

Selected Guidelines to Digitizing Text Collections

Digitized texts are generally delivered to users as navigable, properly sequenced multi-page objects. When moving from source “text” (microfilm, machine-printed text, handwritten manuscripts) to target digital object, determine how many of the following product types you will need to produce:

- facsimile page images (b/w, grayscale, and/or color),
- page images, with hidden text (often uncorrected OCR) for indexing and searching,
- tagged and structured machine readable text (with or without page images) for indexing and display.

Guides to Planning Text Digitization Projects and Workflows

In keeping with the [NISO Framework of Guidance for Building Good Digital Collections](#) the following publications consider audience and use requirements first, as well as attributes of source materials, to define specifications for successful digitization.

Chapman, Stephen. “Levels of Service for Text Digitization, Techniques for Creating Sustainable Digital Collections,” [ALA TechSource](#), Library Technology Reports, Vol. 40, No. 5, September/October 2004.

Overview of baseline, low-, medium-, and high-effort strategies associated with creating digital objects of varying formats, quality, and functionality. Discussion of tools and workflows to produce page images, text, structural and administrative metadata, and object packaging (e.g., for transfer and deposit to digital repository).

Morrison, Alan, Michael Popham and Karen Wikander. [Creating and Documenting Electronic Texts: A Guide to Good Practice](#), AHDS Guides to Good Practice. Oxford: Arts and Humanities Data Service, 2000.

Drawn from authors' understanding of what constitutes good practice from almost twenty-five years of experience running the [Oxford Text Archive](#), this Guide presents a comprehensive and highly readable overview of procedures for planning, document analysis, digitization and markup, metadata and quality control.

[NINCH Guide] Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and NINCH. “[Digitization and Encoding of Text](#),” *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*, National Initiative for a Networked Cultural Heritage, Washington, DC, November 2002.

Good high-level overview of options to create fully searchable text (full-text) resources. Includes definitions of markup and encoding; advantages and disadvantages of SGML and XML, and definitions for the many acronyms that must be mastered to manage text conversion workflows: ASCII, CIMI, EAD, OCR, METS, and TEI.

Selected Technical Documents

Buckley, Robert and Roger Sam, "[JPEG 2000 Profile for the National Digital Newspaper Program](#)," April 27, 2006.

Specifications for JPEG2000 production masters for Phase 1 of the National Digital Newspaper Program. Includes full details for "profile" covering the codestream, image coding parameters, and metadata. Also includes section on tools used in the development of the profile.

University of Michigan Digital Library Production Service, "[Digital Conversion Services: Text Encoding](#)."

Overview of DLPS production practices, with links to products and projects.

Structural Metadata – METS

McDonough, Jerome, Merrilee Proffitt and MacKenzie Smith. "[Structural, technical, and administrative metadata standards. A discussion document](#)," Digital Library Federations, December 2000.

Planning document that framed development of METS at a DLF-sponsored workshop convened in February, 2001.

Metadata Encoding & Transmission Standard (METS), [METS Implementation Registry](#).

Includes descriptions of projects using METS for management and delivery of page-turned objects.

UKOLN. "[Metadata Encoding and Transmission Standard \(METS\)](#)," UKOLN Metadata Resources, last updated 26-Oct-2001.

Overview of METS schema and purposes it serves.