Astronomical Data Analysis Software and Systems IX
ASP Conference Series, Vol. 216, 2000
N. Manset, C. Veillet, and D. Crabtree, eds.

Distributed Data Systems and Services for Astronomy and the Space Sciences

R. J. Hanisch

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218 e-mail: hanisch@stsci.edu

Abstract. Scientific advances do not necessarily follow strict research discipline boundaries. In the area of astronomy and space science, data from multiple missions and observatories operating in various parts of electromagnetic spectrum are necessary in order to answer fundamental scientific questions. However, a researcher attempting to locate, understand, and use data from a variety of sources now faces serious difficulties. Many datasets are available on the Internet but finding the ones of relevance, especially outside of one's immediate field of expertise, is difficult. The metadata used to annotate datasets in different fields can be unfamiliar and obscure, even though the same basic data attributes are being described. Once data has been located, it is often in different and incompatible formats.

We are now developing a distributed information service for the space sciences—ISAIA—which covers many different subdisciplines with datasets of common interest and which will improve the researcher's abilities to locate and use data from a wide variety of on-line resources. This service builds upon experience in implementing a data location service for astronomy, Astrobrowse. The key to implementing such services is the concept of *profiles.* Profiles are a generalization of data dictionaries: they define the metadata labels and content for information resources and provide mappings of those labels onto site or service-specific terms and query protocols. In ISAIA profiles may be hierarchical—general at the highest levels, with sub-profiles for certain disciplines or types of instrumentation.

1. What Are Distributed Data Systems?

Many of the data systems and services that we use today in astronomy are *distributed systems.* The key properties that distinguish distributed systems or services are that the databases, data, archives, and documents being made available to the user need not be located at a single physical site; user queries are passed to, and responses are returned from, multiple underlying services via a common protocol; and responses or results from the underlying services are presented to the user in an integrated fashion, as if the resources were local and of similar structure. Indeed, many of the WWW facilities we use every day are distributed services to some extent, with the popular WWW search engines being good examples. In such systems there is a centralized index of documents, but *only* the index is centralized; the actual documents remain on their host systems. The reliability of such systems depends critically on maintaining up-to-date indexes.

1.1. Why Do We Need Them in Astronomy?

Distributed data systems are important in astronomy for a number of reasons. First, no one site can hold all information (e.g., telescope image archives are already in the multi-TB range, and promise to grow larger quickly with the increasing size of digital detectors and the advent of new all-sky surveys such as GSC-II (McLean et al. 1998), the Sloan Digital Sky Survey (cf. Szalay 1998), and 2MASS (cf. Skrutskie et al. 1997). Second, much of the information is dynamic. Static catalogs and indexes quickly become obsolete. Third, astronomers use multiple types of data: images, spectra, time series; catalogs, journal tables; and journal articles. All should be easily located and easily accessed with query terms and syntax natural to the discipline. Fourth, astronomers need to know the provenance of the data they are using, and the data must be managed and supported with a thorough knowledge of the nature of the instrumentation used to collect it. No one data center or service is able to have expertise in the wide range of astronomical instrumentation and data sets. And finally, single entry points for multiple resources aids users in locating and using data quickly, allowing them to focus on the science rather than on the search for information.

1.2. Existing Systems and Services

Distributed data systems provide several types of services levels for users:

Information Discovery. Several services are primarily data or information location tools. The Astronomical Software and Documentation Service, for example, allows one to locate software for specific applications or telescope and instrumentation manuals (Payne, Hanisch, & Warnock 1996). Astrobrowse (Heikkila, McGlynn, & White 1999; McGlynn, Scollick, & White 1998) provides a single interface for locating data of interest in over 1000 on-line, WWW-based resources.

Information Retrieval. The catalog services of the Astronomical Data Center (NASA/GSFC, Roman 1996) and the Centre de Données astronomiques de Strasbourg (CDS, VizieR; Genova et al. 1998) and on-line archives (MAST; Imhoff et al. 1999), HEASARC (cf. Angelini et al. 1999), and IPAC/IRSA are primarily data delivery facilities. The ADC External Query (AEQ) facility is notable as it serves as a component of other distributed systems (such as the catalog cross-correlation tools at MAST).

Information Integration. The more sophisticated and versatile distributed services are those that support interactive integration and intercomparison of data from multiple sources (catalog cross-correlation, image and graphical overlays, and intelligent query and response management). These include IMPReSS (Shaya et al. 1997) and AMASE (Cheung et al. 1999) from the ADC, SkyView (HEASARC; McGlynn, Scollick, & White 1998), SkyCat (ESO, ST-ECF, CADC; Albrecht et al. 1997), and Aladin (CDS; Bonnarel et al. 1999). These services are increasingly providing facilities to interlink with the astronomical literature (on-line journals) and the bibliographic reference systems (SIMBAD, NED, ADS) (Boyce 1998).

2. Distributed Information Services for Space Science

Even though there are many excellent sources of space science data available on the web, there are substantial barriers to the interoperability of these data:

- Some important data sets, especially those at high time resolution, remain unavailable in electronic form;
- Data are in incompatible formats;
- Data frequently have inconsistent or incomplete metadata, and each discipline tends to have its own metadata vocabulary;
- Good data may be available, but it can be difficult to find for a given object, class of objects, or type of phenomenon.

It is important to try to remove, or at least reduce, these barriers in order to realize the maximum scientific return from these data and to enable new, cross-cutting research using data from traditionally independent disciplines. For example, research on comets requires data from all space science disciplines: ground-based and space-based imaging and spectroscopy, in situ measurements of the solar wind (wind speed, density, and magnetic field strength), and in situ measurements from space probes targeted at comets (Giotto, the International Cometary Explorer, future missions) in order to build a complete model. Similarly, investigation of auroral phenomena around other planets requires information about low-frequency radio emissions, data from UV imaging and spectroscopy, planetary magnetosphere data, and solar wind data.

On the other hand, expertise in data is frequently associated with a specific wavelength region or phenomenon, as the data collection apparatus, calibration processes, and analysis procedures are connected to the physics of the measurement itself. Any information service that provides integrated access to data of diverse origins must also provide links to the documentation needed to understand and use the data properly and to the scientists having direct experience working with the data.

2.1. NASA's Space Science Data Management Structure

In considering the development of a distributed data system for providing access to space science data it is helpful to understand the structure of NASA's Space Science enterprise (Figure 1). Space Science is divided into four science themes: *Planetary Exploration* (covering space missions to other bodies in the solar system), *Search for Origins* (covering primarily UV, Optical, and IR astronomy missions), *Structure and Evolution of the Universe* (primarily high energy astrophysics missions), and *Sun-Earth Connection* (solar physics, solar wind, the earth's ionosphere, thermosphere, and mesosphere, cosmic rays, etc.). The four themes cover three traditionally separate disciplines: solar system exploration, astronomy and astrophysics, and space physics. The three disciplines have managed their data systems differently. The planetary science community as a hierarchical data system called PDS which maintains and distributes data from NASA planetary missions. PDS is composed of a central node and a number of topical nodes, each of which takes responsibility for data in a different area of expertise. The Space Physics Data System—SPDS—is a collection of on-line resources from numerous space physics missions. The National Space Science Data Center—NSSDC—is also a major provider of on-line data sets and

Space Science Data Systems

Planetary Data System	**Astrophysics**	**SEC Data System**
PDS Management Council	Astrophysics Data Centers Coordinating Council	SECDS Coordinating Council
Central Node (JPL) Discipline Nodes: • Atmospheres (NMSU) • Geosciences (WU) • Imaging (JPL) • Navigation & Ancillary Info (JPL) • Planetary Plasma Interactions (UCLA) • Rings (ARC) • Small Bodies (UMd)	Datatype-Based Data Centers: • IRSA (IPAC) – Infrared, Interferometry • MAST (STScI) – EUV, UV, Opt, Near-IR • HEASARC (GSFC) – EUV, X-ray, γ-ray • ADC (GSFC) – Catalogs • ADS (SAO) – Abstracts • Chandra (SAO)	Management Office Service Groups: • Solar Physics • Terrestrial Environment Imagery • In Situ Space Physics Data Providers

Planetary Exploration	**Search for Origins**	**Structure and Evolution of the Universe**	**Sun-Earth Connection**

Science Themes

Figure 1. The relationship between NASA's Office of Space Science data systems and the OSS science themes.

information services for space physics. Solar physics data can be found at the Solar Data Analysis Center (SDAC) at Goddard Space Flight Center and at sites referenced there. The NASA astronomy community has recently organized its data services under an ad hoc Astrophysics Data Centers Coordinating Council. It is the primary goal of NASA's Space Science Data System (SSDS) initiative to interlink and provide interoperability among these various data systems and services.

2.2. Models for a Space Science Information System

Three different models for integrated access to distributed information sources can be characterized as "good," "better," and "best." These are distinguished by how much work the end-user has to do to find and use data of interest. These are shown schematically in Figure 2. The goal is to provide a "best" system for integrated access to space science data.

The World Wide Web: "Good." The situation that generally exists in space science today is certainly "good." As mentioned above, there are literally thousands of on-line resources available. Discipline-oriented document and resource collections such as AstroWeb and SPDS allow researchers to browse these services, though the researcher may be faced with searching many tens of sites in order to find data of interest. An example from astronomy highlights the

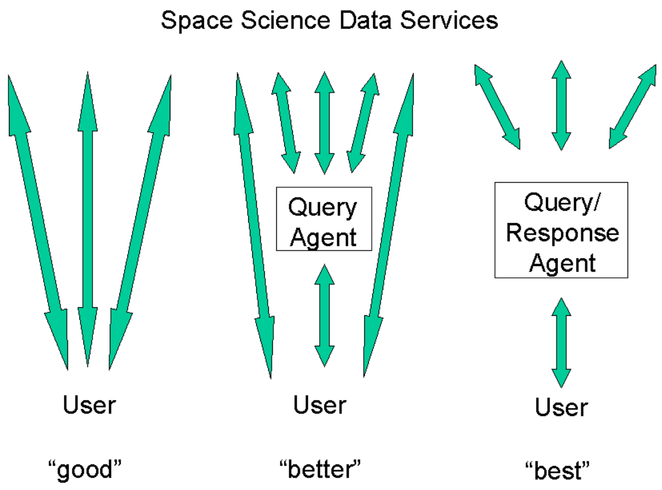

Figure 2. Three models for integrated access to distributed information resources. *Left:* The user interacts with each information resource individually, e.g., discipline-specific collections of WWW resources (AstroWeb, SPDS, etc.). *Middle:* The user sends a single query for information to the query agent. The agent knows what resources are available and what kind of information they contain, and forwards the query only to the appropriate resources. The user then interacts with the resources one by one. *Right:* A query/response agent both knows what resources are available and how responses from those resources are formatted. It is possible to reformat the responses into a uniform presentation for the user.

shortcomings of this situation: suppose someone wants to find images in many parts of the spectrum (from radio to x-ray) of an object they have just observed with the Hubble Space Telescope in order to make comparisons between features in different bandpasses. In general this researcher would first have to be familiar with all of the sites that might contain such data, and then go to them one by one to submit a query for the object of interest. The exact form of these queries is likely to differ from site to site and service to service, even though essentially the same type of information is being specified.

Query Agents: "Better." A query agent is a web service that knows about a class of information resources and their general attributes, i.e., what type of information is available from each resource. For example, in astronomy a query agent would know what types of data (images, spectra, time series, catalogs) in what spectral bandpasses (x-ray, ultraviolet, optical, infrared, radio) are available in each service. The user interested in ultraviolet spectra of certain objects would have queries directed to just those sites holding such data. The query agent passes the user's query along to each site, formatting it as required. The user then interacts further, site by site, to refine the query and request specific data sets. Astrobrowse is such a service.

Query/Response Agents: "Best." A query/response agent is a web service that provides all of the services of a query agent, but in addition is able to collect the responses from the information services and integrate them together for presentation to the user. This integration may require conversion of metadata from one set of units to another and other reformatting. The query/response agent presents the user with metadata from and links to the data at various distributed services. In principle the query/response agent could also become a data agent, collecting data from various services, converting data to a user-specified format, and even performing such complex functions as spatial or spectral rebinning to facilitate comparison. However, even within the discipline of astronomy and astrophysics, data is taken in many different ways, using instruments of widely varying intrinsic resolution and sensitivity. As a result, automated conversions and comparisons are likely to be meaningless. It is critical for the user to understand the source and "pedigree" for the data in order to make proper use of it, and in order to know who to contact for further information about it. Thus, it is not clear that a general data agent function for space science is either practical or in the best interest of the user.

3. ISAIA

We have begun work on a successor to Astrobrowse called ISAIA (**I**nteroperable **S**ystems for **A**rchival **I**nformation **A**ccess). ISAIA will be a query/response agent, performing the tasks of Astrobrowse for locating data services, but also handling the responses from those data services. ISAIA will also be broader in scope than Astrobrowse, encompassing information services in the space sciences and complementary sources of data from ground-based facilities. Wherever possible ISAIA will be layered upon existing systems, using resource, query, and response profiles to build a common interface. ISAIA development will consist of four major phases: definition of the profiles, implementation of a generalized query agent, implementation of a query/response agent, and development of an integrator.

3.1. Profiles

To bridge the differences between the various information services and databases, we require a common language for expressing queries that can at some stage be translated into the specific languages used by each of the databases. For a complex system, it is helpful to consider this language in two parts: the *profile* and the *protocol*. The profile defines the concepts that can be expressed within queries and responses. The protocol defines how the concepts are represented and passed between the clients that pose the queries and the databases that answer them. This division is advantageous because the protocol is closely tied to the technology that is used to implement the system, while the profile is technology independent.

The primary component of a profile is the definition of the set of standard query attributes—the concepts that can be used to state selection criteria. In the space sciences these attributes will include concepts such as sky position or location in space, frequency or band-pass, data type, time of observation, mission or observatory, and other key identifiers.

The profile also defines the operations that a user may invoke on these criteria, such as "equals" and "greater than." It is sometimes appropriate to define a controlled vocabulary associated with an attribute; for example, recognized data types might include "image," "catalog entry," "reference," "spectrum," etc. The profile can also define boolean operators and extended options that allow one to express complex queries.

There are three profile components: *resource profiles*, *query profiles*, and *response profiles. Resource profiles* characterize the data holdings in a given service. The primary purpose of the resource profile is to allow the query agent to determine which sites and services to send queries to. The resource profile functions as a filter and avoids having queries sent across the network to possibly hundreds or thousands of services which do not have data of interest to the user. *Query profiles* provide a detailed mapping of generic query terms to site or service specific terms. For example, in astrophysics the celestial coordinates of right ascension and declination are fundamental for giving the position of an object on the sky. One service may label these coordinates as "RA" and "DEC" in its database, and another may label them "right_ascension" and "declination." They may also be expressed in different units in different databases (decimal degrees, radians, or character strings in sexagesimal notation: hh:mm:ss.s, dd:mm:ss.s) or may given given for different epochs. In space physics data, time is often a key search criterion. Different resources specify observation times in different systems, units, and notations. *Response profiles* label the metadata that comes with a query response in order to facilitate integration of the results from different services. Response profiles are also used to define the type of information being returned (plain text, HTML, or other structured records).

Although the term definitions in each of these profiles are likely to be the same or similar, it is useful to make the distinction between these from the perspective of implementation. For example, information service providers could choose to support only the resource profile and the query profile. This would make their data services visible and queryable in an integrated system, but would not allow their query responses to be handled automatically. Further information on profiles is given by Plante et al. (1997) and Plante (1997).

3.2. The Integrator

The goal of the ISAIA program is to build systems which can integrate results from queries against many different resources and to allow remote resources to appear to be local to a given user. ISAIA's data integrator must regularize the information returned in response to queries, ensuring consistency in units, projections, formats, etc. The integrator must maintain links so that the user can understand the provenance of the data.

The role of the ISAIA integrator is to enable use of results from multiple sources effectively. Using information in the profiles with the returned results, the integrator must be able to:

- Convert units, e.g., to convert from wavelength to frequency;
- Translate coordinate systems, e.g., convert from astronomical Galactic to Equatorial coordinates or from heliocentric to geocentric positions;
- Regularize formats. Multiple tables often display the same type of metadata in different formats—possibly in several different fields. An example is time, which may be broken into day and time of day in one table but provided in a single field in another;

- Reorganize tables. Even when tables provide the same information, comparisons can be difficult if the data are sorted in different order or if the columns to be compared are not conveniently highlighted;
- Store intermediate results. Scientists must be able to interrupt and return to sessions without having to reenter the queries that established their current state;
- Cross-correlate tables. In many cases the role of the integrator is to hide the existence of multiple tables from the user. Users can ask questions of the form "what targets have been observed by both X and Y?" or "what targets have been observed by X and not by Y?"
- Transparently access archival data. While the integrator itself works primarily with the tables and metadata that index archives, it must also provide users with convenient access to archive information which corresponds to the rows that users select. Users will be able to see at least minimal description of the kinds of archive information available. Where they exist, quick look data products such as GIFs or other kinds of preview images, will be highlighted;
- Give explicit credit to the data provider and maintain data integrity. The system must indicate when the metadata are reprocessed by the integrator and must allow users to access data directly from the original provider.

The integrator is the controller that takes user's requests, queries remote sites, translates the results and feeds a local database. It sends outputs to the user interface for display to the user. The integrator can also take data from the local database and invert the process: providing ISAIA-compatible records.

Initially the integrator provides the user interface with descriptions of the options available which the integrator determines by examining its database of space science resources. After the user submits the request and the integrator has marshaled the results, the integrator provides the user interface with descriptions of the results as well as the regularized metadata. The integrator incorporates a parser which analyzes the profile metadata and determines the steps needed to transform the incoming data streams into the formats desired by the user. The integrator uses an internal database system to manage state, to perform cross-correlations among results obtained from multiple sources, and to provide storage for intermediate results.

Acknowledgments. This paper draws upon the experience and knowledge of all members of the ISAIA project team: T. McGlynn, C. Heikkila, and N. White (NASA/GSFC, HEASARC), J. King (NASA/GSFC, NSSDC), R. A. White, C. Cheung, and E. Shaya (NASA/GSFC, ADC), R. Plante and R. McGrath (NCSA/UIUC), J. Mazzarella (IPAC/Caltech), A. Rots (SAO), S. McMahon and S. Hughes (JPL/PDS), M. A'Hearn (U. Md.), R. Beebe (NMSU), F. Genova (CDS), and P. Giommi (BSDC). The work on ISAIA profiles has benefited greatly from discussions with colleagues at the Observatoire Astronomique de Strasbourg, Centre de Données astronomiques de Strasbourg (CDS) during my visit there in the summer of 1999: F. Ochsenbein, M. Wenger, P. Fernique, P. Dubois, and F. Genova. I am most appreciative of the support of the CDS and OAS during this period, and thank CDS director F. Genova and OAS director D. Egret for their hospitality.

The ISAIA project is supported by NASA's Applied Information Systems Research Program under grants to the Space Telescope Science Institute, the National Center for Supercomputing Applications/University of Illinois, and the Goddard Space Flight Center.

URLs

ADC	`http://adc.gsfc.nasa.gov/`
AEQ	`http://tarantella.gsfc.nasa.gov/viewer/AEQdoc.html`
Aladin	`http://aladin.u-strasbg.fr/aladin.gml`
AMASE	`http://amase.gsfc.nasa.gov/`
ASDS	`http://asds.stsci.edu/`
Astrobrowse	`http://heasarc.gsfc.nasa.gov/ab/`
CDS	`http://cdsweb.u-strasbg.fr/CDS.html`
HEASARC	`http://heasarc.gsfc.nasa.gov/`
IMPReSS	`http://tarantella.gsfc.nasa.gov/impress/`
ISAIA	`http://heasarc.gsfc.nasa.gov/isaia/`
IRSA	`http://irsa.ipac.caltech.edu/`
MAST	`http://archive.stsci.edu/mast.html`
NSSDC	`http://nssdc.gsfc.nasa.gov/`
PDS	`http://pds.jpl.nasa.gov/`
SDAC	`http://umbra.nascom.nasa.gov/sdac.html`
SkyCat	`http://archive.eso.org/skycat/`
SkyView	`http://skyview.gsfc.nasa.gov/`
SPDS	`http://spds.gsfc.nasa.gov/`
SSDS	`http://ssds.nasa.gov/`
Starcast	`http://archive.stsci.edu/starcast/`

References

Albrecht, M. A., Brighton, A., Herlin, T., Biereichel, P., & Durand, D. 1997, in ASP Conf. Ser., Vol. 125, Astronomical Data Analysis Software and Systems VI, ed. G. Hunt & H. E. Payne (San Francisco: ASP), 333

Angelini, L., Breedon, L., Garcia, L., Hilton, G., Stollberg, M., & White, N. 1999, AAS Meeting, 194, #83.01

Bonnarel, F. et al. 1999, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. David M. Mehringer, Raymond L. Plante, & Douglas A. Roberts (San Francisco: ASP), 229

Boyce, P. B. 1998, in ASP Conf. Ser. 153, Library and Information Services in Astronomy III, ed. U. Grothkopf, H. Andernach, S. Stevens-Rayburn, & M. Gomez, 107

Cheung, C. Y., Roussopoulos, N., Kelley, S., & Blackwell, J. 1999, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. David M. Mehringer, Raymond L. Plante, & Douglas A. Roberts (San Francisco: ASP), 213

Genova, F., Bartlett, J. G., Bonnarel, F., Dubois, P., Egret, D., Fernique, P., Jasniewicz, G., Lesteven, S., Ochsenbein, F., & Wenger, M. 1998, in ASP

Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 470

Heikkila, C. W., McGlynn, T. A., & White, N. E. 1999, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. David M. Mehringer, Raymond L. Plante, & Douglas A. Roberts (San Francisco: ASP), 221

Imhoff, C., Abney, F., Christian, D., Donahue, M., Hanisch, R., Kimball, T., Levay, K., Padovani, P., Postman, M., Smith, M., & Thompson, R. 1999, AAS Meeting, 194, #83.02

McGlynn, T., Scollick, K., & White, N. 1998, in IAU Symp. 179, New Horizons from Multi-Wavelength Sky Surveys, ed. B. J. McLean, D. A. Golombek, J. J. E. Hayes, & H. E. Payne (Dordrecht: Kluwer), 465

McLean, B., Hawkins, C., Spagna, A., Lattanzi, M., Lasker, B., Jenkner, H., & White, R. 1998, in IAU Symp. 179, New Horizons from Multi-Wavelength Sky Surveys, ed. B. J. McLean, D. A. Golombek, J. J. E. Hayes, & H. E. Payne (Dordrecht: Kluwer), 431

Payne, H. E., Hanisch, R. J., & Warnock, A. 1996, in ASP Conf. Ser., Vol. 101, Astronomical Data Analysis Software and Systems V, ed. G. H. Jacoby & J. Barnes (San Francisco: ASP), 577

Plante, R. L. 1997, `http://monet.astro.uiuc.edu/~rplante/topics/P30/profile.html`

Plante, R. L., McGrath, R. E., & Futrelle, J. 1997, `http://monet.astro.uiuc.edu/~rplante/topics/P30/sysmodel.html`

Roman, N. G. 1996, AAS Meeting, 188, #54.20

Shaya, E., Kargatis, V., Borne, K., & White, R. A. 1997, AAS Meeting, 191, #57.05

Skrutskie, M. F., Schneider, S. E., Stiening, R., Strom, S. E., Weinberg, M. D., Beichman, C., Chester, T., Cutri, R., Lonsdale, C., Elias, J., Elston, R., Capps, R., Carpenter, J., Huchra, J., Liebert, J., Monet, D., Price, S., & Seitzer, P. 1997, in The Impact of Large Scale Near-IR Sky Surveys, ed. F. Garzon et al. (Dordrecht: Kluwer), 25

Szalay, A. 1998, in ASSL No. 231, The Evolving Universe, (Dordrecht: Kluwer), 277