Enhanced Content Technology Industry Day

Q&A

Format Identification and Translation

- Does GPO expect that the presentation will be maintained when translating from other formats to XML?
  - *Yes. The goal is for FDsys to translate ingested content to XML and to preserve the content, structure, and presentation, when possible. We hope that industry can recommend detailed options for maintaining presentation.*

- Have you developed DTDs or schemas for your products?
  - *Yes, we have developed schemas for a limited number of products.*

- In the Format Translation and Identification scenario, you talk about creating a PDF. Will you use the native file (e.g., FrameMaker) or the XML as a basis for the PDFs?
  - *FDsys requirements state that derivatives will be created from the XML. While the initial public release will likely use native documents to create derivatives, it is hoped that in future releases derivatives will be generated from XML.*

- During the presentation GPO mentioned the need to translate older file formats to more recent versions of the file (e.g., Adobe PDF 1.0 to Adobe PDF 2.0)?
  - *Migration is part of FDsys preservation strategy.*

- Will content be preserved in a database or in the CMS?
  - *Content will be preserved in a repository that is managed by the CMS.*

- Will FDsys support formats other that XML (e.g., locator codes and proprietary formats)?
  - *Yes. While the intent is to be XML based it is understood that FDsys will need to support locator codes as well as other formats.*

- Are there any metrics available for the volume of content that will flow through the translation process?
  - *GPO expects to ingest about 50,000 documents in the first year of FDsys operation. At full operation, GPO expects about a million documents a year. Many of these documents will require format translations.*

Note: There were no questions on Format Identification

Automatic Content Formatting

- Are Style Tools different than Composition Systems Replacement (CSR)?
  - *CSR and FDsys (including style tools) are being managed as separate programs; however, it is possible that CSR may satisfy some style tool requirements.*

- Is the point of style tools to get customers to tag content?
  - *No. The point of Style Tools is to provide customers with an easy to use tool to create content. Style tools will allow customers to focus more on developing content and less on tedious page layout, design and presentation.*

- When a content originator uses style tools to compose a document, will they be able to select different output types (e.g., optimized for print or web)?
  - *Yes. The content originator decides the output type. However, preservation and access files may be created by FDsys to support permanent public access for content in scope for GPO's dissemination programs.*

- Do you have expectations for capturing authenticity metadata or information at this stage and are you trying to capture information about who is creating the publication?
  - *Yes. FDsys will capture authenticity, integrity, and chain of custody information for publications.*

- When customers reuse elements of previous publications in style tools, will changes to the new publication affect the previous publication
  - *No. Content submitted to the archive remains fixed and immutable. However, content in pre-publication storage areas (work in progress) could be linked together so that a change to agency logo, for example, would be reflected throughout their work in progress.*

- Will FDsys have a specific flavor (e.g., defined schema) for XML?
  - *Yes. FDsys will conform to W3C standards: XML 1.0 for XML documents, and XML Schema recommendation of 2001 for schema definitions for the XML documents. XML documents created by XML tools of any flavor (e.g. Apache, MS XML product) that conform to the same standards will be supported by FDsys.*

Near Duplicate Detection

- How do you distinguish critical information from non-critical information? For instance a date or name may change. Will there be manual interventions in managing this process?
    - *GPO will need to determine rules and conventions for distinguishing between critical and non-critical information and the rules for if it constitutes a new publication. There will be manual intervention if required.*

Content Parsing

- How does GPO define granularity?
    - *GPO wants to extract content or graphics based on the natural boundaries of the publication so that the context or intent of the material is maintained. Please refer to the FDsys Requirements Document for specific granularity requirements.*

- There has been little mention of images and your examples have been focused on content or text. Is this intentional?
    - *This is not intentional. FDsys must handle all submitted content including images, audio, and video.*

- What is the purpose of content parsing?
    - *Content parsing will enhance access to content and allow GPO to repurpose content.*

- Will FDsys parse at the page level?
    - *We would like FDsys to be capable of parsing at the page level where necessary, based on the rendition. The required level of granularity (e.g., page, article, paragraph) will be dependent on the natural boundaries of the publication. For example, a Table of Contents could define the natural boundaries of a publication. It is important to maintain the context of parsed content.*

- If users are able to reuse fragments of content in new publications, will an update to a fragment trigger a workflow that updates all the publications that use that content fragment?
    - *Content submitted to the archive remains fixed and immutable. However, content in pre-publication storage areas (work in progress) could be linked together so that a change to agency logo, for example, would be reflected throughout their work in progress.*

- How do parsing and XML translation support search?
    - *Parsing and XML translation allow for more robust search capabilities.*

- When content is parsed, does each logical boundary (such as a paragraph) need to maintain an authentic look and feel?
    - *At least one rendition needs to maintain the look and feel.*

Concept Extraction

- At what level are you looking for concept extraction?
  - *Concepts should be extracted from content at the level of content granularity available in FDsys.*

- Where are extracted concepts persisted?
  - *While it could be stored in the descriptive metadata, we would like you to elaborate in your capability statements.*

- Do you have a list of concepts?
  - *No, this list has not been developed.*

Synthetic Documents

- Do the individual pieces of a synthetic document need to maintain their authenticity?
  - *Yes.*

- Can individual pieces of a synthetic document be from different formats?
  - *There are no policy restrictions. We would like to hear from you regarding technological capabilities.*

- Has GPO identified a primary deliverable file format for a synthetic document?
  - *No. We would like to hear from you regarding technological capabilities.*

Content Authentication

- Do integrity marks have to be in the file or can it be in a separate file?
  - *Integrity marks must be associated with the content and easy for the user to access. We would like to hear from you regarding technological capabilities and best practices.*

- Does authentication have to be down to the paragraph level?
  - *Any granularity level we provide must be authenticated.*

- Where does chain of custody begin for content authentication?
  - *It begins when the final published version of content is submitted to FDsys.*

Thank you for your interest in GPO's FDsys program!