T1 - The Path toward Data System Integration

Raymond J. Walker, UCLA

There has been a persistent dream of a global data environment in which all of the Earth and space science data could be organized in a common way with "one stop shopping" for any data product. In this talk we will consider some practical principles for guiding data system integration. The highest priority in any science data development must be providing the scientific community with access to high quality and well documented data. That can only be accomplished with the active participation of scientists involved with the analysis of the data. This in turn requires that the data system be organized by scientific discipline. In the past decade the sources of scientific data have increased greatly as more spacecraft with more sophisticated instruments have been flown both in the United States and abroad. The advent of the internet has made widely distributed data sources both practical and desirable. Earth and space science data are now distributed around the world. The highest priority in organizing the world's scientific data holdings must now be to integrate these distributed data holdings within each discipline. We will use two examples to demonstrate how this can be accomplished- the International Planetary Data Alliance (IPDA) and the SPASE (Space Physics Archive Search and Extract) consortium. In both cases the key to data system integration is the adoption of common metadata standards. In planetary science the existing metadata standards developed by the Planetary Data System (PDS) are being adapted by IPDA members to serve the global community while the international SPASE consortium is developing world wide standards for space physics data. In planetary science working standards existed before IPDA so they are being extended to serve the emerging international data system. In space physics there are many different data systems already in existence so the metadata standard is being designed as an intermediate language (an "interlingua") to connect the ones already in existence. Further integration across disciplines must involve a careful balance between scientific need and cost.

T2 - Long-Term Preservation of Astronomical Research Results

Robert J. Hanisch, Space Telescope Science Institute

Astronomers are producing and analyzing data at ever more prodigious rates. NASA's Great Observatories, ground-based national observatories, and major survey projects have archive and data distribution systems in place to manage their standard data products, and these are now interlinked through the protocols and metadata standards agreed upon in the Virtual Observatory. However, the digital data associated with peer-reviewed publications is only rarely archived. Most often, astronomers publish graphical representations of their data but not the data themselves. Other astronomers cannot readily inspect the data to either confirm the interpretation presented in a paper or extend the analysis. Highly processed data sets reside on departmental servers and the personal computers of astronomers, and may or may not be available a few years hence. Descriptive metadata, adequate at best for archival collections and associated data discovery services, is often inaccurate or lost once data moves out of pipeline systems into scientists' hand-crafted software.

We are investigating ways to preserve and curate the digital data associated with peer-reviewed journals in astronomy. The technology and standards of the VO provide one component of the necessary technology. A variety of underlying systems can be used to physically host a data repository, and indeed this repository need not be centralized. The repository, however, must be managed and data must be documented through high quality, curated metadata. This curation effort can only partially be automated. Multiple access portals can be available: the original journal, the host data center, the Virtual Observatory, or any number of topically-oriented data services utilizing VO-aware access mechanisms.

I will also briefly discuss metadata management challenges encountered thus far in the implementation of the Virtual Observatory.

T3 - Government-University Collaboration in Long-Term Archiving of Scientific Data

Topic: An institutional framework for long-term stewardship of scientific data.

Authors: Robert S. Chen, Robert R. Downs, and W. Christopher Lenhardt, Center for International Earth Science Information Network (CIESIN), Columbia University

The nation's colleges and universities represent an institutional community with a long-standing commitment to the archiving, preservation, and dissemination of knowledge. Some academic institutions have been in continuous existence longer than the Federal government itself, and many have libraries and archives that have lasted longer than most Federal agencies. However, on more recent time scales, academic institutions have lagged in their ability to archive, preserve, and disseminate digital data. Collaboration between government and academic institutions therefore provides the opportunity to establish a sustainable long-term infrastructure for digital data preservation and services that is not subject to the vagaries of annual government budgets or policy decisions and that meets the rapidly evolving challenges of the digital age.

We report here on an ongoing experiment in collaboration between the NASA Socioeconomic Data and Applications Center (SEDAC), one of the Distributed Active Archive Centers of the Earth Observing System Data and Information System (EOSDIS), and Columbia University's libraries and its Earth Institute to establish a long-term archive (LTA) for SEDAC data. The SEDAC LTA is constituted as a cooperative archive managed by a LTA Board with representatives from SEDAC, the libraries, and the Earth Institute. Although housed at present in the active archive, it is being designed and developed to operate independently of the active archive and to integrate in the long term into the University's emerging digital archiving infrastructure. In particular, the LTA is working to develop and implement a set of policies and procedures for long-term data stewardship of SEDAC's interdisciplinary digital data and information resources and a corresponding technical infrastructure that utilizes available standards and best practices. A major objective is to ensure that the archive is sustainable over time, independent of finite projects or the duration of current funding support, and is useful to present and future scientific and applied users.

This experiment is of high interest to the academic library community because it provides a test case for understanding the likely requirements and potential resources needed for stewardship of the large and rapidly growing quantity and diversity of digital data that universities are generating and responsible for. It is also of interest to the University's scientific faculty and staff, who are looking for efficient and sustainable ways to deal with their burgeoning data and information holdings. Finally, we believe that government agencies concerned about long-term data stewardship would benefit from the institutional longevity, expertise, and resources that the academic community could bring to this issue.

T4 - Rule-based Preservation Systems

Reagan Moore, SDSC

Rule-based preservation systems automate the execution of preservation management policies and assessment of preservation repository trustworthiness.

As the size of scientific collections grow, the management of assertions about the authenticity and integrity of the collections becomes onerous. For scalable preservation systems, an essential requirement is the ability to automate the application of management policies, the validation of assertions about the properties of the preservation environment, and the application of recovery mechanisms when faults are detected. The integrated Rule-Oriented Data System (iRODS) (under development with funding from NARA and NSF) supports the characterization of management policies as rules controlling the execution of remote micro-services.

The iRODS data grid implements the mechanisms needed to make assertions about the authenticity and integrity of records (scientific data). The desired properties of a preservation environment are expressed as assertions that are implemented as management policies controlling the execution of preservation capabilities (ingest, access, archival storage). The iRODS system supports the mapping of preservation capabilities to sets of micro-services, where each micro-service is a set of operations performed at a remote storage location (archival storage). Management policies are mapped to a set of rules, where each rule controls the execution of a set of micro-services, or a set of rules and micro-services. Assertions are mapped to sets of persistent state information that are generated on application of the rules. Examples of preservation capabilities are provided in the NARA Electronic Record Archives capabilities list. Examples of assertions about preservation environments are provided in the RLG/NARA assessment criteria for trusted digital repositories. Mappings of both the ERA capability list and the RLG/NARA assessment criteria have been made to iRODS rules and micro-services.

Data grids implement the concept of infrastructure independence, the ability to manage the properties of the preservation environment independently of the choice of archival storage system or access mechanism. Infrastructure independence enables the migration of the preservation environment onto new technology, while maintaining the integrity and authenticity of the scientific data collections. iRODS implements as micro-services the actions executed by the OAIS archival functional entities, including extraction of metadata, replication of files, validation of checksums, migration of files, management of descriptive information, creation of AIPS, parsing of SIPs, and extraction of DIPs. iRODS also supports the creation of the rules needed to express the management policies that dictate the risk mitigation strategies against data loss, the authenticity requirements for provenance metadata, and the access controls and data transformations. IRODS is available as open source software at http://irods.sdsc.edu.

T5 - Archiving in the Data Environment of Heliophysics at NASA

Aaron Roberts, NASA GSFC and HQ

Topic:
This talk will present the NASA Heliophysics data policy that lays out a framework for the lifecycle of HP mission data; the viewpoint seems representative of the efforts of many communities, and demonstrates an approach to dealing with distributed archives linked by virtual observatories.

Abstract:
A modern data policy governing NASA's Heliophysics data environment is under development. We are evolving today's environment of existing services in order to take advantage new computer and internet technologies and at the same time respond to our evolving mission set and community research needs. A strong governing principle is that the HP data environment requires science participation in all levels of data management. We will extend the use of peer-review processes to assist in managing the elements of the environment. We will continue to insist that all data produced by the HP missions are open and are to be made available as soon as is practical. The environment will continue to be distributed and at the same time we are implementing data integration capabilities through the creation of discipline-based virtual observatories. In the case of the Virtual Solar Observatory, this architecture is already permitting the selective inclusion of essential data sets from non-NASA sources. Gurman's "Right Amount of Glue" sets the philosophy [J.B. Gurman: Fall 2002 AGU, SH52C-03] for the environment, a key component of which is a standard of behavior - share one's data with everyone. We are in the process of implementing Resident Archives and the processes to manage these archives which will hold and serve mission data after the active production of mission data terminates. NASA HQ is leading the implementation of this data policy which blends 'bottoms-up' implementation approaches with a 'top-down' vision for an integrated data environment.

By providing an end-to-end guide to the data lifecycle, the Data Policy should make archiving issues considerably easier. We are developing a general language for the description of our data (the "SPASE data model") that will provide much of what is needed for description of archival products. A "Mission Archive Plan" will be required of the missions to avoid, as much as possible, the common situation of having many loose ends at mission termination. The Resident Archives are intended to use the distributed nature of current archives to better serve data with continued expert support. The Data Policy emphasizes throughout the need for adequate documentation to assure the independent useablility that is required by users and thus by archives.

T6 - NASA Planetary Data System -Structure, Mission Interfaces and Distribution

Reta Beebe, New Mexico State University

Topic - The structure and operation of planetary archiving within the NASA Planetary Science Division

The Planetary Data System (PDS) is a distributed system of discipline and support nodes. The discipline nodes represent traditional areas of scientific investigation and provide assistance for both data providers and users while the support nodes provide expertise and ancillary information that is used across the system. The Program Manager, Program Scientist and Management Council, composed of representatives from all nodes, share management duties.

To satisfy the goals of the Planetary Science Division, NASA utilizes a range of missions of different complexity. Flagships, PI led missions, and missions in the Mars Program present a wide range of challenges that are amplified by NASA's policy to compete missions or instruments, generating a changing community and wide range of expertise among data providers. Combinations of NASA funded teams based at individual institutions, internationally supplied instruments and the ensuing variable funding and physical locations of the teams further complicate data pipeline planning and development. As this process has developed, it has become apparent that imbedding PDS personnel within a mission is an effective way to insure that an efficient pipeline is developed. Many of the PI led proposals involve a PDS based individual and other missions negotiate this involvement.

The PDS is faced with ongoing challenges. Within the anticipated funding levels the PDS must: interface with more than 20 missions at a time in various development phases; ingest data from increasingly complex instruments with rapidly expanding data volume; add non-mission data from individuals, laboratories and observatories; educate data providers; provide useful links to related national and international data; respond to an increasing spectrum of demands from a growing community of users; incorporate new storage/distribution technologies and adapt to new modes of data presentation. Although many of these challenges are planetary oriented, the last two, in the form of reliability of archive media and an appropriate format for science animations, are topics that involve many of the groups represented at this workshop.

T7 - **Evolving a Ten Year Old Data Archive**

Jeanne Behnke, NASA GSFC

The Earth Observing System (EOS) Data and Information System (EOSDIS) is a comprehensive distributed system designed to support NASA's Earth Science missions. Designed in the early 1990's, EOSDIS has been archiving, managing, and distributing Earth science data since 1994. Over the life of EOSDIS, an on-going process of technology updates and improvements in user access, distribution mechanisms, and archive management has attempted to keep the system current. However, data volumes have grown rapidly and the science community has gained experience and capability in processing and analyzing their data. The result is a growing desire to re-examine the current operations for gains and improvements in a variety of areas. In 2004, NASA Headquarters (HQ) chartered a study, called the Evolution of EOSDIS Elements Study, to look at how change could be made to the EOSDIS system. The charter established an independent Study Team to provide recommendations and offer guidance. This team consisted of members from academia and NASA research organizations with experience in using or managing data systems for Earth Sciences data. NASA HQ also convened a Technical Team to develop an approach and implementation plan that would put EOSDIS on an evolution path for the future. This Technical Team consisted of senior staff from the project managing the EOSDIS science system (i.e., the Earth Science Data and Information System (ESDIS) Project) and representatives from the various system elements performing the science data processing, archiving and distribution functions of EOSDIS. The result of the study team was to determine a vision for the evolution of the EOSDIS system and then develop an approach to achieve the vision. The objectives of this evolution of EOSDIS are to: increase end-to-end data system efficiency while decreasing operations costs, increase data interoperability and usability by the science research, application, and modeling communities, improve data access and processing, and ensure safe stewardship. Step 1 of the EOSDIS Evolution was proposed to NASA HQ in November 2005.

As of March 2007, EOSDIS has completed several of the objective proposed in Evolution and plans to be complete by the end of the year. In this presentation, we will describe the underlying goals of EOSDIS evolution, the study and analysis process used to facilitate that evolution, the resulting phased implementation approach with a focus on the initial step to be implemented in the immediate future, and the expected benefits.

T8 - An Overview of Selected ISO Standards Applicable to Digital Archives

Donald M. Sawyer   Lou Reich   John Garrett
NASA GSFC   CSC / NASA GSFC   Raytheon / NASA GSFC

Possible Topic: Emerging archival standards and technologies

The National Space Science Data Center (NSSDC) has been participating in the development of standards applicable to the role of digital archives within the space science community, and beyond, for the better part of 2 decades. Some of these standards have been, or are being, developed through processes that are essentially ad hoc, although somewhat international in scope, such as the Space Physics Archive, Search, and Exchange (SPASE) data model. Some standards have been, or are being, developed through codified local processes that are also international in scope, such as the Flexible Image Transport Standard (FITS). A third category, which is the main subject of this overview, includes those standards that have been, or are being, developed by a codified process that is recognized by an international standards setting organization such as the Consultative Committee for Space Data Systems (CCSDS) and the Organization for International Standardization (ISO). The first two categories have typically been used by specific space science sub-disciplines, while the latter category is intended to be applicable not only across space science, but across many other disciplines as well.

This presentation will touch on five standards that have either become ISO standards, or are in development to become ISO standards. These are the 'Reference Model for an Open Archival Information System (OAIS)' that provides conceptual model for archival system, the 'Producer-Archive Interface Methodology Abstract Standard (PAIMAS)' that provides a methodology for the interaction between data providers and the OAIS archive, the 'XML Formatted Data Unit (XFDU)' standard that provides a general packaging mechanism for the exchange of data and metadata, the 'Producer-Archive Interface Standard (PAIS) that addresses a formal description of data and its packaging into a Submission Information Package (SIP) to facilitate checking and validation upon receipt by the archive, and the 'Repository, Audit and Certification' standard that is expected to support self and third party certification of archives. The presentation will address their potential roles within an archival context, their status, the extent to which they have been adopted, and for those still in development, a brief discussion of issues to be resolved.

T9 - Towards an International standard for Audit and Certification of Digital Repositories

Dr David Giaretta, Rutherford Appleton Laboratory

Topic: This presentation will provide an update of the process which aims at developing an ISO standard on which a full audit and certification of digital repositories can be based.

Abstract: Certification of the ability of digital repositories to preserve digitally encoded information over the long-term has been sought for more than a decade. The OAIS Reference Model (section 1.5: ROAD MAP FOR DEVELOPMENT OF RELATED STANDARDS) included an item "standard(s) for accreditation of archives", reflecting the long-standing demand for a standard against which Repositories of digital information may be audited and on which an international accreditation and certification process may be based. It was agreed that RLG and NARA take a lead on this follow-on standard. This they did, forming a closed panel. The final version of the RLG/NARA work has been released as "Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)" available from http://www.crl.edu/PDF/trac.pdf. This document covers much ground and yet there is still a significant amount of work to be done. Other documents have been produced in this area, including the NESTOR "Catalogue of Criteria for Trusted Digital Long-term Repostories" which is available from
http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf.
The latest step in this process has been the creation of a "Birds of a Feather (BoF)" group within the Consultative Committee for Space Data Systems (CCSDS), which aims, as was done with the OAIS Reference Model, at setting up an open international workshops to debate and edit the document which will be submitted via CCSDS into ISO, with full open international review. The significant point is the openness of the process under which this work is being undertaken – see
http://wiki.digitalrepositoryauditandcertification.org
This presentation will provide details of the activities so far, describe the challenges which lie ahead, and discuss the potential significance of this effort.

T10 - Usability Issues Facing 21st Century Data Archives
Making data archives more useful and easier to maintain for providers, users, and management.

Joey Mukherjee and David Winningham, Southwest Research Institute

The typical sequence of events in data archiving is with a mission getting flown; the data being returned to and analyzed by the team; the data being cleaned up; and finally, the data being archived. It is our belief that this chain of events is fundamentally flawed since the act of archiving data is often not an integral part of the mission. Furthermore, the people most knowledgeable about the data have generally moved on to the next mission and the archiving of data is an afterthought, if even thought about at all. Cleaning and validating data are cost-prohibitive as well as resource intensive in terms of labor and computing hardware. If done at the end of missions, money is often so tight that there is little incentive to continue with an archive. Furthermore, space physicists have their own ways of handling data resulting in difficulty for interested observers outside the mission.

These dilemmas can be resolved by making the data used by the instrument team identical to the one that is to be archived. To achieve this, we must develop an ideal data format that is: usable by an instrument team, usable by an archival site, and usable by future scientists analyzing the data. In our framework terminology, what the provider uses for their needs should be identical to what is packaged as a submission information package (SIP) and submitted to archival storage.

Standardizing on a single data format can be achieved by first agreeing to a standard meta-data such as what the SPASE group is doing. The next step is for the space science community to develop a new standard data storage format to which others can agree to use.

The advantages of this approach allow others to develop value added services such as what the virtual observatories are doing as well as allowing instrument teams to continue creating valuable data products throughout the life of the mission that are immediately able to be archived. With this approach, the data is immediately usable to both the instrument team and the general space science community, and will continue to be useful past the life of the mission.

Every instrument group is already defining and distributing data products. However, using a standards based approach, more value added services will become available to the community, thus making the incentive to work with standards easier. As a result, there will be a rich set of tools available to aid the data investigator and data manager alike.

The real challenge is to define what this ideal data storage format would look like. At SwRI, we have developed a data storage format and know some of the challenges of developing a generic data storage format. We know now what we would do differently and what types of things we feel have been done correctly. We plan to discuss what steps can be taken in the future to ensure quality data archives for the entire space science community.

T11 - Associating Persistent Identifiers between Trustworthy Repositories

Topic: Role and implementation of persistent identifiers

Arnold Rots, Alberto Accomazzi, Guenther Eichhorn
Harvard-Smithsonian Center for Astrophysics

Under the auspices of the Astrophysics Datacenters Executive Committee (ADEC), a standard was developed that allows professional papers to link to original datasets used in those papers, and archived datasets to link to professional papers based on those datasets. The agreement was brokered between the datacenters, the Astrophysics Data System (ADS), and the editors of the AAS journals. It consisted of a standard for the definition of persistent identifiers for the archives' datasets (AIPs) and pairing these identifiers with the "bibcodes" that are in use in the ADS.

The identifiers conform to the International Virtual Observatory Alliance's standard for identifiers (IVOA); i.e., they are placed in the IVO domain, under the authority of the ADS and using the ADS's publishers' designations. The precise definition and meaning of the identifiers is left to the publisher, but the two main requirements are that it will lead the user unambiguously to an AIP from which a DIP can be retrieved, and that the identifier will be maintained in perpetuity, even if the repository were to move. The identifiers are currently being inserted into manuscripts on a modest scale and a harvesting tool is still under development.

Two future developments are of great interest in connection with this standard: certification and association.

The OCLC recently published the Criteria and Checklist for Trustworthy Repositories Audit & Certification (TRAC). It would be valuable for the datacenters represented in ADEC to pursue TRAC status, as this would provide a basis upon which indefinite persistence of the dataset identifiers could be built.

It would be natural for the ADS's bibcodes to also be molded into IVOA identifiers. Consequently, the association between papers and datasets becomes a simple association between identifiers and the harvesting of dataset identifier/bibcode pairs becomes a generalized harvesting of associated identifier pairs. This simple generalization offers immediately a mechanism to link datasets between repositories and to harvest such associations. It has become more and more popular to carry out coordinated or associated observations at multiple observatories (HST, Chandra, Spitzer, XMM, Integral, VLA, MMT, Gemini, AT, etc.), but there is no good mechanism to tie the different components together. Being able to harvest dataset identifier pairs in the manner described here would provide this functionality and thus close a significant gap in information management.

T12 - Science archives need to communicate more than data : the example of AMDA at CDPP

V. Génot, C. Jacquey, E. Budnik, R. Hitier, M. Bouchemit, M. Gangloff, and C. Harvey, CDPP/CESR
R. Conseil, D. Heulet, and C. Huc, CNES

Topic : the CDPP (French Plasma Physics Data Centre) has developed a new service to help multi-dataset analysis and in which event lists are at the core of archive communication.

The growing size of science archives and the need for cross-disciplinary approaches foster the demand for a new type of archive communication objects, beyond ordinary data exchange. Indeed as each archive remains close to its data and retains the required expertise to use them properly, the need for complementary communication objects appears : they are added-value concepts and hold a high level of genericity. For example reduced parameters, standardized images, or event lists fall into this category. Tools are therefore needed to produce, ingest and manage these objects in archives. In the context of Space Physics, AMDA (Automated Multi-Dataset Analysis) is such a service developed at CDPP. In this presentation, we shall detail AMDA, its concept and functionalities, and some scientific results obtained in studies based on its use. In AMDA system information circulate by means of two core entities : the parameter (e.g : 'electron density', 'magnetic field vector', 'plasma beta') at the data level, and the event list (e.g : magnetopause crossings, substorm onsets) at a higher level. Around these entities, AMDA is built as a suite of integrated tools allowing to perform massive scientific processing of the content of multi-datasets. This service is designed to work with a local database or (via web-services) distant data sources. The functionalities offered in AMDA include (i) user edited visualization browse, (ii) semi-automated event search (by visual identification), (iii) automated event search (based on user edited mathematical criteria applied to the data content), (iv) data extraction, (v) user edited parameter computation, (vi) functions and models (coordinates transformations, magnetic field/magnetopause/shock models) and (vii) access to data stored in distant archives. Recently, AMDA has been used for statistical studies in the magnetosheah with data from the four CLUSTER spacecraft, and in the magnetotail with a set of data from GEOTAIL, IMP-8, ISEE, AMPTE, INTERBALL and CLUSTER missions. In the field of heliophysics it will highly facilitate the combined analysis of data from ULYSSES, STEREO, CLUSTER and THEMIS and, hopefully, of planetary data. Finally, the experience gained in designing and building AMDA sharpened CDPP look on database management in the Virtual Observatories era, and we will present the standardization needs we feel urgent to complement archive communication regarding descriptive information of data, numerical simulation and model outputs, and event lists.

T13 - ESA Scientific Archives and Virtual Observatory systems

Author
Christophe Arviset, ESA - European Space Astronomy Centre

Co-authors
Inaki Ortiz, Pedro Osuna, Jesus Salgado, ESA - European Space Astronomy Centre

ABSTRACT

The European Space Astronomy Centre located near Madrid in Spain hosts most of ESA astronomy and planetary missionsí archives. That currently includes ISO, XMM-Newton, Integral and ESA Planetary Science Archive (regrouping data from Rosetta, Mars Express, Huygens and Giotto for the time being). In the future, Herschel, Planck, Smart-1, Venus Express, Soho and Gaia will also have their archives located at ESAC.

All these archives have been developed and are operated by a single Science Archives Team at ESAC and are using a common, modular and flexible 3-tier architecture, with a Java Object Oriented approach and XML. Data are saved on magnetic disks and metadata is extracted into a relational database. An "application server" ensures transparent access to the archive data and metadata for the front-end clients, like the powerful and easy-to-use graphical user interface, but also the scriptable interface for more expert users and the access to these data holdings through the Virtual Observatory (VO).

From the graphical user interface, users can also access data from other archives outside ESA, related to the data items found in the ESA archives. Additionally, external applications (eg Aladin, VOSpec) can be launched from the ESA archives to offer additional display and manipulate facilities. Link between the data and their related publications has been present for several years in the ESA archives, offering the possibility to view the articles abstracts from the archive and to get directly to the ESA data from the outside publications archive.
Apart from access to "static" data, on-the-fly reprocessing capabilities have been developed for some of the ESA astronomy archives, allowing the scientific user to get the best data processed with the latest version of the calibration software.

The Virtual Observatory (VO) is a world-wide initiative in astronomy which aim is to allow astronomers to perform new science by providing them with a "federation of astronomical archives and databases around the world, together with analysis tools and computational services, all linked into an integrated facility". Since the beginning, ESAC Science Archives Team has been heavily involved in the VO initiatives, both at European and international levels.

The open architecture of the ESA Scientific Archives has allowed easy integration of ESA astronomy archives into the VO without requiring changes in the way data and metadata were stored for each project.
By using translation layers, existing services have been adapted to meet the VO interoperability standards. By defining registry of resources, data access protocols and data models, new VO tools could be developed which give seamless access to astronomy resources worldwide enabling astronomers to perform new type of science in a more efficient and productive manner.
Re-using this experience on the astronomy side, similar systems have also been developed for the ESA Planetary Science Archive to ensure interoperability with the NASA Planetary Data Systems in the USA, in particular for image data.

T14 - Accessing Diverse Data Sets at the PDS Rings Node

Mark Showalter, Carl Sagan Center, SETI Institute

Topic: Meeting User Needs: We describe some of the challenges associated with providing access to diverse planetary data sets, and present a solution now in development at the Rings Node of the Planetary Data System.

The Planetary Data System (PDS) archives data from numerous spacecraft, telescopes, and instruments. We handle many types of data: images, spectra, cubes, time series, tables, etc. We support diverse users, who might wish to access the same data products for entirely different purposes. The PDS was designed as a distributed archive so that each user community has a primary point of contact, enabling users to access their discipline's key data sets via the most relevant selection criteria.

Each data set is delivered to the PDS with its own unique collection of descriptive information or "metadata". Some concepts are common to all data products (e.g., observation time), some are common to specific data types (e.g., filter names for images), and some are unique to one instrument. Additional types of metadata (e.g., viewing and lighting geometry) are different for each discipline.

Users weaned on modern Internet search engines have high (often unrealistic) expectations of what NASA's archives can deliver. Nevertheless, as NASA's data sets grow, users legitimately face the "needle-in-a-haystack" problem of zeroing in on the finite set of data products most relevant to a particular scientific question. They expect to be able to query the archive using a rich set of reliable metadata, without necessarily knowing in advance which data set(s) might hold the answers, or even whether an answer can be found within the archive's holdings.

The Rings Node is developing a web-based search engine that appears to be capable of meeting many of these challenges. The user can start a query at a very high level (e.g., images of Saturn) and "drill down" through all the available holdings to reach the desired data products (e.g., infrared images of Saturn's F ring at low phase angle and resolution finer than 5 km per pixel). The interface responds immediately to each new constraint entered by the user, removing irrelevant options and adding new ones as appropriate. For example, if it is determined that only Cassini images fit the constraints entered so far, then a page of Cassini-related constraints becomes available but options for occultation profiles and Hubble data are hidden. At each step, the engine displays a live tally of how many available products match the user's constraints. Using Ajax technology, caching of previous results, and a highly optimized database, the system is quick and responsive in spite of the millions of database records that must be searched.

However, we are only beginning to address a critical component of this system---generating the geometric parameters that are a fundamental part of most searches. We have identified the need to generate all such metadata in-house, via SPICE tools now in development. Unfortunately, we have found that this critical information is rarely included by data providers in their submission information packages.

T15 - Integrating a ACE Science Data Center and SAMPEX Resident Archive
into the Emerging Virtual Observatory System: Practical experience and perspectives

Andrew J. Davis, Caltech

Topic: Emerging archival standards and technologies
AND/OR
Meeting user needs: metadata, ontologies, natural language and archive exploration

The SAMPEX Resident Archive is currently under construction,
and is co-hosted at Caltech with the ACE Science Center.
With SAMPEX in low earth orbit, and ACE at L1, and a suite of
instruments on each spacecraft, the combined data cover a very
broad range in species, energy, location, and time. The data include
solar wind, solar energetic particle, and galactic cosmic ray intensity
and composition data, as well as solar wind and magnetic field parameters
on a variety of time scales.

Recent efforts to provide enhanced access to these data via the
emerging virtual observatory system will be described,
including work with the Space Physics Archive Search and Extract
(SPASE) Consortium to ensure that the ACE and SAMPEX data
can be adequately described using the SPASE data model,
development of a SOAP web services interface between the
ACE Science Center and the virtual observatories, and ideas for
combining the ACE and SAMPEX data in useful ways.

Issues and questions that have arisen as a result of these efforts
will be presented for discussion and, hopefully, mutual enlightenment.

T16 - Best Practices in Ingestion and Data Access at the Infrared Processing and Analysis Center

G. Bruce Berriman, California Institute of Technology

Topic: Meeting provider needs - ingestion of data; Meeting user needs - fast access to large data sets

The Infrared Processing and Analysis Center (IPAC) at Caltech hosts the NASA/IPAC Infrared Science Archive (IRSA) and the Michelson Science Center (MSC) Archive. IRSA is the steward of the scientific data sets of NASA's Infrared missions, and the MSC facilitates NASA's planet-finding and exo-planet science program, including multi-mission archives. Together they serve nearly 30-TB of data across the entire electromagnetic spectrum from 17 missions and projects. They share a common hardware and software architecture. This presentation describes their best practices in the areas of ingestion and user access.

Ingestion: While some providers are large missions, others are small groups of astronomers inexperienced in delivering products. Provision of standards and interface specifications for data delivery within a Submission Information Package are necessary for ingestion but have proven insufficient. Communication with the provider starts at the beginning of the project, and the provider is asked to deliver draft products for inspection before their pipelines have entered production. On-line validation tools, whose design has been driven by common mistakes in data delivery, have proven a powerful aid to providers. Functionality offered includes validation of the structure and content of catalogs; generation of the documentation of the attributes of catalogs; registration of images on the sky; and the syntax, content and astrometric accuracy of astronomical images.

Access: The archives must return in real time subsets of large data sets (catalogs, images and spectra) that it will curate for the indefinite future. The archive is optimized for efficient access, maintainability, portability and is highly fault tolerant. Catalogs and are housed in flat tables on a high-end EMC disk farm configured as RAID 0+1. An Informix DBMS offers dynamic parallelization of queries, but indexing for spatial queries is resident in memory outside the database. There are no stored procedures in the DBMS. All queries are composed through "thin" interfaces that sit atop a component based architecture of re-usable ANI-C modules that are "plugged" together for easy development of new applications. This architecture enables cost-effective deployment of new access services, such as those provided for NASA Stellar and Exo-planet Database, the Cosmic Evolution Survey Archive and the Keck Observatory Archive.

T17 - Applying Submission Agreements to Long Existing Data Flows – A NOAA story.

Dan Kowal, National Geophysical Data Center/NOAA

This topic covers the application of submission agreements in an institution with long standing relationships with data providers from the pre and incipient stages of the digital age.

The National Geophysical Data Center (NGDC) is one of three of NOAA's data centers under the National Environmental Satellite, Data, and Information Service (NESDIS). Compared with the other two data centers who focus purely on climatic and ocean data, NGDC handles heterogeneous data streams that are from multi-disciplines where the data sources range from satellites and ships to ground-based observatories.

NGDC is in a state of transition when it comes to archiving. For over a year now, the data center has moved towards using the OAIS reference model as a framework for defining the relationship between data provider and the archive. For new datasets, there is good support for engaging both sides in a formal submission agreement (SA) process as advocated in the PRODUCER-ARCHIVE INTERFACE METHODOLOGY ABSTRACT STANDARD (CCSDS 651.0-B-1 Blue Book). The archive presents the prospective data providers with a SA, and they accept it as a standard business operating procedure.

However, in the case of instituting the SA on existing data streams with a long time archival arrangement at NGDC, the process is not always so simple. Applying the OAIS concepts to existing data flows presents interesting challenges for well established datasets where minimum requirements for metadata, requests for PDI and codifying all aspects of the arrangements between data provider and archive transmissions were few and far between.

This presentation of OAIS infusion at NGDC will highlight some of the aspects of dealing with these well-established datasets. Key points are as follows: a) Brief historical perspective of data ingest; b) The scenario involved with one dataset involving a myriad of data providers; and c) Phased-in approach proposal to incorporate the SA into the data management process.

P1 - Provenance, Production, and Planning

Bruce R. Barkstrom
Science Data Stewardship Project
NOAA's National Climatic Data Center
Asheville, NC

Topics: This paper provides a solution to provenance tracking, data ingest planning, collection organization, and permanent registration of Earth science data collections.

This paper describes a mathematically complete solution to the problem of tracking the complete production history of a given file or file collection of Earth science or astronomical data, as well as a practical implementation of that solution that provides the capability for highly automated data production, production monitoring, and capacity planning. The mathematical solution is based on the fact that most Earth science or astronomical data is produced by discrete (batch) jobs that each accept a finite number of input files and create a finite number of output files. As a result, the production history can be represented mathematically as a graph, in which the files and jobs are nodes (or vertices) and the connections between input files, jobs, and output files are arcs (or edges). In our solution, the file identifiers are maintained in one list, the job identifiers are maintained in a second, and the connectivity between these two lists is maintained in a third. While this triple list structure is readily created using relational database tables, the recursive nature of graph traversals appears to require use of iterative query mechanisms that lie outside the usual SQL approaches. A key element of our approach lies in creating a template structure that allows automated creation of the production graph instances (representing the file-to-job connections in a single production job) for many jobs of the same kind. Once the complete production graph has been created, it is relatively easy to traverse the directed acyclic graph back from a file or collection of files to produce a list of all of the jobs that went into its creation, as well as a list of all the files that were used in that process. By traversing the graph in the opposite direction, starting with a job or small collection of jobs, it is relatively easy to produce a list of the jobs and files that will generate the complete list of child objects. By doing a standard topological sort on the list of planned jobs and files, it is also easy to create a production schedule and a list of the files that must be available to start the production process. This approach leads directly to a very highly automated approach to both data production and to capacity planning (assuming that the file sizes and CPU requirements of the jobs are known). In addition to the assistance this approach provides for production and capacity planning, when it is combined with the fact that almost all Earth science data and much astronomical data are readily viewed as being time sequences, it is possible to organize data collections based on files into a linear hierarchy (or tree). At the lowest level collection, the data files produced by a particular version of the production software can be viewed as a Data Set Version that can be assigned a "catalog identifier" based on sequencing the files in temporal order. Changing the source code, the coefficients of the algorithms, the input data, or the production graph topology will lead to generating a new Data Set Version. This catalog indexing approach leads to a workable permanent registration scheme that can help organize and manage large, federated archive collections.

P2 - LSST: Preparing for the Data Avalanche through Partitioning,
Parallelization, and Provenance

Kirk Borne, Perot Systems Corporation / NASA GSFC

TOPIC: We discuss the application of database partitioning, parallelization,
and provenance for massive dynamic archive ingest, management, and access.

ABSTRACT: The Large Synoptic Survey Telescope (LSST) project will produce
30 terabytes of data daily for 10 years, resulting in a 65-petabyte final
image data archive and a 70-petabyte final catalog (metadata) database.
This large telescope will begin operations in 2013 at Cerro Pachon in Chile.
It will operate with the largest camera in use in astronomical research:
3 gigapixels, covering 10 square degrees, roughly 1000 times the coverage
of one Hubble Space Telescope image. Two pairs of 6-gigabyte images are
acquired, processed, and ingested every 60 seconds. Within those 60 seconds,
notification alerts for all objects that are dynamically changing (in time
or location) are sent to astronomers around the world. We expect roughly
100,000 such events every night. Each spot on the available sky will be
re-imaged in pairs approximately every 3 days, resulting in about 2000 images
per sky location after 10 years of operations (2013-2023). The processing,
ingest, storage, replication, query, access, archival, and retrieval functions
for this dynamic data avalanche are currently being designed and developed
by the LSST Data Management (DM) team, under contract from the NSF. Key
challenges to success include: the processing of this enormous data volume,
real-time database updates and alert generation, the dynamic nature of every
entry in the object database, the complexity of the processing and schema,
the requirements for high availability and fast access, spatial-plus-temporal
indexing of all database entries, and the creation and maintenance of multiple
versions and data releases. To address these challenges, the LSST DM team is
implementing solutions that include database partitioning, parallelization,
and provenance (generation and tracking). The prototype LSST database schema
currently has over 100 tables, including catalogs for sources, objects,
moving objects, image metadata, calibration and configuration metadata, and
provenance. Techniques for managing this database must satisfy intensive
scaling and performance requirements. These techniques include data and
index partitioning, query partitioning, parallel ingest, replication of
hot-data, horizontal scaling, and automated fail-over. In the area of
provenance, the LSST database will capture all information that is needed to
reproduce any result ever published. Provenance-related data include:
telescope/camera instrument configuration; software configuration (software
versions, policies used); and hardware setup (configuration of nodes used to
run LSST software). Provenance is very dynamic, in the sense that the
metadata to be captured change frequently. The schema has to be flexible to
allow that. In our current design, over 30% of the schema is dedicated to
provenance. Our philosophy is this: (1) minimize the impact of reconfiguration
by avoiding tight coupling between data and provenance: hardware and software
configurations are correlated with data via a single ProcessingHistory_id;
and (2) minimize the volume of provenance information by grouping together
objects with identical processing history.

P3 - Science Archives in the 21st Century: a NASA LAMBDA report.

Paul Butterworth, ADNET / NASA GSFC

ONE SENTENCE TOPIC:
The framework and goals of the meeting will be discussed from the perspective
of the NASA Legacy Archive for Microwave Background Data Analysis (LAMBDA).

ABSTRACT:
Lambda is a thematic data center that focuses on serving the cosmic
microwave background (CMB) research community. LAMBDA is an active archive for
NASA's COsmic Background Explorer (COBE) and Wilkinson Microwave Anisotropy
Probe (WMAP) mission data sets. In addition, LAMBDA provides analysis
software, on-line tools, relevant ancillary data and important web links.
LAMBDA also tries to preserve the most important ground-based and suborbital
CMB data sets. CMB data is unlike other astrophysical data, consisting of
intrinsically diffuse surface brightness photometry with a signal contrast of
the order 1 part in 100,000 relative to the uniform background. Because of the
extremely faint signal levels, the signal-to-noise ratio is relatively low and
detailed instrument-specific knowledge of the data is essential. While the
number of data sets being produced is not especially large, those data sets are
becoming large and complex. That tendency will increase when the many
polarization experiments currently being deployed begin producing data. The
LAMBDA experience supports many aspects of the NASA data archive model
developed informally over the last ten years - that small focused data centers
are often more effective than larger more ambitious collections, for example;
that data centers are usually best run by active scientists; that it can
be particularly advantageous if those scientists are leaders in the use of the
archived data sets; etc. LAMBDA has done some things so well that they might
provide lessons for other archives. A lot of effort has been devoted to
developing a simple and consistent interface to data sets, for example;
and serving all the documentation required via simple 'more' pages and longer
explanatory supplements. Many of the problems faced by LAMBDA will also not
surprise anyone trying to manage other space science data. These range from
persuading mission scientists to provide their data as quickly as possible,
to dealing with a high volume of nuisance (spam) messages. Because so many
data center problems and solutions are common across individual data centers
and disciplines it would be very valuable to establish some new systems of
communication - such as informal email lists for administrators and
developers. But resources are very limited, so new time-consuming and
inefficient mechanisms - like too-frequent and too-structured meetings -
should be avoided. Although there are great advantages to being small, agile
and independent, there are also some areas where science data centers
within and without NASA could be better coordinated - for the assignment
of persistent identifiers; to encourage the early adoption of useful
standards and technologies; etc. Some super-structure to facilitate
such coordination might be beneficial as long as it doesn't begin to
control the other work of the archives, and become a 'methodology
police'. In this respect the CCSDS 'Reference Model for an Open Archive
Information System' is a little worrying. It may be that the closer a data
center gets to following such a detailed prescription, the less effective it
will become. It is much better to have an informal coordination process
than a bureaucratic straight-jacket.

P4 - Developing the International Planetary Data Alliance

Daniel Crichton, Jet Propulsion Laboratory

Reta Beebe, New Mexico State University

TOPIC: The development of an international alliance for sharing data from planetary science archives.

Solar System exploration at the beginning of the 21st Century involves complex missions that host instruments that are developed and managed by the international community. Over the past decade, the European and United States space agencies have collaborated to share resources. Japan, India and China, now planning and executing robotic exploration missions of the solar system, have joined in collaboration to develop international standards. The collective results will yield an unprecedented volume of data. At the same time, resources from any one agency are scarce, requiring agencies to leverage existing standards and tools, where possible.

In 2006, ESA and NASA along with JAXA/ISAS, CNSA and RAS/RKA formed the International Planetary Data Alliance (IPDA) and held its first meeting at the ESA Technical and Engineering Center (ESTEC) in Noordwijk, Netherlands in November. A Steering Committee with representatives from the agencies has been formed. Recently members from ISRO, BNSC, CNES, ASI, and DLR have been added.

The purpose of the IPDA is two pronged. First, it is to develop a set of data standards that drive IPDA-compliant data systems and allow for sharing scientific data products across international agencies and missions. Second, it is to develop a set of technical information system standards to allow for interoperability between agency data systems.

At the ESTEC meeting, the IPDA developed an operations approach and identified critical projects, including the identification and development of a core set of data standards based on the NASA Planetary Data System (PDS). The planned IPDA data standards include a data model, data dictionary, a set of data formats, standard grammar and archive organization. The IPDA Steering Committee plans to meet yearly to review progress of the projects with its next meeting scheduled for July 2007 at Caltech in Pasadena, California. A critical focus of this meeting will be to review the progress in developing a core set of data standards for the IPDA.

In addition to development of the IPDA organization and a set of projects, the IPDA is also working to ensure alignment with the international scientific community. As a result, the IPDA is proposing a session at the 37th Committee on Space Research (COSPAR) Scientific assembly in July 2008 in Montreal. The meeting will focus on introducing the scientific community to the IPDA structure, standards and resources.

P5 - Scientific Satellite Data Archives at JAXA

Ken Ebisawa, Japan Aerospace Exploration Agency

Abstract: We will introduce scientific satellite data archiving
activities in Japan, being carried out primarily at JAXA/ISAS. Since
1970's, ISAS has been launching small or medium-size astronomy,
solar-terrestrial physics (STP), and solar physics satellites. In 1997,
ISAS has started scientific satellite data archives "DARTS" (Data
Archives and Transmission System; http://darts.isas.jaxa.jp). At DARTS,
we archive most ISAS satellite data taken after late 1980's.
Currently, ISAS's three STP satellites, two astronomical satellites and
one solar satellite are operational, and these data are being archived
at DARTS. In 2003, ISAS and two other space agencies (NASDA and NAL)
merged, and Japan's sole space agency "JAXA" has established. JAXA is
capable of launching larger scientific satellites such as the lunar orbiting
mission SELENE (SELenological and ENgineering Explorer), which is
planned to be launched in August 2007. The lunar and planetary
exploration is going to be one of the main projects of JAXA. We will
present our plan of archiving SELENE and other future lunar and
planetary data at JAXA.

P6 - Role of a Permanent Archive in the evolving NASA space science environment

E. Grayzeck, NASA GSFC

Topic: archival policies and their implementation

The requirement for scientific archiving of past, actual and future scientific missions is indisputable. NSSDC provides a vital service as NASA's only permanent multi-disciplinary Space Science archives. NSSDC's resources are focused on preserving as a proper steward of potentially highly valuable data. At the same time, the evolution in the research communities expectations for rapid, seamless, access to data are not being neglected as more of NSSDC's data are going online for use under web services

As a general policy, NSSDC establishes an MOU with and acquires data from the Space Science (Heliophysics, Planetary Science, and Astrophysics) Active Archives for long term curation, and it provides it back to them when requested. NSSDC acquires data from projects and researchers for long term curation when those data are not germane to other AAs, and it makes such data available to researchers and the general public. So that in the changing NASA environment, these MOUs maybe with projects, missions, heliophysics Resident Archives, the Planetary Data System , and astrophysics SARCs.

The author will take one example of this new mode, the Resident Archive, to show the current process to establis guidelines for standards and how that will allow the best practices to be used to provide the scientific user high quality data in a reasonable search time. Furthermore, I will explore the possible evolution of multiple RAs given the lessons learned and dynamics of similar mission oriented interactions such as the PDS data nodes. From that experience, the author will outline the evolving role for NSSDC for RAs, possible future interactions with PDS data nodes, astrophysics SARCs, and sister archives in EOS.

P7 - Approaches for Archiving and Distributing Science Data from Planetary Missions

Edward A. Guinness, Thomas C. Stein, and Susan Slavney, Washington University

Topic — Creation of high quality science archives and approaches for meeting user needs for access to data.

For nearly two decades the Geosciences Node, a node of the Planetary Data System (PDS), has led the archiving of science data for NASA's missions to the terrestrial planets, with emphasis on Mars Exploration Program. We have also collaborated with several international planetary missions on data archiving. Our objectives are to facilitate the creation of quality data archives and to make our data holdings available to science users by efficient and useful means. To accomplish these objectives, we have developed a set of standard practices for interfacing with both data providers (i.e., instrument teams) and data users. These practices include working with a mission early in its development phase to develop an archive plan and schedule and to form an archive working group to oversee archive planning and operations. We have found that archive quality can be enhanced and archive production can operate more smoothly if a number of steps are followed by the instrument teams. It is essential that instrument teams assign a representative to be responsible for handling archiving issues and interfacing with the PDS; that early peer-review of archive plans by potential users are conducted so that modifications can be made before initial data release; and that end-to-end archive delivery tests are conducted before beginning full-scale archive operations. The MRO mission and PDS recently completed four archiving tests that identified and corrected several issues before the first scheduled data delivery to PDS in June 2007.

Our user community varies in expertise and has a diversity of interests for using data from our holdings. This diversity is important for data providers to consider in designing archives. For example, there are users with limited technical resources that prefer data products formatted for easy importing into commonly available software tools. Some users want raw data so that they can apply their own calibrations and corrections, whereas others are interested in highly derived data from multiple data sets. Complete and detailed documentation and calibration information are essential as part of an archive, and we encourage data providers to publish papers about their instruments and data in the scientific literature. Tools available to users for finding subsets of data are becoming more important as data sets increase in number, size, and complexity. A common set of standards is essential for building a data system to support cross-instrument and cross-mission searches. Web-based services are needed to minimize system requirements placed on the user. These services should not only allow users to find and access data, but should also provide the context in which data were acquired. For example, we have developed and operated the MER Analyst's Notebook to support access to data returned by the Mars Exploration Rovers, which allows users to understand why particular observations were acquired.

P8 - AND Archives: Freeing Ourselves from the "Tyranny of the OR"

Ted Habermann, NOAA National Geophysical Data Center

Topic: Either Meeting user needs: metadata, ontologies, natural language and archive exploration or Emerging archival standards and technologies

In some parts of the scientific archive community, data users have been divided into two groups: science users and GIS users, and systems are built that serve one group OR the other. We might call these "OR Archives". This approach creates obstacles to sharing data and limits the knowledge that can be applied to many important problems. It is now possible to build systems that support science users AND GIS users. One approach is to use a geospatial database as a front-end to the archive. Data are ingested into the database and served directly using geospatial standards and tools. Standard products can also be created from the database. The data are also written to the archive file system and served to scientific users from there. I will explore this idea using some recent work at NGDC with Level 2 Sea Surface Temperature observations from NESDIS. These data blur some traditional distinctions between GIS and satellite data.

P9 - An application of CCSDS archival standards to meet both submitter and archive needs during data ingest.

H. Kent Hills   Donald M. Sawyer   Patrick McCaslin
Perot Government Systems   NASA GSFC   Perot Government Systems
(NASA GSFC)   (NASA GSFC)

The National Space Science Data Center (NSSDC) is currently in the process of implementing new highly automated procedures for ingest of data into the archive, following the general guidelines of the CCSDS document "Producer-Archive Interface Methodology Abstract Standard (PAIMAS)', for the producer-archive interactions.

The main objective of the current development is to meet the concerns of our staff scientists about the difficulty of using the previous ingest process. We are actually in transition now, moving from our previous process to a new one, and various changes to our main database system will result. We need to make the acquisition of ingest information flow as smoothly as possible,and we need to effectively use as much automation as possible, to be cost-efficient in our operations, as well as to meet current and upcoming archival standards.

This presentation will briefly describe the evolution of the NSSDC archival ingest operations and will provide an overview of the development of the current process of gathering the necessary information relative to the automated ingest of data into the archive. The process is defined as a multi-phase one, as discussed in PAIMAS; in this presentation, we will consider mainly the Preliminary Definition Phase and the Formal Definition Phase.

P10 - Automated determination of astrometric metadata for interoperability
and collaboration

David W. Hogg, New York University

The proper functioning and interoperability of astrophysics archives
for scientific investigations require that every archived image have
astrometric metadata that are precise, accurate, standards-compliant,
and consistent across archives. This is required for users with
specific science goals, most of whom want to match sources across
archives, and for collaboration and discovery, where users benefit
enormously from access to overlapping and matched data.

Right now, astronomy archives do not have consistent astrometric
metadata. Astrometry is currently a show-stopper for generic,
unsupervised or virtual-observatory-mediated cross-archive
investigations, even with modern, space-based data, but especially
with ground-based and legacy data.

We have developed a system (astrometry.net) to automatically generate
and record standards-compliant and consistent astrometric metadata for
images, in any state of archival disarray, by directly matching
detected sources in the images to astrometric standards catalogs. The
system has been shown to work on images over a wide range of
wavelengths and a wide range of technologies, from commercial digital
camera images to scanned photographic plates to space-based telescope
images. The system can produce consistent, standards-compliant
astrometric metadata for all science archives, it can make amateur
astronomers' data interoperable with professional programs, and it can
bring ancient and legacy accessible for the study of long time
baselines. Automated maintainance of astrometric (and other!)
metadata will be absolutely indispensible for a functioning
"federation" of archives or the utopia of interoperability imagined in
a future virtual observatory.

P11 - FRBR in a Scientific Data Context

Joe Hourcle, NASA GSFC

Topic : Standardization in terminology and classifications used to
        describe data granularity and data collections to facilitate
        interoperability across archives.

Abstract

Data can be catalogued at many different levels of granularity -- data
granules, data products, data sets and data collections. Unfortunately, one
discipline's data product is another discipline's data set. The inconsistency
of terms creates difficulty in interfacing the archives -- if one archive
generates metadata records at the data granule level, while another describes
their data at the data product level, there will be confusion in merging
records from a federated search.

Although the OAIS reference model (CCSDS 2002) discusses the concepts of
Collection Descriptions and Representation Networks, it does not discuss the
granularity of the data being described, other than as an Archive Information
Package (AIP), or an Archive Information Collection (AIC). Unfortunately, the
amount of data that makes up an AIC for one archive may be an AIP for another.
This poses a problem when archives return different granularity in requests.

Some Active Archives may provide different versions of the same content, be it
differing Editions, or alternative packaging for different Designated
Communities. Some may return records for Derived AIPs interspersed with their
source AIPs. This may be desired by some users, but it can confuse and
overwhelm users who are not part of the Local Community for that archive.

Since the scientific community lacks coordination in terminology for these
aspects of cataloging, we examine the concepts of the Functional Requirements
for Bibliographic Records [FRBR] (Munchen 1998) developed by the library
community for best practices in this field. We discuss the applicability of
FRBR concepts to scientific data, and the need for a similarly purposed model
as the "glue" necessary to hold together any virtual observatory or other
federated search system for scientific data.

References

CCSDS (2002). "Reference Model for an Open Archival Information System (OAIS)"
<http://public.ccsds.org/publications/archive/650x0b1.pdf>

Munchen, K.G.S. (1998). "Functional Requirements for Bibliographic Records,
Final Report" <http://www.ifla.org/VII/s13/frbr/frbr.pdf>

P12 - The Application of Semantic Technologies to Scientific Archives

J. Steven Hughes and Daniel J. Crichton, Jet Propulsion Laboratory

Topic: Semantic technologies significantly benefit the development, management, and use of scientific knowledge.

Semantic technologies allow information to be read and consumed by computer software. For example, where the HyperText Markup Language (HTML) makes information more easily consumable by humans, the Resource Description Framework (RDF), RDF Schema (RDFS), Web Ontology Language (OWL) and their XML implementations allow software to process and reason about the information.

Semantic web applications that provide powerful browsing capabilities can be developed with little effort by reformatting information contained in traditional database applications. The extracted information is reformatted into a semantic language and then configured for access through off-the-shelf metadata browsers. These web applications provide combined text- and facet-based search and unabridged navigation through both the data and metadata in the resulting knowledge-base.

Ontology modeling tools are powerful tools for managing information models. They allow an information architect to build, modify, analyze, and even populate an information model and subsequently export the resulting knowledge-base to Metadata Interchange (XMI) or another semantic language. They provide single source maintenance of the information model independent of any implementation choice and are able to subsequently drive a variety of implementation and documentation choices, often automatically. In addition, semantic web applications can process and reason about the model.

In 2005 the Planetary Data System (PDS) developed a prototype semantic web application for browsing the science information extracted from its main catalog. The information extracted from the catalog database was reformatted into RDFS/XML and RDF/XML and then configured for an off-the-shelf metadata browser. This prototype successfully demonstrated the ease with which sophisticated search and browse capability could be provided to the planetary science community. Follow-on work has focused on re-capturing and documenting the PDS information model using an ontology tool.

In early 2007, the newly formed International Planetary Data Alliance (IPDA) initiated a project to identify and develop a core set of data standards based on the PDS. The PDS information model, as captured in the ontology modeling tool, will be used to create the initial draft of the IPDA information model and its documentation. The ontology modeling tool provides a means for single source maintenance of the information model, export of the model to a variety of formal modeling specifications, and the ability to test the model. It also allows the designers to focus on information modeling issues in a formal yet flexible environment free from the particularities and limitations of an implementation language.

P13 - NASA Datasets Management Using Process Libraries and Electronic Handbooks [Where Shakespeare Meets Freud]*

Dr. Barry E. Jacobs, NASA GSFC

This talk focuses on the NASA Datasets Management methodologies called Process Libraries and Electronic Handbooks.

This talk focuses on the problem of developing Internet-based tools to support the paperless documentation and management of complex distributed processes such as NASA Datasets Management. The complexity of this problem is illustrated by the fact that there are lots of subprocesses in NASA Datasets Management. For example, Dataset Realization subprocesses include: Datasets Development Facility Solicitation Development, Designing Datasets, Building Datasets, Using Datasets, Improving Datasets, Revising Datasets, Closing Datasets, and Post-Closeout. Dataset Distribution subprocesses include: Distribution Facility Guidelines Document Development, Problem Submissions Problem Review and Selection, Negotiations, Administration, Closeout, and Post-Closeout The complexity is further enhanced by the fact that hundreds of different organizations provide different views of these subprocesses.

This talk focuses on the methodologies called Process Libraries and Electronic Handbooks. Process Libraries (PLs) and Electronic Handbooks (EHBs) are where Shakespeare meets Freud. In Process Libraries (PLs), subprocesses are represented as "plays" where "actors" communicate thru the Internet. Each organization puts on its own "productions". For each role, Electronic Handbooks (EHBs) guide actors thru their parts. [Shakespearean] Organizations are represented as "teams" having "multiple personalities". Subprocess "plays" provide communication vehicles between members of the same team, different teams, and teams from different processes. [Freudian] These methodologies have been in use over the past 15 years throughout NASA and other Federal Agencies.

The Datasets Management Process Library is a subpart of the Flight Projects Process Libraries which can be found at *http://ehbs.org/pls.*

P14 - "Show Me The Data"

Nathan James, NASA GSFC

Topic: What Real Users Want from a Data Archive
(User Requirements for Archival Access)

Abstract:
The National Space Science Data Center (NSSDC) serves as the permanent archive for NASA's Office of Space Science (OSS). A major component of its mission is to ensure future data accessibility and usability. NSSDC also provides current data access, complementary to the efforts of other NASA/OSS "active archives," in support of the NASA and international astrophysics and space physics research enterprises.

The presenter seeks to highlight the results gathered from interviewing science users of these archives concerning their access requirements. The intent of this poster presentation is to generate discussion around the following:

1) The top 3 things that users say they need when retrieving archived data;
2) The common barriers to users getting to the data they want;
3) Tools/techniques that have proven to be most helpful in finding/accessing/retrieving archived data; and
4) New technologies being developed with proven user-friendly tools/techniques in mind

P15 - Implementing a Virtual Observatory: Models, Frameworks and Tools

Todd King, Raymond Walker, UCLA
Jan Merka, Jim Thieman, Aaron Roberts, NASA GSFC

A discussion of the path taken to build a virtual observatory from concept to completion.

The emergence of Virtual Observatories is a recent step taken by NASA to provide new and perhaps revolutionary services to aid scientific research. The goal of a Virtual Observatory is to provide single point, transparent access to distributed resources. The vision for such a such a system dates back to the past century. More recently the World Wide Web has demonstrated the feasibility and effectiveness of such a concept. The recent confluence of community interest, foundational efforts and enabling technology has made Virtual Observatories possible. We discuss the essential elements of implementing a Virtual Observatory which includes the development of data models; standardized expressions of the metadata; frameworks for inter-connecting of resources and services; and the tools required to populate and operate a system. The discussion will include insights gained while participating in the development of the Space Physics Archive Search and Extract (SPASE) data model, building up the Virtual Magnetospheric Observatory (VMO) and the long-term involvement with the design, implementation and evolution of the Planetary Data System (PDS).

P16 - Whither Physical Media?

Mike Martin, PDS Consultant

Topic: PDS examines the use of CD and DVD recordable media for science archives

In mid-2006 the Planetary Data System (PDS) Physical Media Working Group (PMWG) polled its discipline node representatives regarding their experiences with CD and DVD recording. The PDS receives some of its data deliveries on CD and DVD media and also uses it internally for archival storage. The environments for recording CD's and DVD's turn out to be remarkably varied, including a mix of UNIX workstations, pc's and mac's. It came as a shock that half of the nodes were consistently having problems recording CD-R's and nearly every site (9 of 10) was having problems recording DVD-R's. This is a major problem given that the nodes project the use of nearly twenty thousand volumes of recordable media over the next several years. The respondents did not have a good idea where the problems were originating. Initial contacts with the National Institutes of Science and Technology (NIST) and other experts in the field indicated that DVD recording and to a lesser extent CD recording have has become more difficult as the industry has become more competitive. A perfect combination of recorder and compatible high-quality media; latest firmware for optimal power control; recording speed matched to media speed; appropriate software for the recorder; and a host system capable of providing the throughput for high speed recording are all required to produce a successful product. Trying to establish a successful recording environment is exacerbated by the endless parade of new models of recorders and by a flood of poorly manufactured (and sometimes counterfeit) media.

The group also performed an evaluation of a set of archival CD and DVD products to appraise the current state of the archive. A secondary goal of this effort was to develop a testing methodology for CD and DVD discs that could be used if further testing was required. Discs were tested on industrial strength test devices (CD and DVD CATS) by the National Institute of Standards and Technology. These test devices provide both error statistics and detailed measurements of the physical parameters of each disc. Nearly every disc failed in the CATS tests on one parameter or another. However, many of these errors were related to artifacts of the disc writing process found in the lead-in and lead-out areas of the discs and not related to the data stored on the disc. New tests need to be done which skip these areas of the discs. In general terms, most of the CD scans looked very solid, while the DVD scans showed flaws on many discs. Discs were also scanned on three different DVD/CD drives at the Planetary Plasma Interactions (PPI) Node using the "CD Speed" program. This program provides a number of different tests to evaluate both the drive and the media. Each CD disc was scanned at 24X and 40X and each DVD at 4X, 8X and 16X. For the CD's there was little correlation of scan results between the different drives or at different speeds. We believe this could be due to the low sampling rate used by CD Speed when scanning CD's. For the DVD's, there was some correlation between scan results at 4X and 8X and between drives at these low speeds. We feel that the CD Speed program is useful for providing gross quality assessments of CD or DVD media. We are looking for a better scan tool for CD media.

Out of nine CD-R's tested, seven look stable, one is marginal and one disc is

flawed and needs to be copied to new media. Out of nine DVD-R's tested, two look stable, three discs are marginal and four discs are flawed and need to be copied to new media. We do not see any evidence of deterioration due to age. Most flawed discs show evidence that they were written improperly. All data on all discs was fully recoverable, though only by using multiple readers. Based on our research, we recommend migration from CD-R and DVD-R archives to on-line storage systems with high-density (DLT) backup. This is due to dramatic cost reductions for on-line storage, the relatively low storage capacity of CD and DVD media, the difficulty of establishing a successful recording environment and the difficulty in ascertaining the quality of recorded media.

P17 - Use of Archive Information Packages at the NSSDC

Patrick McCaslin, Perot Systems Corporation / NASA GSFC

Possible Topic: Long-term preservation of understanding of scientific data

The National Space Science Data Center (NSSDC) has adopted concepts from
the "Reference Model for an Open Archival Information System (OAIS)" as
a framework for the evolution of its systems and processes. A key element of
the OAIS model is the Archive Information Package (AIP) which consists of
content information, the information to be preserved, and Preservation
Description Information (PDI), the information which is necessary for adequate
preservation of the Content Information. NSSDC's implementation of the AIP
packages digital Content Information and PDI into a single file with standard
pointers to registered, supporting, documentation. An AIP can contain
individual digital files, multiple files contained in a single directories, or
entire directory trees. Future AIP forms may define AIPs for non-digital data.

A set of software utilities called the Multi-file Packager and
Analyzer (MPGA) is used for the creation and validation of AIPs and the
restoration of data in their original format from AIPs. NSSDC uses MPGA to
package all electronically delivered data into AIPs and makes MPGA available to
data providers in order to create AIPs at their sites. Such remote packaging
of data is strongly encouraged. Benefits to data delivery via AIP include:

- Integrity - Fixity information is captured at the provider's site
and incorporated into the AIP affording the maximum possible assurance
that the data is preserved at the NSSDC exactly as the provider intended.

- Long-term usability - Critical low-level attributes, required for long-
term preservation and use of the data, are extracted automatically from
the data at the provider's site and incorporated into the AIP.

- Automation - AIP delivery and ingest is automated, eliminating
human errors and delays in the process at both the Provider's site and the
NSSDC.

- Multiple packaging options - Contents of AIPs may be selected in several ways
(individual file, multiple files, directory structures) allowing the provider
to bundle related data in the preservation package.

- Large capacity - AIPs can accommodate large data volumes (currently tested to
150 GB), again facilitating the bundling of related data in the preservation
package.

Experience with this approach and some issues will be described.

P18 - Replication Policies for Distributed Digital Preservation Environments

Robert H. McDonald, SDSC

Topic: Data Replication across trusted digital preservation environments will be a key component of the 21st century e-science archive.

In recent years there has been much speculation about the infrastructure needed to support distributed or federated preservation environments. This data infrastructure known variously as a component of e-science or cyberinfrastructure would exist to organize, preserve, and make accessible over time the intellectual capital that is being created via research in the sciences, engineering, and humanities disciplines.

The San Diego Supercomputer Center (SDSC) along with the U.C. San Diego Libraries (UCSDL), the National Center for Atmospheric Research (NCAR), and the University of Maryland Institute for Advanced Computer Science have formed a collaborative partnership called Chronopolis. The underlying goal of the Chronopolis digital preservation environment will be to curate this intellectual capital at a national-scale level in order to preserve it for the next generation of scientists and scholars.

As a first step process toward developing the Chronopolis working prototype SDSC and NCAR have signed a memorandum of understanding (MOU) to facilitate cross-replication of collections between the two sites. In its initial stages this redundancy serves as merely storage swap. However, each institution is developing its own policies for collection replication with the goal of creating policies that can interact independently as well as cross-institutionally for the replication of critical data. This iterative process is a first step-towards a model cross-institutional strategy that will eventually extend to all partners working in the Chronopolis preservation environment.

This session will compare the framework of the MOU between SDSC and NCAR and look at policies specific to SDSC collections stored in the replication swap. The analysis of this will provide a framework from which to develop similar MOUs with future repository partners and will offer guidelines for developing replication policies between data repositories. Included will be a best practices summary for data replication as well as a comparison of data management policies pertinent to issues with data swaps versus best practice policies for shared preservation environments such as Chronopolis.

P19 - Data Preservation and Data Reuse in Archive Design and Implementation

Thomas McGlynn, NASA GSFC

Two of the primary goals for data archives are preservation of data and the promotion of data re-use in archival research. There can be tension between these goals. This poster explores how this tension can be seen in many areas in the design and operation of modern archives. It explores how archives have answered questions like: Do archives keep a complete -- and sometimes confusing -- version history, or simply present the latest version -- at the cost of being able to recover earlier data. Can one recover earlier datasets when data retrievals invoke dynamic processing using changing software? When releases are staged must each release be available independently? Are software and calibration archives coordinated with the data archives? Does the growth of virtual archives presage a split where the virtual observatories focus on the reuse functions while the primary archives work to ensure preservation? Do archives need to make primary datasets available, or can they function effectively when users can only get derived products? The paper surveys NASA's astronomy archives and shows how different organizations have come to different conclusions based upon the nature of their datasets, the resources they have available, and communities they serve.

P20 - Guidance for Science Data Centers through Understanding Metrics

John F. Moses, Carol L. Boquist, NASA GSFC

EOSDIS has built a multi-year set of metrics about the evolving, broad collection of earth science products and a diverse set of users.

These metrics include dataset ingest and distribution transactions, user access characteristics and measures of user satisfaction. They are used to determine trends that can be the basis for understanding the results of cross-cutting initiatives and for management decisions about future strategies.

The information is available through two complementary methods: metrics collected regularly from the major science data centers, and user satisfaction information collected through the American Customer Satisfaction Index survey. The combination provides the basis to understand utilization trends in the research community.

This poster will present metrics from NASA EOSDIS efforts, lessons learned, planned improvements and innovations, and best practices. The poster will highlight metrics collection as a best practice and discuss how we use metrics to evaluate our progress and guide planning.

Authors:
Ani Thakar (presenter), Alex Szalay, Maria Nieto-Santisteban, Nolan Li
(The Johns Hopkins University), and Jim Gray (Microsoft Research)

Topic:
Emerging archival standards and technologies for very large datasets in
Astronomy and other sciences

Abstract:
The Sloan Digital Sky Survey (SDSS) has been serving a multi-Terabyte
catalog dataset to the astronomical community for a few years now. By
the beginning of the next decade, the Large Synoptic Survey Telescope
(LSST) will be acquiring data at the rate of one SDSS every 3-4 nights
and serving a Petabyte-scale dataset to the community by 2015 or so. I
will discuss the SDSS data ingest, archival storage and data management
strategies and how we are building on them to go from Terabyte- to
Petabyte-scale data access. The Web services-based SDSS access
functions - SkyServer, CasJobs and ImgCutout - have worked very well in
serving this large dataset to a wide user community. However, as we
look ahead to successors of SDSS and even bigger archives like
PAN-STARRS (Panoramic Survey Telescope and Rapid Response System) and
LSST, we are experimenting with several existing and new technologies
to deal with the prodigious predicted data volumes. These include
automated data ingest, asynchronous query execution, data partitioning
for parallel data access and on-demand cross-matching between federated
archives, and bringing the analysis to the data and not vice-versa to
minimize movement of large amounts of data.

I will also describe the extensive traffic and usage tracking system
that logs every single SDSS Web hit and SQL query, thereby greatly
facilitating preservation planning, data management tuning and
administration. Since 2001, the SDSS catalog archive has logged
nearly a quarter billion Web hits and 30 million SQL queries. We are
collaborating with the JHU Digital Libraries group in an effort to
provide long-term preservation of SDSS and other datasets.

I will treat these topics in the larger context of the international
Virtual Observatory (VO) effort that seeks to bring these technologies
together so that all astronomical data is federated and accessible
seamlessly and efficiently. JHU is a major participant in the projects
that I will discuss - SDSS, PAN-STARRS, LSST and the VO - and is also
building a 100-Terabyte data analysis facility as a stepping stone to
Petabyte-scale data analysis. The Digital Laboratory for Multiscale
Science (DLMS) will analyze data a Terabyte at a time from large-scale
turbulence simulations.

Tradeoffs in the Development of the SPASE Data Model

Topic: Facilitating access to and retrieval of data across widely-distributed, disparate, science archives.

James R. Thieman1, Todd King2 and D. Aaron Roberts3
1National Space Science Data Center, NASA/Goddard Space Flight Center
2Institute of Geophysics and Planetary Physics, UCLA
3Heliospheric Physics Laboratory, NASA/Goddard Space Flight Center

SPASE is the Space Physics Archive Search and Extract project. This project is intended to provide a common basis for locating and retrieving data of interest for the space and solar physics community across multiple widely-distributed heliophysics archives and data centers. Common terminology that maps to much of the disparate metadata being used by these data archives around the world enables unified searches and ready comparison of the results to determine time overlaps, data commonalities, applicability for research purposes, etc. The project has developed the SPASE Data Model for the description of heliophysics data sets. The success of this project depends on the wide usage of the Data Model in the community.

In this presentation we will talk about the development of the Model through dedicated international committee work and, in particular, the difficult tradeoffs that must be made. For example, to what level of detail should the model aim to describe data sets? Should just the finding and approach to acquisition of the data be supported through the model usage or should it describe data sets in sufficient detail that the user will obtain knowledge of all the parameters measured and details of resolution, cadence, etc. for the parameters? What tools will be needed to enable archives and data centers to map their existing metadata to the SPASE metadata in a reasonably automated manner? Version 1.2 of the Data Model is available presently and will soon be "frozen" for usage in a stable environment. (See http://www.spase-group.org) The Model will evolve as the needs of the community dictate. We invite participation and feedback in the evolution as we make the decisions that affect the future of heliophysics science archiving.

Bio: Dr. James Thieman is a scientist/data systems manager who leads the international Space Physics Archive Search and Extract (SPASE) program. He also manages the continuing development and maintenance of the National Space Science Data Center's information systems. These systems provide access to NASA space science data for scientists and general public users around the world.

P23 - Science Archives over the Past Centuries. What can we learn?
J. Zender, European Space Agency, ESTEC

Introduction:
Within natural science – as opposed to social science and formal science – human kind studies the universe mostly using data from available sensors. Natural science is as old as history. The sensor development from ears and eyes to complex micro-technology instrumentation is rapidly evolving and nano-technology instrumentation for scientific measurements will be available soon.

I go back in time studying, if 'archiving' was seen as a task in previous generations and how they have solved it.

Cometary and Meteor Science:
Within astronomy, the observations of comets and meteors – individuals and meteor showers – go back in history for at least 2000 years. Several others have studied the available material, see [1], [2] and [3], and came up with predictions and models. Within the cometary and meteor science field, enough literature is available to study the following questions:

- Were 'scientists' in previous centuries aware of 'archiving'?
- What was the approach towards 'archiving', in respect to technology and preservation?
- What was the environment, in which the 'archiving' took place, e.g. who did finance this work?
- Can we deduce information from the past for the 'archiving' work in natural sciences today?

Study Approach:
To answer to above questions, the nomenclature used within archiving will be clarified, especially when comparing activities from several centuries. Then, I will define parameters that are of importance for archiving. Each parameter will then be compared to the individuals or institutes over the past. The outcome of the study is not yet know as of writing this abstract.

References:
[1] "Comets - A Chronological History of Observation, Science, Myth, and Folklore" D. Yeomans, Wiley, New York, 1991,
[2] "Cometography: A Catalog of Comets: Ancient – 1799", Volume 1, Gary W. Kronk
[3] "The History of Meteors and Meteor Showers", Vistas in Astrono. 26, 325-345