

ABOUT THE EVENT

The format of the ECT event will consist of a very brief presentation by the FDsys Program Management Office (PMO) – one (1) presentation for each of the ECTs identified in the RFI – followed by a 20 minute Q&A Session for each ECT. The PMO presentation will include a scenario to support each ECT.

3/20/07 - AGENDA:

8:00 – 8:30	Arrive GPO	Vendors
8:35 – 8:40	Objectives for the Session	Scott Stovall
8:40 – 9:00	Introduction and Overview Welcome and background	Mike Wash
9:00 – 9:30	Format Identification (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Kate Zwaard
9:30 – 10:00	Format Translation (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Magan Fleetwood
10:00 – 10:30	Automatic Content Formatting (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Magan Fleetwood
10:30 – 11:00	Near Duplicate Detection (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Matt Landgraf
11:00 – 11:30	Content Parsing (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Lisa LaPlant
11:30 – 12:00	Concept Extraction (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Lisa LaPlant
12:00 – 12:30	Synthetic Documents (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Lisa LaPlant
12:30 – 1:00	Content Authentication (Includes Q&A) Brief presentation of a scenario that describes FDsys needs.	Lisa LaPlant
12:30 – 1:00	Close – Next Steps	Scott Stovall

AFTER THE EVENT WHAT SHOULD YOU DO?

1) Write Capability Statements and Assessment Documents

Capability Statements: Keeping GPO's goals in mind please outline how your products/services would support any or all of the ECTs. Capability statements should be kept to 10 pages or less - and no marketing material - per ECT.

Assessment Documentation: In addition, vendors are hereby encouraged to submit assessment documents (e.g., white papers or other documentation - no more than 50 pages total) that address any gaps or concerns with GPO's ECTs or ECT requirements. Feedback should include specific comments and suggestions for refining or clarifying ECTs and should reflect clear guidance from industry and/or industry best practices.

Budgetary Estimates: Vendors are also encouraged to provide budgetary estimates on each of the ECTs based on the information available.

NOTE: Please also list company points of contact and GSA Schedule number (if applicable).

2) Questions for the FDSys PMO Team

Questions for the PMO team or questions regarding capability statements, white papers and other documentation must be submitted via email (with the subject line "**Questions about ECT Industry Day**") to Herb Jackson at hjackson@gpo.gov. Questions must be submitted by noon on 26 March 2007. Questions submitted after this date and time will not be answered.

3) Submitting Material

Capability statements and/or other white papers and documentation related to ECT Industry Day must be submitted by noon on 6 April 2007 to int@gpo.gov. Documents must be submitted as **Adobe PDF or Microsoft Word files (2000 or higher)**. Faxed copies are not acceptable.

Notice

The industry day event and related RFI are requests for information only and do not obligate the GPO in any way. The event and subsequent dialog are not a request for proposal and the GPO will not pay for any information submitted or for any expenses associated with providing information. Any information submitted by respondents to the event and the RFI is strictly voluntary. Material submitted will be deemed proprietary to the extent permitted by applicable laws and regulations if so marked by the respondent.

U.S. Government Printing Office

Session Objectives

ECT Industry Day
March 20, 2007

Objectives

- **Session:**
 - Clearly describe and explain ECT's identified by GPO
 - Provide an open forum for GPO and industry to discuss these ECT's and FDsys challenges
 - Provide industry with enough data and understanding
- **TO SUPPORT**
- **Overall:**
 - To obtain industry feedback on ECT's
 - Assessment of materials
 - Include white papers, and other industry best practices
 - Budgetary
 - To understand possible technology roadmaps for each ECT – through capability statements.
 - Is COTS support available?
 - How soon will COTS support be available?

Next Steps – Submit Materials

- **Capability Statements:**
 - How does your product/service support any/all ECT
 - 10 pages or less per ECT
 - No marketing material
- **Assessment Documentation:**
 - Gap assessment of ECT or ECT requirements
 - Comments, suggestions, refinements, guidance, white papers or industry best practices that are relevant to ECTs.
 - No more than 50 pages total.
- **Budgetary Estimates:**
 - Based on available information
 - Include company points of contact and GSA Schedule number (if applicable).
- **Submit by 4/6/07**

Additional Information

Program documentation is available on the
GPO website:

www.gpo.gov/projects/fdsys.htm

Scott Stovall
Deputy Chief Technical Officer
[sstovall@gpo.gov](mailto:ssstovall@gpo.gov)

A world class system for managing official Government content, which will verify and track versions, track versions, assure authenticity, preserve content, and provide assure authenticity, preserve content, and provide permanent access. A world class system for managing official Government content, which will verify and track versions, assure authenticity, preserve content, and provide permanent access. FDsys an integrated digital content management system

U.S. Government Printing Office

Content-centric Solution to Support Printing, Publishing and Permanent Public Access

ECT Industry Day
March 20, 2007



GPO Mission

- To provide the three branches of the Federal Government with expert publishing and printing services.
- To provide perpetual, free, and ready public access to the printed and electronic information published by the Federal Government, in partnership with Federal Depository Libraries.
- To distribute, on a cost recovery basis, printed and electronic copies of information published by the Federal Government.

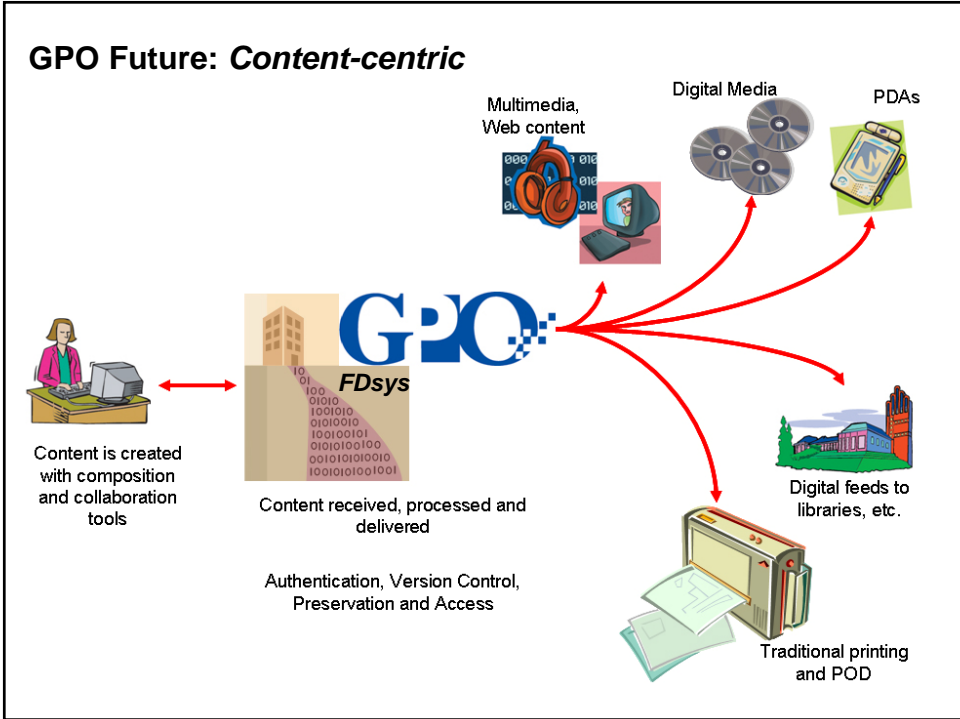
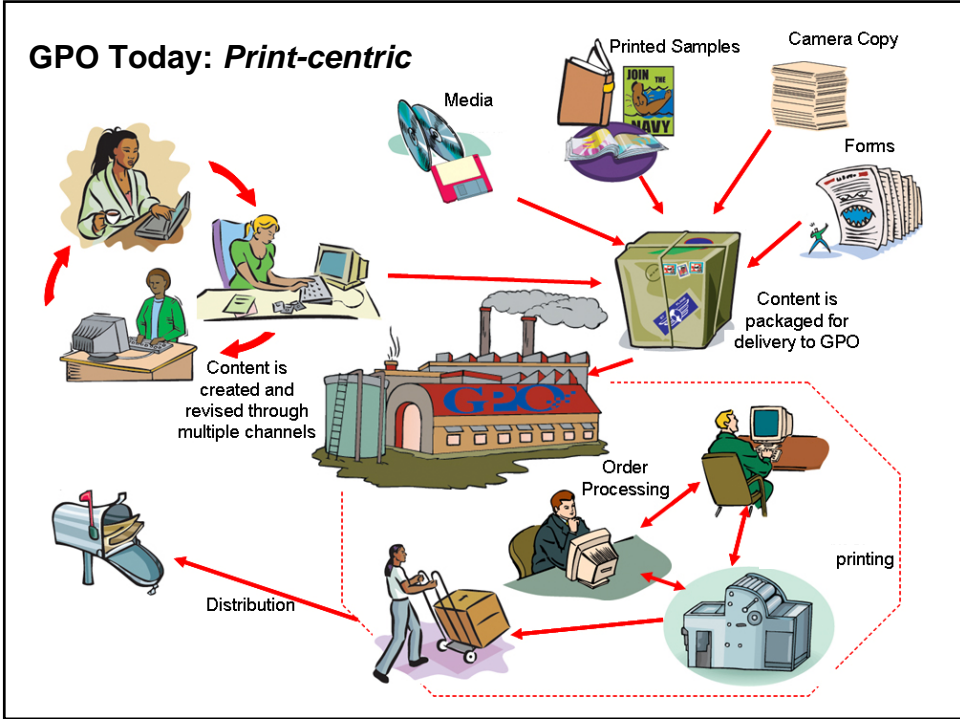


Challenges to meeting the Mission

- Access to government published information is now widely expected to be electronic.
- Digital information needs to be authentic and verified to be the correct version.
- Digital information needs to be available for access almost immediately.
- Information needs to be preserved, making it available for generations to come.

Transformations at GPO

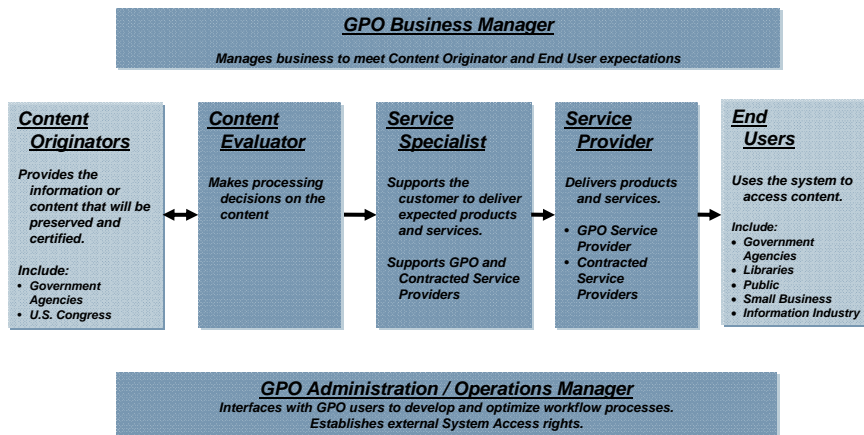
1. Deploying new information technology systems
 - A new information infrastructure is required to replace legacy systems and applications.
2. Restructuring around business units
3. Moving to a content-centric model
 - GPO has historically been print-centric. To effectively achieve our mission today and in the future, we need to be content-centric.



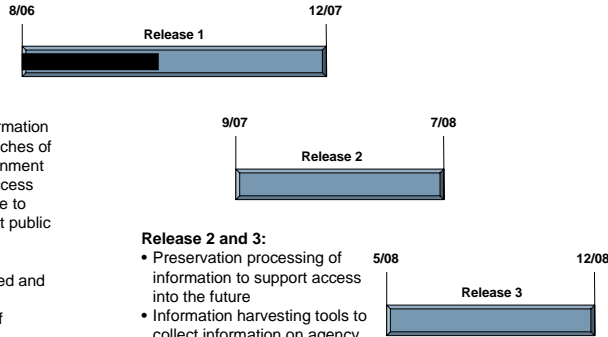
Content-centric Solution – FDsys

- FDsys will automate the collection and dissemination of electronic information from all three branches of government.
- Information will be:
 - permanently available in electronic format
 - authenticated and versioned
 - accessible for Web searching, viewing, downloading and printing
 - available for conventional and on-demand printing

GPO User Class Model



FDsys Release Schedule



Release 1:

- Accept digital information from all three branches of the Federal Government
- Replaces GPO Access
- Information storage to support permanent public access
- Authentication of information received and delivered
- Version tracking of publications

Release 2 and 3:

- Preservation processing of information to support access into the future
- Information harvesting tools to collect information on agency websites
- Automated version tools
- Extended storage to support harvested and legacy publications as required.
- Enhanced information authentication

ECT Release Schedule

ECT	Base Functionality	Full Enhanced Functionality
Automatic Formatting of Text Based Content: Please refer to FDsys Style Tools concept and requirements.	N/R	R 3 - End '08
Synthetic/Hybrid/Dynamic Text Based Documents: Combining content from multiple publications or a single publication at varying levels of granularity (e.g., page, paragraph, section, complete publication) to create a new document. This may be done manually or based on rules to automate the process of creating these synthetic/hybrid/dynamic documents.	N/R	R 3 - End '08
Content Authentication: Adding integrity marks (e.g., digital signatures, watermarks) to non-PDF content including XML, audio, and video at multiple levels of granularity.	R 1C - End '07	R 3 - End '08
Format Translation (All Formats): Capability to translate a variety of content formats to XML without any loss of content integrity. Capability to translate a variety of content formats to formats other than XML without any loss of content integrity. Capability to translate system metadata into registered XML metadata. Capability to translate one schema to another schema (e.g., MODS to Dublin Core).	R 1C - End '07	R 2 - Mid '08
Format Identification (All Formats): Format identification is a technology for identifying the formats of ingested files without depending upon extension or MIME type.	R 1C - End '07	R 2 - Mid '08
Content Parsing (All Formats and Content Types): Separating digital content within a publication into smaller components. Some examples include separating content into separate smaller pieces of content at logical boundaries (e.g., paragraphs or sections), creating tags within publications to identify smaller components, and creating references which indicate the separation of smaller components.	R 1C - End '07	R 3 - End '08
Concept Extraction: Analyzing content for the purpose of extracting concepts and suggesting related concepts that can be used for categorization and search (e.g., search for "World Series" returns articles on the Red Sox).	R 1C - End '07	R 2 - Mid '08
Near Duplicate Detection: Capability to detect near duplicate documents. This includes tools that can detect slight variations between documents, such as a few different sentences or paragraphs. This could be used to detect when a newer or older version of a document that is already in the archive is being ingested.	N/R	R 3 - End '08

ECT Release Schedule

ECT	Base Functionality	Full Enhanced Functionality
Automatic Formatting of Text Based Content: Please refer to FDsys Style Tools concept and requirements.	N/R	R 3 - End '08
Synthetic/Hybrid/Dynamic Text Based Documents: Combining content from multiple publications or a single publication at varying levels of granularity (e.g., page, paragraph, section, complete publication) to create a new document. This may be done manually or based on rules to automate the process of creating these synthetic/hybrid/dynamic documents.	N/R	R 3 - End '08
Content Authentication: Adding integrity marks (e.g., digital signatures, watermarks) to non-PDF content including XML, audio, and video at multiple levels of granularity.	R 1C - End '07	R 3 - End '08
Format Translation (All Formats): Capability to translate a variety of content formats to XML without any loss of content integrity. Capability to translate a variety of content formats to formats other than XML without any loss of content integrity. Capability to translate system metadata into registered XML metadata. Capability to translate one schema to another schema (e.g., MODS to Dublin Core).	R 1C - End '07	R 2 - Mid '08
Format Identification (All Formats): Format identification is a technology for identifying the formats of ingested files without depending upon extension or MIME type.	N/R	R 1C - End '07
Content Parsing (All Formats and Content Types): Separating digital content within a publication into smaller components. Some examples include separating content into separate smaller pieces of content at logical boundaries (e.g., paragraphs or sections), creating tags within publications to identify smaller components, and creating references which indicate the separation of smaller components.	R 1C - End '07	R 3 - End '08
Concept Extraction: Analyzing content for the purpose of extracting concepts and suggesting related concepts that can be used for categorization and search (e.g., search for "World Series" returns articles on the Red Sox).	R 1C - End '07	R 2 - Mid '08
Near Duplicate Detection: Capability to detect near duplicate documents. This includes tools that can detect slight variations between documents, such as a few different sentences or paragraphs. This could be used to detect when a newer or older version of a document that is already in the archive is being ingested.	N/R	R 3 - End '08

Current Status

- Design for Release 1B is complete
- Component selection for 1B is complete
- Market research is being conducted for remaining releases
 - Design is in process
 - Components have not been selected

Release 1B - FDsys Internal Prototype

Vendor	Component
EMC/Documentum	CMS
FAST	Search
Software AG	ESB
Oracle	Web Server
Oracle	App Server
Oracle	Relational Database
Oracle	Identity & Access Manager
LDAP	Open LDAP
Trend Micro	Virus Detection
NTT Verio	Managed Host
Ex Libris	Integrated Library System (ILS)

About R1B:

- Purpose is to test concepts and design for key system components (e.g., Packaging)
- Selected components are for R1B
- Design leverages existing GPO EA infrastructure where appropriate
- Support 50 internal and approximately 10 external users
- Will have a small storage footprint (< 1 TB)
- Will consist of a very simple hardware platform
- Will connect to an existing ILS instance
- Will support beta and end user acceptance test.

Release 1B - FDsys Internal Prototype

Vendor	Component
EMC/Documentum	CMS
FAST	Search
Software AG	ESB
Oracle	Web Server
Oracle	App Server
Oracle	Relational Database
Oracle	Identity & Access Manager
LDAP	Open LDAP
Trend Micro	Virus Detection
NTT Verio	Managed Host
Ex Libris	Integrated Library System (ILS)

About R1B:

- Purpose is to test concepts and design for key system components (e.g., Packaging)
- Selected components are for R1B
- Design leverages existing GPO EA infrastructure where appropriate
- Support 50 internal and approximately 10 external users
- Will have a small storage footprint (< 1 TB)
- Will consist of a very simple hardware platform
- Will connect to an existing ILS instance
- Will support beta and end user acceptance test.

Summary of FDsys

- A comprehensive content management system to support GPO's mission to *Keep America Informed* is being developed
- Content will be managed around the themes of *authentication, version control, preservation and access*
- The new system will be developed and deployed without a disruption in operations at the agency
- Core functionality is expected to be complete in 2007

Additional Information

Program documentation is available on the GPO website:

www.gpo.gov/projects/fdsys.htm

Mike Wash
Chief Technical Officer

mwash@gpo.gov

U.S. Government Printing Office

Format Identification and Translation

Enhanced Content Technologies Industry Day

March 20, 2007

Definitions

- **Format Identification**

- The ability to identify file formats without using MIME type

- **Format Translation (All Formats)**

- Capability to translate a variety of content formats to XML without any loss of content integrity.
- Capability to translate a variety of content formats to formats other than XML without any loss of content integrity.
- Capability to translate system metadata into registered XML metadata.
- Capability to translate one schema to another schema (e.g., MODS to Dublin Core).

Scenario

- **An Adobe FrameMaker 5.0 document, fonts, and TIF, GIF, and EPS images are uploaded to FDsys**
 - FDsys correctly identifies all file formats
 - FDsys creates an XML file that retains the presentation of the document
 - FDsys creates an ASCII file of the FrameMaker text and image descriptions for 508 compliant access
 - FDsys creates a print optimized PDF for print-on-demand
 - FDsys creates a FrameMaker 7.2 document

Requirements

- **Most format translation requirements can be found under section 3.2.4.3.2.2 Preservation Processing in the FDsys Requirements Document v3.0**

Discussion Points

- **What technologies currently exist for content identification? Content translation?**
- **Success rates for migration of content to newer versions (e.g., Word 97 to Word 2003) without loss of content integrity**
- **Challenges in translating one schema to another (e.g., MODS to Dublin Core)**

U.S. Government Printing Office

Automatic Formatting of Text Based Content

Enhanced Content Technologies Industry Day
March 20, 2007

Definition

- **Style Tools**
 - Content composition: tools to compose content into appropriate layouts based on user requirements
 - Content collaboration: tools to enable concurrent users at separated workstations to work on a single project
 - Content approval: tools for authorized users to approve changes made by collaborators, final content, and final content presentation
- **Goal: to move GPO upstream in the content origination process**

Scenario 1

- **Content Originator accesses FDsys style tools and indicates that they want to create a new scientific manual for printed output.**
- **Content Originator enters text and uploads graphics. The document is composed automatically.**
- **Content Originator adjusts the position of a graphic and the text reflows automatically.**
- **Content Originator changes the output to electronic handheld devices and the content is automatically recomposed.**

Scenario 2

- **Content Originator uses FDsys style tools to create a custom publication using parsed content that is stored in the ACP and/or WIP.**
- **Example: A small Government commission accesses their annual report for 2005 and re-uses a paragraph about their mission in their 2007 annual report.**

Requirements

- **3.2.6.5.2 Requirements for Style Tools in the FDsys Requirements Document v3.0**

Discussion Points

- **Other than templates, what automated content composition technologies are currently available? At what level of maturity?**
- **Delivery mechanisms to be supported by style tools**
 - Hard copy
 - Electronic presentation
 - Digital media

U.S. Government Printing Office

Near Duplicate Detection

Enhanced Content Technology (ECT) Industry Day

March 20, 2007

Description

- Detection of:
 - The same document with a small set of changes from an existing document (e.g., a few sentences or paragraphs are different).
 - The same document with the same text but different formatting.
 - The same document in different file formats.
- Very close relationship to Version Control.
 - Will be useful in detecting when a newer or older version of a document that will be ingested is already in the archive.

Description (cont.)

- Technology must be able to characterize the nature of the difference between near duplicate documents.
 - Examples:
 - A line of text or paragraph was added.
 - Blank spaces or characters were added
 - A graphic or equation within a document was re-positioned.
- GPO will use this information to determine whether the changes constitute a new version.

Key Requirements

- 3.2.4.1.1.1.8.1.4: The system shall have the capability to detect near duplicate documents.
- 3.2.4.1.1.1.8.2.2: The system shall notify users when near duplicate content is detected.

Scenarios

- An EPA report on hazardous waste already exists in the system (deposited content from the Content Originator).
- GPO harvests the same report from the EPA Web site. The title and version are the same, but the harvested version has a different date and two sentences were added to the “Dedications” page of the report.

Discussion Points

- Requirements for near duplicate detection of non-text elements and granular portions of publications.
- Using near duplicate detection technologies detect different renditions (e.g., the same document in different formats).
- What technologies and methodologies are available in the market now?
- Are they mature, or are they in the research and development phase?
- For publications with minor changes (e.g., a few sentences) what constitutes an entire new version?

U.S. Government Printing Office

Content Parsing

Enhanced Content Technology (ECT) Industry Day

March 20, 2007

Description

- Content Parsing refers to the process of separating digital content within a publication (or audio/visual equivalent) into smaller components. Examples include the following:
 - Separating content into smaller pieces of content at logical boundaries (e.g., paragraphs or sections).
 - Creating tags within publications to identify smaller components.
 - Creating references which indicate the separation of smaller components.
 - Physically segmenting content.
- Formats include text-based formats such as PDF, ASCII text, HTML, and XML; audio formats; and video formats.

Key Requirements

- 3.2.4.4.2.1.0.1 - The system shall support granularity of any content based on the natural granularity boundaries of that content.
- 3.2.4.4.2.1.0.1.2 - The system shall support granularity of GPO Access content referenced in 3.2.7.4.2.2.4.2 based on the natural granularity boundaries of that content.
- 3.2.4.4.2.1.1.0.1.1 - The system shall have the capability for a user to apply multiple levels of granularity to a publication (e.g. the whole publication can be found, every paragraph in the publication can be found but images can not be separately found).
- 3.2.4.4.2.1.1.0.2 - The system shall allow elements to be retrieved by at all levels of granularity.

Key Requirements (cont.)

- 3.2.4.4.2.1.1.2 - The system shall support granularity down to the level of any paragraph in a publication.
- 3.2.4.4.2.1.1.3 - The system shall support granularity down to the level of any individual graphic.
- 3.2.4.4.2.1.1.4 - The system shall support granularity down to the level of any embedded graphical element in a publication.
- 3.2.4.4.2.1.1.6 - The system shall support granularity down to the level of any frame of a video.
- 3.2.4.4.2.1.1.8 - The system shall support granularity of audio down to smallest segment of time the audio encoding allows.

Scenario

- End User accesses authentic Federal Government content at a level of granularity that is less than a publication. End User also accesses the entire publication and the complete content package.

Discussion Points

- What technologies currently exist for parsing and segmenting content in a variety of formats including text-based formats such as PDF, ASCII text, HTML, and XML; audio formats; and video formats? How mature are these technologies?

U.S. Government Printing Office

Concept Extraction

Enhanced Content Technology (ECT) Industry Day

March 20, 2007

Description

- Analyzing content for the purpose of extracting concepts and suggesting related concepts to be used by cataloging in addition to categorization and search.

Key Requirements

- 3.2.7.6.2.1.6 - The system shall support the extraction of metadata from content.
- 3.2.7.4.2.2.6 - The system shall allow users to perform a search for conceptually related terms (e.g., search for "World Series" returns articles on the Red Sox).
- 3.2.7.4.2.2.6.1- The system shall allow authorized users to manage concept relationships.
- 3.2.7.4.2.2.6.1.4 - The system shall suggest new concept relationships based on ingested content.
- 3.2.7.4.2.2.6.1.4.1 - The system shall automatically create new concept relationships based on an authorized users acceptance of suggested new concept relationships.
- 3.2.7.4.2.2.6.1.5 - The system shall use new concepts without requiring previously indexed content is re-indexed.

Scenarios

- Metadata is extracted from content and is used to populate system metadata.
- New concept relationships are created for content as it is ingested into FDsys.

Discussion Points

- What technologies currently exist for extracting metadata from content? How mature are these technologies?
- How mature are technologies for concept creation?

U.S. Government Printing Office

Synthetic Documents

Enhanced Content Technology (ECT) Industry Day

March 20, 2007

Description

- Combining content from multiple publications or a single publication at varying levels of granularity (e.g., page, paragraph, section, complete publication) to create a new document containing authentic content. This may be done manually or based on rules to automate the process of synthetic document creation.

Key Requirements

- 3.2.7.5.2.4.16 - The system shall provide the capability for users to select level of granularity from available options (e.g., title, part, section, paragraph, graphic, page).
- 3.2.7.5.2.4.22.2 - The system shall provide the capability for authorized users to preview publications that have been created from custom composition and content formatting.
- 3.2.7.5.2.4.23 - The system shall have the capability to support custom composition and content formatting from available options (e.g., 2 columns, cover stock, font).

Scenarios

- An authorized user requests custom publication creation based on rules. The user should not have to manually select each item for inclusion in their custom publication.
- An End User uses a composition engine to manually select content and create a single custom publication.

Discussion Points

- What technologies currently exist for creating synthetic documents? How mature are these technologies?
- Are there preferred formats for the creation of synthetic documents?
- Are technologies available to create synthetic documents that maintain the authenticity, style, and formatting of the source documents?

U.S. Government Printing Office

Content Authentication

Enhanced Content Technology (ECT) Industry Day

March 20, 2007

Description

- Adding integrity marks (e.g., digital signatures, watermarks) to content including text-based content, such as PDF and XML, audio content, and video content at multiple levels of granularity.

Key Requirements

- 3.2.4.6.2.4.4.1 -The system shall have the capability to apply a cryptographic digital signature, in accordance with IETF RFC 3447, to content delivered from the system.
- 3.2.4.6.2.6.3 - Integrity marks shall employ widely accepted information security mechanisms (e.g., public key cryptography, digital certificates, digital signatures, XML signatures, digital watermarks, or traditional watermarks).
- 3.2.4.6.2.7.1 - The system shall provide the capability for users to validate the authenticity, integrity, and official status of the content packages that are delivered from the system.

Key Requirements (cont.)

- 3.2.4.6.2.7.2.1 - The system shall enable GPO to add integrity marks to FDsys content that is delivered to End Users in the form of electronic presentation.
- 3.2.4.6.2.7.2.2 - The system shall enable GPO to add integrity marks to FDsys content that is delivered to End Users in the form of hard copy output.
- 3.2.4.6.2.7.2.3 - The system shall enable GPO to add integrity marks to FDsys content that is delivered to End Users in the form of digital media.

Scenarios

- GPO is able to customized the application of digital signatures (or other integrity marks) on a per publication per rendition (e.g. format) basis.
- An authenticated publication is delivered to an End User in the form of electronic presentation, digital media, or hard copy.
- Content authenticity and integrity are maintained at all levels of granularity available in the system.

Discussion Points

- What technologies currently exist for adding integrity marks to publications (or their audio/visual equivalent)?
- How mature are these technologies and are they being implemented in situations that involve dissemination of publications (or documents) to the general public?

A world class system for managing official Government content, which will verify and track versions, track versions, assure authenticity, preserve content, and provide assure authenticity, preserve content, and provide permanent access. A world class system for managing official Government content, which will verify and track versions, assure authenticity, preserve content, and provide permanent access.

FDsys
an integrated digital content management system

U.S. Government Printing Office

Next Steps and Close

ECT Industry Day
March 20, 2007

G2O

FDsys

Next Steps – Q&A

- Submit Written Questions
 - By 3/26
- GPO will respond in writing
 - By 3/30

G2O

1

Next Steps – Submit Materials

- **Capability Statements:**
 - How does your product/service support any/all ECT
 - 10 pages or less per ECT
 - No marketing material
- **Assessment Documentation:**
 - Gap assessment of ECT or ECT requirements
 - Comments, suggestions, refinements, guidance, white papers or industry best practices that are relevant to ECTs.
 - No more than 50 pages total.
- **Budgetary Estimates:**
 - Based on available information
 - Include company points of contact and GSA Schedule number (if applicable).
- **Submit by 4/6/07**

FAQ's

- Can we have a face-to-face meeting with GPO to describe our products/services more clearly
 - To be determined after receipt of capability statements and other materials.
- When will you begin acquiring ECT?
 - See Release schedule handout.

Additional Information

Program documentation is available on the
GPO website:

www.gpo.gov/projects/fdsys.htm

Scott Stovall
Deputy Chief Technical Officer
sstovall@gpo.gov