Capabilities Document for the Simple, Scalable Script-based Science Processor Archive (S4PA)

Version 0.3

10/6/06

What's New

Version 0.3 Additions

- What's New section
- Description of how re-ordering of missed or lost subscription data is accomplished.
- Description of the S4PA directory structures.
- Description of cross-strapping subscriptions to prevent data loss before the data are backed up to tape.
- Information on integrity checking, including a note on how often a file gets checked.
- Acknowledgements section.

Acknowledgements

I would like to thank the S4PA development teams, both software and hardware, for contributing to the architecture itself and implementing it in a timely fashion. Several also provided information to this document. Also, I thank the AIRS, Aura, GOLDS, A-Train and TADS evolution teams for contributing requirements and design ideas, as well as actually deploying the S4PA instances to realize the capabilities described in this document.

Also, a special thank you to the projects and teams whose external review has provided valuable critical input on S4PA's capabilities, often leading to changes in design or addition of useful features:

- AIRS SCF
- AVDC
- ESDIS
- HIRDLS SCF
- MLS SIPS and SCF
- ODPS
- OMI SIPS

Table of Contents

1. Introduction

1.1 Identification

This is the S4PA Capabilities document of the Goddard Earth Sciences Data and Information Services Center (GES DISC). S4PA is the successor to the Version 0, Version 1 (commonly known as the TRMM Support System) and Version 2 system, commonly known as the EOSDIS Core System (ECS).

1.2 Scope

This document describes the functional capabilities of the Simple, Scalable, Script-Based Science Processor Archive system (S4PA), developed by the GES DISC.

1.3 Purpose

This document's purpose is to communicate the capabilities and architecture of S4PA to the various stakeholders:

- Atmospheric Infrared Sounder (AIRS) Science Team and Support Personnel
- High-Resolution Dynamics Limb Sounder (HIRDLS) Science Team and Support Personnel
- Global Modeling and Assimilation Office (GMAO)
- Microwave Limb Sounder (MLS) Science Team and Support Personnel
- Ozone Monitoring Instrument (OMI) Science Team and Support Personnel
- Solar Radiation and Climate Experiment (SORCE) Science Team and Support Personnel
- Earth Science Data and Information System (ESDIS) project
- NASA Headquarters
- Future missions, such as Glory and GPM

1.4 Status and Schedule

This is the intial draft version of the document, compiled from several other source documents. As such, it is likely to contain some inconsistencies, but does give the overall picture of S4PA capabilities.

The released version is scheduled for 15 November 2006.

1.5 Organization

Section 1 identifies the document.

Section 2 lists relevant documents.

Section 3 provides the overall approach, including operations concept and high-level architecture

Section 4 provides a more detailed listing of specific S4PA capabilities.

Appendix A gives the DTD for S4PA granule-level metadata.

Appendix B describes GES DISC conventions for submitting Directory Interchange Format information to the Global Change Master Directory.

2. References

2.1 Parent Documents

The parent documents are the documents from which this Implementation Plan's scope and content are derived.

423-10-69 Archiving, Distribution and User Services Requirements Document

2.2 Applicable Documents

The following documents are applicable:

423-35-01 ICD between EMSn and ECS Elements

423-41-56 ICD between ECS and DAS

423-41-57 ICD between the EOSDIS Core System (ECS) and the Science Investigator-Led Processing Systems (SIPS) Volume 0, Interface Mechanisms

423-41-57-8 ICD between ECS and SIPS, Volume 08 MLS Data Flows

423-41-57-10 ICD between ECS and SIPS, Volume 10, TES Data Flows

423-41-57-12 ICD between ECS and SIPS, Volume 12, HIRDLS

423-41-57-13 ICD between ECS and SIPS, Volume 13, OMI

423-41-57-14 ICD between ECS and SIPS, Volume 14 SORCE

423-41-57-15 ICD between ECS and SIPS, Volume 15, ODPS ECS Data Flow

423-45-02 ICD between ECS and ECHO for Metadata Inventory and Ordering

423-ICD-EDOS/EGS
ICD between the Earth Observing System Data and Operations System and the EOS Ground System Elements.

505-41-33 ICD between ECS and SCFs

505-14-34 ICD between ECS and ASTER GDS

505-41-39 ICD between ECS and LaRC DAAC

505-41-40 ICD between ECS and GES DAAC

2.3 Reference Documents

423-41-45 ICD between ECS and NSIDC DAAC

3. Overall Approach

3.1 S4PA Approach

S4PA was developed originally to replace Version 0 tape-based systems whose hardware was nearing end-of-life. Rather than address the porting to the next generation of tape libaries, S4PA was designed from the start to be a disk-based archives, take advantage of trends in disk capacity and processing power. The result has been significant savings in maintenance and operations. Based on this experience, the EOSDIS Evolution Study team accepted a proposal by the GES DISC to replace the expensive EOSDIS Core System with a less-expensive S4PA-based disk archive. (That replacement is currently underway.)

However, simply implementing a cheaper version of the older system would be an incomplete evolutionary step at best. Rather, the system must be adaptable to new technologies and new science drivers as evolution proceeds. By using the new generation of reliable, inexpensive commodity hardware and a core of robust operational system components, the GES DISC system will be able to adapt to changing requirements and technology even as the first two years of evolution proceed.

3.2 Operations Concept

3.2.1 Overview

The S4PA approach to EOSDIS Evolution is designed to ingest science data, store it, produce higher level products, and allow user access to the data. However, the architecture is significantly different than current architectures such as the ECS, which allows for radical simplification and consequent elimination of some more detailed requirements.

3.2.2 Ingest

In the current system, the Ingest function consists of three main components:

1. Data Acquisition
2. Data Reformatting
3. Data Cataloguing

3.2.2.1 Data Acquisition

The data acquisition interfaces with data providers will remain unchanged, in order to minimize the effect on data providers of the evolution effort. Most of these interfaces are various flavors of the so-called Science Investigator-led Processing System (SIPS) interface, which itself is based on the Product Delivery Record (PDR) interface. The PDR interface is currently implemented in S4PA for FTP transfers, with secure (using both SSH and bbFTP) transfers to follow. Likewise, the EDOS interface, based on signal files, will also remain unchanged, except with respect to actual machine addresses.

3.2.2.2 Data Reformatting

In general, S4PA does not reformat data. Rather, it delegates that task to a partner data processing system using the Simple, Scalable, Script-based Science Processor for Measurements (S4PM). S4PM actually predates S4PA by several years and is the system that has been used since 2001 for operational processing of EOS data at the GES DISC. S4PM and S4PA are designed around the same kernel of software and share external interfaces, so they fit well together.

3.2.2.3 Data Cataloguing

S4PA stores its metadata in separate files alongside the data files. These are XML files following a schema similar to the Data Pool and ECHO. Metadata that are received in ODL format will be reformatted into XML. However, the ODL metadata is encapsulated within the XML file for later distribution to customers who still require this legacy format. If no metadata are supplied, they are extracted from the data files by dataset-specific metadata extraction software.

It is important to note that S4PA is not designed with an integrated search function. Instead, it relies on publishing metadata to external clearinghouses or search applications. For instance, metadata will be published to the EOS Clearinghouse (ECHO), as well as any local GES DISC search interfaces (e.g., Web Hierarchical Ordering Mechanism, Mirador).

As a result, the metadata that it needs to manage each data file is restricted to a few very basic fields such as dataset short name, version identifier, begin date/time and end/date time, plus any additional metadata that are useful for search applications. Thus the data providers have a significant degree of latitude in specifying the metadata for a given dataset. Furthermore, should it later become apparent that additional fields would be useful, the additional fields can be extracted or inferred from the data itself, because *they are all online*, all the time. Thus, there is no need to supply an excess of attributes, just because they may be useful at some point in the future.

On the other hand, in addition to the required fields, the data provider can supply as much or as little additional metadata as they like. S4PA will not read these metadata, but rather simply serve them up to users as received from the data producer. Note, however, that there is a significant chance, indeed likelihood, that users will download only the data files, and not the associated metadata. Even those who do may lose the metadata file later.) Thus, it is strongly recommended that the data file itself include any metadata needed to identify, use and understand the data.

3.2.3 Archive

The S4PA-based archives will be completely disk-based. As a result, a major change in operations concept is the elimination of the robotic tape libraries. This has a set of cascading simplifications.

Firstly, archive management shifts from a primarily hardware and media-oriented effort to one that is software-based. A number of archive maintenance tasks become unnecessary, such as volume group closure, tape compacting and volume group migration. Troubleshooting bad

tapes, tape drives and robotic arms is no longer necessary. Instead, archive management consists of running and analyzing the results of auditing software that checks the data record for completeness and integrity.

Secondly, the disk-based archive eliminates the need for a database mapping files to tape libraries and individual tapes, with the attendant elimination of database administration requirements.

3.2.3.1 Data Backups

The S4PA system supports backups of data, but implemented in a fashion almost identical to the standard disk backups conducted by system administrators. This is done by dividing up the disk into tape-sized disk partitions. When the partition fills, the system administrator makes a backup, which is then stored across campus. In cases where data rate is low (less than the size of a partition per day), incremental backup is performed in addition to a full backup when the partition fills up.

3.2.3.2 Recovery from Data Loss

Most data will be recoverable from the tape backups noted above. However, in rare cases, data may be lost before the disk partition fills and the backup is made. This will be data recently produced by the data provider, and so will be requested for retransmission (or reprocessing if necessary). Level 0 data will be requested from EDOS, which may still have recent data locally. If not, EDOS will provide data from its White Sands backup archive.

3.2.4 Production

As mentioned earlier, production is delegated to the partner S4PM systems.

3.2.5 User Access

Of all the operations concept changes, those affecting User Access are the most profound. In the current system, the load and response time associated with the tape libraries result in significant queues (and waits) to obtain data from the archive. As a result, the system accepts orders, stages the data for the user, then notifies the user that the data are ready. However, in the S4PA-based system, all data are accessible instantly online through FTP. As a result, there is no need for a user to place an order for the data: it is already "staged". Consequently, orders will not be supported. In its place, the GES DISC user interface supports the construction of a batch FTP script to download the contents of a user's shopping cart.

The demise of orders results in more cascading simplification. There are no order transactions to track. The lack of order transactions in turn makes user profiles superfluous. Also, there are no metrics regarding order fulfillment times. More importantly, however, this synchronous data access paradigm allows more services to be offered to the user community. In addition to instant data access, the S4PA-based system will be able to supply more on-the-fly operations to the data. These include subsetting, interactive data mining, and online data analysis (through Giovanni).

A typical user scenario for basic data access is given below using the Mirador search tool:

1. User enters search criteria

2. Mirador returns datasets and estimated hits for specified time and spatial constraints
3. User selects datasets (or drills down to granules)
4. User adds selections to Shopping Cart
5. User goes to Checkout
6. User selects "Batch Download for Data and XML"
7. Browser returns batch download script
8. User executes batch script from local workstation

3.2.6 Operations Monitoring

Most operations monitoring in S4P based systems is conducted using a simple interface called "*tkstat*" (Fig. 1).

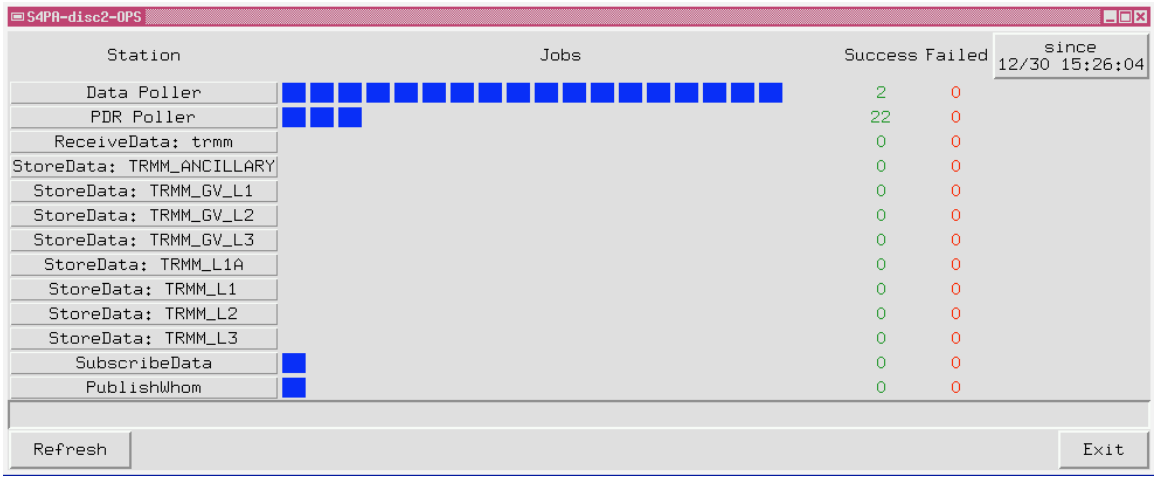


Figure 1. Operator interface (tkstat) for monitoring an S4PA instance.

In addition to tkstat, an interface called *tkargus* is provided to look for trouble spots in all S4PA or S4PM instances across multiple machines (Fig. 2).

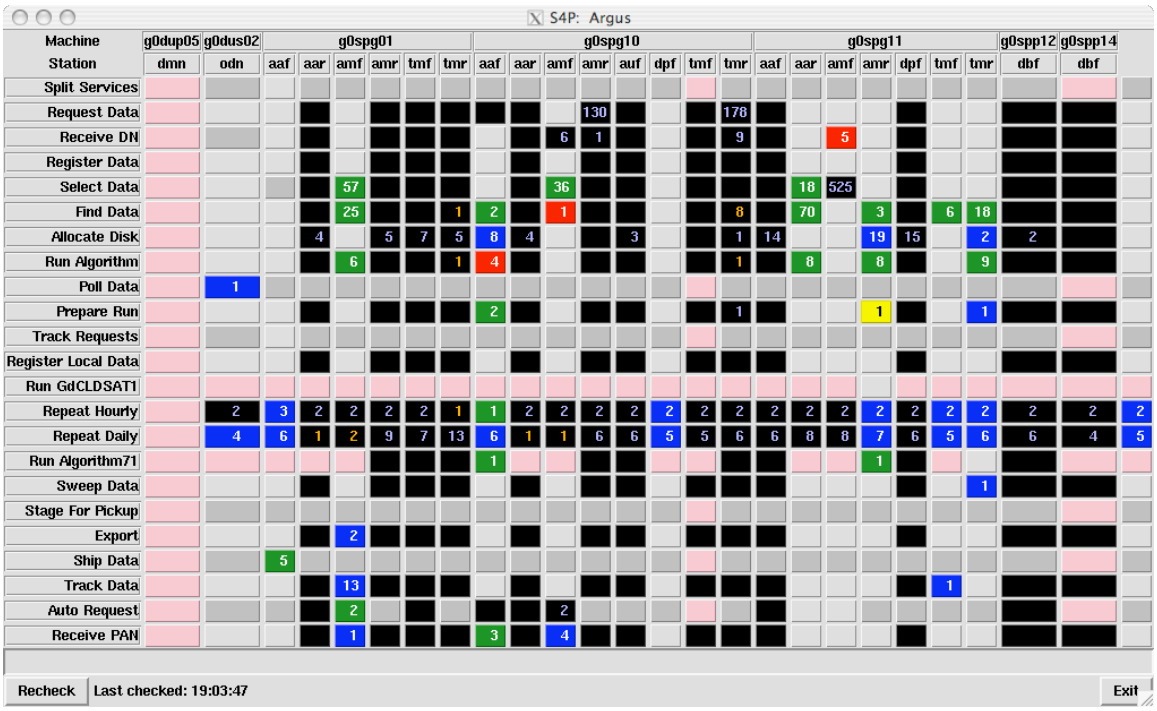


Figure 2. Operator interface (tkargus) for monitoring several S4PM instances on multiple computers at once.

This and other S4PA interfaces are focused on identifying trouble spots to the operators. They can also be configured with failure handlers that can be executed by the operators at a simple push of the button.

It is also important to note that there will be fewer components to monitor in the future. Essentially the overall system will consist of about a dozen S4PM instances for processing, plus about the same number of S4PA-based archives. Hardware will be relatively uniform, i.e., Linux platforms with RAID disk arrays. There will be no tape silos or attached machines and no orders to manage, thus resulting in decreased operations costs.

3.2.6.1 Off-hours Failures

There is no need to oversize S4PA computers to protect against off-hours software failures, since very little processing is done within the system. Thus, it will be easy to catch up. However, a cross-strapping failover capability will be implemented between the Aura and AIRS Level 0 machines, allowing the Aura L0 flow to be redirected through the latter and vice versa should a machine failure threaten a long-lasting outage (24 hours or more.)

The partner S4PM instances will be oversized from that needed to meet baseline processing requirements. This will allow the operations staff to monitor them on an 8x5 basis, using the extra power to catch up should a serious failure occur in off-hours. However, special instances of S4PM that have significant downstream consequences, i.e., DPREP, will be mirrored on a second machine as a failover.

3.2.6.2 Problem Resolution Process

Trouble tickets are written by GES DISC staff using the Bugzilla system, which has automated capabilities for routing S4PA trouble tickets to the proper staff on entry. They are also reviewed daily in the operations meeting.

3.2.7 Administration

Systems Administration will be essentially unchanged from today's administration mode, except for the more homogeneous platform mix. However, database administration will undergo a major change: with the obviation of order tracking and the complicated mapping of files to tape volumes, there will be no need for transaction processing in the evolved system. Thus the role of relational databases will be reduced to that of simply supporting individual applications where necessary, such as search interfaces. In this role, accounts are very limited and schemas are highly simplified. Also, the need for high-reliability journaling is reduced: if the database gets out of sync, it can simply be repopulated from scratch. As a result, database administration activities will be significantly reduced from the current database-centric environment.

3.2.8 Testing

Testing of system enhancements is done in three different environments:

Initial small scale testing of S4PA is conducted on the development machine.

Full-scale regression testing is then conducted on a standalone test system, configured with all GES DISC datatypes and interfaces. The software is now ready for release and deployment on the target machines.

Finally, in many cases testing will be conducted in a separate mode (either TS2 or TS1) on the target machine before installation into Operations, much the same way as test currently happens at the DAAC.

Full-up performance testing is accomplished by exercising a given system in OPS mode before they are commissioned for full operations. This allows full resources to be deployed for this testing, but without affecting the user community. Note that the architecture (many small systems holding parts of the data) is designed so that hardware enhancement is done by adding standalone systems, rather than by enhancing existing ones. This allows performance testing to be conducted on the enhancements to the overall architecture without impacting running systems.

Testing of Off-the-Shelf software (OTS) will be performed on the development machine followed by the standalone test machine. In rare cases, a pre-operational system may be used for further testing. If no such system is available, the upgrade will be applied directly to affected systems, but in a rolling fashion (i.e., one at a time, with periods in between.

Fortunately, OTS is essentially limited to the perl scripting language plus subsidiary modules. Furthermore, Perl installations enforce fairly stringent version separation on installation, allowing more than one version to be installed on a machine at a time. This allows us to switch Perl versions through simple symbolic links, with an equally quick rollback.

Operating system upgrades will also be tested in the same way: on the development system first, a separate integration test next, and a rolling upgrade across the operational systems with rollbacks whenever problems are encountered.

3.2.9 Configuration Management

Modifications to operational systems are approved by a Configuration Control Board. Configuration Change Requests (CCR) are submitted to a Web-based system. Concurrences must be given by all cognizant parties for a given CCR before it is placed before the board chairman, who then signs off on the CCR. A trouble ticket is generated by the configuration management for each CCR in order to accomplish the work and closed by the CCR originator when the work is complete.

Version tracking is managed by maintaining version numbers within software modules, promulgating a description of changes and features of each module version, and publishing a matrix of module versions versus host installations.

3.2.10 Metrics Collection and Reporting

Ingest/Archive metrics will continue to be reported via the EDGRS interface (at least as long as ESDIS maintains that interface). However, one important impact of the above operations concept is that all user access metrics will be in the form of FTP and HTTP download statistics. There will no longer be any order statistics as such.

3.3 System Architecture

The S4PA System Architecture represents a significant departure from the current architecture in a number of respects. Firstly, the architecture is based on magnetic disk storage rather than tape storage. While more expensive on a per-gigabyte basis, this allows dramatic simplification of the software and hardware needed to support the data, along with a commensurate reduction in operations costs. Secondly, the software system to support this, the Simple, Scalable, Script-based Science Processor Archive (S4PA) is radically simplified from the current EOSDIS Core System. This simplification is enabled by the disk-based support of the data. Thirdly, rather than support all missions and instruments in a single system, the S4PA architecture actually consists of several small standalone systems, each of them supporting a particular mission or set of measurements. This is now possible because we need no longer rely on a bank of large tape silos.

3.3.1 Overall Architecture and Data Flows

Fig. 3 shows the overall architecture of S4PA systems supporting EOS at a system level. In actuality, the GES DISC "system" is not a single system, but rather several standalone systems. Data from the three Aura missions are stored together in the Atmospheric Chemistry Data and Information Services Center (ACDISC) because they form together a set of closely related measurements of atmospheric chemistry. The ACDISC provides not only basic anonymous FTP access to the data, but also other online services, including subsetting and online analysis using Giovanni.

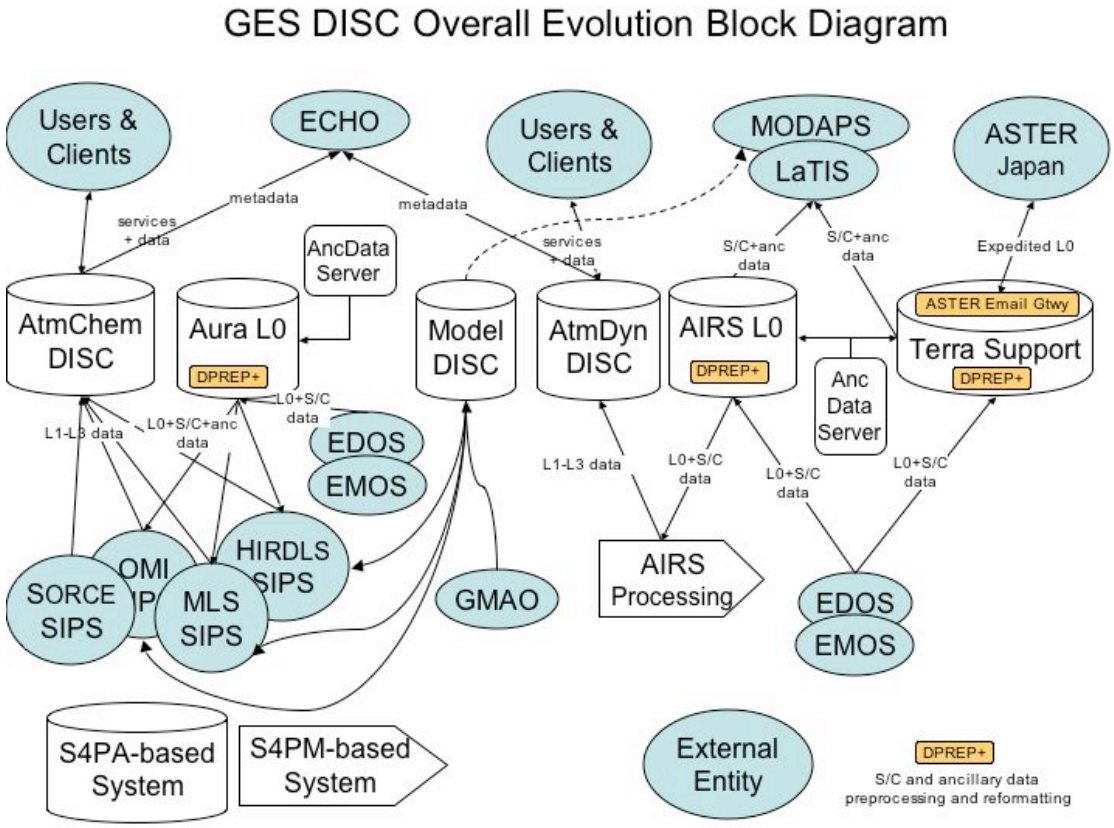


Figure 3. Block diagram of overall S4PA architecture.

The Aura Level 0 data are stored in a separate S4PA-based archive. This allows us to tailor the available services and permissions differently for the Level 0 data, which have very few services beyond restricted distribution to the Science Investigator-Led Processing Systems (SIPS). Also it makes it possible to site the Level 0 server on the same EOSDIS network as EDOS without similarly constraining the systems hosting the Level 1 through Level 3 data. The sole role of the Level 0 archive is as a pass-through and backup to the Level 0 archives at the respective SIPS.

Similarly, the AIRS Level 0 data are stored in a separate S4PA-based archive from the Level 1-3 data, which are stored in the Atmospheric Dynamics archive. The Atmospheric Dynamics system provides not only archive and distribution services, but also includes the data processing function.

The Model DISC system is designed for storing Level 4 model output from the Global Modeling and Assimilation Office. It also does some processing to create subsets for instrument teams and to support server-side processes on behalf of the user community. The Terra Support system consists of a small S4PA archive and an S4PM processing system to process ancillary and ephemeris data, as well as reformat ancillary data from external sources.

3.3.2 Disk-Based Archive

The key feature of the evolved GES DISC system is that all data are stored on disk, rather than in tape silos. In general, the disk-based archives are managed by the S4PA software system. This

is a system of Perl scripts for managing data ingest and ongoing curation functions. The latter include checking data integrity (i.e., verifying checksums), determining the completeness of the data record, and handling data file replacement, among others. Since the data are on disk, these checks can be both complete and repetitive, resulting in an increased confidence in the safety of the data record, and faster detection of corruption than for data on tape.

3.3.3 Data Access

Data access for unrestricted data is through anonymous FTP. Note that this is similar to the current predominant access method (FTP Pull), with the exception that the data need not be staged before the user can begin downloading. The data are already on disk (all the time) and ready for immediate download. As a result, there is no need to support "orders", in which a user asks for data to be made available for download. This in turn relieves us of having to track order fulfillment (a complicated and time consuming process), or even of having to register users and their profiles. As a result the database needs of the system are dramatically simplified, if not completely eliminated.

A few datasets will be restricted to instrument teams or during a validation period. These data will be available via the Hypertext Transfer Protocol (HTTP), which supports a simple password-based mechanism for user authentication. (Since the main purpose of the restriction is to prevent other scientists from inadvertently using unvalidated data in research, this basic mechanism provides sufficient protection.)

3.3.4 Data Search

Data search across these standalone systems is achieved by the publication of file-level metadata from them to various clearinghouses and search clients. For example, the S4PA systems export metadata to both of the GES DISC search interfaces, the Web Hierarchical Ordering Mechanism (WHOM) and Mirador, a simple form-based search based on the Google mini-appliance and a PostGres spatial database. Both of these interfaces support the ability of the user to generate a list of desired files which is transformed into an FTP batch download script, thus avoiding the need to click on hyperlinks for each file.

S4PA includes the capability of exporting its XML metadata to the EOS Clearinghouse (ECHO), with a minor transformation effected through XSLT. This will allow users to search the S4PA-based systems through the Warehouse Inventory Search Tool (WIST). (Note that WIST does not yet include the batch download capability, but talks have begun between the GES DISC and the ECHO/WIST team on adding it.)

3.3.5 Hardware Architecture

The hardware architecture is built around commodity hardware: Linux servers with relatively low-cost drives in a Redundant Array of Inexpensive Disks (RAID) configuration. This provides some basic protection against data loss through disk failure from day to day. (In addition, tape drives are used for backups providing an additional layer of protection.) These servers represent standalone systems, with data flowing between them (when necessary) using protocols such as FTP, HTTP and Open-source Project for a Network Data Access Protocol (OPeNDAP). Some care is taken to allocate datasets to systems to minimize this flow; the rest is handled via high-speed (Gigabit Ethernet) networks.

We recognize that a Storage Area Network (SAN) would eliminate these issues. However, past experience of the GES DISC with two separate SAN technologies as a data cache points to some significant risks. (These are magnified when one considers putting data for archive on a SAN.) Some SAN failure modes (say, at the fabric switch) can bring down all the archives at once; this is an issue even for preventive maintenance activities. Also corruption by the SAN software layer or in the SAN switch has the potential to affect data files across the SAN, forcing time-consuming recovery efforts simply to identify the corrupted data. On the other hand, use of standalone data servers isolates corruption problems and reduces the overall risk to the archive. Also, it allows rolling downtimes for maintenance. As a result, it is unlikely that the entire GES DISC would be down at any time. Rather, one system will be down or inaccessible, while the rest remain up and running.

3.3.6 External Interfaces

External interfaces designed to remain *as similar as possible* to the current interfaces in order to minimize the impact to the other parties, such as the instrument teams. Table 2 shows the other parties and the interfaces to be used.

The most common ingest mechanism is through structured files called Product Delivery Records (PDRs) and Product Acceptance Notices (PANs), which are already a feature of S4PA. However, the handshake with EDOS also uses a file-by-file signal file handshake which has been added to S4PA. A PAN response to EDOS will also be added. Distribution is typically done through FTP or Secure FTP push, with a Distribution Notification (DN) emailed or pushed as the handshake.

Organization or System	**Ingest/ Distribution**	**Handshake**	**Transfer**
EDOS	Ingest	Signal files (.xfr)/PAN	ftp push
EMOS	Ingest	Signal files (.xfr)	ftp push
GMAO	Ingest	PDR/PAN	bbftp pull
HIRDLS	Ingest	PDR/PAN	sftp pull
MLS	Ingest	PDR/PAN	sftp pull
OMIDAPS	Ingest	PDR/PAN	sftp pull
NOAA	Ingest	polling	ftp pull
SORCE	Ingest	PDR/PAN	sftp pull
AIRS SCF	Distribution	DN	ftp push
HIRDLS	Distribution	DN	sftp push
MODAPS	Distribution	DN	ftp push
OMIDAPS	Distribution	DN	sftp push
ODPS	Distribution	DN	sftp push
MLS	Distribution	DN	sftp push

One exception to the support of current interfaces is the ECS Machine-to-machine gateway. The functionality provided by this tool will be provided by a variety of mechanisms, including WIST/ECHO search of restricted datasets, and an operator script to automate re-transmission of data required by the external interface.

4. S4PA Capabilities

The evolved systems will comprise a single software system, S4PA, instantiated on multiple hardware systems. These instantiations will be drawn from a single source baseline, but will typically be configured to support the interfaces appropriate to the functions served by that hardware system. For example, the Level 0 archive systems will be configured to ingest data from EDOS using the EDOS signal file mechanism over FTP, while the Atmospheric Chemistry DISC will be configured to ingest data from Aura data producers using the PDR mechanism over SCP.

Note: items that are in progress or planned for future implementation as part of the evolution implementation are noted as such.

4.1 Data Acquisition

4.1.1 Handshake Protocols

4.1.1.1 Science-Investigator-Led Processing System (SIPS)

The SIPS handshake protocol begins with the data provider placing a Product Delivery Record in a location that can be accessed by the archive. S4PA polls the location for new delivery records, compares them against an "oldlist" to see if any new PDRs are available, and then pulls the PDR. This then identifies the data and metadata files to be transferred. If a checksum is supplied by the provider, it will be checked to ensure that the transfer did not corrupt the data (*in progress*). Supported checksum types are CRC32 (aka Unix cksum) and MD5. On successful completion of data transfer, a Product Acceptance Notice is returned with information on the success or failure of the transfer.

A full description of this protocol is given in ESDIS document 423-41-57, *ICD between the EOSDIS Core System (ECS) and the Science Investigator-Led Processing Systems (SIPS) Volume 0, Interface Mechanisms*.

4.1.1.2 EOS Data Operations System (EDOS)

The EDOS handshake protocol also begins with EDOS placing a Product Delivery Record in a location that can be accessed by the archive. However, the format of this PDR is somewhat different. EDOS also transmits signal files with information about the checksums as each file transfer finishes. Again, S4PA polls the location for new delivery records, compares them against an "oldlist" to see if any new PDRs are available, and then pulls the PDR. This then identifies the data and metadata files to be transferred. On successful completion of data transfer, a Product Acceptance Notice is returned with information on the success or failure of the transfer.

A full description of this protocol is given in the ESDIS document *ICD between the Earth Observing System Data and Operations System and the EOS Ground System Elements*.

4.1.1.3 TRMM Science Data and Information System (TSDIS)

A mechanism similar to the SIPS PDR mechanism is also used for TSDIS data transfers. However, this is a file-based implementation of the old socket-based Data Availability Notice mechanism.

4.1.1.4 Polling

For data providers that do not wish to go to the extra effort of handshake protocols, S4PA can simply poll a designated area for new data.

4.1.2 Transfer Protocols

The protocols for actually moving the data are based on standard Internet protocols, FTP and SFTP.

4.1.2.1 FTP Pull

S4PA can pull the data files via FTP. (For security reasons, it is recommended that this be anonymous FTP if crossing network boundaries to avoid the risk of passing a sensitive password in the clear.)

4.1.2.2 SFTP Pull

For secure transfers across network boundaries, Secure FTP, based on Secure Shell, is supported.

4.1.2.3 Local Copy

For data transfers on the same machine (e.g. partner S4PM systems), a simple local copy is used.

4.1.2.4 Utcpole/leapsec

TBS.

4.2 Data Cataloguing

Cataloguing is the process of identifying the data and obtaining or generating the metadata needed for data management and data search support within the archive. S4PA supports several ways for the data provider to supply metadata: either XML, ODL, or internally within the data file. Also, note that there are no restrictions placed on what metadata should be included inside the data file, except S4PA required metadata when no external metadata file is present. (Indeed, it is recommended that both the metadata supplied to S4PA and additional "use" metadata be included in the data file to provide users with a self-describing data file.)

4.2.1 XML Format Metadata Files

S4PA's metadata architecture is based on XML. Thus, providers can supply metadata as an external metadata file following the S4PA DTD for granule-level metadata (Appendix A).

4.2.2 ODL Format Metadata Files

In order to preserve existing interfaces for EOS data providers, the legacy ODL format using the ECS metadata model is also accepted by S4PA. In this case, a metadata transformation program is used in conjunction with templates in order to extract metadata from the ODL and reformat it

in ODL. In this case, the original ODL is encapsulated inside the XML as ProviderMetadata, allowing it to be redistributed to those users that still require the ODL format metadata.

4.2.3 Internal Metadata

Metadata can also be supplied inside the data file. In this case, the GES DISC writes a datatset-specific metadata extractor script to read the data file and create the XML metadata file.

4.3 Data Management

4.3.1 Directory Structure

There are actually two directory structures of significance in S4PA. From the user's perspective, the data are organized according to data groups that are designed by the S4PA instance owner to be intuitive for that user community (Fig. 4). Thus, gridded data may be kept in a separate group from swath data, or data from one instrument might be grouped separately from another. Beneath each group is a directory for each data set (aka data collection). Below that, data are organized by time, with each year in a subdirectory and day of year (or month for monthly data) in a subdirectory below that.

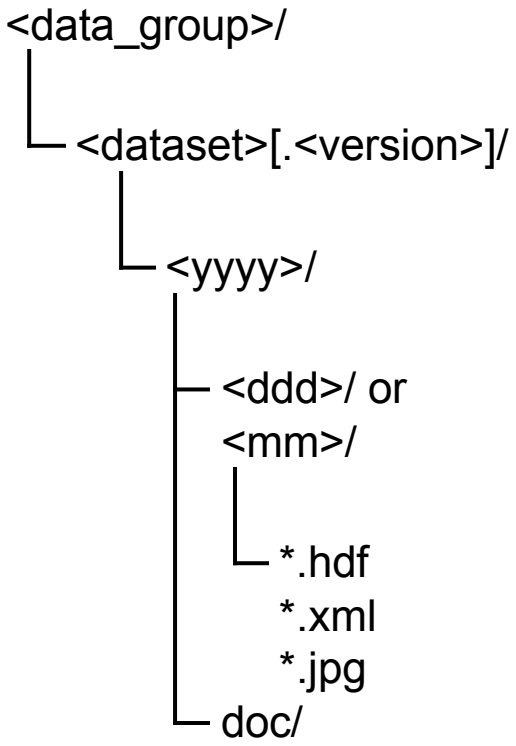


Figure 4. User view of S4PA Directory Structure.

The user view of the data is actually implemented by symbolic links to the physical locations of the data. The directory structure of the physical locations is designed around data management needs. In this structure, data classes are organized around what kind of metadata extraction is required or other data management activities that are done using common settings for a set of

datasets. Datasets may also be spread across multiple filesystems. However, this complexity is hidden from the user via the symbolic links in the user view of the data.

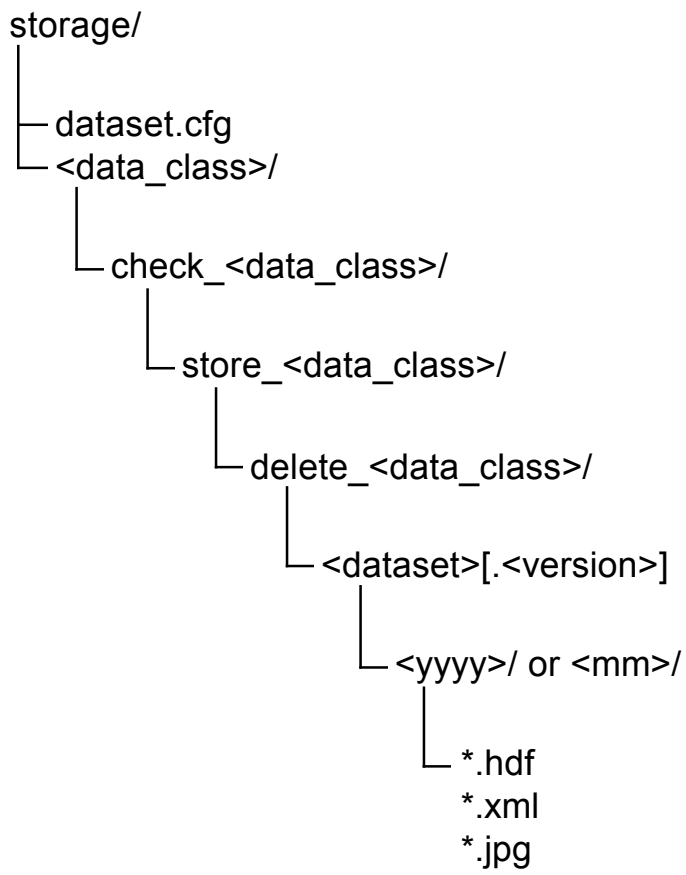


Figure 5. Physical storage directory structure.

4.3.2 Access Restrictions

Data may need to be restricted from public view either during validation or due to MOU constraints with other agencies. Such data are restricted by restricting directory permissions to prevent anonymous FTP users from accessing them. Access to these data are then provided through HTTP, with username/password required.

Note that this is different from simply not advertising data that are not restricted per se, but of little use to the average user, such as Level 0 data. For performance reasons, these are stored in the anonymous FTP area, but the metadata are not published to locations such as ECHO. To mitigate against general FTP browsers chancing across them, the directory paths are also randomized and the parent directories read-protected.

4.3.3 Versioning and Replacement

S4PA supports a rich set of automatic file replacement rules. Replacement may occur within a given version of a dataset, or replacement of a file in one version by a granule in another (usually later) version. In the latter case, the replaced data files are retained for six months after the

newer file is received. However, the replaced file is withdrawn from metadata publication sites such as ECHO, WHOM and Mirador.

Replacement rules are typically time-based (i.e., replace the 2006-08-02T22:05:00 granule). However, the rules also allow for time tolerances (e.g. 2006-08-02T22:05:00 +/- 30 seconds) as well as any arbitrary set of product-specific attributes (e.g. StationName=DARW).

4.3.4 Data Integrity and Completeness

4.3.4.1 Backups

All data are backed up to Super DLT tapes on a regular basis. The tapes are stored in another building on GSFC's campus.

For the period between data archive and the data backup, two solutions will be implemented. Since Level 0 data is held at EDOS for about a month, data will be re-requested from them in the event that a disk crash loses recently archived data.

For data provided by the SIPS the solution will consist of:

(1) setting up a subscription to send the data to another machine, and

(2) a simple cron/find -delete script on the temporary mirror to clean data older than N days, where N is the typical backup lag X 2.

For example, airscal1 would send to airscal2, which would send to airspar1, which would send to airscal1.

4.3.4.2 Checksums

Data are checked after initial transfer, and on any push distribution (before the push) to a user. Checksums are included in the metadata allowing pull users to verify the checksum after they acquire data. Checksums are also included in legacy-format distribution notices. In addition, ongoing data integrity checks are performed by scripts on configurable time intervals. The frequency that a given file is checked varies based on configuration and the size of the filesystem, but will typically be on the order of once every few days. (This is a significant improvement over the current system which takes many months to cycle all the way through.)

4.3.4.3 Completeness Checker

A data completeness tool can be run to check whether the data record is complete. This is configurable to account for the vagaries of data records for different datasets.

4.3.4.4 Dot Charts

The operators also have a graphical tool for verifying such data completeness (known locally as the "Dot Chart"). (*In progress*.)

4.3.5 EMS Metrics

The S4PA system will provide both archive and distribution metrics to the EOSDIS Metrics system. In addition, the ancillary search tools and web pages providing services will be instrumented to support the EMS NetTracker application.

4.4 Data Search

4.4.1 ECHO and WIST

Metadata for all publicly accessible data products will be published to ECHO, which will in turn make them available to WIST and other search clients.

Metadata are reconciled on a regular basis (e.g., weekly) to ensure that ECHO's picture of the S4PA inventory matches the actual inventory. (*In progress*.)

4.4.2 WHOM

Metadata are also published to the Web Hierarchical Ordering Mechanism, one of the GES DISC's local interfaces. WHOM provides a directory- or navigation-style interface.

4.4.3 Mirador

Mirador is the GES DISC's new search tool to provide searching in a form to common search engines today. It supports keyword searching using Google, together with spatial and temporal searching using PostGRES. In addition, it has both a geographic gazetteer and an "event" gazetteer, allowing the user to search based on the space-time trajectory of a geophysical event such as a tropical storm or dust storm (*future*).

4.5 Subscriptions

A mechanism similar to the ECS mechanism of subscriptions is provided by S4PA in order to minimize the change on the provider's end. This is not completely identical, as the construction of Granule Identifiers is different in S4PA than in ECS, and because orders and requests do not exist. However, every attempt is being made to match both the distribution protocols and the format of distribution notices to the legacy ECS format.

4.5.1 Distribution Protocols

4.5.1.1 FTP Push

Subscriptions can be pushed to a user using FTP Push. When going across networks, anonymous FTP is recommended to avoid transfer of passwords in the clear.

4.5.1.2 SFTP Push

For secure transfers, SFTP push is supported.

4.5.1.3 FTP Pull

FTP Pull is supported. However, in this case, only the distribution notice is sent to the end user. The user must then pull the data using anonymous FTP based on the information in the distribution notice.

4.5.2 Distribution Notices

Distribution notices are available in three formats, a legacy ECS format, a simplified URL format and a Product Delivery Record format. In addition, the transmission protocol is configurable to either email, FTP Push or SFTP Push.

4.5.2.1 Legacy Format

A legacy format is available to support existing interfaces with ECS subscribers. This follows the ECS DN format exactly, albeit with different values for the GranuleUR, where a URL is used, and ORDERID and REQUESTID, which are filled in with system time and process_id information simply to ensure uniqueness.

USERSTRING is preserved as a feature of the leegacy format distribution notice. This is set on a per-subscription basis in the subscription configuration file as "label", e.g.:

```
%cfg_subscriptions = (
  "FTP-Push" => {
     urlRoot => {
      public => 'ftp://disc2.nascom.nasa.gov/data/s4pa/',
     },
     notify => {
       address => 'mailto: hegde@daac.gsfc.nasa.gov',
       format => 'LEGACY'
     },
     destination => "ftp:discette.gsfc.nasa.gov/private/s4pa/push",
     label => "FTP-Push",
     "MOD021KM..*" => {
        validator => [ 'true' ]
     }
  }, ...
```

4.5.2.2 URL Format

The URL format simply lists the URLs of the distributed data.

4.5.2.3 PDR Format

The PDR format is the same as that used for SIPS ingest. However, note in this case that no PAN processing is done.

4.5.3 Subscription Qualifiers

Most subscriptions are currently unqualified. However, in some case, qualifiers may be needed, particularly to confine the data received to a particular time period (say, current data only, not reprocessed). To support this, S4PA supports the insertion of dataset-specific qualification scripts. These are run on the metadata from each data insert, exiting 0 in the case of a match, and non-zero for no match. Matches then trigger the subscription. As a result, virtually any

combination of metadata can be used for this triggering. Currently, the only qualification script is one that supports time-based qualifiers. Other qualification scripts can be as needed, however.

4.5.4 Redistributions

In the event that a subscriber loses or does not receive data shipped via subscription, a script will be used to redistribute the data to the subscriber *(In progress*). Since this will be a complete redistribution, it will include a Distribution Notice, whether the distribution is a Push or Pull.

The redistribution request protocol will be a simple HTTP GET request. This will support both access via a simple Web interface (to be provided) or from the command line using a common tool such as wget or curl.

4.5.5 ODL Metadata Distribution

In order to match existing interfaces as closely as possible, a "filter" mechanism has been provided to distribute metadata in ODL format to some users. Essentially, the filter extracts the ODL from the provider stored in the ProviderMetaData element (see Appendix A) and distributes that. Note that this means that this capability is not available for datasets that are provided with metadata in XML format, or internal to the data file.

4.6 ASTER Email Gateway

TBS.

4.7 Metadata Architecture

4.7.1 Granule-level Metadata

S4PA's metadata requirements for data granules are primarily targeted toward data management tasks and supporting data services such as search or subsetting. It is expected that additional "use" metadata will be included by the provider inside each data file so that the data are self-describing to users. As a result the metadata model is slightly less complex than its predecessor ECS metadata model. Otherwise, the S4PA metadata model is patterned after the ECS model as well as the ECHO metadata model. As a result, very little modification is needed to the metadata in order to publish them to ECHO. Appendix A shows the Document Type Definitiaion for granule-level metadata.

4.7.2 Collection-level Metadata

On the other hand, the metadata model for collection-level information is that of the Global Change Master Directory. Indeed, S4PA delegates to GCMD the role of the "master" repository of dataset-level metadata. These metadata are extracted from GCMD and transformed using XSLT to populate other dataset-level repositories, such as the S4PA directory system, the Mirador search tool, and the ECHO metadata repository (Fig. 6).

The GMD Data Interchange Format documents are prepared based on documentation from the Science Team and submitted using the GCMD's online DIF submission editor. (This editor can support collaborative DIF development if agreed upon with the Science team.) These DIFs are

then extracted by S4PA for transformation and publication to other repositories of dataset-level information. In order to support the variety of other repository needs, an internal GES DISC standard has been developed to ensure that the DIFs are populated with enough information. A draft version of this is included in Appendix B.

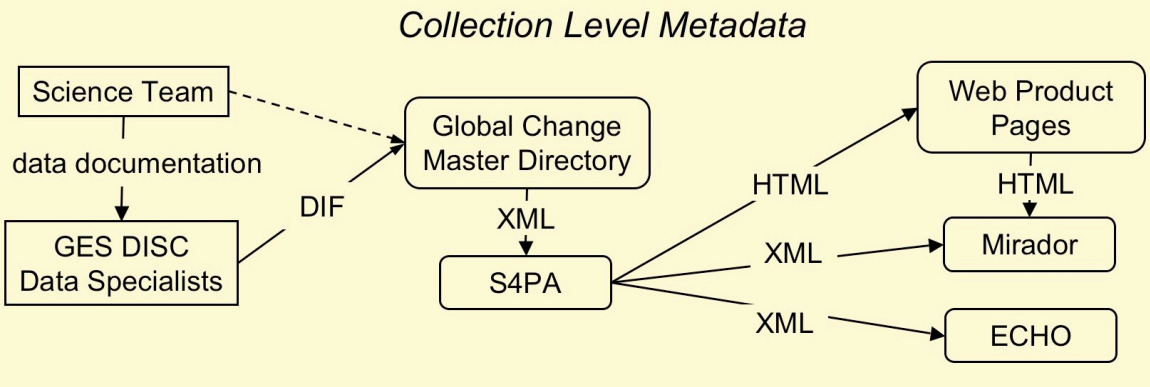


Figure 6. Dataset-level metadata architecture.

4.8 S4PA Hardware

The following is an initial specification of the hardware systems to support the various archives and processing systems. More detailed requirements specifications and hardware phasing are still underway, in an attempt to defer as much of the hardware purchase to FY07 as possible.

Hardware Systems	Type of System	Number of Systems	Amount of Disk (Total)
AIRS Level 0 Archive	4-CPU Linux	23	25 TB
Aura Level 0 Archive	4-CPU Linux	21	10 TB
AIRS Processing8-CPU Linux3Atmosperhic Dynamics Archive (AIRS)	4-CPU Linux/8-CPU Linux	2/1	70 TB
Atmospheric Dynamics Archive (GEOS-5 Forward)	4-CPU Linux	2	30 TB
Atmospheric Chemistry Archive	4-CPU Linux	3	30 TB
Terra/Ancillary Support	4-CPU Linux	1	5 TB

Appendix A. Document Type Definitions for S4PA XML for Metadata Files

```
<?xml version="1.0" encoding="UTF-8" ?>
- <!--
  This schema is a placeholder for Granule level, S4PA
  -->
- <!--
  meta-data elements
  -->
- <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    elementFormDefault="qualified" attributeFormDefault="unqualified">
  <xs:include schemaLocation="S4paCommonTypes.xsd" />
- <xs:element name="CollectionMetaData">
- <xs:annotation>
  <xs:documentation xml:lang="en">Collection Meta-Data Information. This contains the
    product ShortName and LongName as well as the VersionID. These elements are
    associated with the Collection (S4paCollection.xsd) level
    schema.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="Description" minOccurs="0" />
  <xs:element ref="LongName" minOccurs="0" />
  <xs:element ref="ShortName" />
  <xs:element ref="VersionID" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="DataGranule">
- <xs:annotation>
  <xs:documentation xml:lang="en">Meta-data specific to the granule. This currently
    includes six elements, including granulID (ID for the granule, which is usually the
    filename), Format (format of data e.g., HDF-EOS), CheckSum,
    SizeBytesDataGranule (size of the granule in bytes), InsertDateTime (date and
    time granule was inserted into S4PA), ProductionDateTime (date and time the
    granule was produced).</xs:documentation>
    </xs:annotation>
```

```
- <xs:complexType>
- <xs:all>
  <xs:element ref="GranuleID" minOccurs="0" />
  <xs:element ref="Format" minOccurs="0" />
  <xs:element ref="CheckSum" minOccurs="0" />
  <xs:element name="SizeBytesDataGranule" minOccurs="0" />
  <xs:element ref="InsertDateTime" minOccurs="0" />
  <xs:element ref="ProductionDateTime" minOccurs="0" />
  <xs:element ref="PGEVersionClass" minOccurs="0" />
  <xs:element ref="Granulits" minOccurs="0" />
  <xs:element ref="DayNightFlag" minOccurs="0" />
    </xs:all>
    </xs:complexType>
    </xs:element>
- <xs:element name="PGEVersionClass">
- <xs:annotation>
  <xs:documentation xml:lang="en">Lists the PGE version that was used to process the
    data</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="PGEVersion" minOccurs="0" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="PGEVersion">
- <xs:annotation>
  <xs:documentation xml:lang="en" />
    </xs:annotation>
- <xs:simpleType>
  <xs:restriction base="xs:string" />
    </xs:simpleType>
    </xs:element>
- <xs:element name="SizeBytesDataGranule">
```

```
- <xs:simpleType>
- <xs:list>
- <xs:simpleType>
  <xs:union memberTypes="xs:unsignedLong" minOccurs="0" />
    </xs:simpleType>
    </xs:list>
    </xs:simpleType>
    </xs:element>
- <xs:element name="InsertDateTime">
- <xs:simpleType>
- <xs:list>
- <xs:simpleType>
  <xs:union memberTypes="xs:date xs:time" />
    </xs:simpleType>
    </xs:list>
    </xs:simpleType>
    </xs:element>
- <xs:element name="ProductionDateTime">
- <xs:simpleType>
- <xs:list>
- <xs:simpleType>
  <xs:union memberTypes="xs:string" />
    </xs:simpleType>
    </xs:list>
    </xs:simpleType>
    </xs:element>
- <xs:element name="GranuleID">
- <xs:annotation>
  <xs:documentation>The unique identifier for each granule. In most cases this is the
    file name.</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="180" />
```

```
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="Format">
- <xs:annotation>
  <xs:documentation>The format of the data. This type is unique to
    S4PA.</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="32" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="CheckSum">
- <xs:annotation>
  <xs:documentation xml:lang="en">Checksum, which is a unique number, based on the
    content of the file. There are two elements, CheckSumValue (value of the
    checksum) and CheckSumType (the type of checksum e.g., 64bit
    long).</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="CheckSumType" minOccurs="0" />
  <xs:element name="CheckSumValue" type="xs:string" minOccurs="0" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="CheckSumType">
- <xs:annotation>
  <xs:documentation xml:lang="en" />
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="12" />
```

```
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="Granulits">
- <xs:annotation>
  <xs:documentation xml:lang="en">This container is for multi-file granules or
    granulits.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:all>
  <xs:element name="Granulit" minOccurs="0" />
    </xs:all>
    </xs:complexType>
    </xs:element>
- <xs:element name="Granulit">
- <xs:annotation>
  <xs:documentation xml:lang="en">The container is for individual files, or granulit, in
    multi-file granules or granulits.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:all>
  <xs:element ref="GranulitID" />
  <xs:element ref="FileName" />
  <xs:element ref="CheckSum" />
  <xs:element name="FileSize" type="xs:unsignedLong" />
    </xs:all>
    </xs:complexType>
    </xs:element>
- <xs:element name="GranulitID">
- <xs:annotation>
  <xs:documentation>The unique identifier for each granule. In most cases this is the
    file name.</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
```

```
<xs:maxLength value="8" />
  </xs:restriction>
  </xs:simpleType>
  </xs:element>
- <xs:element name="FileName">
- <xs:annotation>
<xs:documentation>The unique identifier for each granule. In most cases this is the
  file name.</xs:documentation>
  </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
<xs:maxLength value="180" />
  </xs:restriction>
  </xs:simpleType>
  </xs:element>
- <xs:element name="DayNightFlag">
- <xs:simpleType>
- <xs:list>
- <xs:simpleType>
<xs:union memberTypes="xs:string" minOccurs="0" />
  </xs:simpleType>
  </xs:list>
  </xs:simpleType>
  </xs:element>
- <xs:element name="RangeDateTime">
- <xs:annotation>
<xs:documentation xml:lang="en">The date and time range of the granule. This
  includes four elements, RangeEndingTime, RangeEndingDate,
  RangeBeginningTime and RangeBeginningDate.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:all>
<xs:element name="RangeEndingDate" type="xs:date" />
<xs:element name="RangeEndingTime" type="xs:time" />
<xs:element name="RangeBeginningDate" type="xs:date" />
```

```
<xs:element name="RangeBeginningTime" type="xs:time" />
  </xs:all>
  </xs:complexType>
  </xs:element>
- <xs:element name="SpatialDomainContainer">
- <xs:annotation>
<xs:documentation xml:lang="en">Spatial information on granule. Currently contains
  two element, VerticalSpatialDomain and HorizontalSpatialDomainContainer, which
  provide information on Vertical fields, and the following horizontal fields:
  GPolygon and BoundingRectangle lat/lon coordinates for the
  granule.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:sequence>
<xs:element name="LocalityValue" minOccurs="0" />
<xs:element name="ZoneIdentifier" minOccurs="0" />
<xs:element ref="VerticalSpatialDomain" minOccurs="0" />
<xs:element ref="HorizontalSpatialDomainContainer" minOccurs="0" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
- <xs:element name="VerticalSpatialDomain">
- <xs:annotation>
<xs:documentation xml:lang="en">Contains vertical spatial fields.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:sequence>
<xs:element ref="VerticalSpatialDomainContainer" minOccurs="0"
  maxOccurs="unbounded" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
- <xs:element name="VerticalSpatialDomainContainer">
- <xs:annotation>
<xs:documentation xml:lang="en">Vertical spatial information on the granule. Contains
  two elements, a specific vertical spatial type and value.</xs:documentation>
```

```
    </xs:annotation>
- <xs:complexType>
- <xs:all>
  <xs:element ref="VerticalSpatialDomainType" />
  <xs:element ref="VerticalSpatialDomainValue" />
    </xs:all>
    </xs:complexType>
    </xs:element>
- <xs:element name="VerticalSpatialDomainType">
- <xs:annotation>
  <xs:documentation>Contains specific vertical spatial type.</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="80" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="VerticalSpatialDomainValue">
- <xs:annotation>
  <xs:documentation>Contains specific vertical spatial value.</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="80" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="HorizontalSpatialDomainContainer">
- <xs:annotation>
  <xs:documentation xml:lang="en">Horizontal spatial information on the granule.
    Currently contains two elements, GPolygon and
    BoundingRectangle.</xs:documentation>
    </xs:annotation>
```

- <xs:complexType>
- <xs:all>
<xs:element ref="**GPolygon**" minOccurs="**0**" />
<xs:element ref="**BoundingRectangle**" minOccurs="**0**" />
</xs:all>
</xs:complexType>
</xs:element>
- <xs:element name="**GPolygon**">
- <xs:annotation>
<xs:documentation xml:lang="**en**">**Horizontal spatial information, on the granule, in the GPolygon form.**</xs:documentation>
</xs:annotation>
- <xs:complexType>
- <xs:all>
<xs:element ref="**Boundary**" />
</xs:all>
</xs:complexType>
</xs:element>
- <xs:element name="**Boundary**">
- <xs:annotation>
<xs:documentation xml:lang="**en**">**Boundary of the GPloygon, consisting of a series of points.**</xs:documentation>
</xs:annotation>
- <xs:complexType>
- <xs:sequence>
<xs:element ref="**Point**" minOccurs="**0**" maxOccurs="**unbounded**" />
</xs:sequence>
</xs:complexType>
</xs:element>
- <xs:element name="**Point**">
- <xs:annotation>
<xs:documentation xml:lang="**en**">**A GPolygon pint, consisting of a longitude and latitude coordinate.**</xs:documentation>
</xs:annotation>
- <xs:complexType>

```
- <xs:sequence>
  <xs:element ref="ExclusionFlag" minOccurs="0" maxOccurs="1" />
  <xs:element ref="PointLongitude" />
  <xs:element ref="PointLatitude" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="ExclusionFlag">
- <xs:annotation>
  <xs:documentation>Indicates whether the polygon should be included or
    excluded</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
  <xs:restriction base="xs:string" />
    </xs:simpleType>
    </xs:element>
- <xs:element name="PointLatitude">
- <xs:annotation>
  <xs:documentation>decimal degrees (+ = north (default), - =
    south)</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:minInclusive value="-90.0000" />
  <xs:maxInclusive value="+90.0000" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="PointLongitude">
- <xs:annotation>
  <xs:documentation>decimal degrees (+ = east (default), - = west)</xs:documentation>
    </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:string">
```

```
<xs:minInclusive value="-180.0000" />
<xs:maxInclusive value="+180.0000" />
  </xs:restriction>
  </xs:simpleType>
  </xs:element>
- <xs:element name="BoundingRectangle">
- <xs:annotation>
<xs:documentation xml:lang="en">Horizontal spatial information, on the granule, in
  the BoundingRectangle form.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:sequence>
<xs:element ref="WestBoundingCoordinate" />
<xs:element ref="NorthBoundingCoordinate" />
<xs:element ref="EastBoundingCoordinate" />
<xs:element ref="SouthBoundingCoordinate" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
- <xs:element name="OrbitCalculatedSpatialDomain">
- <xs:annotation>
<xs:documentation xml:lang="en">Orbit spatial temporal information on the
  granule.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:sequence>
<xs:element ref="OrbitCalculatedSpatialDomainContainer" minOccurs="0"
  maxOccurs="unbounded" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
- <xs:element name="OrbitCalculatedSpatialDomainContainer">
- <xs:annotation>
```

```
<xs:documentation xml:lang="en">Orbital Information on the granule. Includes,
  OrbitNumber, EquatorCrossingLongitude, EquatorCrossingDate and
  EquatorCrossingTime.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:sequence>
<xs:element name="OrbitNumber" type="xs:unsignedLong" minOccurs="0" />
<xs:element name="StartOrbitNumber" minOccurs="0" />
<xs:element name="StopOrbitNumber" minOccurs="0" />
<xs:element ref="EquatorCrossingLongitude" />
<xs:element name="EquatorCrossingDate" type="xs:date" />
<xs:element name="EquatorCrossingTime" type="xs:time" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
- <xs:element name="EquatorCrossingLongitude">
- <xs:annotation>
<xs:documentation>decimal degrees (+ = east (default), - = west)</xs:documentation>
  </xs:annotation>
- <xs:simpleType>
- <xs:restriction base="xs:float">
<xs:minInclusive value="-180.0000" />
<xs:maxInclusive value="+180.0000" />
  </xs:restriction>
  </xs:simpleType>
  </xs:element>
- <xs:element name="Platform">
- <xs:annotation>
<xs:documentation xml:lang="en">The platform (short name) the instrument was
  mounted on. Includes reference to the instrument container element. The platform
  container structure is the same in ECHOs schema. The S4PA Collection level meta-
  data uses a platform container, which has the same structure in GCMD-DIF
  schema.</xs:documentation>
  </xs:annotation>
- <xs:complexType>
- <xs:sequence>
```

```
  <xs:element ref="PlatformShortName" />
  <xs:element ref="Instrument" minOccurs="0" maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="Instrument">
- <xs:annotation>
  <xs:documentation xml:lang="en">The instrument associated with the sensor. Includes
    an InstrumentShortName and reference to sensor container
    element.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
- <xs:element name="InstrumentShortName">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="20" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
  <xs:element ref="Sensor" minOccurs="0" maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="Sensor">
- <xs:annotation>
  <xs:documentation xml:lang="en">Sensor (short name) used to measure
    data.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:all>
  <xs:element ref="SensorShortName" />
    </xs:all>
    </xs:complexType>
    </xs:element>
```

```
- <xs:element name="PSAs">
- <xs:annotation>
  <xs:documentation xml:lang="en">PI specific attributes.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="PSA" minOccurs="0" maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="PSA">
- <xs:annotation>
  <xs:documentation xml:lang="en">PI specific attribute. Includes, PSAName and
    PSAValue.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="PSAName" />
  <xs:element ref="PSAValue" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="PSAName">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="PSAValue">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="500" />
    </xs:restriction>
```

```
    </xs:simpleType>
    </xs:element>
- <xs:element name="MeasuredParameters">
- <xs:annotation>
  <xs:documentation xml:lang="en">PI specific attributes.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="MeasuredParameter" minOccurs="0" maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
- <xs:element name="MeasuredParameter">
- <xs:annotation>
  <xs:documentation xml:lang="en">PI specific attribute. Includes, ParameterName,
    QAPercentMissing, QAPercentOutofBounds, QAInterpolatedData,
    AutomaticQualityFlag, AutomaticQualityFlagExplanation, OperationalQualityFlag,
    OperationalQualityFlagExplanation, ScienceQualityFlag,
    ScienceQualityFlagExplanation.</xs:documentation>
    </xs:annotation>
- <xs:complexType>
- <xs:sequence>
  <xs:element ref="ParameterName" minOccurs="0" />
  <xs:element ref="QAPercentMissing" minOccurs="0" />
  <xs:element ref="QAPercentOutofBounds" minOccurs="0" />
  <xs:element ref="QAPercentInterpolatedData" minOccurs="0" />
  <xs:element ref="QAPercentCloudCover" minOccurs="0" />
  <xs:element ref="AutomaticQualityFlag" minOccurs="0" />
  <xs:element ref="AutomaticQualityFlagExplanation" minOccurs="0" />
  <xs:element ref="OperationalQualityFlag" minOccurs="0" />
  <xs:element ref="OperationalQualityFlagExplanation" minOccurs="0" />
  <xs:element ref="ScienceQualityFlag" minOccurs="0" />
  <xs:element ref="ScienceQualityFlagExplanation" minOccurs="0" />
    </xs:sequence>
    </xs:complexType>
```

```
    </xs:element>
- <xs:element name="ParameterName">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="QAPercentMissing">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="QAPercentOutofBounds">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="QAPercentInterpolatedData">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="QAPercentCloudCover">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
```

```
    </xs:simpleType>
    </xs:element>
- <xs:element name="AutomaticQualityFlag">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="AutomaticQualityFlagExplanation">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="500" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="OperationalQualityFlag">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="OperationalQualityFlagExplanation">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="500" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="ScienceQualityFlag">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="50" />
```

```
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="ScienceQualityFlagExplanation">
- <xs:simpleType>
- <xs:restriction base="xs:string">
  <xs:maxLength value="500" />
    </xs:restriction>
    </xs:simpleType>
    </xs:element>
- <xs:element name="ProducersMetaData">
- <xs:annotation>
  <xs:documentation xml:lang="en">This section will contain a copy of the ODL file in a
    text output format.</xs:documentation>
    </xs:annotation>
    </xs:element>
    </xs:schema>
```

Appendix B. Draft Convention for GCMD DIF Contents and Formatting at the GES DISC

1. General Background of the Metadata at GES DISC

The current state of the management of dataset-level metadata at the GES DISC is complex and heterogeneous, with dataset-level metadata stored in a variety of locations: ECS, ECHO, S4PA, Giovanni, WHOM, PILOT, Mirador and GCMD. As a result, metadata are sometimes conflicting, missing or difficult to find. With evolution the GES DISC is moving to store dataset-level metadata as much as possible in a single place, namely GCMD.

GCMD already combines an easy-to-use web interface with a sophisticated database and a variety of access modes. Also, storing dataset-level metadata in GCMD is an EOSDIS requirement. However, in order to be useful to the various GES DISC and EOSDIS applications, some conventions for GES DISC metadata must be established to get the full benefit.

The goal of this paper is to develop standardization for Global Change Master Directory (GCMD) Directory Interchange Formats (DIFs) use for Goddard Earth Science Data and Information Services Center (GES DISC) Data Products.

The following table shows the GCMD Fields and Sub-Fields used by the different GES DISC Data Clients:

GCMD Fields and Sub-Fields	**S4PA**	**Mirador**	**Giovanni**	**ECHO**
Entry_ID	X	X	X	X
Entry_Title	X	X	X	X
Dataset_Creator		X		X
Dataset_Title				
Dataset_Series_Name				
Dataset_Release_Date				
Dataset_Release_Place				

Dataset_Publisher				
Version	X	X	X	X
Issue_Identification				
Data_Presentation_Form				
Other_Citation_Details		X		X
Online_Resource		X		X
Category		X	X	X
Topic		X	X	X
Term		X	X	X
VariableDetailed_Variable		X		X
ISO_Topic_Category				
Keyword		X		
Sensor_Name:[ShortName]>[LongName]	X	X	X	X
Source_Name:[ShortName]>[LongName]	X	X	X	X
Start_Date	X	X	X	X
Stopt_Date	X	X	X	X
Data_Set_Progress		X		X
Southernmost_Latitude	X	X	X	X
Northernmost_Latitude	X	X	X	X
Westernmost_Longitude	X	X	X	X
Easternmost_Longitude	X	X	X	X
Minimum_Altitude				
Maximum_Altitude				
Minimum_Depth				

Maximum_Depth				
Location		X		X
Latitude_Resolution		X		X
Longitude_Resolution		X		X
Horizontal_Resolution_Range		X		X
Vertical_Resolution	X	X		X
Vertical_Resolution_Range	X	X		X
Temporal_Resolution		X		X
Temporal_Resolution_Range		X		X
Project:[ShortName]>[LongName]				X
Quality	X	X		X
Access_Constraints		X		X
Use_Constraints		X		X
Originating_Center		X		X
Data_Center_Name:[ShortName]>[LongName]		X		X
Data_Center_URL		X		X
Role		X		X
Last_Name		X		X
Email		X		X
Phone		X		X
Fax		X		X
Address		X		X
City		X		X
Province_Or_State		X		X

Postal_Code		X		X
Country		X		X
Distribution_Media				X
Distribution_Size				X
Distribution_Format		X		X
Fees				X
Multimedia_Sample>File				
Multimedia_Sample>URL				X
Multimedia_Sample>Format				
Multimedia_Sample>Caption				X
Summary		X		X
Related_URL> URL_Content_Type				
Related_URL> URL				X
Related_URL> Description				X
Parent_DIF				X
Metadata_Name				
Metadata_Version				
DIF_Creation_Date				
Last_DIF_Revision_Date				X
DIF_Revision_History				X

2. Collection Metadata registered in GCMD

All GES DISC Collection Metadata will need to be registered in NASA's Global Change Master Directory (GCMD). The GCMD will be the main repository for all public GES DISC Collection Metadata. The GES DISC will generate and maintain the Directory Interchange Format (DIF) for Collection Metadata. A description of the GCMD DIF can be found at http://gcmd.nasa.gov/User/difguide/whatisadif.html.

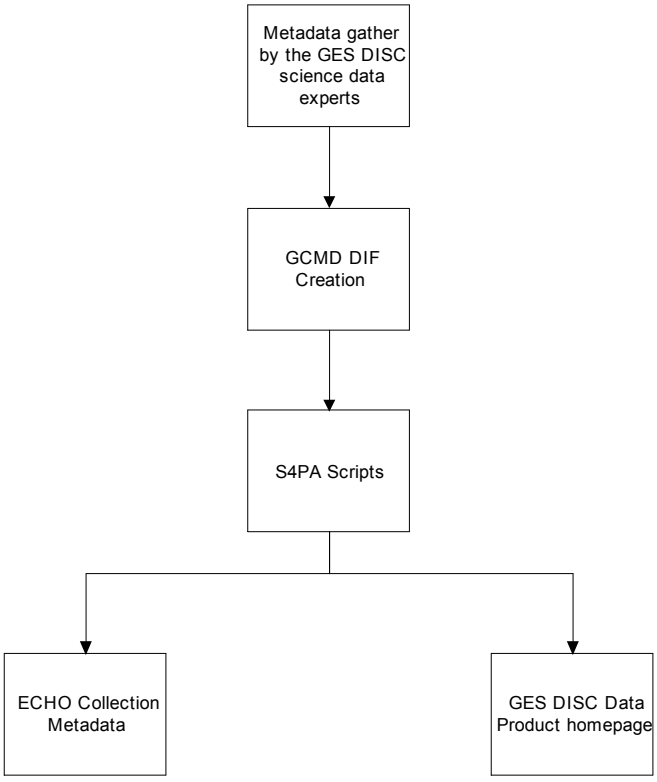
The GCMD DIF Metadata will be used for ECHO Collection Metadata for GES DISC data products and the GES DISC Data Product Overview homepages. (*see Figure 1*: Metadata Flow). Also, Mirador's Keyword search is base on the GCMD Parameter Keywords in the data sets' DIFs.

To create or modified a DIF go to http://gcmd.nasa.gov/User/authoring.html or for information on creating the DIF go to http://gcmd.nasa.gov/User/difguide/difman.html.

Or

The Developer can create multiple XML files using the GCMD DIF XML Schema and then send them to metadata@gcmd.nasa.gsfc.gov. The GCMD DIF XML Schema is located at http://gcmd.nasa.gov/User/difguide/difman.html.

Figure 1: Metadata Flow



3. Guidelines for define DIFs

The GCMD only requires 8 attribute fields but GES DISC DIFs should have more because the metadata will be reformatted for ECHO and the multiple GES DISC Data Clients. The extra fields will help with Data Discovery of the data sets using the different data clients.

The GCMD has guidelines on developing DIFs for their repository:

- The DIFS should not have 2 different processing levels of a product join together in one DIF, such as, combining a level 1A product with the Level 1B product. *(Example: Visible and Infrared Scanner (VIRS) Level 1 Raw and Calibrated Radiance Products (TRMM Products 1A01/1B01))*.
- DIFs should not combine two different versions of a Collection together. *(Example: MODIS/Terra Aerosol 5-Min L2 Swath 10k(MOD04_L24))*. http://gcmd.nasa.gov/getdif.htm?MOD04_L24
- Finally, the DIF developer should not create DIFs with too little metadata or generic metadata used to define more then one data collection: It will not help users discover the right data sets for their research.

The following are examples of good DIFs from the GCMD:

http://gcmd.nasa.gov/getdif.htm?USGS_NED

http://gcmd.nasa.gov/getdif.htm?NASA_MODIS_RAPID_RESPONSE

4. GES DISC GCMD DIF Standard Attributes

DIFs can easily be generated using the GCMD DocBUILDER tool at http://gcmd.nasa.gov/User/authoring.html.

When creating a DIF for the GES DISC the following fields should be used (if applicable):

Entry_ID	Entry_Title	Parameters (Science_Keywords)
ISO_Topic_Category	Data_Center	Summary
Metadata_Name	Metadata_Version	Temporal_Coverage
Data_Set_Progress	Spatial_Coverage	Location
Data_Resolution	Data_Set_Citation	Instrument
Platform	Project	Quality
Access_Constraints	Use_Constraints	Distribution
Related_URL	Keyword (Ancillary Keyword)	Originating_Center
Multimedia_Sample	Reference	DIF_Creation_Date
Last_DIF_Revision_Date	DIF_Revision_History	Parent_DIF (If applicable)

Description of the different fields, sub-fields, formats and examples:

Entry_ID

The field value should be the GES DISC plus the ECS or S4PA ShortName and VersionID of the data set. The GCMD limit is 80 characters for the field.

Format: GES DISC + ShortName + VersionID

Example: GES_DISC_MOD021KM_v4 or GES_DISC_TRMM_3A46_v6.

Entry_Title

The Entry_Title could be the same as the ECS LongName or the S4PA LongName with a 220 character limit. It should be a short description of the product.

Format: GES DISC + Sensor/Platform + Data subject or parameter + Processing level + granule temporal resolution (5-min, daily, weekly, etc) + Spatial resolution + version (example: (v2)) + (ShortName). (The variables between Sensor/Platform and Spatial resolution can be order differently within the format.)

Example: GES DISC MODIS/Terra Aerosol 5-Min L2 Swath 10km (v2) (MOD04_L2)

Parameters (Science_Keywords)

The Science keywords are very important for data discovery in the GCMD. The DIF developer should place as many as possible keyword groupings in a DIF. A list of the keywords can be found at http://gcmd.nasa.gov/Resources/valids/gcmd_parameters.html.

Category: [EARTH SCIENCE]

Topic: [Topic keyword]

Term: [Term keyword]

Variable: [Variable keyword]

Detailed_Variable: [Free text] – In this field the DIF Developer can place very specific keyword values about the data.

Detailed_Variable_Description (Future: Fall/2006) [Free text]: In the future the DIF Developer will be able to precisely describe the Detailed_Variable keyword values with as much detail as the want to add to the field value.

ISO_Topic_Category

This field is a keyword valid for ISO standard. A list of valids for this field can be found at http://gcmd.nasa.gov/User/difguide/iso_topics.html.

Example: ISO_Topic_Category: Climatology/Meteorology/Atmosphere

Data_Center

The Data_Center field should be filled-out with the following values:

Data_Center_Name: NASA/GSFC/ESED/GCDC/GES-DISC/DAAC >Distributed Active Archive Center, Goddard Earth Sciences, Data and Information Services Center, Global Change Data Center, Earth-Sun Exploration Division, Goddard Space Flight Center, NASA

Data_Center_URL: http://disc.gsfc.nasa.gov/

Role: DATA CENTER CONTACT

Last_Name: GES DISC Help Desk Support Group

Email: help-disc@listserv.gsfc.nasa.gov

Phone: 301-614-5224

Fax: 301-614-5268

Address (more than one): Distributed Active Archive Center Global Change Data Center, Code 610.2, NASA Goddard Space Flight Center

City: Greenbelt

Province_Or_State: MD

Postal_Code: 20771

Country: USA

Summary

This field is a description of the Data Collection with a limit of 80 characters per line but no line limit. The DIF developer should keep in mind that the DIFs will also be used to create the ECHO Collection Metadata, which has a 4000 character limit. The DIF developer should have the most important information about the data set within this 4000 character limit. Then any additional information should be added to the summary entry after the limit. The Summary should not be a generic description to cover a group of data products but a specific description of one data product.

Metadata_Name

The current standard for the DIF is known (Metadata_Name: CEOS IDN DIF) and is automatically populated when using online authoring tools.

Metadata_Version

The current standard for the DIF is known (Metadata_Version: 9.0) and is automatically populated when using online authoring tools.

Temporal_Coverage

Represents the Start and Stop dates for a data set from the first granule to the last granule.

Format: Start_Date: yyyy-mm-dd

Stop_Date: yyyy-mm-dd

In cases were datasets are still in progress the Stop Date should be omitted from the Temporal entry.

Data_Set_Progress

The production status of the data set. The Status can be set to:

Planned – Future data set

In Work – currently undergoing production or continuously being update.

Complete – not farther updates or data collection will be made.

Spatial_Coverage

The field represents the (horizontal) geographic coordinates and the vertical profile (if applicable) of the data set. **Sub-fields**:

Southernmost_Latitude: The southern most limit of the Data Collection less than or equal to +90 decimal degrees and greater than or equal to -90 decimal degrees.

Northernmost_Latitude: The Northern most limit of the Data Collection less than or equal to +90 decimal degrees and greater than or equal to -90 decimal degrees.

Westernmost_Longitude: The Western most limit of the Data Collection less than or equal to +180 decimal degrees and greater than or equal to -180 decimal degrees.

Easternmost_Longitude: The Eastern most limit of the Data Collection less than or equal to +180 decimal degrees and greater than or equal to -180 decimal degrees.

Minimum_Altitude [Not Required]: The Altitude lower level limit value of the Data Collection.

Maximum_Altitude [Not Required]: The Altitude Higher level limit value of the Data Collection.

Minimum_Depth [Not Required]: The upper-most depth of the Data Collection.

Maximum_Depth [Not Required]: The lowest-most depth of the Data Collection.

Examples of units for Altitude and Depth are: m (meters), km (kilometers), ft (feet), hPS (hecto Pascals) and mb (millibars).

Location

The value(s) for this field is/are the Geographic or Atmosphere location name of area **Example**: Ireland.

The valids for this field are available at http://gcmd.nasa.gov/Resources/valids/location.html.

Data_Resolution

The geographic, vertical, or time resolution values of a data set. Some of the attributes may not apply to the data set. **Sub-fields**:

Latitude_Resolution: Examples of units: 1 meter, 1 km, 5 minute or 1 degree.

Longitude_Resolution: Examples of units: 1 meter, 1 km, 5 minute or 1 degree.

Vertical_Resolution: Examples of units: 5 meters or 20 kPa for this attribute.

Temporal_Resolution: Examples of units: 5 minute, Daily, Weekly or Monthly.

Data_Set_Citation

This field supplies information to the user for properly citing the data and data creator.

Example:

Dataset_Creator: AIRS project at NASA JPL

Dataset_Title: AIRS/Aqua FINAL AIRS Level 2 Cloud Clear Radiance Product

Dataset_Release_Date: 2002-11-04

Dataset_Release_Place: Greenbelt, MD, USA

Dataset_Publisher: Goddard Earth Science Data and Information Services Center (GES DISC)

Version: Version 4

Issue_Identification: AIRI2CCF24

Data_Presentation_Form: Digital Science Data

Online_Resource: http://disc.gsfc.nasa.gov/AIRS/airsL2_CC.shtml

Instrument

Name of the device used to measure or record the data.

Format: Sensor_Name: Sensor ShortName > Sensor LongName

Example: Sensor_Name: PR > Tropical Precipition Radar

Platform

The name of the spacecraft, ship, or craft that the Sensor is mounted on to collected the data.

Format: Source_Name: Source ShortName > Source LongName

Example: Source_Name: UARS > Upper Atmosphere Research Satellite

Project

The field valid represents scientific endeavor encompassing data.

Format: Project: Project ShortName > Project LongName

Example: Project: TRMM > Tropical RainFall Measuring Mission Project

Quality

The field represents the information about the accuracy of the data.

Example: Group: Quality

> The quality of this version 004 product is established as 'Validated'. This means its accuracy has been assessed over a widely distributed set of locations and time periods via several ground-truth and validation efforts. Results are peer-reviewed and published at scientific conferences and workshops.For the latest information please see: http://modis-atmos.gsfc.nasa.gov/products_calendar_overview.html.

End_Group

Access_Constraints

Any restrictions and/or legal prerequisites for accessing the collection.

Example: Group: Access_Constraints

> None

End_Group

Use_Constraints

Any restrictions and/or legal prerequisites for using the collection.

Example: Group: Use_Contraints

> This Data set (version 002/first public release) is not fully validated yet. Before using it in any publication please contact algorithm team lead (Dr. Pawan K. Bhartia, Pawan.K.Bhartia@nasa.gov)for the current known problems and updates.

End_Group

Distribution

Information about the data set's Media, Size and Format. **Sub-fields**:

Distribution_Media: FTP
Distribution_Size: Approx 35 MB per file
Distribution_Format: HDF-EOS5
Fees: None

Related_URL

This field provides an URL link to internet sites that contain information about the data, such as project home pages, data archive, data description, quality, and web services. **Sub-fields**:

Content Type: GET DATA
URL: http://disc.sci.gsfc.nasa.gov/data/datapool/OMI/Level2G/OMTO3G/
Description: Access OMI Level-2G, the Global Gridded (0.25x0.25 deg grids) Ozone (O3) Total Column product OMTO3G data from the data pool.

Content Type: GET SERVICE
URL: http://Giovanni.gsfc.nasa.gov/
Description: This interface is designed for visualization and analysis of the Aura OMI Level 2G Daily Global Products (Experimental). Users can select area, generate plots or ASCII Output for area average (Area Plot), time series (Time Plot), and Hovmoller diagram. The animation is also available.

Content Type: VIEW RELATED INFORMATION
URL: http://disc.gsfc.nasa.gov/Aura/OMI/omto3.shtml
Description: OMTO3 product page at the GES DISC.

Content Type: VIEW PROJECT HOME PAGE
URL: http://www.knmi.nl/omi/research/news/index.html
URL: http://toms.gsfc.nasa.gov/omi/
Description: OMI Home Page and OMTO3G Algorithm Page.

Content Type: GET RELATED DATA SET METADATA (DIF)
URL: http://gcmd.nasa.gov/getdif.htm?OMTO3
Description: Metadata describing the GES-DISC Interactive Online Visualization and Analysis Infrastructure (Giovanni).

Content Type: GET RELATED SERVICE METADATA (SERF)
URL: http://gcmd.nasa.gov/getserf.htm?Giovanni
Description: Metadata describing the GES-DISC Interactive Online Visualization and Analysis Infrastructure (Giovanni).

Order:

For the Related URLs:

GET DATA (Data Pool FTP site, WHOM access, Mirador site, etc.),

GET SERVICE (Data subsetting service or any other data modification service),

GET MAP SERVICE (OGC),

VIEW RELATED INFORMATION (dataset documentation and the general information site for the collection),

VIEW PROJECT HOME PAGE,

GET RELATED DATA SET METADATA (DIF), and

GET RELATED SERVICE METADATA (SERF).

Keyword (Ancillary Keyword)

Any words that can farther describe the Data Set should been add in this field.

Examples: Atmospheric Chemistry, Total Column Ozone, Ozone Layer, Air Quality, Ozone Depletion, Ozone Hole, Aerosol Index and UV based Effective Surface Reflectivity

Originating_Center

The center or organization that originally generated the data set.

Example: Originating_Center: GES DISC or MODAPS

Multimedia_Sample

Sample data image

Example:

Group: Multimedia_Sample

URL:
http://disc.sci.gsfc.nasa.gov/data/datapool/TRMM/01_Data_Products/02_Gridded/01_Monthly_Tmi_Prod_3A_11/sample.gif

Format: GIF

Caption: 3A11 TMI oceanic monthly accumulated surface rainfall, 2005/11/01-2005/12/01

Group: Description

Sample image for TRMM data product 3A11, TMI oceanic monthly accumulated surface rainfall, 2005/11/01-2005/12/01

End_ Group: Description

End_ Group: Multimedia_Sample

Reference

Bibliographic references pertaining to the data set described in the DIF.

Example:

Group: Reference

Wilheit, T.T., A.T.C Chang and L.S. Chiu, 1991: Retrieval of Monthly Rainfall Indices from Microwave Radiometric Measurements Using Probability Distribution Functions. J. Atmos. Oceanic Tech., 8, 118-136.

Interface Control Specification Between the Tropical Rainfall Measuring Mission Science Data and Information System (TSDIS) and the TSDIS Science User (TSU). Volume 4: File Specifications for TSDIS Products-Level 2 and Level 3. NASA Goddard Space Flight Center, March 5, 1999.

End_Group: Reference

Parent_DIF (If applicable)

The field allows an association between generalized DIF (parent) and specific DIFs (Children). A link will be supplied to the parent DIF from the associated (child) DIF.

Format: Parent_DIF: [Entry_ID of the Parent_DIF]

Example: Parent_DIF: AIRIBRAD2

DIF_Creation_Date

The date the DIF was created for the data set.

Format: DIF_Creation_Date: [yyyy-mm-dd]

Last_DIF_Revision_Date

The date of the last revision of the data set.

Format: Last_DIF_Revision_Date: [yyyy-mm-dd]

DIF_Revision_History

A log of changes made to the DIF over time.

Example:

Group: DIF_Revision_History

1999-09-13: Added "1997" to first paragraph of "Summary."

2000-09-15: Updated broken link in Reference Field.

2003-01-23: Added new location valid, "TROPICS"

2006-07-13: Added "_V6" to entry ID to prevent overwriting during reprocessing

End_Group: DIF_Revision_History