Content Packaging Approach for a Large OAIS Repository

Gil Baldwin, Matthew Landgraf, Kate Zwaard; U.S. Government Printing Office; Washington, D.C.; John Faure, Harris Corporation; Melbourne, FL

Abstract

The Future Digital System (FDsys) is a digital archive being developed by the U.S. Government Printing Office and Harris Corporation. It will provide permanent public electronic access to authentic government publications via a web interface. The FDsys is being designed based on the Open Archival Information Systems (OAIS) reference model. The OAIS model incorporates many archiving best practices and increases the chance for developing a successful archive that will protect and preserve content over the long term. Applying the OAIS reference model requires making many engineering decisions about how to apply the model to a particular archive's mission and operations. This paper describes the specifics of how the information packaging concepts of the OAIS model are being applied to FDsys to address the archiving problems of storing complex (compound) digital objects, managing multiple renditions of the same publication, preserving publications over time, managing access rights to publications, and maintaining a record of the history of a publication within the archive. How information is packaged within a digital archive is critical to the success of the archive. This paper provides a look at the design of a large digital archive and a concrete example of how some important aspects of OAIS are being applied.

Introduction

The U.S. Government Printing Office (GPO) Future Digital System (FDsys) will ingest, authenticate, provide version control, preserve and provide access to digital content from all three branches of the U.S. Government. FDsys, which as of May 2007 is in beta testing, is a comprehensive, systematic, and dynamic means for preserving digital content free from dependence on specific hardware or software. The system will automate many lifecycle processes for digital content and make it easier to deliver content in formats suited to customers' evolving needs.

Digital content managed in FDsys will be composed of deposited, converted, and harvested content. Deposited content is intentionally submitted to GPO by U.S. Federal agency content creators. Deposited content will include the digital object received from the content creator as well as corresponding processing requirements and additional metadata. The system will ingest all common text, graphical, audio, and video formats used in government publishing, and will accept deposited content that is furnished in a wide variety of formats and media. Harvested content, with accompanying metadata, within the scope of GPO's dissemination programs is gathered from Federal agency Web sites. Converted content is digital content created from a tangible product, typically by scanning legacy print publications.

The long term focus of the Future Digital System will be on deposited content. As the amount of deposited content submitted to GPO increases, there will be a gradual decrease in the need for legacy conversion and harvesting. Regardless of its source, content that is ingested into the FDsys will be managed as an information package. The content's source will be recorded in metadata, but the package structure for deposited, harvested, and converted content is identical.

In the early years of production, FDsys is expected to manage about six million Federal government publications, including the digitized national collection, totaling more than 3,000 terabytes. In later years, GPO expects to ingest about a million publications a year.

Digital Content Management Goals

In order to meet GPO's strategic goals, FDsys should be able to accomplish the following:

- Support GPO's business processes and enable improvements to their efficiency, quality, effectiveness, and timeliness;
- Ingest content in all popular digital content creation formats;
- Store and manage content in a manner that is independent of any particular hardware and software component over long periods of time;
- Provide content storage scalability sufficient to hold a complete collection of in scope publications, including converted hard copy, web harvested, and deposited content;
- Accommodate future digital formats;
- Preserve digital content for future use;
- Ensure the authenticity of the content that GPO preserves;
- Provide access to content descriptions and metadata;
- Provide access to the content in a manner that is consistent with current technology and the changing expectations of GPO's diverse user communities.

OAIS Reference Model

The system design is based on the Reference Model for an Open Archival Information System (OAIS) developed by Consultative Committee on Space Data Systems (CCSDS) with broad input from other communities. It was issued by International Standards Organization (ISO) in 2003 as standard ISO 14721:2003: Space data and information transfer systems -- Open archival information system -- Reference model. OAIS is a domain-neutral reference model with characteristics broadly applicable to the management of any information over time. The OAIS model has been adapted and used in other research collaborations and institutional settings and provides a framework for achieving the scalability, extensibility, and interoperability required for a system of FDsys' magnitude. This model does not prescribe an implementation. Using the OAIS as a reference model is the start of the process of defining what is necessary to achieve GPO's strategic objectives for FDsys. It is recommended that the system be an integrated system that provides OAIS foundation services such as content ingest, storage of digital content in the

form of information packages, content preservation, and the ability to provide access to the content from anywhere on demand.

Implementing OAIS

While OAIS is a valuable high level reference model, it does not serve as a design specification. Consequently, the FDsys team has gone through an intensive design process following the IEEE system design model. Implicit in this process is the need to make basic decisions about how content and metadata are arranged. Early on, the team identified a design based on content packages, a concept derived from the Warwick Framework, first expressed as an outcome of the 2nd Dublin Core Metadata Workshop, held at Warwick University in the U.K. in April 1996. The Warwick Framework (WF) is a container-package approach in which discrete packages of metadata can be aggregated in conceptual containers.

WF architecture is based upon two main components: the container and the package. The container is considered a first-class object and can be managed as any other digital resource, stored in servers, and accessed using a URI. The package contains actual metadata sets, including MODS elements, MARC 21, Dublin Core, ONIX, PREMIS, and others.

FDsys Package Lifecycle

In the FDsys design, an information package contains metadata packages, various renditions of the content, and the components, related by a binding XML file. These information packages serve different purposes and exist for varying durations. A key simplifying assumption in FDsys is that an information package contains one edition of a single publication. The team has found this to be a very good design decision for an archive that is based on storing publications. This removes a level of complexity from interpreting an information package that makes processing, indexing, and retrieving the package simpler. This information package convention would probably not work well for an archive designed for storing records in which there may be hundreds of instances of the same record type produced by the same organization on a monthly or yearly basis. In FDsys, the information packages for various published editions or versions of a publication are linked together so that a user may find the specific version of the publication they wish to use.

In the high level view of FDsys processing, the Submission Information Package (SIP) is a transitory object which, once ingested, becomes the Archival Information Package (AIP). From the SIP, an Access Content Package (ACP) is also derived, which includes such renditions of the content which will facilitate efficient and effective public access. The ACP is a GPO innovation that is designed to separate the read only, delivery oriented renditions (for example, screen or print optimized PDF and HTML formats) of a publication from their long term archival renditions (which includes the original and preservation formats). The ACP usually requires ten times less storage than the AIP, allowing it to be kept in fast access storage. This supports a paradigm of separating content creation from content publishing, which improves archive security and scalability. The AIP is stored in a trusted repository environment and insulated from daily use. User requests for access are directed to the ACP, from which a Dissemination Information Package (DIP) is created to satisfy a user request for a specific rendition or presentation of the content. The DIP is also transitory in nature, and is not retained in FDsys.

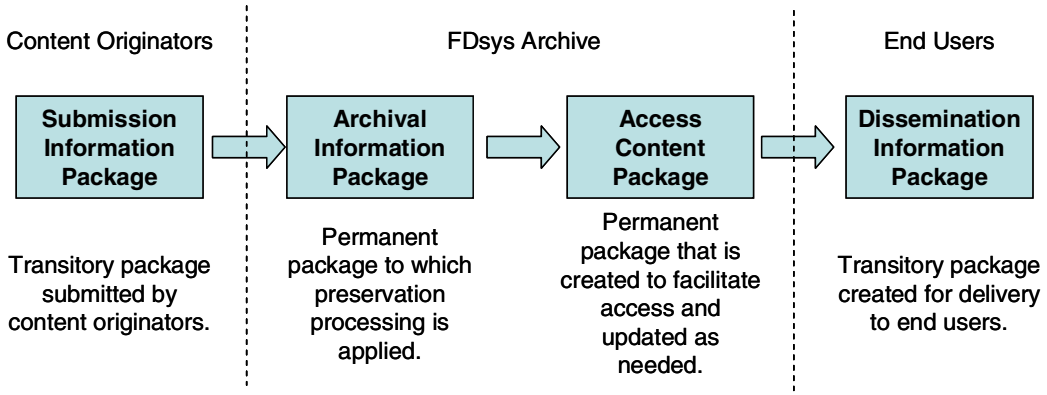

Figure 1. *Information Package Lifecycle in FDsys*

FDsys Package Structure

Each of the types of content package (i.e., SIP, AIP, ACP, DIP) share a common structure. For example, an AIP is composed of content digital object(s) and metadata about the digital object(s), and a binding metadata file (aip.xml) that relates the digital objects and metadata together to form a system-compliant AIP. GPO has adopted version 3.2 of the Metadata Encoding and Transmission Standard (METS) as the encoding standard for the aip.xml file, and GPO will specify profiles and extension schema for METS as required. The specifications for the GPO information packages will be published at a later date.

Related digital objects within a package comprise a rendition, which may be thought of as a digital document expressed in a particular format. For example, a single package may include the document and image files for a page layout program as a rendition (in which the document was created), a press-optimized PostScript rendition, and a screen-optimized PDF rendition. The files and folders making up each rendition are stored in their own folder within the package structure. The folder structure of the rendition is preserved within the information package so that the document may be successfully opened using the original software in the future. Metadata specific to each rendition is stored in its own parallel folder within the information package. Each metadata file

is tied to the whole document, a single rendition, or a single digital object in a rendition. Descriptive metadata may be tied to the publication as a whole. Format information may be tied to a rendition. Conversion information may be tied to a single conversion file. The FDsys information package structure is illustrated in Figure 2.

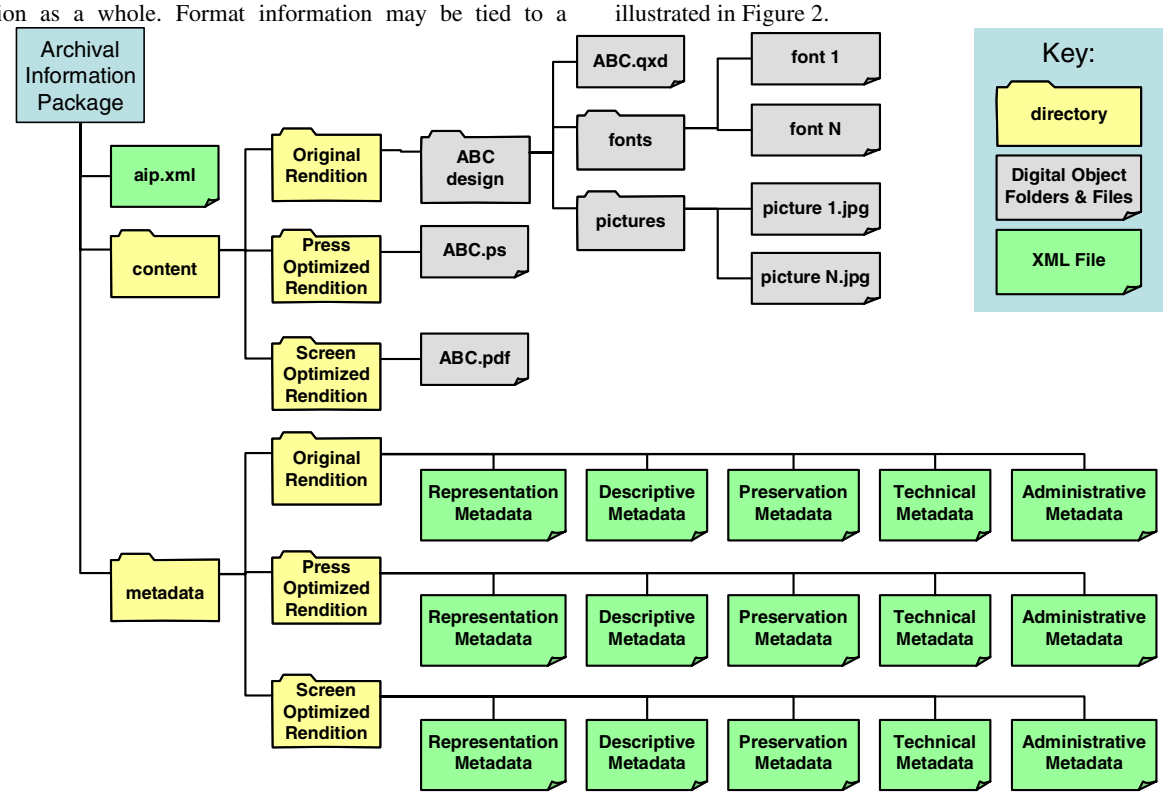

Figure 2. *Representative AIP Package Structure*

Associated with each rendition are several metadata files expressed in XML.

- Representation Information - original metadata and other information, objects, or application names and versions that are required to render the publication at a specific level of accuracy. The representation information should also include enough metadata for certification, version control, access, and preservation to take place.
- Descriptive - metadata that allows users to discover the content, such as information about the author, publication date, version, etc.
- Preservation - metadata needed to accurately describe the content, verify its fixity, and provide an understanding of the environment (context) in which the content was created.
- Technical – metadata required to accurately render the content, include file attributes, format, etc. For example, an AIP may contain an electronic journal, information about how the journal is structured (by pages or sections) and formatted.
- Administrative – metadata about the origin of the content (provenance), dates when system processes occurred, and access rights for viewing the content

The aip.xml file expresses in METS the structure of the renditions and the relationships between the content and the metadata.

How Packages are Created and Preserved

FDsys verifies the completeness and correctness of a SIP prior to ingesting it into the archive. This is critical to ensure that invalid, unofficial, and duplicate content do not enter the archive. The pre-ingest processing functions are

- Detect and eliminate malicious content such as viruses,
- Verify the authenticity of the source of the document,
- Verify content completeness and correctness checking (documents can be opened, no missing graphics or fonts),
- Verifying metadata files comply with required XML schemas,
- Determining if this is an (identical) duplicate of an existing publication,
- Determining if this is a new version of an existing publication, and
- Verifying the content originator's approval to publish.

Ingest Processing

FDsys ingest processing takes a verified SIP and moves it into the archive. An AIP is created from the SIP for permanent preservation within the archive. The AIP includes FDsys assigned

unique identifiers for each piece of the SIP. Preservation processing is carried out on the AIP. An ACP is created from the AIP that includes dissemination oriented renditions of the publication. It is expected that the formats stored in the ACP will change as new formats become popular for viewing documents. GPO staff are involved in the ingest workflow as needed to enhance or create metadata or resolve issues with the package. The ingest processing functions are

- Accept verified SIPs,
- Create AIPs from SIPs,
- Apply a digital time stamp to each content file for fixity verification, and
- Create initial ACPs from SIPs.

Preservation Processing

AIPs are stored in a secure environment and acted upon by FDsys preservation processes. FDsys preservation processes ensure that content does not degrade, is not deliberately altered, and remains useable by commonly available software. Periodic sampling of the archive will be performed to detect storage related degradation. Conversion to more current or permanent formats (and other preservation techniques) is performed when necessary. New dissemination formats are created in the ACP when necessary. Only content in scope for GPO's dissemination programs is accepted into FDsys archival storage and managed by preservation processes.

Summary

The OAIS reference model describes, at a high level, an archival system dedicated to preserving digital information and making it available over the long-term. However, the OAIS model is descriptive rather than prescriptive. Institutions planning and building OAIS-compliant archives face a series of design decisions, not the least of which is the structure and relationship of content and metadata. GPO's package design is faithful to the OAIS reference model, consistent with emerging best practices in the digital preservation field, and will enable the GPO to ensure permanent public access.

References

[1] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," (2002).

[2] International Organization for Standardization, "ISO 17421:2003 Space Data and Information Transfer Systems -- Open Archival Information System -- Reference Model," (2003).

[3] Carl Lagoze, "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata." Cornell University - D-Lib Magazine. (July/August 1996).

[5] Brian Lavoie, "The Open Archival Information System Reference Model: Introductory Guide." (2004).

[6] United States Government Printing Office, "Concept of Operations for the Future Digital System V2.0," (2005).

Author Biography

Gil Baldwin is an associate director in the Program Management Office planning and implementing GPO's Future Digital System, FDsys, with special emphasis on digital preservation, library services, and version control. A native Virginian, Baldwin received the B.A. in American History from the College of William and Mary, a Masters of Library and Information Science from Florida State University, and pursued additional postgraduate work at the Catholic University of America.

Matt Landgraf has worked for the U.S. Government Printing Office since July 2000. His recent work with GPO's Program Management Office has focused mainly on the development and implementation of GPO's Future Digital System (FDsys), specializing in the areas of content submission, Web harvesting, and content packaging. He received a B.S. in Business Administration from Towson University in 2000.

Kate Zwaard is a program manager on the FDsys team concentrating in digital preservation, metadata management, data mining, and system sizing. She was graduated from the University of Maryland in 2002 with a double major in political science and journalism with specialties in public opinion and statistics.

John Faure received his BS in Computer Science from the Ohio State University (1985). He has worked as a software engineer and later as a software architect on many large information systems, including space launch, electric power generation control, satellite weather systems, and weather data archiving. For the last four years he has worked exclusively on the issues and technologies needed for reliable and cost effective long term archiving of digital content at Harris Corporation.