A Metadata Element Set for Project Documentation

Gail Hodge
Information International Associates, USA
gailhodge@aol.com

Clay Templeton
NASA Goddard Space Flight Center, USA
Thomas.C.Templeton@nasa.gov

Robert B. Allen
College of Information Studies, University of Maryland, USA
rba@umd.edu

Abstract

NASA Goddard Space Flight Center is a large engineering enterprise with many projects. We describe our efforts to develop standard metadata sets across project documentation which we term the "Goddard Core". We also address broader issues for project management metadata.

***Keywords**: Project Documentation, Knowledge Management*

1. Introduction

The NASA/Goddard Space Flight Center (NASA GSFC) carries out extensive programs of research in the earth and space sciences. The bulk of the research performed at the Center relies on data collected by unmanned spacecraft designed and built at the Center.

Therefore, in addition to producing spacecraft, the GSFC produces a tremendous amount of knowledge. This knowledge includes not only the output of the Center's scientific activities but also knowledge about the unique engineering tasks involved in designing, implementing, launching and maintaining spacecraft.

Much of the latter category of knowledge is captured in the ephemeral objects generated in the course of a project. Our goal is to provide an infrastructure that facilitates access to the various information objects produced by projects at Goddard. These information objects may be recorded on diverse media, located in disparate repositories, and created to serve diverse functions. Thus, we wish to unify the discovery function across project libraries, object types, and media.

2. Projects and Project Libraries

Projects are a "significant activity ... characterized as having defined goals, objectives, requirements, Life-Cycle-Costs (LCC), a beginning, and an end" [8] --- for example, the design, construction, and launch of a satellite. There are no inherent limitations, in our view, on the genre or format of a project document; the primary component of our definition is that they reflect some activity carried out as part of a project.

Goddard's projects are organized into programs. We have worked with knowledge management efforts within one such program, the Earth Observing Systems (EOS) program, in formulating the early versions of our Goddard Core descriptive metadata set. The EOS program consists of approximately ten projects that produce a significant volume of project documents. Of these, four participate in the program-wide EOS Program Library.

2.1 Project Libraries

The EOS Program Library collects both paper and digital information objects; the latter are stored in Docushare. More than thirty metadata elements are manually input for all objects, paper and digital, and then ingested into the Program Library. This metadata is input and maintained in a Microsoft-Access-based Web application.

The other six EOS projects use idiosyncratic document management techniques. Three of these six use Docushare as their primary document management system, two use Oracle-based systems, and one uses MesaVista. Even among projects that use similar software, the metadata provided varies widely.

The diverse approaches of the EOS projects to document management in present challenges in conducting document management in a heterogeneous, project-based organization. Organizational entities often employ more or less unique software to manipulate their digital objects and more or less unique semantics to describe them.

Our challenge is to provide a means of discovering diverse objects across multiple distinct and autonomous repositories. To meet this challenge, it is necessary to provide, at the least, searchable descriptive metadata about the objects and a means of determining their locations.

However, projects will want to maintain complete control over their content while they are active. Therefore, it will be necessary for us to provide for discovery of resources that are not under our control. To this end, our data management functions will be accessible independently of our archival storage. Through the use of a persistent identifier (PID) server/resolver, descriptive metadata records in Data Management will be associable with objects existing outside of the Library's own Archival Storage, provided that they are registered with the PID server.

The Data Management function of our archive will thus be powerful enough to provide descriptive metadata for all of Goddard's project documentation in one place, regardless of where the documentation actually resides.

2.2 Media

The scheme described in Section 2.1 above ensures that our Data Management function will be able to facilitate discovery of non-digital resources, provided they are registered with the PID server. In addition to providing pointers to and descriptions of information objects that we do not have control of or that exist in non-digital form, we would like to be able to offer projects and other organizations the option of depositing digital information objects into an Archival Storage system which we are developing concurrently.

2.3 Object Types

The Library is already archiving videos, images, and web pages that have relevance to NASA's mission. Each of these digital object types requires a specific digital context to be rendered meaningfully to human users. However, it is efficient to be able to treat them as homogenous in dealing with them as archived objects. Furthermore, we wish to be able to accommodate additional object types that are at present unspecified. To accommodate the homogenous treatment of digital information objects of all types, we use a metadata wrapper such as the Metadata Encoding and Transmission Standard (METS).

3. "Goddard Core" Elements for Project Documentation

3.1 Methodology

To facilitate the use of our metadata repository by other entities within the organization, we are specifying a single set of descriptive metadata, the Goddard Core, to be populated consistently for all objects referenced by metadata sets in Data Management. Metadata records conforming to this specification will be populated either through the use of mappings from existing metadata sets or, in the case of objects supported by insufficient metadata, through generation. The use of a single metadata set for discovery will facilitate the development of effective tools that will interact with the metadata server through an API.

In formulating the Goddard Core, we drew from sources within the EOS program as well as from metadata sets already in use by the GSFC library in its provision of project-related videos, images, and Web sites. As the Core developed, we established mappings from metadata sets used by the EOS project libraries to help ensure that the Goddard

Core was developing in a direction that would accommodate as much useful descriptive information as possible without becoming unwieldy. These mappings were prototypical of a process that will be an integral part of the Goddard Digital Archive.

Other metadata sets in use at Goddard are being collected for evaluation as additional examples of heterogeneous project libraries. To achieve this, a Metadata Review Committee has been established with representatives from major programs with repositories of datasets, project documentation, and images. This Metadata Review Committee also includes metadata experts from external organizations that are involved in building institutional repositories or with which Goddard may want to interact in the future.

To support the multiplicity of metadata sets and the mapping required to the Goddard Core, a Metadata Registry is being developed. Initially, the disparate element sets and their mappings were collected in a simple spreadsheet. However, as this grows a more standardized mechanism is needed. A database following the ISO 11179 Metadata Registry standard is being developed, using lessons learned from EPA and the Jet Propulsion Laboratory.

The ISO 11179 Metadata Registry standard identifies the information that should be stored to describe an element. It allows for mapping between and among element sets or to a standard set such as the Goddard Core. It also allows documentation of domain values, validation rules or rules for processing at input or output.

Owners of metadata sets from throughout the Center and the archive staff will be able to populate this database and make connections to both the Goddard Core and from one element set to another. This database will allow computer processing to facilitate interoperability of element sets and automatic production of metadata output from one element set to another.

3.2 Requirements

In addition to the architectural requirements addressed above, several overarching requirements have emerged for a scheme of descriptive metadata elements in the management of project related digital objects.

Projects often produce multiple versions of the same piece of documentation. It is essential to have the capabilities of distinguishing between different versions and of gathering together all existing versions of a single document.. Therefore, the metadata set must include robust mechanisms both for distinguishing between and linking versions of a work. In addition, in active projects, the availability and format of documentation frequently changes. It is desirable, then, not only to provide information about the various formats available and their location, but also to include a dynamic mechanism of determining availability as close to the time of search as possible.

Moreover project documentation is often considered proprietary or "in progress". This means that a system must provide flexibility of terms and conditions of use over the life of the project.

A third requirement is that the metadata elements defined have significance for description over time and in a variety of situations. As part of Goddard's knowledge management environment, project-related digital objects may have a variety of uses and users. Therefore, the elements need to be descriptive over time, but with enough consistency to make retrieval efficient. For this reason, we are applying both a controlled subject category element and free-text keywords. The former will allow groups such as EOS to select terms from a specific taxonomy that will support the needs of that program for development of an EOS portal. The latter will allow more specificity and provide for new areas of interest to be highlighted.

3.3 Framework for definition of elements

To address these special requirements, we have applied the distinctions between work, expression, manifestation, and item made in the *Functional Requirements for Bibliographic Records* [5]. We have interpreted the FRBR terms to meet our particular needs:

- *Item* - a concrete entity that exists in one place at any time
- *Manifestation* - all and only those items that express a particular content and are produced to the same technical specification format

- *Expression* - all and only those manifestations which contain the same content and differ in form only version
- *Work* - all and only those expressions whose intellectual content springs from the same creative act

The Goddard Core operates primarily at the level of the expression. The persistent identifier located in the Identifier element can be considered to refer to one expression of a work. The expression referred to is linked to its manifestations through the resolution of the persistent identifier, as well as through the Format element. It is also linked to its items through the former. An expression is linked to the work of which it is an expression through the Title, Creator, Subject, and most directly, Relation elements.

At present the Goddard Core consists of the standard Simple Dublin Core elements, augmented by eight extension elements. Of the sixteen Simple Dublin Core elements, special attention has been given to eleven. These elements are being treated essentially as defined in the DCMI Elements and Element Refinements List. A discussion of the elements whose specific use in the Goddard Core bears comment follows. This discussion relies on the terms of the FRBR as defined above.

3.4 Interpretation of Base Elements

The **Title** element contains the name of the work of which the Identifier identifies one expression.

The **Creator** element contains the name of a creator or developer of the work of which the identifier identifies one expression.

The **Contributor** element contains the name of a person who has participated in an effort that has resulted in the expression identified by the Identifier element, but did not necessarily participate in the creation of the work of which it is an expression.

The **Subject** element contains a characterization of the subject matter of the work of which the Identifier element specifies one expression. The terms allowed in this element are uncontrolled.

The **Description** element contains information useful in evaluating the relevance of the expression to the user's need. This information includes "abstracts, table(s) of contents, reference to a graphical representation of content or a free-text account of the content".

The simple **Date** element is used in the Goddard Core to identify the date at which the particular expression referred to by the Identifier was first made widely available in some manifestation.

The **Type** element contains an indication of the intellectual genre of the work of which the object specified in the Identifier element is an expression, taken from a taxonomy of content types.

The **Format** element contains the name(s) of the medium(s) on which items of the expression specified by the Identifier element exist.

The **Identifier** element contains a persistent identifier that resolves to physical and/or virtual locations of all known items of a unique expression.

Each **Relation** element contains persistent identifiers each of which resolves to physical and/or virtual locations of all known items of an expression that is derived from the same work as the expression specified in the Identifier element.

The **Rights** element contains a formal statement of copyright and access controls that apply to the expression identified in the Identifier element. These do not include idiosyncratic restrictions imposed by individual repositories in which manifestations of the expression may exist.

3.5 Extensions

In addition to the Simple Dublin Core elements, the Goddard Core employs several extensions. Brief descriptions of these elements follow:

The **Controlled Subject** element contains terms that characterize the subject of the expression specified in the Identifier element and that are drawn from a controlled vocabulary.

The **Code** element contains a Goddard organizational code under which the expression originated in whole or in part.

The **Contract** element contains a value that identifies the contract under which the expression originated in whole or in part.

The **Project** element contains a value that identifies the project under which the expression originated in whole or in part.

The **Organization** element contains a value that identifies the organization under whose auspices the expression was produced in whole or in part.

The **Project Phase** element specifies the phase that the project specified in the Project element was in when the expression was completed.

The **Instrument** element contains a value that identifies a piece of equipment to which the expression pertains.

1. Example

Using the definitions above, a sample metadata record for a Goddard project might be:

Title: Extreme Ultraviolet Flight Explorer to Explorer Platform Interface Control Document

Creator: Frank J. Cepollina
Creator: Thomas Sorensen
Creator: Roger Malina

Contributor: George Hogan

Subject: Spacecraft
Subject: Science payload module
Subject: Explorer Platform
Subject: Mission Equipment Deck
Subject: Interface Control
Subject: Astrophysics
Subject: Extreme Ultraviolet
Subject: Astronomy
Subject: Electromagnetic Spectrum
Subject: Radiation

Description: defines the interfaces between the Mission Equipment Deck (MED) of the explorer platform and the Extreme Ultraviolet Explorer (EUVE) payload.

Date: 01-01-1986

Type: Project Document.Specification

Format: PDF

Identifier: doi:10.1000/7545

Relation: doi:10.1000/4589
Relation: doi: 10.1000/2561

Rights: unrestricted/uncopyrighted

Code: 672
Project: XTE

Organization: NASA/GSFC
Organization: Center for Extreme Ultraviolet Astrophysics

Project Phase: Formulation

Instrument: Extreme Ultraviolet Explorer
Instrument: Explorer Platform

2. Domain Values

Several elements, including the Controlled Subject, Project Phase, and Code can be identified by specific domain value sets. In the pilot project with the EOS Libraries, the Controlled Subject element is limited to a taxonomy that is being developed by the EOS Pilot Project. There may be a need to coordinate with the NASA-Wide Taxonomy [4] which is under development.

Another area where domain values are important is that of **Type**. While some of the types of interest in project documentation are the same as the general types defined by the DC Genre Working Group, there is the need for other genres that are specific to the NASA Project Documentation.

To ensure interchange with others using a less specific set while retaining the specificity required for Goddard's needs, the specific Goddard genre types are being merged with the more general Dublin Core genre types. The more specific genre type will be assigned, but if required, up-posting can be used to inherit the broader term from the genre hierarchy.

3. Future Work

This work is ongoing and there are several additional areas of investigation, including the mark- up of the internal structure of project documents, project management metadata beyond project documentation, end-to-end management of metadata including the addition of metadata for preservation, and the development of context.

6.1 Describing the Internal Structure of Project Documents

In cooperation with the Goddard XML Working Group, we may extend the metadata mark-up into the project documents themselves. The Working Group wants to drive the markup down to mark up within the actual documents. So, there would need to be some coordination between the metadata and XML tags in project documentation mark up because title, project name, etc. would appear in both.

The aspect that the Working Group is trying to identify is similar to a project, to characterize the data in terms of "lessons learned". They are attempting to mark up and extract technological and management lessons.

6.2 Preservation Metadata

Ultimately, the goal of the Goddard Archive is to provide long-term preservation and access to project information. To this end, the Metadata Encoding and Transfer Standard (METS) [2], developed by the Library of Congress for interchange between digital libraries is being investigated, as is an alternative framework from the CCSDS.

Additional elements are needed to track the provenance, validate the success or document the loss caused by a particular migration event, and to support the rendering and reuse of the project object in the future. In addition to the Goddard Core for discovery and evaluation, elements will be identified to support the preservation of the formats and media of significance. This will involve issues such as significant characteristics and technology assessment. Members of the Goddard technical group are involved with various national and international activities in this area, including the NISO Preservation Metadata Working Group, which is working to extend the work of the OCLC/Research Libraries Group effort in this area.

6.3 Project Management Metadata Beyond Project Documentation

We can envision an integrated project management information environment which goes well beyond the metadata requirements for individual documents. Thus our archive could include a wide variety of project materials beyond summary documents – there could descriptions of project leaders, schedules, and spreadsheets [7]. We believe that a consistent metadata should be developed for all aspects of a project. With such a over-arching metadata set, the role of project documentation could be better understood in context. Indeed, this is consistent with recent discussions in the archival community (e.g., [3]).

As one example, Allen and Templeton [1] have developed an XML schema for capturing and transferring the semantics of Role Activity Diagrams (RADs) [6]. Additionally, they have developed functionality to render instances of this schema as a RAD (see Figure 1). It is hoped that preserving a representation of the workflows within which a document was generated and used will serve not only to aid in evaluating an object's integrity, but also as a mechanism of discovery. A RAD editor is currently under development.

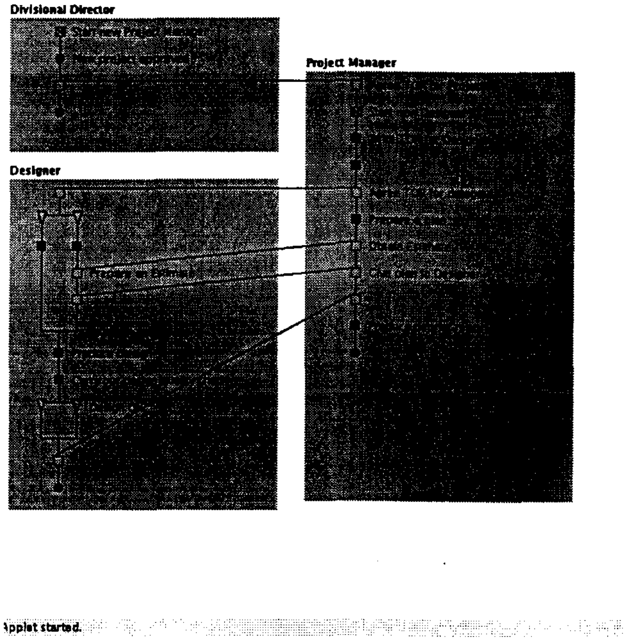

Figure 1. Role Activity Diagram generated from XML markup that describes organizational structures and processes (from [AT]).

Acknowledgments

We thank Janet Ormes and Members of the Goddard Metadata Committee.

Clay Templeton is also affiliated with the College of Information Studies of the University of Maryland.

Gail Hodge serves the NASA/Goddard Space Flight Center under contract NAS5-01161.

References

[1] Allen, R.B. & Templeton, C., Digital Preservation and Organizational Context. in preparation.

[2] Beaubien, R., METS: An Overview and Tutorial, http://www.loc.gov/standards/mets/METSOverview.html.

[3] Cook, T., *Fashionable Nonsense or Professional Rebirth: Postmodernism and the Practice of Archives.* Archivaria *51.* Spring 2001. 14 - 35.

[4] Dutra, J. & Busch, J. Taxonomy Development with NASA, *DC 2003*, submitted.

[5] IFLA, *Functional Requirements for Bibliographic Records.* IFLANET, 1997. http://www.ifla.org/VII/s13/frbr/frbr.htm

[6] Ould, M.A. *Business Processes - Modeling and analysis for re-engineering and improvement,* Wiley & Sons, Chichester, 1995.

[7] Project Management Institute. *Practice Standard for Work Breakdown Structures,* 2001.

[8] NASA Program and Project Management Processes and Requirements (7120.5B)