EXAMPLES OF EXISTING ARCHIVES

(This annex is **not** part of the Recommendation.)

The following are copied from OAIS v1 (2002) and may be out of date

1 Planetary Data System Archive

1.1 DOMAIN

Domain and Consumers. The Planetary Data System (PDS) is chartered to provide data archiving services, data access and expert help to the NASA-funded planetary science community. The PDS is a distributed system with a Central node at the Jet Propulsion Laboratory and discipline nodes (imaging, geosciences, atmospheres, planetary plasma interactions, small bodies, rings) located at universities around the country. The early focus has been on restoring historical mission data and has produced several hundred CD-ROM volumes containing about 80 per cent of the important planetary data archives. There has been an increased emphasis on providing access to the general public for educational outreach over the past several years.

Data Producers. Planetary data sets originate with NASA flight project data management and science teams (new data, some restorations), individual scientists (newly processed or value-added data) or via the PDS discipline nodes (restorations and value-added data). At least 50 percent of the PDS resources have been devoted to restorations over the past seven years, with several more years of work needed to capture all historical data.

1.2 INGEST PROCESS AND INGEST INTERFACE

The PDS has developed a very formal interface with the major data producers (flight projects). This interface is documented in the Data Preparation Workbook and involves substantial interaction between node personnel, data engineers and project representatives. A Project Data Management Plan, signed by the PDS project manager provides the basic project data description and agreement to deliver to PDS. Since about 1993 all NASA announcements for Planetary investigations or analysis require that all data generated be delivered to PDS in conformance with PDS standards.

Submission Agreements

Projects provide a Project Data Management Plan (PDMP). Sometimes a more specific document, the Archive and Transfer Plan, supplements the PDMP, providing extended product documentation and a schedule of deliveries.

Individual scientists can propose to be ‘data nodes’ and receive funds from a PDS discipline node for preparing restored or value-added data sets for inclusion in the archive. There is no formal submission agreement for data nodes.

The PDS discipline nodes each maintain a list of outstanding restorations. These are worked-off based on their priority within discipline. At some point this list will be completed and only

new project or data node data sets will be ingested into PDS. There is no formal agreement associated with discipline data restorations.

Each data set that is identified for ingest in PDS is assigned to a Central node data engineer. It is the responsibility of the data engineer to see that all archiving steps are completed. The archiving steps are called out in the PDS Data Preparation Workbook.

Typical Data Delivery Session. Typically a delivery session will consist of a single data set contained on one or more volumes of CD-ROM or CD-Recordable media. A data set is defined within PDS to be a group of homogenous data granules at the same data level (raw, decalibrated, reduced) which differ only in time of acquisition and major category of target body. For example, the images of Jupiter taken by both Voyager spacecraft comprise a single data set. The standard process includes up-front negotiations between PDS and the provider; the production of test products which are evaluated in the peer review; revised final test products which are validated by the data engineering staff at the central node; approval and production of CD-ROM volumes; distribution by the appropriated discipline node or the central node; entry of the data set into the PDS central catalog; and entry of the data set into the NSSDC ordering system.

Transformation Process. In most cases the original data formats are maintained when data is brought into PDS. This allows existing software tools to continue to be used with the data. Much of the data preparation involves carefully documenting the data format and preparing metadata (granule labels, index files and catalog templates).

Validation. Validation is generally performed as part of the peer review of a product or by using validation tools. In some cases (for example, Magellan), the project develops its own internal validation process. The main validation tool of the PDS is the Volume Verifier. This program is run by the Central Node data engineers on each product delivered from a project or a data restoration. It validates the format and content of all product labels, and validates data files using checksums.

Security. The only area where any special security issues exist involves the receipt of proprietary data. Some projects have one-year proprietary periods before data in released to the science community. The PDS policy is to avoid receipt of any proprietary data sets during the proprietary period.

1.3 INTERNAL FORMS

The PDS has developed standards for documenting data sets (templates) and individual data products (PDS labels) using a keyword=value label system called the Object Description Language (ODL). Recommendations are also provided for volume organization and data product formatting to optimize the utility of resulting data products.

The PDS standards are specified in the PDS Standards Document. Standard documentation requirements include templates describing the data set, instrument, mission, etc. These templates are included on data volumes and also entered in the PDS high-level catalog. Standard terminology is maintained in the Planetary Science Data Dictionary (PSDD), which is jointly maintained by the PDS and the multi-mission ground data system. The metadata values for new data products are carefully compared with the PSDD and existing values used wherever possible. Additions are made to the PSDD to add new standard values to

accommodate new data sets and when justified new keywords are added to the PSDD. Data products can have specialized metadata values which are not cataloged in the PSDD.

The PDS product labeling system is flexible enough to allow nearly any data structure to be described. Labels can be attached to the beginning of the data file or detached in a stand-alone text file which points to the data file. In some cases a single label file is used to describe multiple data files. Detached labels can be used to describe data stored in other formats (FITS or HDF, for example). In cases where complicated raw telemetry formats are stored the Software Interface Specification (SIS) for the product is included in lieu of descriptive labels.

Archive Volume Components

An archive quality data set is required to contain the following components:

- AAREADME.TXT - Text summary of data contents;
- VOLDESC.SFD - Standard volume label;
- VOLINFO.TXT - Text description of data contents;
- CATALOG - DATASET.CAT, MISSION.CAT, INST.CAT;
- INDEX - ASCII index for each granule on the volume;
- SOFTWARE - Software needed to interpret/display the data;
- CALIB - Calibration data sets;
- BROWSE - Browse products for this volume.

Peer Review. All restoration and data node-produced data sets are required to undergo a peer review before acceptance as archive products. Products produced by flight projects do not go through a formal peer review process. In general, there is ongoing negotiation between the data engineer or the discipline node staff and the data producer. The peer review team consists of a number of scientists familiar with the data set, the discipline node leader and one or more data engineers. All product documentation and sample products and software are supplied to the peer review group for evaluation. The peer review group determines the adequacy of documentation and quality of the data products and either approves the product or provides a set of liens which must be fixed prior to approval. The PDS nodes and data engineers have access to a Volume Verifier tool which aids in validating the quality of an archive volume. The volume verifier checks internal checksums, verifies that the index contains entries for all data products and validates the volume templates as well as the descriptive keywords supplied for each product.

Delivery Media. Discipline restorations and data node products are recorded on CD-ROM or CD-recordable media as a standard practice. Flight projects are urged to provide archive quality products on CD media but may not be able to due to funding constraints. Products delivered to PDS on magnetic tape media are assigned to the PDS restoration queue. It is the goal of PDS to convert all data sets to CD-ROM or CD-recordable media which is replicated at a separate geographic facility. This separate facility is generally the National Space Science Data Center (NSSDC) at Goddard Space Flight Center.

1.4 ACCESS

Nearly all access to PDS data sets is via the CD-ROM volumes which are distributed to the entire research community. Large discipline node data collections including a substantial volume of CD-ROM data are accessible via the Internet. Several of the discipline nodes have developed on-line retrieval systems customized to meet the needs of their discipline scientists.

Finding Aids. The Pilot PDS devoted substantial resources to designing a central catalog system and distributed query and processing capabilities at the discipline nodes. These efforts were largely dropped as the Planetary Data System focused on data restoration rather than data access. In general, most of the user community already had home grown tools for data analysis and were most concerned with getting access to the data sets. The growth of the user community due to Internet and increased usage of CD-ROM readers has spurred to prototype a more consistent finding aid. The PDS Navigator has been developed for selecting images from the Clementine mission. It includes three components, a forms-based traditional database retrieval capability, an image-based retrieval and a text-based retrieval.

Security. The high-level PDS catalog can be accessed via a group account. Most of the data access services at the discipline nodes require the user to obtain a valid account on the node computer. All PDS data sets are certified "General Technical Data Available" by the Department of Commerce and are distributable worldwide.

Customer Service/Support. The order function of the PDS is distributed. Data inventories are kept at NSSDC, the PDS central node and at each discipline node. In general each site serves a special group of users:

- discipline node - members of the NASA funded discipline;
- central node - other NASA scientists and engineers, other agencies;
- NSSDC - other scientists, agencies, public and foreign users.

The PDS Operator at the central node handles requests for PDS documentation or standard data products. The discipline nodes handle data requests from within their discipline and also provide expert help in the utilization and interpretation of the data. Access to tools is also provided.

The vast majority of data dissemination is done via CD-ROM disc. Several hundred copies of over 500 titles have been distributed to date.

Event Based Orders. Nearly all PDS distribution is done via event based orders (i.e., subscriptions or standing distribution lists). It is the responsibility of each discipline node to maintain a distribution list for its discipline scientists. This list determines the order amounts for most CD-ROM titles. The central node maintains a distribution list for engineering and management personnel and for other external recipients (reciprocal distribution, software developers).

Media/Network Use. Nearly all final products are delivered to the user community on CD-ROM. Archival products that need not be widely distributed are stored on CD-Recordable media, with a duplicate copy provided to the NSSDC. Most PDS data is available for downloading via anonymous ftp connection to a large CD-ROM jukeboxes at the central node and the imaging node.

Data Manipulation. Each discipline has a suite of government developed analysis tools which can be applied to the discipline data sets. These software packages are available for UNIX workstations or VAX VMS platforms. Several nodes provide the user a menu of processing functions that can be performed on selected data and will carry out requested processing and provide the results electronically or via media. The most widely used commercial tool is IDL.

Pricing Policy. The PDS distributes data to legitimate NASA researchers for no charge. There are no charges for on-line computer usage or data processing to NASA researchers. The NSSDC distributes CD-ROMs for $10 per volume.

1.5 SPECIAL CHARACTERISTICS

PDS has invested a substantial engineering effort in its common data dictionary, data standards and procedures for preparing archival quality data sets. By having these standards in place the PDS is able to demand better quality data sets of its data providers.

2 National Archives and Records Administration's ELECTRONIC AND SPECIAL MEDIA RECORDS SERVICES DIVISION

2.1 DOMAIN

Domain and Consumers

The Electronic and Special Media Records Services Division is the organization within the U. S. National Archives and Records Administration (NARA) that appraises, accessions, preserves, and provides access to federal records in a format designed for computer processing. NARA serves as the archives for the records of the United States federal government. Consumers for this data are as diverse as the electronic records they seek to access and range from individuals seeking to assert their rights to other government agencies to academic researchers, private consultants, media personnel, and a wide variety of other users.

Data Producers

Originally this data is produced (created or received) by agencies of the U.S. federal government (producers). The data may concern virtually any area or subject in which the government is involved. They may come from a variety of computer application such as data processing, word processing, computer modeling, or geographic information systems. They can include records made directly by government employees or indirectly through government grants and contracts.

Special Features

The most noted special feature of NARA's Electronic Records program is the diversity of the collection of more than two billion logical data records in over 129,000 data sets from more than 100 bureaus, departments, and other components of executive branch agencies and their contractors and from the Congress, the Courts, the Executive Office of the President, and numerous Presidential commissions. A small portion of the data originally were created as early as World War II. An even smaller portion contains information from the nineteenth century that has been converted to an electronic format. Most of the data, however, has been created since the 1960s. The major types of holdings and subject areas include agricultural data, attitudinal data, demographic data, economic and financial statistics, education data, environmental data, health and social services data, international data, and military data.

Scientific and technological data already transferred to NARA include the National Register of Scientific and Technical Personnel; the National Engineers Register; the 1971 Survey of Scientists and Engineers; major portions of the National Ocean Survey's Nautical Chart Data Base; numerous Environmental Protection Agency series relating to pesticide use, hazardous

wastes, and pollution abatement; the Nuclear Regulatory Commission's Radiation Exposure Information Reporting System; biometric data sets and epidemiological studies (such as the National Collaborative Perinatal Project) from the National Institutes of Health, the Centers for Disease Control, and the National Center for Health Statistics; and text from presidential commissions on Three Mile Island, coal, and the Space Shuttle Challenger Accident. NARA recently ingested the e-mail of the Executive Office of the President, including the White House Office and the Office of the Vice President for the period from 1986 through January 20, 2001. While NARA's scientific and medical holdings are rich and varied, they do not fully reflect the extent and diversity of federal activity in this area.

2.2 INGEST

The ingest process begins with producers (records managers and records creators in federal agencies) inventorying all electronic records and determining how long to retain the records for current agency business. The next step in the process is for the producer and NARA to develop a *Request for Records Disposition Authority*, Standard Form 115 (SF 115), the formal submission agreement for all federal records. Here information on the content, retention and disposition, and the availability and extent of documentation and related reports is listed in the context of the producer's business needs for the information. Data with continuing value are listed as permanent and the timing and frequency of their transfer to NARA is established. The producer submits the SF 115 to NARA for its review and appraisal. NARA appraises electronic records items on all SF 115s. Identifying permanently valuable electronic records for retention by NARA's Electronic and Special Media Records Services Division involves cooperation between NARA and the producers. Through the process of scheduling and appraisal, NARA identifies and selects the electronic records it judges to have enduring value. NARA evaluates electronic records in terms of their evidential, legal, and informational value and their long-term research potential. Some of the factors in this appraisal evaluation include estimation of past, present, and probable future research value within the context of the data's origin and current use and its impact on federal programs and policy. Administrative and legal value, as well as the potential for linkage with other data, may bear on the decision. Unaggregated microlevel data sometimes has the greatest potential for future secondary analysis. Once NARA determines that the records have enduring value, it then determines whether the records should be preserved in electronic format.

Submission Agreements

The actual Submission Information Package (SIP) between NARA and the agency that creates or receives the data is a *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States,* Standard Form 258 (SF 258) accompanied by the data object(s) and sufficient documentation and descriptive information to use the data. The SIP transfers physical and legal custody of the electronic records from the producer to NARA. This agreement is the end product of the ingest process described above. The SF 258 also contains any restrictions on access to the data that conform with exemptions listed in the U.S. Freedom of Information Act (FOIA). NARA enforces all legitimate restrictions on access. At the same time NARA also works with the producer to determine if any "disclosure-free" version of the data can be produced for consumers.

Typical Delivery Session

This inventorying, scheduling, and appraisal process specifies the data object(s) and related metadata and documentation to be transferred, and establishes the timing and frequency of

submissions. Specific instructions for how the data are to be organized and when they should be submitted are established in the *Code of Federal Regulations* (36 CFR 1228.188). All data should be transferred on either open reel magnetic tape, tape cartridges, or CD-ROM. NARA negotiates acceptance of other forms of magnetic media such as class 3590E or DLT, with producers. The CFR sets the specific technical requirements in terms of format, block size, and extraneous characters. While the current regulations also require that all SIPs should be transferred in a software-independent format, NARA staff recognize that the research potential and utility of some data would be significantly reduced if they were transferred in such a format. In such cases NARA works with the producers to determine the best mode of transfer.

What are the Information Objects that are Delivered? Producers typically will transfer a series consisting of one or more data sets with the related documentation which minimally should include the record layout and codes, methodology statements, technical information about the data including number of records and size. Ideally, the SIP also includes associated analyses and reports. Increasingly agency-created metadata also is included. The majority of electronic records come as flat files of data; increasingly, however, text files and output from data base management systems, and geographic information systems also are transferred.

What are Collections? NARA organizes all Archival Information Collections (AIC) on the basis of Provenance and Original Order. Provenance maintains the identity of an Archival Information Package (AIP) or an AIC and preserves as much information as possible about its origins and custodial history. Within NARA this is accomplished through the use of Record Groups which reflect the structure of the federal government and subgroups and sub-subgroups which place the AIPs and AICs within the producer's place within its agency. Original order argues for maintaining the contents of an AIP or AIC in the order developed and used by the producer. This helps reveal the producer's organization and how it used the data objects and can provide additional information to consumers. For electronic records, "original order" is expressed in the logical structure of files and databases and in the indexing which the producer used. Within NARA the basic unit for arrangement and description is the series which is an AIC that can include a number of related AIPs.

What Descriptive Information is Provided? The extent and quality of the descriptive information provided by the producer varies from quite sketchy to extremely detailed. NARA staff attempt to flesh out the producer-created descriptors with AIC level descriptions, title list entries, abstracts, and Dissemination Information Packages (DIP) and to provide the descriptive information in a variety of formats to reach different consumers.

What sorts of Validation Objects are Provided? Producers are required to transfer metadata and descriptors adequate to access, process, and interpret electronic records. For formatted data files the DIP must include a record layout with appropriate field definitions and codes. It frequently also includes methodology statements, input documents, data entry instructions, processing directions, sample outputs, reports and analyses of the information and system manuals.

What Transformation Processes are Performed Prior to Storage

What Metadata is Created? The most extensive metadata product created by NARA is the DIP. It includes an Introduction which can discuss the origin, creation, and administrative uses of the data object(s), list related objects that are or will be available, and discuss characteristics of the data that could cause problems for consumers based on initial validation processes. The

DIP also can include sample printouts of the data and tables and reports related to computer verification of the data. NARA also captures metadata on record layouts, domains, ranges, and links between files in a metadata database as a byproduct of the automated verification process. Other metadata created by NARA staff include AIC descriptions, formatted abstracts, title line entries, and collective descriptions which place the data in a broader context. Increasingly, producer-created metadata is part of the SIP transferred to NARA.

What Validation is Performed? NARA's initial ingest procedures include creating a new preservation master and backup copy of each data object on new certified media to ensure the best physical media for long-term storage. This procedure includes a 100% byte for byte comparison between the SIP media and the AIC media. At this time staff perform automated verification of the data contents with the record layout and codes, and of the physical structure including the number of records, blocks, and bytes. Staff also perfect the DIP to facilitate secondary use of the data.

Security

All AICs are maintained off-line. Consumers access only DIPs. The AIC preservation master and backup copies are maintained in separate secure stacks at two different physical locations. AICs that require additional security measures, for example Bureau of the Census data subject to restrictions imposed under Title 13 of the *United States Code* and national security classified information restricted under Executive Order, are afforded the appropriate level of protection. NARA is moving to provide enhanced access to selected data onsite by providing reference copies on a wider variety of media and by providing a broader range of services and output products. This may include use of vendors who can provide enhanced access to the holdings utilizing "value-added" services.

2.3 INTERNAL FORMS

How do you Store your Data? All preservation master and backup copies of AICs are stored on newly certified class 3480 magnetic tape cartridges. Some of the holdings have not yet been migrated from nine-track, 6250 bpi open-reel magnetic tape. Data are received and stored temporarily on other media including diskettes, 4mm, 8mm, CD-ROM, DLT, and various removable hard drives, although not all of these media conform with regulatory requirements.

Migration (Data). Based on recommendations from the media manufacturers, the National Technology Alliance, the National Institute of Standards and Technology, and various standards organizations, NARA has been migrating its AICs to new class 3480 magnetic tape cartridge when each media unit is ten-years old. NARA continues to reassess storage media. NARA anticipates storing larger AICs on class 3590E and/or DLT cartridges as appropriate.

Migration (Metadata). Metadata has been stored in a variety of formats depending on the original format transferred with each AIC. Traditionally most metadata existed in textual format. The metadata captured in the verification process is maintained in a relational database. There are no current plans for migrating from this format, although the metadata can be exported in flat file format. NARA encourages data producers to create and transfer metadata in electronic form. In the near future NARA will begin scanning and digitally converting metadata so it can be preserved and provided in an electronic format along with the data.

Migration (Format). The *Code of Federal Regulations* requires data producers to transfer all data in ASCII or EBCDIC with all extraneous characters removed from the data except record

length indicators or tape marks and blocked at no higher than 32,760 bytes per block for open-reel and 37,871 bytes for class 3480 magnetic tape cartridge. When CD-ROM is used they must conform to ISO 9660 standard and the data must be in discrete files containing only the permanent data. Additional software files and temporary files may be included on the CD-ROM. The CFR also requires all electronic records to be transferred in a software-independent format. NARA works with data producers who cannot meet those requirements to determine the most appropriate transfer and storage formats.

2.4 ACCESS

What Finding Aids are Provided?

Information about the holdings are available in multiple levels of detail and by multiple sources as a way to provide various consumers with information about NARA's holdings. The least specific detail is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States* where AICs are described in the context of the larger holdings from a producer. Other collective descriptions include *Information About Electronic Records in the National Archives for Perspective Researchers,* General Information Leaflet 37, which also is available on the Division's homepage (http://www.nara.gov/nara/electronic), and a title list of data sets available on the Division's homepage and as a printout. Specific AIC descriptions were created as formatted metadata for a portion of the Division's holdings for inclusion in a proposed automated description data base which has not been implemented. The most detailed description for any data set is the DIP. Each DIP may contain a narrative describing the data file(s), the record layout and codes for the data, a methodology, sample input forms and questionnaires, annotations regarding the data validity, and a bibliography. The Division also has established an email site (cer@nara.gov) for queries regarding its holdings and services.

Security.

All NARA holdings are maintained in environmentally controlled closed stacks which are accessible only by NARA staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. The Division's national security classified data sets are in separate environmentally controlled stacks approved for the storage of classified information. All processing is performed in limited access processing rooms at NARA or at other government computer centers. Computer processing is done on closed systems which require both a registered logon and personal identification number or password to access the system. Researchers do not have direct access to any AIC. Presently they access copies of the data that they have purchased for their own use.

Customer Service/Support.

The Division has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The staff responds to both general and specific inquiries by telephone, letter, email, or in-person visit and fills orders for copies of specific data and their DIPs. For a limited number of AIPs the staff also provides information from records to respond to researcher requests. The staff also functions as a filter between researchers and the data producers when problems develop in understanding or interpreting the data. The staff develop

a variety of informational material about NARA's holdings and services, much of which is available online.

Do You Support Subscriptions?

NARA accepts standing orders (subscription) for electronic records that it receives on a regular, periodic basis from producers of the Federal government. Under current NARA regulations all subscriptions must be prepaid prior to shipment.

What Media/formats do you use?

Currently NARA provides DIPs on a variety of magnetic media including nine-track open-reel magnetic tape or class 3480 magnetic tape cartridges encoded in ASCII or EBCDIC, labeled or unlabeled and written to the maximum block size requested, diskettes for smaller data sets and CD-ROM . NARA also can provide an exact copy of records in nonstandard formats, if the agency transferred them this way, but it cannot validate or verify the contents of these files. In the past these other formats have included packed decimal, zone-decimal, binary, National Information Processing System (NIPS), Statistical Analysis Software (SAS), Statistical Package for the Social Sciences (SPSS), or OSIRSIS. On-line transfer of SIPs to NARA via File Transfer protocol (FTP) was implemented in 2001; providing DIPs via FTP may occur as early as 2002.

What Transformation (Value Added) is Provided?

NARA currently preserves AICs as received from the producers; it does not routinely provide customized DIPs or other value-added services beyond computer verification of the AIC contents and enhanced documentation. Planned enhancements include value-added services such as custom DIPs and data transformation.

Pricing Policies.

NARA uses a cost-recovery fee schedule developed by the National Archives Trust Fund. The 2001 charges include a basic order handling fee of $89.00 with an additional fee of $9.00 for each file. Media costs range from $2.50 for diskette to $22.50 for a 9-track open reel. Paper reproductions cost $10.00 for a minimum order of up to 20 pages; additional pages are $0.50 per page. If the documentation is on microfiche, copies are $2.50 per fiche.

Dissemination Security.

The same security considerations developed in relation to Access apply to Dissemination. NARA's national security classified data is made available only to researchers who have both the appropriate security clearances and the appropriate need-to-know. Other restricted data are made available only with prior written approval of the creating agency or under the terms of the restrictions which must be supported as a legitimate exemption under the Freedom of Information Act.

2.5 SPECIAL CHARACTERISTICS

NARA's Electronic and Special Media Records Services Division has a diverse collection which reflects the diverse activities of the federal government. The staff shape the holdings through the process of scheduling, appraisal and accessioning. Currently, NARA acquires less than one percent of all federal records created in an electronic format. The timing of the transfer

of electronic records from the creating federal agency to NARA is negotiated with the creator to ensure that the records are available for agency use for as long as necessary for current business and that they are transferred to NARA as soon as practicable to ensure their long term preservation for secondary use. NARA is the only federal agency with an explicit archival mandate for Federal records and thus the only Federal agency that preserves and provides access to a wide range of historically valuable records for the indefinite future. As such it is an archives of last resort for the electronic records of some federal agencies which undertake an active data dissemination function while there is a researcher interest in the data but whose mandate ceases or may cease once the demand wanes or ceases.

3 LIFE SCIENCES DATA ARCHIVE

3.1 DOMAIN

What is the domain and who are the customers of the Archive and who are the producers of the data? What are the special features of this archive?

The Life Sciences Data Archive (LSDA) is responsible for collecting, cataloging, storing and making accessible the data of NASA funded Life Sciences space flight investigations. There are two general goals for NASA space life science research; one, to find counter measures to problems encountered by human bodies as a result of space flight, and two, to broaden the understanding of the effect of gravity on living systems. The LSDA's designated consumer is the life sciences research community, but it is also used by students, educators and the general public. The data archived in the LSDA is produced by both intramural and extramural investigators funded to perform flight experiments through NASA grants. It is anticipated that the archive may grow to include data from investigations which are completely ground based.

The LSDA is a distributed archive with responsibilities distributed to LSDA Nodes at NASA Centers with life sciences activities. The LSDA Project Nodes at Ames Research Center (ARC), Kennedy Space Center (KSC), and Johnson Space Center (JSC) are responsible for the collection and cataloging of data. The LSDA Central Node is responsible for importing the data from the Project Nodes, integrating and providing access to the data via the World Wide Web. The Central Node is also responsible for maintaining the LSDA Data Dictionary and coordinating development and maintenance of the archiving system.

The LSDA contains animal, plant, and human space flight data. This archive is notable in that it contains a unique collection of data describing, in considerable detail, biology experiments carried out in space by NASA over the past thirty years. The nature of the data is highly varied and spans many life science disciplines.

The LSDA is also unique in that it provides both digital and non-digital information. The non-digital data may be either reproducible or non-reproducible. Examples of reproducible, non-digital data are video and audio tape. An example of non-reproducible, non-digital data is a biomedical sample.

3.2 INGEST PROCESS

Submission Agreements

There are two types of data producers to the LSDA: the NASA Flight Project offices that designs the hardware and flies the experiment, and the NASA-funded Principal Investigator (PI).

To acquire data from the NASA Flight Project offices, the LSDA Project Nodes work closely with them to acquire data during flight operations. The LSDA assists the NASA Flight Project Offices in distributing this data to the Principal Investigators and gathering it as an archival product. As the LSDA is relatively new (1993) there is also archiving of past missions being done on a funding available basis.

To acquire data from the NASA funded Principal Investigator, there are a couple of methods of data collection currently being used depending on the 'age' of the experiment. For previously flown experiments (prior to 1994) there is an informal submission agreement between the LSDA and the PIs that is based on cooperation, and is not binding. For experiments being selected for flight (after 1994) the funding agreements include a contractual stipulation that the Principal Investigator must supply the LSDA with raw data, analyzed data and a final science report.

These funding agreements are finalized when proposed investigations are selected for flight. At this time the PIs are sent a letter informing them, that upon acceptance of funding they will be responsible for delivering the data collected as part of their investigation in a form usable by the sciences community after their one year post flight proprietary period.

Submission of data to the LSDA begins one year post flight. To assist in data submission, the LSDA Project nodes send the Principal Investigator a Data Inventory package including forms and instructions. The Principal Investigator fills out the data inventory forms and returns them to the LSDA Project Node. The Project Node then contacts the PI to begin data submission. In order to clarify the 'usable form' requirement throughout the entire LSDA project, the LSDA is in the process of developing a post flight data reporting handbook which explains exactly how the data should be provided to the archive.

Typical Submission Session

A typical Submission Information Package (SIP) consists of two parts: 1) the Data Inventory forms, and 2) actual data or Content Information. The inventory forms are made up of Preservation Description Information (PDI) (i.e., treatments, parameters measured, research subjects and IDs, date/period of collection, collection location, analysis phase, comments) and Descriptive Information (i.e., title, description, keywords). The Content Information consists of physical samples, spreadsheets, final science reports, published articles, procedural documents, crew logs, photographs, video tapes, analog tapes, digital or printed images, and other types of digital data files (e.g., HRM).

Upon receipt by the LSDA the SIP will be cataloged with Descriptive and Packaging Information added, including; experiment and mission ID, Principal Investigator and Co-Investigators name, and other Descriptive information.

Collections

Several SIPs will then be combined to comprise an Archival Information Package (AIP) of one experiment. The Descriptive Information and PDI are entered into a database comprised of LSDA-approved fields and use valid values whenever possible. The Preservation Description Information is developed by the LSDA personnel at the LSDA Project Node responsible for obtaining the data. The Preservation Description Information provides layers of metadata for

the data collection that describe the experiment, mission, hardware, personnel, sessions, biospecimen, and research subjects from which the data was collected.

It is anticipated that future uses of the archive will involve the creation of Archival Information Collections (AIC), combining several AIPs based on discipline or measured parameters.

Transformation Processes

In most cases the Content Information is kept in its original submitted form. Exceptions to this case include data submitted on outdated media requiring transfer to current media. As little transformation as possible is performed on the data at ingest in order to keep costs down and to insure the integrity of the data. There are some instances where the data has been collected in an application format that is not widely available and in this case the Project Node will transform the data into a more commonly accessible format. (e.g., spreadsheets created in Supernova are migrated to MS Excel).

After the AIP is created, the information goes through a validation process. This post-entry validation is accomplished by a second check of the data by the LSDA Project Node Manager. AIP validation is further ensured by sending the completed catalog entries to the data producer (Principal Investigator, Flight Project Offices) for verification. The data producer reviews the information, makes corrections or additions and sends the information back to the Project Node. Edits are then made to the records and the information is once again printed and sent to the data producer. This process is repeated until the Principal Investigator is satisfied that the experiment data is accurately represented. At this point the data producer signs and returns a verification letter to the Project Node. The AIP is now ready for review by the LSDA Project Scientist and LSDA Change Control Board before it is placed in the public record. The Projects upload the data to the Central Node, it is integrated into an intranet server available only to the LSDA CCB and this group, including the LSDA Project Scientist, reviews the data for overall form and cogency. After a two week review period the data is moved to the public web site unless change requests are logged by the LSDA CCB. If CRs are issued they are resolved before any of the data is made available to the public.

3.3 INTERNAL FORMS

Storage

The LSDA back-up and storage procedures vary between LSDA Node types. The LSDA Master Catalog and on-line data reside on a Microsoft SQL Server. These are backed up daily to tape. At the LSDA Project Nodes most of LSDA's data and metadata are stored on magnetic disks and backed up to tape. Long term storage is provided on CD-ROM. Biospecimens are stored in -80 degree freezers.

AIPs are stored as a piece of Content Information (a spreadsheet, word processing document, strip chart or biospecimen) which is linked to the Descriptive Information which is stored in a database record. These AIPs are linked, through the database, into AICs via an Experiment Number. A space life sciences experiment is, in this sense, an AIC. It is a collection of tens or hundreds of AIPs.

Migration

The LSDA migration procedures are still in a developmental phase, but there is some ongoing data migration. LSDA Project Nodes are in the process of converting information on outdated media (RA60s, RL02s) to CD-ROM format.

Migration of application formats (e.g., MS-Excel) and in particular, version changes, is an area of concern. The cost of continually updating all LSDA spreadsheets to the current versions is prohibitive, and storing and making available the application is also expensive and complicated. A universal read-only format such as Adobe PDF might be the solution, but it is a proprietary format and as such, its life span is uncertain.

3.4 ACCESS

Finding Aids. Users access the LSDA through the World Wide Web (WWW) (http://lsda.jsc.nasa.gov/) where they search and retrieve information via the Master Catalog. The Master Catalog is a relational database with a WWW forms interface, and it allows users to search Descriptive Information across experiments and Missions to find data that meets their search criteria. Users can search within ten information groups; Experiments, Missions, Data Sets, Hardware, Documents, Personnel, Specimen or Subjects, Data Collection Sessions, Biospecimens, and Images.

There is currently no method available for searching data at a 'sub-AIP' level. The AIP record contains a considerable amount of detailed, searchable data so that a collection of AIPs could be found for a particular manual sub-search.

Event Based Order. The LSDA does not support event based orders (or subscriptions) since the Master Catalog is accessible to all users. Most data is made accessible to the user through links in the catalog to an anonymous FTP site from which the data is downloaded. If data are in non-digital format, but are reproducible (i.e., hardcopy documents, or log books), users may request them through on-line ordering forms available in the Master Catalog. The requested information is reproduced via photocopying and shipped US Mail to the requester. There is an e-mail update notification service offered to users.

Media/Formats. The LSDA provides some data on CD-ROM format, but otherwise most data is provided on-line. The LSDA also contains unique non-reproducible pieces of data such as microscope slides and space flight biospecimens. These unique resources are provided to a requester after a scientific proposal has successfully undergone peer review. Biospecimens are used to produce original data which is then ingested into the archive.

Pricing Policies. LSDA data that has been verified and cleared for release is available to the public, free of cost, through the Internet. If significant requests are generated for hardcopy documents, a processing fee for copying the document may be charged. As yet this has not been determined. In the future CD-ROMs may be generated for providing large data sets. These CDs would be priced for cost recovery only.

3.5 COMMON SERVICES

Customer Service/Support. The LSDA provides user support for questions and problems concerning the Master Catalog (on-line data request system) and for science questions about the data being provided. The primary means of user feedback and support is through the LSDA Central Node. Questions are addressed to the LSDA through on-line 'What do you think?' links located throughout the system. From these links a WWW forms interface allows users to submit questions. Specific questions about the Content Information are currently answered by

the NASA Life Sciences Acquisition Scientist and the LSDA Program Scientist. Questions which can not be answered by these individuals are forwarded to the LSDA Project Node which collects the data. In some instances questions are forwarded to the Principal Investigator or NASA Flight Project Office.

Security. Overall security procedures stipulate that all digital data are backed up on a daily basis with off-site storage. Data on magneto-optical disks are stored in a locked file cabinet in a cipher locked room. Access to on-line servers is controlled through the use of password and/or address port filtering. Only data that is fully validated and approved for release is placed on publicly accessible servers.

The LSDA does not have any special security concerns for access to the Central Node. It is available to anyone with access to the WWW.

The LSDA has strict security measures for data from human subjects which require sensitivity and secure handling due to the Human Data Privacy Act. Only mean-pooled human data is made available to the public.

4 NATIONAL COLLABORATIVE PERINATEL PROJECT (NCPP) 1959-1974

4.1 DOMAIN

Domain and Customers

The National Collaborative Perinatal Project was a multi-institutional, multi-year study of pregnant women and the children born from those pregnancies to provide baseline information useful for later determining the causes of neurological diseases which appeared in a portion of the studied population. The data came from medical histories, examinations, and observations. The records also contain socioeconomic, family history, and family health information. The data are used by a variety of medical and other researchers.

Data Producers

The predecessor to the U.S. National Institutes of Health's National Institute of Neurological Disorders and Strokes (NINDS) began the National Collaborative Perinatal Project in 1958. Fourteen university-affiliated medical centers across the United States participated in the study. Between 1959 and 1965 each cooperating medical center collected information on between 300 and 2000 pregnancies each year for a total of 55,908 pregnant women utilizing their clinic services. This represented between 14% and 100% of the women utilizing these services depending on the sampling rate employed at each clinic. The final population was reduced to 39,215 due to miscarriages prior to twenty weeks, 445 multiple births, exclusion of subsequent or repeat pregnancies, and deletion of incomplete records due to women withdrawing from the study prior to its completion. The children were given neonatal examinations and follow-up examinations through eight years of age. The last examinations were conducted in 1974. The computer data files resulting from the research that NINDS transferred to the National Archives and Records Administration (NARA) consist of approximately 6,200,000 records organized into a Master File, a variable file, and eighteen work files, one of which consists of thirteen distinct data files.

Special Features

The Collaborative Perinatal Project was a longitudinal multi-disciplinary research effort which sought to relate the events, conditions, and abnormalities of pregnancy, labor, and delivery to the neurological and mental status of the children of these pregnancies and their siblings through eight years of age. The study sought to link any later appearance of cerebral palsy, mental retardation, learning disorders, congenital malfunctions, minimal brain dysfunction, convulsive disorders, visual abnormality, or communicative disorders to patterns during the perinatal period in order to develop strategies for prevention and intervention. The sample population is large enough so that statistically significant numbers of such disorders would appear in the children. Study of the records relating those children could result in the development of predictive factors and possible preventive care or intervention actions which could reduce future incidence rates.

The data are available in two formats; microfilm of the individual case files for the mother and child of approximately 270 pages per case file, and the computer data files. Access to the microfilm and two of the computer data files is restricted because they contain personal identifiers. The National Archives has created a public use file for the Master File and Work File 16: Serum Specimen Inventory.

4.2 INGEST

The ingest process for transferring any federal agency records to NARA begins with the agency identifying the records and assessing their potential evidential, legal or research value. The next step is for the agency to develop a Submission Agreement (Standard Form 115, *Request for Records Disposition Authority),* and submit it to NARA. NARA staff then appraise the records in terms of their evidential, legal, and informational value and their long-term research potential. NARA and the creator then establish a transfer date, negotiate any restrictions on access, and initiate the ingest process.

Ingest for the NCPP computer data was a two phase process. In Phase One, from 1958 through 1974, NIH's NINDS funded the project and the cooperating institutions conducted the research. Contractors accumulated the original examination records, created the consolidated case files, microfilmed the records, normalized the data, and developed the Master File, an extract file of frequently used variables, and special files such as 'refined diagnoses'. The data were stored on 23 reels of magnetic computer tape recorded at 1600 bpi. Prior to 1980 the data were available only to NINDS, the cooperating hospitals, and selected government researchers.

In Phase Two NINDS developed the documentation necessary for more generalized use of the data and negotiated a submission agreement, including access provisions, with NARA. Since the data which could not be released could be made anonymous through creation of a Public Use File, the producer and NARA worked on transferring the data files first.

Submission Agreements

NARA and NIH executed the U.S. Government's standard transfer form, Standard Form 258, *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States,* in mid-1985. This transferred legal custody and preservation responsibility to NARA. A similar agreement for the microfilm examination records was executed in 1990 after NARA and NIH resolved the privacy and access concerns and NARA developed a statistical research form.

Delivery Session

The delivery session was a single transaction in which NIH provided NARA with copies of the 23 reels of magnetic tape containing the NCPP and the related documentation consisting of seven volumes containing the background of the study, the sample, data collection and data processing overviews, record layouts and coding for each variable, sample forms, and a bibliography of all published research through 1985. NINDS transferred the 8000 rolls of microfilm containing the examination records in 1990.

Transformation Process

NARA has maintained and preserved the original data format. The data are in a hardware and software independent EBCDIC format which facilitates wide researcher access. All data were copied to new nine-track open-reel magnetic tape when received in 1985 and are migrated to new media every ten years to ensure Long Term Preservation. The more than 7000 pages of documentation are available in both a page format and in a microfiche format on 75 fiche. The documentation has not been scanned or digitized.

Validation

Validation was performed as part of quality control throughout the life of the NCPP. Extensive use of the data during the life of the project (1958-1974) and its use by NIH approved researchers (1958-1985) provided a second de facto validation. NARA also validated sample portions of the data at the time of ingest. Continuing researcher use also validates the data contents.

Security

The computer data is maintained in environmentally-controlled closed stacks which are accessible only to NARA staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. NARA has created Public Use Files of the restricted data files to prevent unauthorized access to personal and medical data.

What Descriptive Information is Provided?

NARA has prepared multiple levels of descriptive information for the NCPP. These range from entries for each data file in the Electronic and Special Media Records Services Division Title List, an abstract entry for the series, a series description, a full documentation package, to a series-level entry in the three-volume *Guide to Federal Records in the National Archives.*

What Validation Objects are Provided?

During its active life the NCPP established and used elaborate data collection, input and verification procedures. Extensive use also validates the information. NARA's routine transfer and storage procedures also validated the data. The extensive seven-volume documentation includes record layouts and codes, methodology statements, data analyses, and a bibliography of research use.

What Transformation Processes are Performed Prior to Storage?

What Metadata is Created? NARA staff supplemented the documentation with an abstract and introduction discussing the origin, creation, and uses of the data, including an explanation of restrictions on access and the characteristics of the Public Use File.

What Validation is Performed? NARA's accessioning and storage procedures included creating a new master and backup copy on new certified magnetic media and creating a Public Use File of the two restricted data files. Sample portions of each data set also were verified against the documentation.

4.3 INTERNAL FORMS

Storage. NARA maintains separate sets of the master and backup copies of the data and the Public Use Files on newly certified 3480 class magnetic tape cartridges.

Migration (Data). NARA migrated the NCPP from 23 nine-track, 1600 bpi open-reel magnetic tapes it received in 1985 and stored the data on seven nine-track, 6250 bpi open-reel magnetic tape. NARA migrated the data to four 3480 class magnetic cartridge when the media was ten years old.

Migration (Metadata). NCPP metadata is available in textual (7000+ pages) and microfiche (75 fiche) forms. There are no plans to scan or digitize the text.

Migration (Format). The data currently are encoded in EBCDIC with all extraneous characters removed. There are no plans to migrate the format at this time.

4.4 ACCESS

What Finding Aids are Provided?

Information (of varying detail) about the NCPP is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States*, where the records are described in the context of the larger holdings of the National Institutes of Health; in *Information About the Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37; in *Title List: A Preliminary ad Partial Listing of Data Files in the National Archives and Records Administration*; and in the documentation package for NCPP. Much of this information is available on the Division's homepage (http://www.nara.gov/nara/electronic) or by posting an enquiry to the Division's e-mail site (cer@nara.gov).

Security

NCPP, like all of NARA's holdings, is maintained in environmentally-controlled closed stacks which are accessible only by authorized staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. Researchers do not have direct access to the data. Presently they acquire copies of the data on a cost-recovery basis permitting indefinite use of the data for their own purposes.

Customer Service/Support

The Division has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The reference staff responds to inquiries by telephone, mail

correspondence, e-mail, or in-person visits. They fill orders for copies of all or part NCPP and the relevant documentation. The staff also function as a filter between researchers and the NINDS when problems develop in understanding or interpreting the data.

Do You Support Event Based Orders?

NARA's Trust Fund is willing to establish accounts that allow researchers to acquire data that is transferred on a recurring basis. Since the NCPP stopped collection data in 1974 there is no need for a subscription for this data.

What Media/Format do you use?

Copies of the 32 data sets comprising the NCPP are available on seven nine-track open-reel magnetic tapes, six 3480 class magnetic tape cartridges, or two CD-ROM.

What Transformation (Value Added) is Provided?

The NCPP is provided as received from NINDS. NARA has created Public Use Files for the two data files containing personal identifiers in conformance with the Freedom of Information Act and NARA restrictions on access to records whose release might result in unwarranted invasion of personal privacy.

Pricing Policies

Electronic data sets are available on a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently the charge for an exact copy of all NCPP data on a storage reel or 3480 class cartridge is $80.75 when copied to a 3480 class magnetic tape cartridge and $90.00 when copied to a nine-track open-reel magnetic tape. Copies on CD-ROM are $90.00 for the first file and $24.50 for each additional file written to the CD-ROM. Paper reproductions cost $10.00 for the first 20 pages and $5.00 for each additional block of 20 pages. Microfiche reproductions cost $2.10 per fiche.

4.5 SPECIAL CHARACTERISTICS

The National Collaborative Perinatal Project was a prospective study. NINDS expended more than $200 million over two decades to collect information on more than 58,000 pregnant women and their children at fourteen cooperating institutions. It is unlikely that a study of this duration and magnitude will be repeated. The data continue to constitute an important resource for biomedical and behavioral research in many areas of obstetrics, perinatalogy, pediatrics, developmental psychology and other fields.

5 ARCHIVE SCENARIO FOR THE *CENTRE DES DONNEES DE LA PHYSIQUE DES PLASMAS* (CDPP)

5.1 DOMAIN AND CUSTOMERS

The CDPP (*Centre des Données de la Physique des Plasmas* - Center for Data on Plasma Physics) is a new service currently being set up. It has been developed to ensure the long-term conservation and availability of natural Plasma Physics data (magnetospheric plasma, planetary plasma, etc.) for the international scientific community. More specifically, the data concerned is from either ground-based or space-flown experiments in which France has participated or wholly directed. The CDPP is designed around two principal components:

– Technical Activity segment, located on the premises of the French space agency, CNES, mainly in charge of developing and maintaining the archive system. The latter has the following functions: addition of data and metadata to the system, preservation of data and metadata, organization of search and product ordering facilities, and dissemination.

– Scientific Activity segment, located at the CESR (*Centre d'Etudes Spatiales des Rayonnements* - Center for the Study of Space Radiation), a science laboratory near CNES. The CESR is in charge of all aspects relating to scientific knowledge of the data: validating data with its producers, ensuring that the data is useable by the scientific community, setting up added-value services, etc. This Center is also responsible for developing a WWW server to present CDPP services, supplying educational information on Plasma Physics to the general public, and guiding users to access and dissemination functions.

The two complementary segments work closely together.

A number of associated laboratories will be able to join the two main components of the CDPP provided they offer a service (data dissemination or information) relating to natural plasma physics.

The archive system is currently being validated. The service is planned to be made available to the scientific community on September 1999.

Data Producers. Data producers are mainly either current or future experiments, or projects concerned with rehabilitating existing data. Ongoing experiments include, for example, the French experiments flown aboard Russian satellites (INTERBALL), aboard the US satellite (WIND), aboard the future European satellites (CLUSTER2), or even some data from the EISCAT radars. The projects to rehabilitate existing data cover a many French experiments performed since 1975, mostly flown on European, US and Soviet satellites or probes.

5.2 INGEST

The CDPP has drawn up a specification for deliverable data products. The specification defines the characteristics (either mandatory or optional) that the data and metadata to be delivered to the CDPP must exhibit. It defines the rules systematically applied with respect to:

– file structure, data encoding and standardization of times and dates;
– orbit or trajectory data;
– the minimum content and format of catalogues;
– complementary information needed to use or interpret the data.

The CDPP provides technical support in order to apply this specification to each data-producing project.

As far as future projects are concerned, the authorities empowered to make decisions on projects will make the drawing up of an obligatory data management plan. The plan must define exactly which data will be archived (physical values, raw data, etc.), how the data will be organized, and when the data will be delivered to the CDPP.

One particular service within the CDPP is the G2ID (*Groupe de Gestion des Informations et des Données* - Service for Managing Information and Data), in charge of the interfaces with

data-producing projects and the formatting of some metadata before its delivery to the archive system.

Submission Agreements. As far as future projects are concerned, the submission agreement shall be constituted by the project Data Management Plan, to be approved by both the project and the CDPP. As far as existing data to be rehabilitated is concerned, the framework is less formal; there is normally no project team left and no longer a budget specific to that project. Rehabilitation is thus the responsibility of a team of engineers from CNES and those of the Principal Investigator or members of his team. The CDPP suggests priorities for the work to be completed. It also influences the choices and compromises to be made with regard to the level of data to be archived.

Delivery Session

Data delivery. Data-producing projects must normally store the data produced before delivery. They do so using the facilities offered by the STAF, a multi-mission storage service at CNES. The main function of the STAF (*Service de Transfert et d'Archivage des Fichiers* - Service for Transferring and Archiving Files) is the long-term physical storage of files. The interface is stable and therefore the technologies and storage media can thus be replaced or changed in-house without affecting the interface. The STAF also monitors and renews the media used.

From a user project viewpoint, the STAF appears as a virtual tree structure in which files may be stored. When all the data to be delivered has been produced, the delivery process merely amounts to a change of ownership of the STAF directories in which the data is stored. There is no actual physical movement of data.

Delivery of metadata. Metadata generally takes up less space than data. A delivery disk space is set up by the CDPP and the data-producing project has the right to write onto this space. When all the data and metadata has been delivered, the G2ID can begin its checking and formatting (see below). This process is valid for a complete set of data, a partial delivery or an update of previously delivered metadata.

Transformation Process

The format of experiment data is not altered during the delivery process. On the other hand, metadata delivered will be subject to a kind of packing (without changing the contents), and new metadata will be created by the G2ID. To give some examples:

- The archive system manages the descriptions of both data collections and objects, browse data and documentary information in the form of graphs on collections and objects. The delivery of a new collection results in the creation of a new node in the data description graph and logical links with existing collections. The creation of this information, granting a global and consistent view of all the data and metadata available, is not within the domain of the data producer.
- When the Principal Investigator delivers a Microsoft Word document describing an experiment, he places the corresponding file in the delivery disk space. The G2ID will then use this file to create a documentary object descriptor giving the document title, author, publishing body, language, associated keywords, stating the existence of an abstract, etc.

The insertion of metadata in the archive system is mostly based on use of Parameter Value Language (PVL) and a Data Entity Dictionary (DED), which is configuration managed. One of the roles of the G2ID will thus be to create this new metadata and construct the PVL structure describing it. Generally speaking, metadata appears as an extremely heterogeneous set of information objects. Using PVL means that these heterogeneous objects may be delivered in both a homogeneous and standard format.

Validation

The G2ID is responsible for ensuring that the deliverable product specifications for each data set have been respected. It also performs a number of coherence checks, such as checking coherence between catalogue data and the files containing experiment data.

Once these checks have been completed, the results, together with all the metadata, are presented at a formal peer review whose purpose is to decide whether the CDPP can accept the data set and issue recommendations in this field. Once accepted, the CDPP becomes the guarantor of the data set. This review brings in scientists from outside both the CDPP and the Principal Investigator team.

Despite the various checks carried out, the scientific validity of the experiment data delivered to the CDPP remains the responsibility of the Principal Investigator or data-producing project.

Security. The delivery process for both data and metadata takes place within a dedicated environment, accessible only by the data producer and the CDPP.

5.3 INTERNAL FORMS

Storage. The STAF multi-mission storage service (see above) takes charge of the data and metadata. This service currently uses several StorageTek silos with high capacity Reedwood cartridges (10 and 50 Gigabytes compressed). The objects archived by this service are files. There are several different layers of service with regard to file retrieval time and file duplication. The STAF is in charge of all data migration involved when changing from old to new media or to a new technology medium. They do not affect the upper layers of the system.

Formats. The format of data stored must be independent of all operating systems. In practice, experiment data is usually in IEEE or ASCII code and divided up into sequential files. The application of CCSDS encoding for times and dates is compulsory for all record structure files. The syntax and semantics of each file must be described with EAST and a DED unless self-descriptive structures such as FITS or NCAR are used. As far as documentary information is concerned, no reference standard for the internal representation of documents has yet been applied.

Data Management

Data management revolves around use of a graph describing data collections and objects. For the purposes of simplification, this graph is usually known as a data graph. It is oriented and non-cyclical. The relations associating a node with its descending nodes are (from an object-oriented viewpoint) inheritance and composition relations. A data set, also known as a terminal collection, thus inherits the characteristics of all the collections above it.

Documentary information, browse data and event tables are also managed through graphs which are nonetheless distinct from the data graph. The graphs contain either explicit metadata or references to external files or documents.

5.4 ACCESS

Access facilities are seen by the user through a WWW server. These facilities include aids to search for data collections and objects, means of retrieving certain metadata (such as documents and catalogues) immediately, ways of ordering data products which include special protective mechanisms for data not made public, and finally, generation and delivery of these data products.

Finding Aids

The aids to search data of interest to the user are based on navigation within the different graphs: the experiment data collection and object graph, the browse data graph, the documentary object graph and the events table graph. These graphs are independent but a certain number of links are used to move from one to another. Navigation within the graphs is, depending on the case, through criteria such as a keyword (parameter measured, location of measurements, etc.), time or other types of criteria.

The data object and collection graph grants several views of the data, and the final objects may be selected after several navigations within the graph.

The events table graph may be used to make indirect selections over time, such as selecting only data corresponding to a given instrument operating mode, or data corresponding to the periods during which a particular type of magnetospheric event was observed, etc.

These aids may be used to select data which is stored either on the main archive site (at CNES) or at an associated laboratory.

Security

Without exception, metadata is visible and accessible to the general public without any prior authentication. On the other hand, data may only be ordered by a user previously authorized by the CDPP, as it normally implies the consumption of resources. The user makes his request for authorization by a form available on-line, indicating his name, e-mail address, the name of the laboratory he belongs to and the reasons for his request. Once the user has received authorization to order products, he must authenticate his request (name and password) before ordering.

Data archived by the CDPP is usually public in nature, but in the case of recent data, data ordering may be temporarily restricted to one particular user group. The system must therefore be capable of handling access rights to the service (for ordering data) independently from access rights to the data itself.

Finally, the system is designed and has a number of protective measures such that any accidental or deliberate modifications to the data stored in the Center may be avoided.

Customer Service/Support

The system can handle profiles peculiar to each user, taking into account in particular the capability of the network linking him to Internet and the laboratory to which he belongs (laboratories directly supported by CNES, laboratories involved in cooperative projects with French laboratories, other laboratories, etc.).

The CDPP has a customer support team able to reply to technical questions (how to use the system, read data, etc.). This team can also direct the users to the Principal Investigator or data producers.

Data Transformation before DIP delivery

The data objects distributed to scientific users are not necessarily identical to the data objects stored in the system. Depending on the standards respected and tools available, a certain number of transformations of archived objects may be requested, in particular:

- Time-related retrieval which provides data corresponding to one (or more) time periods specified by the user. This kind of retrieval is only possible when times and dates have been encoded in compliance with CCSDS recommendations.
- Retrieval of fields, which permits the user to select fields of interest on the basis of an EAST data descriptor.

These transformations are known as 'subsetting services'. Other such transformations are planned for future versions, so as (for example) to be able to deliver data in the user's native machine format, or deliver data as physical values although it is stored as raw values.

Media/Network Use for DIP deliveries

The data from Plasma Physics experiments is often bulky (a data set often contains between ten and several hundred Gigabytes). It is not planned to systematically create pre-defined, widely disseminated products as is often the case for planetary data, particularly as users are often interested in a specific period of time and not the whole data set.

Products may be delivered either over a network or on a variety of media (currently CD-ROM, DAT or Exabytes). The choice between these two types of delivery depends on the capacity of the network between the user and the CDPP at any given time.

As far as network deliveries are concerned, the system proposes the HTTP protocol at the user's initiative or the FTP protocol at the CDPP's initiative, but at a time specified by the user. The latter choice is subject to certain constraints. Deliveries of data via a network offer optional data compression and grouping facilities in the form of .tar files.

Pricing Policy. The pricing policy has not yet been fully determined. It will include an invoice for dissemination of data on an external medium (CD-ROM, DAT, Exabytes).