

10. Metadata Administration

Every science data product generated and archived by the ECS must be described to the system by metadata which are put into an inventory and then used to retrieve and distribute the data to users of the system. The ECS Earth Science Data Model, described in documents 420-TP-015-002 and 420-TP-016-002, organizes the metadata into groups of related attributes and services to be performed on the data products. These "core" attributes are necessary to identify, interpret and perform services on granules and collections. The Data Model also provides for "product-specific" attributes, i.e. attributes which are unique to a specific data product.

The smallest aggregation of data that is independently described and inventoried in ECS is referred to as a data granule. Granules are organized into logical groupings called collections in which the granule metadata varies principally by time or location, called single-type collections. In Release 2 only single-type collections are supported (more complex organizations will be supported in later ECS releases).

Every collection is described by an Earth Science Data Type (ESDT) and is made known to the system by an ESDT descriptor file and associated software code which is built into the Data Server's dynamic link library (DLL) to perform the services.

Metadata administration includes creating and updating ESDTs within ECS. Collections may be modified and updated over time. In addition, quality assurance will be performed after ESDTs have been installed and granules have been generated and stored in the archives. Collection-level metadata can only be updated by installing a revised ESDT. In Release 2, the only granule-level metadata which can be updated manually (i.e. not as a result of an operation such as Subsetting which modifies the science data content of a granule) are the Quality Assurance flags and explanations. Procedures for updating these flags are provided in Section 15.

10.1 ESDT Descriptor Files

The primary task in establishing a collection is providing the core and product-specific metadata attribute values. This is done by creating an Earth Science Data Type (ESDT) descriptor file. The descriptor file is also used to specify the data services that are available for granules that belong to the collection. These services are implemented as methods of a Dynamic Link Library (DLL) containing C++ code to accomplish each service. The descriptor file and the DLL are the means by which a collection is made known to the Science Data Server (SDSRV).

The ESDT descriptor is composed of sections, all in ODL, containing the following information:

- Collection level metadata attributes with values contained in the descriptor.
- Granule level metadata attributes whose values are supplied primarily by the Product Generation Executives (PGEs) during runtime.

- Valid values and permitted ranges for all product-specific attributes.
- List of services for all the granules in the collection and events which trigger responses throughout the system.

The ESDT descriptor file is created by Metadata Works (MDWorks in Figure 10.2-1), discussed in Section 10.3.3.

The services which apply to a collection are specified in the ESDT descriptor file. Metadata Works automatically inserts standard ECS-supplied services such as insert and search into the descriptor file. Product-specific services, such as Subsetting or a product-specific acquire, require executable code to enact those services. This code is contained in the DLL. The DLL is written and tested by either the ECS developers or sustaining engineering personnel at the DAAC.

After the ESDT (both descriptor file and DLL) has been generated it must be installed on the Science Data Server before the first data granule can be inserted. During this installation process, information from the ESDT Descriptor File is propagated to the Data Management and Interoperability subsystems, and to the Subscription Server, which must all be operating during the ESDT installation process. The detailed procedures for ESDT installation into ECS are described in Section 11.3.

10.1.1 Steps in Generating a Descriptor File

ESDTs for Distributable Product

These are the typical steps used in generating a descriptor file:

.Identify desired collection-level metadata attributes

- For permanent and interim files use only the minimum attributes.
- For distributable products identify all applicable attributes. This will involve reading appropriate documentation and interacting with the data provider.

.Identify granule-level attributes

- If a sample metadata configuration file is available from the data provider, use this.

.Check valids for core attributes (write CCR if new valids are required).

.Check PSAs (register PSAs if new).

.Gather metadata into a spreadsheet, or use Metadata Works to enter metadata directly.

.Use custom built scripts to generate the descriptor file from the spreadsheet, if used.

.Verify the descriptor file as outlined below.

.Check descriptor files into ClearCase.

.Notify the DLL Team Lead of the newly prepared descriptor files.

10.1.2 Verifying Descriptor Files

- Run the PERL script "update.pl", following the instructions in the script prologue. (This script makes sure that the inventory metadata attributes are all listed as event qualifiers in the EVENT group.)
- Run the PERL script "esdtQC.pl", following the instructions in the script prologue. Make any necessary corrections in response to errors issued, and rerun. Repeat until there are no errors. (This script checks for more than 30 common descriptor file errors.)
- Run the PERL script "required.pl", following the instructions in the script prologue. Add any missing attributes as indicated.
- Run the "testodl.csh" utility to ensure that there are no errors in the ODL structure for the descriptor file. Make any necessary corrections in response to errors issued, and rerun. Repeat until there are no errors.

10.2 Preparation of Earth Science Data Types

An ESDT goes through pre-operational life cycle steps starting with an analysis of the collection's need and continuing through development and operational installation. This process involves actions by the Data Provider or User in addition to ECS. These procedures are detailed in Project Instruction SO-1-002, "Earth Science Data Type Generation Procedures". The overall workflow is shown in Figure 10.2-1.

DEFINITIONS

Archive - A File Type indicating granules will be inserted to Data Server for long term storage and acquisition for distribution.

Full - A level of metadata coverage intended for data products which are produced within ECS.

Collection - A related group of data granules.

Granule - The smallest data element which is identified in the inventory tables.

Interim - A File Type indicating granules are temporarily stored in support of product generation.

Intermediate - A level of metadata coverage intended for contemporaneous data products which are not produced within ECS.

Limited - A level of metadata coverage intended for heritage data products brought into ECS for distribution

Minimal - A level of metadata coverage sufficient to uniquely identify a collection or granule.

Permanent - A File Type indicating static or semi-static granules which are used only as inputs in product generation.

Product - Attributes that are defined by the data provider in support of searching for specific granules

Valid - An allowable metadata value.

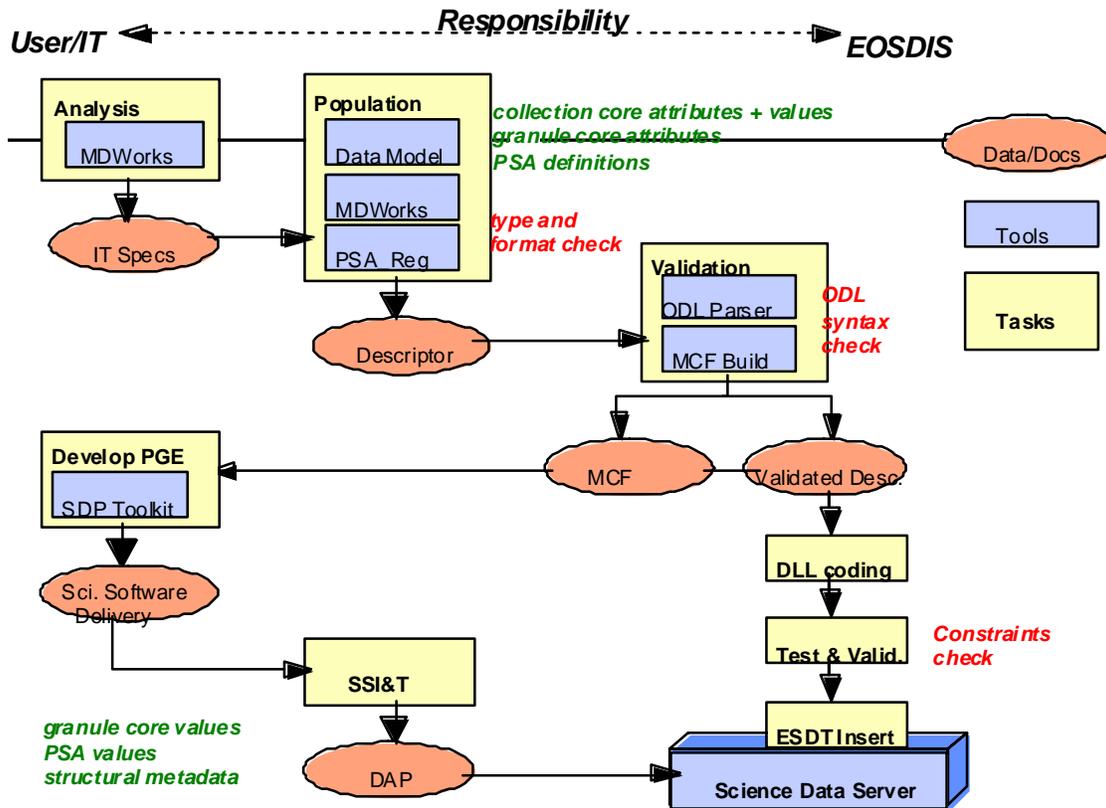


Figure 10.2-1. The ESDT Generation Process

PROCESS:

1. Need Analysis

- The baseline list of science ESDTs and their services is controlled by the ECS CCB. This baseline was established through an analysis of the ECS Functional and Performance Requirements Specification, the Technical Baseline established from inputs from the Ad Hoc Working Group on Production, and meetings with the individual data providers to define the basic requirements of each ESDT.

These basic requirements are:

- Data Provider File Designation,

- File Type (Permanent, Interim, Archive)
- Level of Metadata Coverage (Minimal, Limited, Intermediate, Full)
- For new ESDTs not currently in the development baseline, the result of the Need Analysis forms the basis for approving the inclusion of the ESDT into ECS. This is accomplished through the CCR process, governed by PIs "Configuration Change Request Preparation" [CM-1-003] and "Change Control Board Process" [CM-1-004].

2. ESDT Specification

- This step results in a set of specifications extending the results of the needed analysis and providing the information needed to implement an ESDT. This step is executed only if the ESDT has been included in the baseline. The roles and responsibilities for developing the specification are as above.

The specifications must include:

- ShortName and VersionID of the ESDT
- A list of the metadata attributes needed, valids, and any constraints on attributes. This list is to be drawn from the B.x Descriptor File Template. This template, under CCB control, is based on the attributes defined in "Release-B SDPS Database Design and Database Schema Specifications" [311-CD-008-001] (i.e., DID 311), as modified for B.0 (for example) by "B.0 Earth Science Data Model " [420-TP-015-002].
- A list and specification of the services needed (e.g., specification of the INSERT, SEARCH, ACQUIRE and SUBSCRIPTION semantics).

3. ESDT Generation

- Once the ESDT Specification has been developed and the applicable attributes identified, the necessary metadata has to be gathered, the metadata values checked against the valid values and the product-specific attributes (PSA) need to be checked against the list of PSAs that are already defined (see Fig. 10.2-2).
- Once the collection-level metadata and granule-level attributes have been checked, then the descriptor file is generated, the Dynamic Link Library produced, and testing and validation of the ESDT performed. This process is further elaborated in the sections below.
- For a one-of-a-kind, distributable product with Full metadata coverage, this process can take up to six weeks to accomplish. For a related group of products with identical services, much of the Descriptor File and DLL of the first ESDT can be reused, and the cycle time for preparing subsequent ESDTs in the related group is much less.

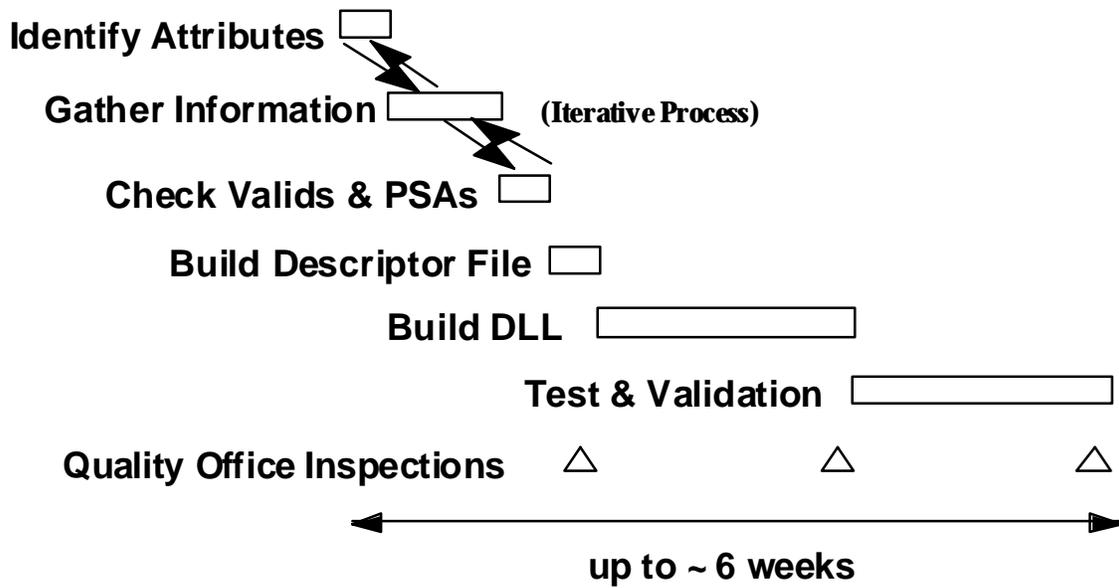


Figure 10.1-2. Steps in ESDT Development

10.3 Tools used in Generating a Descriptor File

Figure 10.3-1 shows the some of the tools that have been developed, and indicates supporting information flows between these tools. Brief descriptions of these tools are provided below.

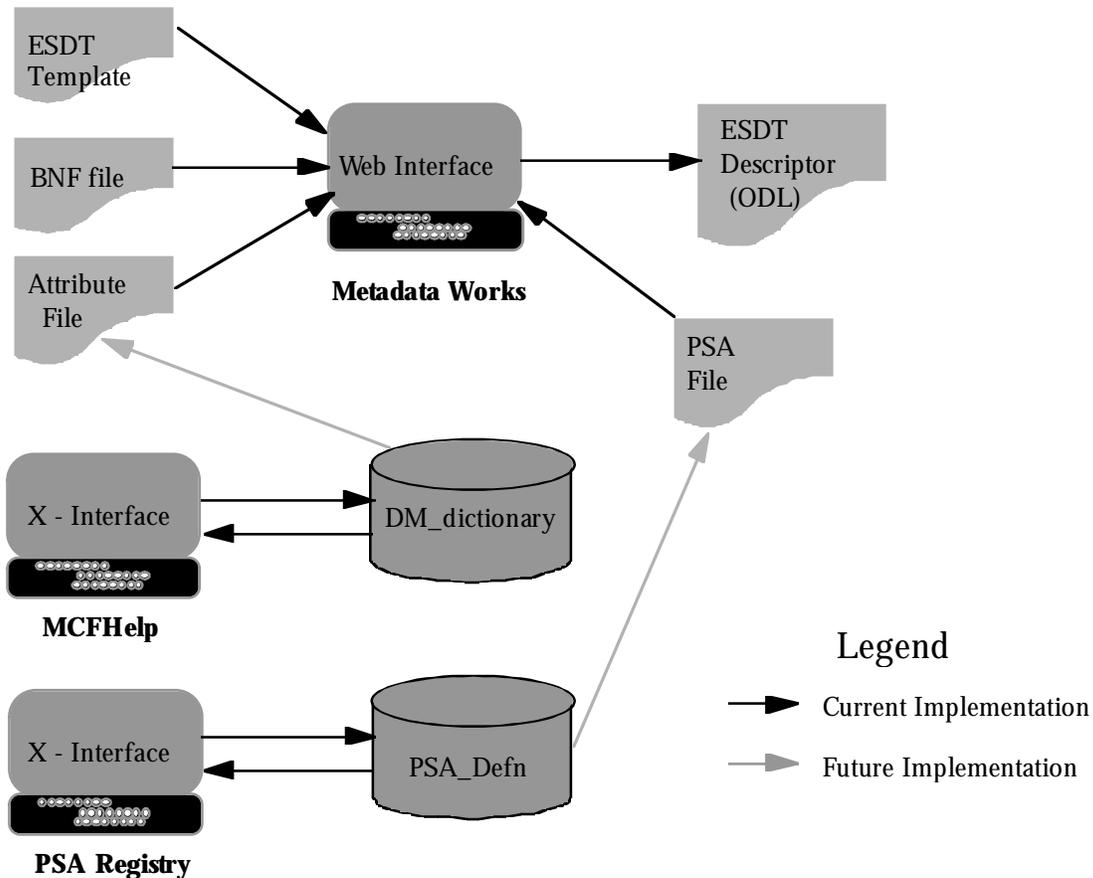


Figure 10.3-1. Tools and Supporting Information Flows

10.3.1 Data Dictionary and Valid Checks

Each proposed new valid must be reviewed by the Data Modeling group in order to validate that they fit within the model. When approved, within a couple days at most, the new valid will be available for use.

10.3.2 PSA Registry

Each Product Specific Attribute must be registered to ensure that the name is uniquely defined across all products and that no two names are applied to the same definition (aliases may be applied at the user's Client and in Data Management, but should not be applied within the inventory). An X-Windows based GUI has been developed to assist in the registration process.

10.3.3. Metadata Works

To support the entry of metadata for ESDTs with Full or Intermediate coverage, an HTML-based GUI has been developed. Metadata Works allows the person entering the metadata to start with a completely empty form, or to define a new descriptor file employing defaults of already defined groups or entire previously defined descriptor files. When the metadata entry is complete for a descriptor file, Metadata Works is then used to generate a complete Descriptor file employing the proper ODL syntax.

10.4 Metadata Population

10.4.1 Collection-Level Metadata

A majority of the attributes in the ECS Data Model apply to all the granules in the collection. These are known as collection-level attributes. There can be both core and product-specific collection-level attributes, defined once prior to establishing the collection.

Collection-level metadata is input via an HTML-forms GUI tool called Metadata Works, described further in Section 10.3.3. Based on the ECS Data Model it is designed for use by the data provider; e.g., an instrument team scientist or other person having extensive knowledge of the data. It can import as defaults the attribute values from a collection that has already been populated. A sequence of screens is presented to the user enabling specification of all required and optional attributes, with a list of permitted values presented where appropriate. Help screens give attribute definitions, data types and other relevant information to assist in the specification.

10.4.2 Granule-Level Metadata

The attributes in the ECS Data Model which can vary on a granule-by-granule basis are known as granule-level attributes. There can be both core and product-specific granule-level attributes.

Granule-level metadata are specified and populated using the Metadata Configuration File (MCF). The MCF is derived from information contained in the ESDT descriptor file and delivered by the Science Data Server for use by the Science Data Processing or Ingest Subsystems. The MCF specifies how the searchable metadata attributes will be populated in the SDSRV database. For data products generated by ECS, the science software or Product Generation Executive (PGE) interacts with the MCF using metadata tools contained in the Science Data Processing Toolkit. Through this process values are set for metadata attributes specified in the "source" MCF, such as the temporal or spatial coverage of each granule. These values are then inserted into a "target" MCF at PGE run time. The MCF is used in a similar manner for data entering ECS through the Ingest subsystem.

Procedures for entering data into ECS through the Ingest subsystem are described in Chapter 16. "Ingest". Procedures for running a PGE are described in Chapters 11.10 "Running a PGE in a Simulated SCF Environment at the DAAC" and 11.13 "PGE Planning Processing and Product Retrieval".

The Inventory Metadata section of Metadata Works is used to capture granule-level metadata specifications. An unofficial MCF may be generated as output from Metadata Works, for testing purposes. Final testing should always be done with the MCF provided by ECS which is guaranteed to be identical to the one delivered by SDSRV at run time.

The actual population of the granule-level attribute values into ECS inventory data bases takes place during the insert of a data granule into the SDSRV. Each data granule consists of one or more physical files. Accompanying each granule is a metadata record; i.e., an ASCII file containing the granule level attributes and their values in ODL. Only one metadata record is allowed per granule, i.e. no. sub-granule records are allowed, and no metadata records are shared between granules.

Procedures used to initiate the running of PGEs are described in Chapter 11.13 "PGE Planning and Processing".

10.4.3 Product-Specific Metadata

Product-specific metadata can be at both the granule-level and the collection-level. Product-specific metadata may (at the data providers election) be contained in the inventory tables in the database, in which case it will be searchable by ECS. There is also a provision to store product-specific metadata within granules that is available only when the granule has been ordered and delivered. This is termed archive metadata and is specified in a separate ODL group in the MCF.

In the granule metadata, the core attribute that is available to store product-specific metadata is called ParameterValue. The metadata describing this attribute is specified by the data provider through the AdditionalAttributes class at the collection-level. The units of measure, range, accuracy and resolution for this is specified in the PhysicalParameterDetails class, also at the collection-level.

Product-specific metadata at the collection level is specified with Metadata Works at the time the other collection level metadata attributes values are defined. At the granule-level, product-specific metadata is defined in the MCF. In both cases, a list of valid values and permitted ranges are specified in the ESDT data dictionary.

10.5 Testing and Validation

Testing and validation involves installation of the ESDT on a Data Server, and subsequent tests of the data services for the ESDT. These tests include insertion of actual or simulated data, search acquire, and other services that may apply and be available and supported under the extant version of ECS. (Section 11.3)

After testing, the ESDT Descriptor File and DLL are promoted to the development CM environment if pre-ECS release, or to the operational environment if after the ECS release is made operational.

This page intentionally left blank.