – the modular design offers ready extension and reuse on many Earth science data sets.

At the core of GENESIS is its scientific workflow engine known as *SciFlo*. SciFlo combines four core ideas to promote software reuse and create a marketplace for science analysis services: loosely-coupled distributed computing using SOAP; exposing scientific analysis operators as SOAP web services; specifying a data processing stream as an XML document; and a dataflow execution engine for a parallel execution plan and load balancing. The SciFlo design grew out of the pressing needs of scientists active in studies with these new sensors. The tools themselves will be co-developed by atmospheric scientists and information technologists from several institutions. At each step the tools will be tested under fire within active investigations, including cross-comparison of spaceborne climate sensors; cloud spectral analysis; upper troposphere-stratosphere water transport; and global climate model testing. The GENESIS tools, eventually to be inserted into routine DAAC operations, will help to inaugurate Earth System Science and will advance a modern data system architecture for realizing the broader vision of NASA's Earth Science Enterprise.

PLANETARY DATA SYSTEMS

The oldest of the NASA data archive systems still in existence today, the Planetary Data System (PDS), was begun in the mid-eighties, as an early prototype of the distributed archives model. Based on recommendations by the National

Academy of Sciences, NASA built an archive system of distributed discipline data centers (called nodes) under central management and development. These PDS nodes conceptually remain in existence today, and are responsible for the curation of the entire suite of data collected by all US and non-US planetary missions, dating back to the early days of the Viking and Voyager missions. Until the advent of the recent Mars orbiting missions, however, the entire planetary science data collection remained small (less than 3 terabytes), and access was easily provided on CD-ROM. However, with the increased number of orbiting missions to Mars, the Galileo and Cassini orbiting missions to Jupiter and Saturn, and the increased resolution of the instrument measurements, the size of the PDS archive now exceeds 300 terabytes. What was once a manageable problem suddenly grew to Earth data system proportions, and the PDS must now provide the same tools and access mechanisms that their Earth science colleagues have come to take for granted. Data must be accessible on-line as collections are too large to span a CD-ROM or two. The PDS has recently developed a new, distributed, on-line access to its data, seamlessly integrating all data from all nodes from all missions. This distributed system resembles the EOSDIS system discussed above.

PDS from the beginning developed a standardized set of "objects", and a standard nomenclature for referring to these objects. They also developed a standard language for describing the binary files containing these objects. This level of standardization works well within the system, but interoperability outside the PDS, with other widely accepted standards, is poor. This limits the software that can be used to manipulate the data, and the task of converting PDS objects into standard formats is left to the user. While PDS itself attempts to provide some rudimentary display tools, there are significant problems in integrating data across missions and across instruments. For, while the data object language is standard, the coordinate systems and projections used for the data are not. Thus the data sets, even from the various Mars missions, must first be transformed into a common system before they can be used together.

A pending effort by NASA is proposing to solve this problem for the terrestrial planets by leveraging off the Earth science work in the geo-spatial world. Extending the OpenGIS model to include planets other than Earth, and extending the standard interface protocols (WMS and WCS) to apply to the Moon and Mars is part of an effort underway at JPL. The end result will be a server of standardized "maps" of all terrestrial planets, usable by all OpenGIS software. The task of co-registering the individual images, and of re-projecting them to a common standard will still need to be done, but once done, these data will be available for all to use.

ASTRONOMY DATA SYSTEMS

Astronomical datasets exist in the form of images, catalogs, spectra, time series, and numerical simulations, which embody our knowledge of the universe at wavelengths that span the electromagnetic spectrum, from radio waves through gamma rays. Rapid technology improvements in detectors, telescopes, computing, communications, and storage, have given rise to all-sky surveys and precipitated exponential growth in the size of these datasets [Bru02]. As an example, in 1983 the InfraRed Astronomical Satellite (IRAS) captured nearly the entire infrared sky in four wavelengths to yield less than 1 GB of imagery. Contrast this with more recent surveys such as the Digitized Palomar Observatory Sky Survey (DPOSS) and Sloan Digital Sky Survey (SDSS) in the visible wavelengths and the Two Micron All Sky Survey (2MASS) in the near-infrared wavelengths, which each host data at the multi-terabyte level, a 4 orders of magnitude leap in data size in little over a decade. Proposed missions like the Large Synoptic Survey Telescope (LSST) will continue this progression to the petabyte level.

The international astronomy community widely recognized that in the 1990's a significant and ever growing gap had emerged between the size and complexity of the datasets being captured and our ability to effectively extract the wealth of information inherent in them, in order to maximize the science impact of our space missions. This is an information technology challenge that is exasperated by the fact that the data are distributed and served via disparate mechanisms, and also a sociological challenge in that community standards are needed for data, query and search mechanisms, and computational services. A number of "Virtual Observatory (VO)" projects are addressing this problem, including the National Virtual Observatory (NVO) in the USA [Djo02], AstroGrid in the UK, and others elsewhere in Europe and Asia. The International Virtual Observatory Alliance (IVOA) is an effort to coordinate the various VO projects and encourage cooperation on issues that they have in common, such as data formats and interfaces between services.

The Flexible Image Transport System (FITS) has long been a lingua franca among the international astronomy community for sharing astronomical imagery. The FITS format encapsulates the image data with a header containing keyword-value pairs that describe the image and how pixels map to the sky. This meta-data information specifies image dimensions, pixel sampling, location on the sky of a specific reference pixel in the image, coordinate system, projection, arbitrary rotations, and others.

A common operation in astronomy is to extract a catalog of celestial objects from an image. An XML format for catalogs called VOTable has been proposed by the NVO, supporting hierarchical representation of metadata using custom tags [Wil02]. One effort to standardize the form of these tags is the Unified Content Descriptor (UCD), a proposal by the IVOA to define a formal vocabulary to express unambiguously the semantics of astronomical concepts in a human- and machine-readable form [Der03].

The NVO has also defined a set of standards for data providers that define how images and catalogs may be searched and served to the community. The Cone Search protocol is a simple mechanism that enables catalog searches based on a simple center location on the sky and a radius. A VOTable is returned containing the objects that fall within the specified bounds. The Simple Image Access Prototype (SIAP) specification is an extension of this search capability for images instead of catalog objects.

One key theme in the virtual observatory is the idea that new technology and standards will enable easy access to multiple datasets that can be used jointly for multi-wavelength science. The science drivers for this include: (i) search for and study of brown dwarfs, which are faint in the visible wavelengths, but may show up in the infrared; (ii) study of pulsars, which radiate strongly, but with dramatically different characteristics, at a wide range of wavelengths from radio through gamma ray; (iii) study of quasars, which may be distinguished from stars by their unique spectral signature across multiple wavelengths; (iv) search for new objects that are so faint as to be indistinguishable from noise in a single wavelength, but may be identified by correlation across wavelengths; (v) identifying similar classes of objects through clustering in multi-wavelength space; and (vi) discovery of entirely new types of objects with unusual spectra by examination of those outlying objects that do not cluster well in this multi-wavelength space [Djo01].

The technical challenges to be overcome in data federation are in how to relate objects in one archive to objects in another, how to handle the situation where objects appear in one archive but not in another, and how to provide uniform access to distributed data archives hosted by different organizations. This entails providing a data federation layer that hides the underlying search and access mechanisms of the multiple distributed data archives. Furthermore, data federation enables queries and searches that span multiple archives.

Since instruments can provide either high spatial resolution or wide area, but not both, an all sky survey typically archives many thousands or even millions of images, each with a different projection corresponding to the pointing of the telescope. Image mosaics are essential in order to enable the study of objects that span multiple images or for study of star formation regions or large-scale structure of the universe. Construction of an image mosaic entails reprojection of the input images to the output coordinate frame, matching the background intensities across the images, and combining the images to produce a single output mosaic. Several image mosaicking projects exist, including JPL's yourSky custom mosaic web portal [Jac02a], a follow-on project at Caltech and JPL called Montage for flux preserving mosaics [Berr02], and SWarp from the French TERAPIX center [Bert03].

The virtual observatory community has widely recognized that grid computing is a natural fit for astronomy because the data, compute resources, and domain expertise are all distributed. The Grist project led by Caltech is architecting a framework for interoperable services for astronomy compliant with NVO, grid, and web services standards. A number of services will be deployed in support of astronomical data mining, including services for data access, mosaicking, extracting source catalogs from images, clustering, catalog manipulation, statistics, and visualization. An interactive workflow system will allow a scientist to use a visual programming interface to link together these distributed services as needed and control service deployment and execution from a desktop computer.

Computational grids are being used for image mosaicking, which is both a compute- and data-intensive operation. The yourSky portal has been extended into yourSkyG, a mosaic service on the Information Power Grid (IPG), NASA's computational grid infrastructure. In addition, the Montage mosaic software has been deployed as a service on the TeraGrid, NSF's 20-teraflop computational grid [Berr04].

A number of standard visualization tools are widely used and suitable for modest size data analysis, including SAOImage DS9 [Joye03], and OASIS [Good03]. For larger datasets, two applications developed at JPL enable large-scale visualization of astronomical images and catalogs using high-end or low-end resources. The Electronic Light Table (ELT) provides high-performance visualization driven by supercomputers connected to multi-screen "Powerwall" displays [Jac00]. Visualization of selected datasets on standard desktop computers is provided by a web-based system called SkyLite [Jac02b]. Key features of these visualization tools include interactive, high-performance pan and zoom

on multi-wavelength and multi-resolution datasets that are larger than typical memory sizes, and the inter-relation of numerous catalog and image layers to each other.

UNDERLYING THEMES AND RECOMMENDATIONS

We can see from the discussions above that, no matter what the science domain or the subject matter, there are common themes that run through the data systems. An analysis of these undercurrents may give us some clues as to how currently separate efforts may be able to share development resources and leverage common solutions. We have seen the benefit of standards *within* a single discipline; we have yet to realize the benefit of standards *across* disciplines. If we consider these from an information technology perspective only, noting the areas where discipline differences may warrant special cases, we may be able to bring about a greater level of cooperation, and therefore, *a higher level of progress*, than before.

Data Access and Utilization

All of the archive systems need to provide better access to data given the increasing volumes that are now on hand to sort through. Faster searches, narrower selections, and speedier transmission of resulting information are on everyone's list of improvements to be made. The relatively "raw" data stored in the archives must be processed to additional constraints (subsetting, re-gridding, re-projecting, etc.) that cannot be a priori specified, as the constraints depend on the particular investigation at hand. However, the notion of the scientific workflow system to execute this string of processing constraints on combinations of data sets and parameters is common throughout. While the individual operators may have to be data set specific, or at least data format specific, the environment for running them can be generic. Web services and grid services facilitate remote computing and distributed access to data. These services are all part of the new efforts underway, and the use of emerging standards for interfacing to these services can lead to far greater cooperation than before. Joint registries of specific grid services that appeal across science domain boundaries would facilitate publishing and discovery of these common services.

The barriers to this level of cooperation are more social than technical in that the difficulty lies in the communication between what tend to be isolated development teams. Cross team discussions are often difficult, but more and more of that is happening at conferences and meetings such as this one. The more this takes place the more leveraging of work can occur. The emergence of the grid-computing paradigm, originally for high-energy physics, has fostered this level of interchange, and discussions of grid computing is catching on and spreading to all the science domains. While not all applications warrant using a high-performance computing grid, the adherence to the grid data standards facilitates its use where it is advantageous.

Intelligent Archives and Data Warehouses

Another paradigm emerging in the Earth science realm is that of the "intelligent archive" [Ramapryian et al]. In this paradigm, the access and utilization tools reside *within* the data archive system and are brought into execution as needed to satisfy a processing request on demand. The archive's "intelligence" is in determining what needs to be invoked without being specifically told. The scientist specifies what he needs; the intelligent archive figures out how to give it to him. The GENESIS system, when resident at a DAAC, is a flavor of this concept. Format translations, subsetting algorithms, higher-level parameter processing, data mining algorithms, all run on the archived data to produce more meaningful data for the scientist, are all examples of this approach. A related concept in astronomy is the "virtual data" idea being explored in projects such as Grist, which refers to intelligent caching mechanisms in a workflow that enable data products to be drawn from a pre-computed collection when available, and dynamically computed only when necessary.

These "on-the-fly" transformations are in contrast to stored "warehouses" of pre-processed data. This concept also has its place in the scientific world. Large stores of data are served out in the geo-spatial world by Web Mapping Servers (WMS) to promote interoperability and layering of widely dispersed data sets. Indeed, the entire field of Geographic Information Systems (GIS) has been revolutionized by the availability of these standard data sets. Raster images from remote sensing instruments and vector data of known geologic features can be overlaid to produce geologic maps.

Earth scientists have long known the value of these standards; the planetary science community is recently realizing its benefits. The advent of planetary rovers has increased the importance of this technology and the use of GIS for terrestrial planets will greatly enhance these missions. In astronomy, relating multiple wavelength images to each other and to catalog objects is a key theme in a number of data analysis and visualization systems, analogous to GIS in the Earth and planetary domains.

Virtual Observatories and Data Federations

Although the term "virtual observatory" originated in the astronomical community, the concept is widespread in the Earth and planetary realms. Put simply, a virtual observatory (VO) is the amassing of data collected and archived by individual groups to yield new results in combination with other such collections. This "federation" of data thus becomes a virtual observatory for the scientist. Unions of data collections span organizations and countries; the secret to interoperability among them lies in the agreed upon standards used to represent the data and used to interface the data to common computing services. Thus there is arising a federation of earthquake data in the Solid Earth Virtual Observatory (SERVO) based on the astronomy example. Virtual Mars, a similar concept in the Planetary Data System of integrating all of the known data about Mars into standard representation, is another application.

Federated Earth science data systems are perhaps the most mature of the data federations. The concept was first prototyped with the Earth System Information Partners, a union of the DAACs and competitively selected Earth science applications to promulgate the data collected by the NASA Earth science missions and render them more amenable to commercial use. The federation still exists on its own, and NASA has recently added a new round of information partners selected in the REASoN awards to further promote research with and application of Earth science data.

CONCLUSION

This paper has examined three NASA science data archive systems, from the Earth, planetary, and astronomy domains respectively, from the point of view of transforming data into knowledge. We have discussed some of the various efforts underway to provide the science communities with not only better access to the data holdings, but also to provide them with the services and tools they need to interpret the data and better understand their physical meaning. We have identified selected themes common to all three domains and suggested ways that common standards, technologies, and even implementations could be leveraged to benefit each. It remains for us as developers of these systems to ensure that we work together to promote this cooperation.

ACKNOWLEDGEMENT

Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

- [Berr02] G. B. Berriman, D. Curkendall, J. Good, J. Jacob, D. S. Katz, T. Prince, and R. Williams, *Montage: An On-Demand Image Mosaic Service for the NVO*, Astronomical Data Analysis Software and Systems (ADASS) XII, October 2002, Astronomical Society of the Pacific Conference Series, eds. H. Payne, R. Jedrzejewski, and R. Hook.
- [Berr04] G. B. Berriman, E. Deelman, J. Good, J. Jacob, D. S. Katz, C. Kesselman, A. Laity, T. A. Prince, G. Singh, and M. H. Su, *Montage: A Grid Enabled Engine for Delivering Custom Science-Grade Mosaics on Demand*, Proceedings of SPIE Vol. 5493, SPIE Astronomical Telescopes and Instrumentation: Optimizing Scientific Return for Astronomy Through Information Technologies Conference, June 2004.
- [Bert03] E. Bertin, *SWarp v2.0 User's Guide*, 2003.
- [Bru02] R. J. Brunner, S. G. Djorgovski, T. A. Prince, and A. S. Szalay, Massive Datasets in Astronomy, in Handbook of Massive Data Sets, Kluwer Academic Publishers, 2002, pp. 931-979.
- [Der03] S. Derriere, F. Ochsenbein, T. Boch, and G. T. Rixon, *Metadata for the VO: The Case of UCDS*, Astronomical Data Analysis Software and Systems XII, 2003.
- [Djo01] S. G. Djorgovski, A. Mahabal, R. Brunner, R. Williams, R. Granat, D. Curkendall, J. Jacob, and P. Stolorz, *Exploration of Parameter Spaces in a Virtual Observatory*, Proc. SPIE Vol. 4477, p. 43-52, Astronomical Data Analysis, Eds. Jean-Luc Starck and Fionn D. Murtagh, November 2001.
- [Djo02] S. G. Djorgovski, et al., National Virtual Observatory Science Definition Team Report: *Towards the National Virtual Observatory*, April 2002.
- [Good03] J. C. Good, M. Kong, and G. B. Berriman, *OASIS: A Data Fusion System Optimized for Access to Distributed Archives*, Astronomical Analysis Software and Systems XII, ASP Conference Series, Vol. 295, eds. H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, 2003.
- [Jac00] J. C. Jacob and L. E. Husman, *Large Scale Visualization of Digital Sky Surveys*, Virtual Observatories of the Future Conference, June 2000, Astronomical Society of the Pacific Conference Series, Volume 225, 2001, eds. R. J. Brunner, S. G. Djorgovski and A. S. Szalay, pp. 291-296.
- [Jac02a] J. Jacob, R. Brunner, D. Curkendall, G. Djorgovski, J. Good, L. Husman, G. Kremenek, and A. Mahabal, *yourSky: Rapid Desktop Access to Custom Sky Image Mosaics*, SPIE Astronomical Telescopes and Instrumentation: Virtual Observatories Conference, August 2002.
- [Jac02b] J. C. Jacob, G. Block, and D. W. Curkendall, *Architecture for All-Sky Browsing of Astronomical Datasets*, Astronomical Data Analysis Software and Systems (ADASS) XII, October 2002, Astronomical Society of the Pacific Conference Series, eds. H. Payne, R. Jedrzejewski, and R. Hook.
- [Joye03] W. A. Joye and E. Mandel, *New Features of SAOImage DS9*, Astronomical Analysis Software and Systems XII, ASP Conference Series, Vol. 295, eds. H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, 2003.
- [Ram02] H. K. Ramapriyan, S. Kempler, C. Lynnes, G. McConaughy, K. McDonald, R. Kiang, S. Calvo, R. L. Roelofs, D. Sun, *Conceptual Study of Intelligent Data Archives of the Future*, 10th NASA Goddard Conference on Mass Storage Systems and Technologies and 19th IEEE Symposium on Mass Storage Systems, April 2002.
- [Wil02] R. Williams, F. Ochsenbein, C. Davenhall, D. Durand, P. Fernique, D. Giaretta, R. Hanisch, T. McGlynn, A. Szalay, and A. Wicenec, *VOTable: A Proposed XML Format for Astronomical Tables*, 2002.
- [Yun04] T. Yunck, B. Wilson, A. Braverman, E. Dobinson and E. Fetzer, *GENESIS: The General Earth Science Investigation Suite*, NASA Earth Science Technology Conference, July 2002.

End of File

