

**Technical Capabilities of the Digital Research Library (DRL)
University Library System, University of Pittsburgh
15 December 2004
Rev. 10 July 2006**

The following will briefly describe the DRL's capabilities to create, index and deliver digital library collections on behalf of the University Library System at the University of Pittsburgh. Three Web servers host DRL content, including the streaming AV material. The DRL has created guidelines and documentation for many of its projects and activities (<http://digital.library.pitt.edu/documentation/>).

Digitization of Source Material

In-House

The DRL maintains equipment and in-house staff to perform several kinds of document imaging. Generally speaking, the DRL can scan photographic prints or graphic material (e.g., manuscripts, music scores, postcards, etc.) up to a size of 11 x 17 inches as well as photographic negatives and 35mm slides. Double-sided loose texts can be scanned using a document-feeding scanner. The DRL can also scan bound books and flat art work (including loose posters, maps, etc.) up to 24.5 x 33 inches.

Outsourcing

The DRL has experience creating RFPs and working with service vendors to manage the digitization of high-volume book scanning (including bound volumes), as well as large-format graphics (e.g. maps, broadsheets), and audio-visual material.

Metadata Services

The DRL has some capability to capture *structural* and *administrative* metadata for digital objects, but collection of *descriptive* metadata is usually performed by content holders or through an arrangement with the University Library System's Technical Services department.

Metadata Scheme Consulting

The DRL maintains an up-to-date knowledge of metadata schemes used by the digital library community. The department can assist partners with selection of the most appropriate metadata scheme to use for a particular project. The department can also assist in creating tools for metadata gathering, such as spreadsheet or database templates.

Data Transformation and Automated Markup

Metadata must often be transformed from one form to another in order to create rich digital objects for online display. The DRL has some capability to transform data formats, and to generate XML markup from pre-collected metadata.

Open Archives Initiative

Metadata from collections hosted by the DRL can be shared using the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH uses the Dublin Core metadata scheme as its baseline record format; the DRL can assist in the mapping process if a different metadata scheme is in use.

Indexing Services

Optical Character Recognition (OCR)

The DRL implements very powerful text recognition software known as PrimeRecognition. It combines the processing efforts of five OCR products to produce highly accurate results. The DRL also utilizes software to recognize only selected portions of a text and to verify the resulting text.

XPAT

The DRL licenses an SGML-aware search engine (XPAT) from the University of Michigan, which supports the indexing and searching of full-text, finding aids, and bibliographic collections.

Online Access to Digital Library Collections

Overview: Class-Based DLXS Middleware

The DRL primarily uses the open source Digital Library eXtension Service (DLXS) middleware, developed and distributed by the University of Michigan's Digital Library Production Service, to deliver digital library collections. The DLXS middleware is organized around four *classes* of material: text, images, finding aids, and bibliographies. Accordingly, the DRL attempts to ensure that its hosted digital library collections fall into one or more of these categories.

Within each class, items must belong to a specific collection; however, collections within the same class can be searched together. For example, all image collections within Image-Class can be searched together at once (e.g., Historic Pittsburgh Image Collections).

Text-Class

Text-Class collections are generally comprised of monographs or serials, although in some circumstances archival records may make appropriate Text-Class collections. At minimum, Text-Class requires a scanned image of every page in an item, and a bibliographic record for the item. However, much of the strength of the middleware depends on a full-text index, which can be created by either extracting text via optical character recognition (OCR), or by otherwise transcribing or capturing the printed text. Text-Class can also accommodate structural metadata, such as chapter or article information. Text-Class collections are made up of XML-encoded texts with corresponding page images (usually bi-tonal TIFF image files).

Examples:

- Historic Pittsburgh Full-Text Collection (<http://digital.library.pitt.edu/fulltext/>)
- 19th Century Schoolbooks Collection (<http://digital.library.pitt.edu/nietz/fulltext/index.html>)
- Publications of the Allegheny Observatory (<http://digital.library.pitt.edu/parallax/>)

Image-Class

Image-Class collections are comprised of photographs or graphic images with accompanying descriptive records. The images are frequently high-resolution grayscale or color images, converted into the JPEG2000 format to allow for custom views online. Descriptive records are searchable, and stored within a simple flat-fielded database format. Image-Class records are not appropriate for representing relational or hierarchically-structured data. The DRL implements a common baseline metadata framework for image metadata: Dublin Core Metadata Initiative.

Examples:

- Historic Pittsburgh Image Collections (<http://images.library.pitt.edu/pghphotos>)
- George Washington Manuscripts (<http://images.library.pitt.edu/g/gwletters>)
- Jack B. Yeats Broadsheets (<http://images.library.pitt.edu/y/yeats>)
- Visuals for Foreign Language Instruction (<http://images.library.pitt.edu/v/visuals>)

Finding Aid-Class

Finding Aid-Class provides for indexing and display of archival finding aids encoded with the Encoded Archival Description (EAD) application of XML.

Examples:

- Historic Pittsburgh Findings Aids (<http://digital.library.pitt.edu/ead>)
- Archives Service Center Finding Aids (<http://digital.library.pitt.edu/cgi-bin/f/findaid/findaid-idx?page=index&c=ascead>)

Bib-Class

Bib-Class collections consist of bibliographic records, often derived from MARC records. However, any flat, fielded descriptive records could be handled by a Bib-Class collection. Bib-Class also forms the basis for DLXS's capabilities as an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) data provider.

Examples:

- Historical Society of Western Pennsylvania's Library & Archives Catalog (<http://digital.library.pitt.edu/hswp>)
- Nietz Old Textbook Collection catalog (<http://digital.library.pitt.edu/nietz/biblio/search.html>)