Automated Metadata Creation to Enhance Search Capabilities in GPO's Federal Digital System

Lisa LaPlant; U.S. Government Printing Office; Washington, D.C.; Blake Edwards; U.S. Government Printing Office; Washington D.C.; Paul Nelson; Search Technologies Corporation; Washington, D.C.

Abstract

GPO's Federal Digital System (FDsys) automates the collection and dissemination of electronic information from all three branches of the United States federal government. Information is submitted directly into FDsys, permanently available in electronic format, authenticated and versioned, and publicly accessible for searching and downloading. FDsys relies on three key system components to accomplish this: 1) A content management system that securely manages digital content to safeguard content integrity and authenticity, 2) A preservation repository that follows archival system standards to ensure long-term preservation and access to digital content, and 3) An advanced search engine combines extensive metadata with modern search technology, allowing the public to find federal government publications.

FDsys is a metadata driven system. It has been developed to utilize multiple metadata creation techniques, including parsing from unstructured or semi-structured text, transformations from structured formats, transformations to apply constants, and manual metadata entry. Using these techniques, GPO is able to create and store metadata to power the search engine on the FDsys.gov web site. This metadata is used to deliver advanced search, field operator search, search result filters (aka faceted navigation), and hierarchical browsing of documents. It is also used to provide downloads of descriptive metadata files in MODS XML. This emphasis on source metadata makes FDsys a dynamic, rich, flexible, powerful, and transparent system for document search and access. This paper will discuss different methods used to create metadata, how FDsys descriptive metadata is stored, and how this metadata is utilized by the FDsys search engine.

Introduction

Dating back to the late 1800s, the United States Government Printing Office (GPO) has had three essential missions: 1) GPO provides the agencies and organizations which make up the three branches of the federal government with expert publishing and printing services, on a cost recovery basis, in order to avoid duplication and waste of government resources. 2) GPO provides, in partnership with federal depository libraries, for nationwide community facilities for the perpetual, free and ready public access to the printed and electronic documents, and other information products, of the federal government. 3) GPO distributes, on a cost recovery basis, copies of printed and electronic documents and other government information products to the general public [1].

In order to continue to meet its missions, GPO's Federal Digital System (FDsys) was developed to ingest, preserve, and provide access to official publications from the legislative, executive, and judicial branches of the United States federal government [2]. The system, which launched in January 2009, relies on three key components: 1) *a content management system* that securely manages digital content to safeguard content integrity and authenticity, 2) *a preservation repository* that follows archival system standards to ensure long-term preservation and access to digital content, and 3) *an advanced search engine* that combines extensive metadata with modern search technology allowing the public to find federal government publications [3].

The system design is based on the Reference Model for an Open Archival Information System (OAIS) developed by the Consultative Committee on Space Data Systems (CCSDS) with broad input from other communities [4]. FDsys is designed around the concept of content packages, derived from the Warwick Framework, an approach in which discrete packages of metadata can be aggregated in conceptual containers [5]. FDsys packages contain metadata files, content files, and a binding Extensible Markup Language (XML) file. A Submission Information Package (SIP) is a transitory package that becomes an Archival Information Package (AIP) upon ingest to the preservation repository. An Access Content Package (ACP), a type of package invented by FDsys to improve archive access, is derived from the SIP, and is designed to separate access renditions in the content management system from long term preservation renditions in the preservation repository. A Dissemination Information Package (DIP) contains content and metadata delivered to users from the search engine through a custom Web application [6].

Content in FDsys is organized into collections. A collection consists of related content that is processed or accessed in a similar manner. Examples include the Congressional Bills collection and the Federal Register collection. Collections contain individual packages. A package is roughly equivalent to one printed publication. Examples include one Congressional bill or one daily issue of the *Federal Register*. Packages contain one or more renditions. A rendition is generally associated with a file format. Examples include a pdf-submitted rendition from a content originator or an html rendition for access that was created from a text-submitted rendition. As appropriate, select content in for-access renditions is further subdivided into granules. A granule is the smallest usefully searchable unit of content. Examples include a proposed rule in the *Federal Register* or a unit of business on the floor of Congress in the *Congressional Record*. Each granule is uniquely processed and described. For example, Portable Document Format (PDF) granules in an access rendition are digitally signed, granules are listed separately in search results and may be represented by multiple renditions, and metadata about each granule is persisted in the Metadata Object and Description Schema (MODS), a MARC-compatible XML schema for encoding descriptive data [7].

 ©2011 Society for Imaging Science and Technology

Data Management Definition

The System Design Document (SDD) for FDsys consists of multiple volumes of individual design documents (63 at last count). The goal of the SDD is to cover the system architecture and design. The design documents contain the high-level architecture as well as separate detailed design documents for each of the major system components.

The SDD also includes separate Data Management Definition documents (DMDs) for each content collection within FDsys [8]. DMDs present a horizontal view of how information flows through FDsys. This is because each collection in FDsys has unique requirements related to the structure and format of files, structure and format of the collection's descriptive metadata, processing that needs to be performed on these files, and how content should be searchable and displayed to the public.

As collections are brought into FDsys, a DMD is used to analyze, describe, and document how content and descriptive metadata should flow through FDsys. The document provides valuable information for developers and stakeholders including the following:

- Background and Description – Provides an overview of the collection.
- Metadata Description – Identifies each metadata element and how that element is represented internally, including the "arity" (multi-valued constraint) and data format of the element.
- Content Processing – This includes information about how files in submitted renditions are packaged and processing steps (format transformations, digital signing, etc.) performed on package content.
- Parsing and Extraction – Instructions for how descriptive metadata should be parsed or extracted from submitted files.
- Index Transform – Describes how metadata is transformed and stored into search engine index fields. This includes prioritizing metadata for cross-collection relevancy ranking.
- Search – Lists all search fields along with the query expressions required to implement each field.
- Results Presentation – Describes how search results are formatted for presentation.
- Browsing and Navigation – Identifies the search queries and index fields required to implement search engine navigators (facets), content landing pages (more information pages), and collection browsing interfaces.
- MODS Mapping – Describes how descriptive metadata is persisted in MODS.

Metadata Creation

Metadata is generated in various ways in FDsys. When a package is submitted to FDsys, metadata could be parsed from unstructured (or semi-structured) text, transformed from structured text (e.g., XML), set as a constant for a collection, generated from content processing steps, or manually entered by a user. When metadata is generated, it is managed in FDsys in an XML file which is stored in the package. A combination of the methods listed above is used to generate metadata for each collection, as specified by the DMD for the collection.

Parsing from Unstructured Text

For the majority of the over forty collections in FDsys, the submitted renditions include PDF and unstructured American Standard Code for Information Interchange (ASCII) text files. Based on thorough analysis of the content in each collection, parsers were developed to use regular expressions in a framework that emphasizes flexibility and fallback rules. Typically, a parser will try one pattern, evaluate its effectiveness, then try a second pattern, et cetera until a pattern is found which is successful. For example, this is the regular expression that is used to identify references to public laws within unstructured content files: "(Public Law|Pub. L.|PL|P. L.) (1[0-9][0-9])-([0-9]+)". When metadata is extracted from content, it is normalized and managed in an FDsys XML metadata file that is associated with each package.

Note that many of the collections in FDsys are "semi structured", in other words, they have common title pages, formatting, or presentation indicators which can be used to extract metadata reliably from document to document, making the job of the parser writer easier – or at least possible. In addition, some elements are extracted from fully unstructured text, such as references to other government documents (e.g., "Public Law 110-32", or "10 CFR 32.3").

Transforming Structured Text

For a limited number of collections, the submitted renditions include Standard Generalized Markup Language (SGML) or XML files. For these collections, an Extensible Stylesheet Language Transformations (XSLT) is applied to the content to transfer data stored in the structured text file to an FDsys XML metadata file. For example, the value for congress number is extracted from the <congress> field within a congressional bill in XML format.

In cases where the value of a field in a structured text file matches the format required for FDsys, the value of the field can be copied directly from the structured text file to an FDsys XML metadata file. In other cases, the data stored in one field may need to be normalized or further parsed in order to acquire the metadata in the format required by FDsys.

Applying Constants

Many FDsys descriptive metadata fields have the same value across all publications within a collection. These values are handled as *collection-based constants* for the collection. Constant values are identified in the DMD. For example, the constant "Regulatory Information" is applied as a top-level category to all documents in the Federal Register and Code of Federal Regulations collections.

Applying Authority Files

As metadata is parsed from unstructured text or transformed from structured text, the values received may not be formatted consistently for every publication or collection. For certain metadata fields, authority files are used to transform the acquired values into standardized formats. For these fields, the original values from parsing are stored in the FDsys XML metadata file. However, additional values based on an authority file are also looked up and stored. For example, the President's name may be parsed simply as Clinton, but based on the authority file it is stored

in the FDsys XML metadata file as both the official name (William J. Clinton) as well as the common name (Bill Clinton).

Processing Content

Content processing performs a variety of actions on content files in FDsys. Some of the actions include creating renditions from other renditions, renaming files, splitting large files into multiple smaller access files (called *granules* in FDsys parlance), generating combined files at various levels of granularity (e.g., chapter files), and applying digital signatures to PDF files. When processing is performed on content files, digital provenance metadata is generated to describe the new renditions and files that are created. This metadata is persisted in a PREMIS XML file that is stored in the package.

Manual Metadata Entry

When the structure of content within a collection is consistent from publication to publication, the automated methods described above can be used to generate metadata. When the structure of content varies from publication to publication within a collection, metadata is entered manually for that publication. Additionally, if automated methods generate metadata but some values are incorrect, metadata can be edited manually through a custom XForms-based editing tool. This is a useful technique for handling rare exception cases without incurring additional parser development expense.

Manual metadata changes are stored in the FDsys XML metadata file and persisted in MODS in the AIP. Changes made through the editing tool also trigger the package to be re-published to the search engine with the new metadata values that have been modified.

Metadata Management

Acquired metadata is managed in an FDsys XML metadata file in the ACP, and metadata is persisted in Metadata Encoding and Transmission Standard (METS), MODS, and PREMIS XML files in the AIP within the preservation repository. Each ACP contains one FDsys XML metadata file that describes the package, the renditions within the package, the files within each rendition, and any granules generated for the package. Information from each FDsys XML metadata file is transformed into a MODS XML file. The MODS file is persisted in the AIP and also made available to the public at multiple levels of granularity (e.g., package level, chapter level, granule level).

Enhanced Search

Metadata generated using the techniques described above enable the FDsys search engine to provide robust search functionality to public end users who are looking for content on the FDsys web site.

Publishing

After a package has completed processing within the content management system component, the package is sent to the FDsys content publisher. The content publisher continuously checks the content management system for new packages, updated packages, or deleted packages. Based on the results of these checks, the publisher sends content and metadata to the search engine for indexing or requests that information be removed from the search engine. Once the content and metadata have been indexed by the search engine, a custom web application uses indexed data to present search results, browse pages, and more information pages to users.

Simple Search

FDsys provides a simple "one box" search on the FDsys homepage where public users can enter a word or phrase and generate relevant search results across all collections of data. In addition, this same search box supports Boolean and proximity operators to construct more complex search queries for precisely targeted queries. A query parser is used to translate a user query into a statement that can be processed by the search engine. Relevancy ranking is controlled by the search engine and can be influenced by the presence of key metadata elements in specific fields. For example, a query on "H.R. 123" will prefer documents with "H.R. 123" in the citation field over documents which reference the bill in the full-text of the document. Note that citation variations are also indexed, so "H.R. 123" will return "HR 123", and "House Bill 123", all with equal relevancy.

Advanced Search

For each collection, select metadata fields and their values are indexed by the search engine. Some of these fields are made available through a web form that allows users to search the metadata fields for each collection. Searching by fields such as date published allows users to retrieve more precise results than they would when using a simple full-text search.

Users are first given the option to search by date. They can then select which collections they would like to search. Next metadata fields common to all selected collections are presented in an advanced search form. Search fields will accept simple values or complex Boolean expressions. The search results will match the field values specified by the user.

Field Operator Search

To reduce user interface clutter, only selected fields are made available for searching on the advanced search page. To provide a comprehensive metadata search, additional indexed fields are available for searching using FDsys field operators. Users can enter a query into the search box on the FDsys home page using the field operator syntax in fieldname:fieldvalue format (e.g., *sponsor:Mikulski*). Complex Boolean expressions (e.g., *sponsor:(mikulski or cardin)*) are also allowed. For fields stored as integers or dates, range searches are available, using the range() operator (e.g., *congress:range(109,110)*). Field operator search syntax can be used for multiple fields in a single query and can be combined with Boolean search operators to provide users with the capability to construct complex search queries.

MODS Search

By transforming FDsys metadata into MODS and indexing the MODS file with each package, FDsys enables users to search directly over MODS descriptive metadata. Users can enter a query into the search box on the FDsys homepage using the MODS search syntax. The syntax for searching across MODS metadata relies upon the user's understanding of the hierarchy of metadata

©2011 Society for Imaging Science and Technology

as it is stored in the MODS XML file. A sample MODS search might look like this:

mods:name:(mikulski and role:roleTerm:speaking)

Filtering Search Results

Once a user has retrieved search results from a simple search, advanced search, field operator search, or MODS search, the user can filter search results. Metadata captured for each publication is used to populate a set of selectable filters (also called facets or navigators). All collections share a common set of filters, however a user can also use collection-specific filters that are displayed after a user has selected a collection. Filtering uses metadata to provide users with the capability to refine their search query in order to retrieve more precise search results.

Retrieve by Citation Form

Many government publications have well known document citations. While simple search, advanced search, field operator search, and MODS search provide users with a wealth of search results to choose from, the functionality to retrieve by citation provides users with the capability to enter metadata values into a web form and receive a single, specific PDF file. The web form guides users through the relevant metadata fields and values required to retrieve a specific document based on a common citation format for that collection.

Browse

FDsys provides a unique browse page to browse the entire contents of each collection. For each browse page, a hierarchical tree structure is generated using metadata. Each node in the tree structure groups documents that have the same value for a particular metadata field. This allows users to drill down into the collection by clicking on the nodes until they are presented with a list of documents. FDsys allows for each tree to be customized to be a natural representation of the documents for the collection. For example, Congressional Bills are browsed with a tree based on Congress Number, Bill Type, and Bill Number range. The Federal Register is browsed by Year, Month, Day, and Agency Name. Once the user drills down to the lowest level of the tree, they are presented with a list of documents which includes the title plus other relevant metadata elements, links to the content for download, and a link to the document's "more information" page.

More Information Pages

Whether a user accesses a document through search results or a browse page, they can either view the document content (typically a PDF representation of the document) or click on a link to access the "more information" page. The more information page contains links to all available public formats for the document, links to MODS and PREMIS XML metadata, and a link to a zip file containing all of the publicly available files for the package as a whole. If the document is just one granule within a larger package, it will be displayed "in context" with the other granules from the package in a package table of contents on the more information page.

Conclusion

With FDsys, GPO has implemented a process to analyze content collections and document the flow of information through the system. The documentation provides a blueprint for the flow of content and metadata through FDsys from packaging to public display. The methods to acquire descriptive metadata are tailored to each collection and are specified in a design document for each collection. Descriptive metadata is currently used to provide public users with a world-class search experience on the FDsys web site, enabling users to find the information they need, and it gives GPO and other organizations the opportunity to build new capabilities in the future to ensure that GPO continues to meet its mission of Keeping America Informed.

References

[1] United States Government Printing Office, "A Strategic Vision for the 21st Century," (2004).

[2] GPO's Federal Digital System (FDsys). <http://www.fdsys.gov>.

[3] Lisa LaPlant and Kate Zwaard. "A Holistic Approach for Establishing Content Authenticity and Maintaining Content Integrity in a Large OAIS Repository." In Proceedings from Archiving 2008, June 2008, pp 109-113.

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

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

[6] Gil Baldwin, Matthew Landgraf, Kate Zwaard, John Faure. "Content Packaging Approach for a Large OAIS Repository." In Proceedings of Archiving 2007, May 2007, pp 44-47.

[7] Guenther, R. and McCallum, S., "New Metadata Standards for Digital Resources: MODS and METS," Bulletin of the American Society for Information Science and Technology, 29, 12-15 (2003).

[8] FDsys System Design Document (SDD). <http://www.gpo.gov/fdsysinfo/designdoc.htm>.

Author Biographies

Lisa LaPlant received her B.A. in Media Arts and Design from James Madison University. She has been working on the design and implementation of FDsys since 2005 as a Lead Program Planner for the U.S. Government Printing Office in Washington, D.C.

Blake Edwards received his B.S. in Graphic Communication from California Polytechnic State University, San Luis Obispo and his M.A. in Educational Technology Leadership from George Washington University. He has been working on the design and implementation of FDsys since 2005 as a Program Planner for the U.S. Government Printing Office in Washington, D.C.

Paul Nelson, the Chief Architect for Search Technologies Corporation, has been working for GPO as the FDsys Search Engine Architect since 2008. He received his B.E.E. in Digital Design from the University of Delaware and his M.S.C.S. in Natural Language Software from The John Hopkins University. Paul is one of the co-founders of ConQuest Software, a search engine company started in 1989 that developed into what eventually became (after several acquisitions) Convera RetrievalWare, a fully-featured search engine that is now owned and supported by Microsoft. He holds a patent for Multimedia Document Retrieval By Application of Multimedia Queries To A Unified Index of Multimedia Data For A Plurality of Multimedia Data Types (United States Patent 6,243,713).