DIGITAL LIBRARY
FEDERATION

Benchmark for Faithful Digital Reproductions of Monographs and Serials

Version 1

December 2002

The Digital Library Federation Benchmark Working Group (2001-2002)

Contents

1. Introduction
2. What is a Faithful Digital Reproduction?
3. Benchmarks for Masters of Page Images and Machine-Readable Text
4. Benchmark Functions: Metadata Requirements and Recommendations
5. Notes

http://www.diglib.org/standards/bmarkfin.htm | http://purl.oclc.org/DLF/digrepo1

1. Introduction

This document defines a minimum benchmark for digital reproductions of printed monographs and serials. The case for such a benchmark is made in an article by Greenstein and George that is available in RLG's *DigiNews* -- http://www.rlg.org/preserv/diginews/featured/

The benchmark grew out of DLF's investigation into the need for and functional specification of a registry of information about the monographs and serials that have been digitally reformatted (see http://www.diglib.org/collections/reg/regpapfunc.htm). Functional requirements for a proposed registry were produced as part of the DLF investigation. The requirements state the importance of ensuring that registry records for digital reproductions include "a description or a pointer…to a description of the technical standards used in creating the Master Copy."[1]

Although the registry is not exclusive (it will record information about materials that are born digital as well as digital reproductions, and about masters that meet agreed benchmarks as well as those that do not), it provides an important opportunity to identify and build consensus around minimum characteristics that might be expected of certain kinds of digital objects.

This benchmark has been prepared and endorsed by the DLF to document the minimum characteristics of digital reproductions — regardless of whether or not they are registered in the DLF or other registries — required to ensure usability, persistence and interoperability. One important objective is to define baseline levels of quality that would minimize or eliminate the need to digitize a work more than once. A *Report* on the initial discussion leading to this document is available from DLF's website -- http://www.diglib.org/standards/presreformatsum.htm

Companion documents may be developed defining benchmarks for other digital reproductions — for example, those that may apply to born digital monographs and serial publications, to manuscript items, or to encoded text reproductions of historic materials.

2. What is a Faithful Digital Reproduction?

Faithful digital reproductions are digital objects that are optimally formatted and described with a view to their *quality* (functionality and use value), *persistence* (long-term access), and *interoperability* (e.g. across platforms and software environments). Faithful reproductions meet these criteria, and are intended to accurately render the underlying source document, with respect to its completeness, appearance of original pages (including tonality and color), and correct (that is, original) sequence of pages. Faithful digital reproductions will support production of legible printed facsimiles when produced in the same size as the originals (that is, 1:1).

In practice, digitizing might yield multiple versions of the digital reproductions:

- **masters**: optimized for longevity and for production of a range of delivery versions (e.g., for screen, for print)
- **deliverables**: optimized to meet defined use requirements

This benchmark defines minimum characteristics for both versions. ***Section 3*** pertains to masters of page images and machine-readable text. ***Section 4*** pertains to functional requirements for delivery that must be supported by structural metadata.

3. Benchmarks for masters of page images and machine-readable text

To meet functional requirements stated above, faithful digital reproductions must include page images of a quality sufficient to produce printed facsimiles.

High-resolution page image masters will meet or exceed the benchmarks presented in the table below. In cases where multiple masters are produced — e.g., an RGB, "archival master," and a CMYK "print master"— at least one version must meet or exceed the benchmark.

This benchmark acknowledges that what ultimately constitutes legibility and fidelity is a subjective decision. In part for this reason, the benchmark refers minimally to file formats and compression, and does not prescribe minimum tone reproduction requirements for non-textual components (e.g., illustrations and covers). It also does not provide production-level guidance, for example on how to deal with missing pages, to "clean up" foxing or blemishes, or to select an appropriate dpi for fonts or source pages of different sizes. Such guidance is available elsewhere or will evolve through experience and may be attached as companion documentation to this benchmark.

Minimum Benchmarks for Page Image Masters		
Black and white For text, and may also be used for line drawings, de-screened halftones.	**Grayscale** For covers and illustrations printed in black and white. Recommended, but not required.	**Color** For covers, and meaningful text or illustrations printed in color. Recommended, but not required.
600 dpi, 1-bit or bitonal TIFF images [2] Images must be sized and saved at 1:1 scale to the dimensions of the original page. Images must be saved uncompressed or with lossless compression. Where images are compressed they must be made available in the Group 4 (ITU-T6) format. The images may be interpolated from 400 optical dpi 8-bit images.	**300 dpi, 8-bit grayscale uncompressed TIFF, or lossless compressed image (e.g. LZW, JPEG2000).** Images must be sized and saved at 1:1 scale to the dimensions of the original page. The dpi specification will relate directly to the font-size and page dimensions of the original source document, and to local definitions of legibility and fidelity. In many cases, 400 dpi will be preferred. Where larger pages are concerned, the lower dpi specification may be required.	**300 dpi, 24-bit color uncompressed TIFF, or lossless compressed images (e.g. LZW, JPEG2000).** Images must be sized and saved at 1:1 scale to the dimensions of the original page. RGB and YCC are the recommended color spaces for masters, particularly when only one master version is produced. The dpi specification will relate directly to the font-size and page dimensions of the original source document, and to local definitions of legibility and fidelity. It may also relate to the perceived artifactual value of the source object or the extent to which its physical characteristics such as foxing, etc., are perceived of as conveying some important information or meaning.

In addition to page images, faithful digital reproductions may also include machine-readable (keyboard or OCR) text. That text may be corrected or uncorrected. If it is corrected to a uniform minimum level, the accuracy level will be specified (e.g. as 99.995%). Such text may be encoded (at any level, e.g. as specified in *TEI Text Encoding in Libraries. Guidelines for Best Encoding Practices. Version 1.0, July 30, 1999*: http://www.diglib.org/standards/tei.htm

4. Benchmark Functions: Metadata Requirements and Recommendations

While the characteristics above are meant to apply to digital masters, the functional requirements below are somewhat different. In order to keep the master viable over time and create new delivery copies as necessary, the metadata needed to meet the functional requirements below must be collected. However, systems may not exist to perform those functions relative to the master copy. The functional requirements are likely to be met, in terms of usable systems, with the delivery copy.

Faithful digital reproductions of monographs and serials must have descriptive, structural and administrative metadata, and the metadata must be made available in well-documented formats. Sufficient metadata must be created to support a number of essential functions, listed in sections A, B, and C below.

These functions will be accomplished through the production of metadata with appropriate richness. No recommendations are made with respect to production practices except for sufficient quality control at least to ensure that benchmark specifications are met.

No recommendations are made with respect to the form the metadata should take or how it should be encoded. It is expected that in order to enable interoperability, metadata and its representation will conform to emerging standards and good practices.

A. Functions required of all digital masters

The following functions are required of all digital masters:

It will be possible to produce, in print or as an online (on-screen) display, a faithful, citable rendering of the physical source including the sequencing of its component parts (pages, volumes, etc.).

It will be possible to navigate sequentially through the physical components (go to next, previous, first, last, or nth sequential page image).

The relationship between component parts of the physical source (pages, volumes, etc.) will be represented.

Images of blank pages (including backs-of-plates) will be included as sequenced components.

It will be possible to associate higher-level descriptive metadata with digital component parts of the object (e.g. for the purposes of citation).

B. Functions required where applicable

The following requirements are distinguished from those cited above (4A) because they cannot be met by all digital masters. For example, pagination can only be faithfully supplied where pages are enumerated in the physical source. Placeholders for missing pages can only be reliably supplied for pages that are known to be missing.

Where possible, masters will support navigation to, between, and among logical structures (e.g. chapters for monographs; volumes, parts, and issues for serials) and significant features (e.g. tables, illustrations, blank pages). Citation of those features will also be supported.

Where applicable and in a manner appropriate for the physical object in question, any enumeration found on pages of the physical object will be represented. Representation will maintain all variations in the enumeration of the physical object's component parts (signature pages, preface, etc.)

Placeholders for known missing pages will be included as sequenced components. In the interest of creating complete digital masters, missing pages and other components should be identified as such in higher-level metadata. Where page images are supplied by third parties, information to that effect should be noted in descriptive metadata.

C. Functions strongly preferred

The following functions are useful and recommended, but not required.

High-level logical structures will be identified (e.g. for the purpose of rendering and navigation).

- For monographs, logical structures may include title pages, tables of contents, lists of illustrations, indexes, chapters, etc.
- For serials, logical structures may include volumes, parts, issues, articles, etc.
- Significant features such as tables, illustrations, blank, missing and supplied pages, maps, etc. will be identified (e.g. for the purpose of rendering and navigation).

For the purposes of citation, etc., it will be possible to support association of higher-level metadata with enumerated pages, logical structures, and features as identified.

Representing page rectos and versos for the purpose of printing faithful codices.

Notes

The Benchmark Working Group (2001-2002) included: Daniel Greenstein (DLF); Anne Kenney (Cornell); John Price Wilkin (University of Michigan); Ron Murray (Library of Congress); Robin Dale (RLG); Eileen Fenton (JSTOR); Carla Montori (University of Michigan) Judith Thomas (University of Virginia); Chris Ruotolo (University of Virginia) Sherry Byrne (University of Chicago); Janet Gertz (Columbia University) Stephen Chapman (Harvard University); Daniel McShane (University of Virginia); David Ruddy (Cornell University); Robin Wendler (Harvard University). Sections 1-3 prepared on July 30, 2001 (rev. December 6, 2002); Sections 4-6 prepared on March 26, 2002 (rev. December 6, 2002). Revisions and publication, December 2002, by Dale Flecker (Harvard University) and David Seaman (DLF).

1. Dale Flecker, "Registry of Digital Reproductions of Paper-based Monographs and Serials: Functional Requirements," DLF, December 2001, http://www.diglib.org/collections/reg/regpapfunc.htm.

2. 600 dpi will capture roman scripts down to 6-point type with the microfilm QI equivalent of 8. Smaller text, scripts with fine lines and small dots and other diacritics (like italics, Arabic, etc.) need higher resolution to be captured completely.