Baseline Requirements for Digital Reformatting and Delivery of Legacy Federal Document Collections

This draft *Baseline Requirements for Digital Reformatting and Delivery of Legacy Federal Documents Collections* was prepared by the Center for Research Libraries (CRL) for the Government Printing Office (GPO). The baseline requirements set minimum requirements governing file formats, interoperability, accessibility, asset management controls, and other aspects of digital reformatting and delivery for print and microform legacy materials. These requirements are CRL recommendations. GPO is not proposing and has not adopted these baseline requirements. This document is to be considered a reference document.

November 29 2004

The systematic conversion and delivery of federal government documents to digital or electronic format is part of the GPO's larger management plan for federal depository library holdings. The goal of the GPO conversion and delivery regimes is not only the preservation of legacy government-produced documents but their persistent and widespread availability as well. Ensuring persistent and widespread availability will require three things:

1. Optimal conversion standards and methodologies
2. Production of adequate metadata for the management of digital objects
3. Sound management regimes and structures

Since digital conversion will provide a means for disseminating and delivering tangible federal documents content, and hence will augment and parallel the delivery of born-digital federal documents, the requirements for digital conversion and presentation must mesh with the GPO's prospective digital asset management activities.

Since digital and network technologies are dynamic, the GPO reserves the right to refine, revise and expand requirements for conversion, metadata, and management of federally-produced documents.

1. Optimal Conversion Standards and Methodologies

Production of digital master and derivative files from federally-produced tangible documents (source documents) will follow the specifications outlined in the Digital Library Federation's *Benchmark for Faithful Digital Reproductions of Monographs and Serials.* The DLF benchmark provides the minimum necessary specifications for digitally reformatted monographs and serials intended to faithfully reproduce the underlying source materials.[1]

---

[1] Digital Library Federation Benchmark Working Group. *Benchmark for Faithful Digital Reproductions of Monographs and Serials. Version 1***.** December 2002. accessed at: http://www.diglib.org/standards/bmarkfin.htm

General Characteristics and File Formats –Master and derivative digital files created from federally-produced documents (source documents) must have certain general characteristics that enable widespread accessibility and display without reliance upon a single proprietary platform or software.

a. Master files (storage files): Master capture and storage files will be in standard file formats that are able to be viewed, copied, edited, stored, managed and transmitted using a range of commercially produced and widely available platforms and operating systems (such as Linux, Oracle, Microsoft Windows, Mac OS); and softwares (such as Adobe Acrobat Reader, Photoshop, Kodak Imaging, Microsoft Word).

b. Derivative files for access (Use files): Use files will be produced in or converted to standard file formats supported by a range of commercially produced and widely available authoring, editing, viewing, and management softwares and platforms and operating systems (such as Microsoft Windows, Mac OS, Palm OS). File formats must support display with uniform results in industry-standard browser software (such as Microsoft Internet Explorer, Netscape). Use files must also be able to be successfully transmitted and received under low-bandwidth conditions (i.e., data transmission speeds of as low as 33kbps), so as not to impose unreasonable technological and cost barriers to users.

2. Production of Adequate Metadata for the Management of Digital Objects

The metadata provided for each converted publication will follow the METS schema, a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation. [2]

The METS schema provides a mechanism for encoding descriptive, administrative, and structural metadata for a digital library object, and for expressing the complex links between these various forms of metadata. It can therefore provide a useful standard for the exchange of digital library objects between repositories. In addition, METS provides the ability to associate a digital object with behaviors or services.

2.1 Descriptive metadata -- The descriptive metadata section of the METS document for the digital object derived from a converted publication consists of one or more elements that provide identifying information about that digital object. Each element may contain a pointer to external metadata, (such as the URL to be used in retrieving a persistent Web document containing the metadata), to internally embedded metadata, or both.

Minimum identifying information includes the following:

- Unique identifying number ID for the digital object
- Superintendent of Documents number of the source document
- Title of the source document and, where appropriate, series title

[2] Library of Congress, Network Development and MARC Standards Office. *METS Metadata Encoding and Transmission Standard,* accessed at: http://www.loc.gov/standards/mets/

2.2 Administrative metadata -- Administrative metadata will address the process of conversion of the source document, rights management information, and other aspects of the source document, and digital provenance information regarding the master and derivative files.

a. Technical metadata includes information regarding the creation, format, and use characteristics of master files:

- Date of capture and conversion of the source document
- Identifying information on individual, organization, or agency responsible for capture and conversion of the document
- Identifying information on "capture" equipment (digital camera, scanner) used in producing the master files for the electronic version of the document.
- Identifying information on scanning and editing softwares and versions
- Master document file format(s)
- Master document file size (in bytes)
- Date and nature of revisions including sharpening, infill, and other editing operations performed on the master file.

b. Intellectual property rights metadata -- includes copyright and license information, when appropriate. This may also include information on underlying third-party rights governing materials incorporated in the source document.

c. Other source metadata - descriptive and administrative metadata pertaining to the source document from which the digital master and derivative derives, to include at minimum:

- Authorship: the identity of authoring individual or organization and issuing agency or department
- Title and, where appropriate, series title
- Date of issuance
- Superintendent of Documents classification number
- Ownership information on copy or copies used as source document

d. Digital provenance metadata - information regarding source/destination relationships between files, including master/derivative relationships between files and information regarding migrations/transformations employed on files between original digitization of an artifact and its current incarnation as a digital library object.

Minimum digital provenance metadata will include the following information about the derivative files:

- Date on which derivative file was generated from master file
- Identifying information on individual, organization, or agency responsible for generating the derivative file
- Nature and date of revisions or conversion of the derivative
- Identifying information on conversion software and version
- Derivative document file format(s)
- Derivative document file size (in bytes)

e. Structural map -- The structural map section of a METS document defines a hierarchical structure that can enable users to navigate through the digital object. The metadata must support preservation or replication in digital form of the following basic structural characteristics of the source document:

- Page format, including size, textual layout and original typographical characteristics
- Pagination of each volume or other physical entities in sequence
- Composite collation of individual volumes or other physical entities
- Sequence of and relationship between multiple related volumes, documents or other physical entities.

3. Sound Management Regimes and Structures

Conversion, hosting, storage, presentation, and management of digital master and derivative files will be undertaken by GPO or by other organizations and entities. Certain minimum safeguards must be put in place to ensure accountability of those entities to the general public, or user community.

3.1 General specifications – The service model for management of and access to digital derivatives of Federal documents must conform to the Open Archival Information System (OAIS) Reference Model, which presents a framework for the long-term preservation of electronic information.[3] It must also be compatible with the technical architecture specified for the National Digital Information Infrastructure and Preservation Program (NDIPP).[4]

Derivative files must be deliverable and able to be made continuously available over a non-proprietary, global network to which all citizens may connect from any location at a reasonable cost and without undue restriction. Derivatives must be able to be made available on "portable" electronic media (such as DVD-ROM, CD-ROM, others).

3.2 Transparency -- In order to ensure "no-fee permanent public access to the official publications of the United States Government," the operations and services of organizations and entities with the primary responsibility for the management and delivery of government documents in electronic form should be transparent. Specifically, those operations should be auditable and/or subject to public disclosure on request.

The managing organization's policies governing digital conversion and management, its methods, and the responsibilities of critical staff must be written and formalized. These policies and responsibilities should be reviewed on a regular basis at the executive or board level of the organization. Moreover, all policies and procedures governing core activities, such as storage, migration, and management of master and derivative files must be published, available on the Web, and archived in a persistent format.

---

[3] The Consultative Committee for Space Data Systems *Reference Model for an Open Archive Information System*, accessed at http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/p2/CCSDS-650.0-R-1.pdf In 2002 the model was formally approved as an ISO Standard (ISO 14721:2002).

[4] Version 0.2 of the Technical Architecture for the National Digital Information Infrastructure and Preservation Program, accessed at http://www.digitalpreservation.gov/repor/NDIIPP_v02.pdf

3.3 Redundancy – In order to minimize the risk of temporary or permanent failure of access to the documents certain measures must be put in place. There must be adequate redundancy or duplication of both the derivative or use files and the dark archives.

a. *Service archive* – The service archive is the site for management of the derivative or use files as well as the software, systems, and hardware necessary to transmit and make them accessible for public display and use. Critical service archive activities, such as hosting, storage and data management, must take place in a facility or facilities located two miles or more from a strategically important potential military or terrorist target (such as a major power plant, dam, agency-level federal government building, reservoir, chemical or munitions manufactory, or military installation) and more than twenty miles from a nuclear power plant.

The facility or facilities should be located in an inherently stable area, and sited a minimum of five feet above and 100 feet from any 100-year flood plain areas, or be protected by an appropriate flood wall that conforms to local or regional building codes. The facility should be located in an inherently stable locale and be reinforced and seismically sound.

b. Secondary service archive (SSA) – There must be a "mirror repository" of the service copy to provide instantaneous and continuous access to all designated constituents when the access copy or service archive is disabled. The SSA must be revised continually to reflect improvements, enhancements, and additions to the service archive.

c. Dark repository –A second "copy" or instance of all master and derivative digital files, data, and underlying enabling code must reside in at least one "dark" repository that is inaccessible to the general public and is under the control of the managing organization or its proxy. The conditions under which the dark repository contents might be accessed or used, and the terms of access to it (including terms under which the data is managed), must be stipulated in a legally binding and publicly accessible written agreement. Dark repository activities must take place in a facility or facilities located two miles or more from a strategically important potential military or terrorist target (such as a major power plant, dam, agency-level federal government building, reservoir, chemical or munitions manufactory, or military installation) and more than twenty miles from a nuclear power plant.

d. S*econdary dark repository (digital) -- At least one* "copy" or additional instance of the dark repository must be maintained separately under the control of the managing organization or its proxy, as assurance against the failure or loss of the original dark repository. The secondary dark repository must provide comprehensive redundancy of content to the original dark repository, and the systems and resources necessary to support access to and management of that content must be both redundant and fully independent of those supporting access to the original dark repository.

3.4 Indemnification of digital assets – When maintained by a party other than the Superintendent of Documents/GPO all digital objects, software, systems, and hardware necessary to manage and preserve the master and derivative digital files must be indemnified by a credible underwriter to meet the full actual cost of replacement.