Digital Libraries

Lecture 14
CS 410/510
Information Retrieval on the Internet

Digital Library

"*A collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization and maintenance of the collection*."

– Akscyn and Witten, quoted in: Witten et al. Greenstone: *A Comprehensive Open-Source Digital Library Software System*. Proceedings of the fifth ACM conference on Digital libraries, San Antonio, TX, pp. 113 –121, 2000.

For more definitions, see: http://ei.cs.vt.edu/~fox/dlib/def.html

Digital Library

- How is a digital library different from a traditional library?
  - Is a digital library the digitization of a library?

Digital Library

- How is a digital library different from the Web?

Outline

- Digital Libraries: what and why
- 5S model
- Document representations
- Finding documents
- Architectures, frameworks, & protocols
- Citation analysis
- Examples of digital libraries

5S Model of Digital Libraries

- Five abstractions to characterize and formalize foundations of DLs
  - Streams
  - Structures
  - Spaces
  - Scenarios
  - Societies

Gonçalves et al. *Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries*. ACM Transactions on Information Systems, Vol 22, April 2004, pp 270-312, 2004

5S Model of Digital Libraries

- Streams (sequences)
  - Content (text and multimedia)
  - Protocols
  - Logs
- Structures: organizing constructs
  - Data structures
  - Databases
  - Inverted files
  - Metadata records

5S Model of Digital Libraries

- Spaces
  - Dimensions – 1D, 2D, 3D
  - Vector spaces, probability spaces, concept spaces
- Scenarios
  - Processes, transformations, services
  - Usage scenarios, HCI specifications
- Societies
  - Users, librarians, annotators

Minimal Digital Library

- Minimal set of components to be a DL
  - A repository
    - Family of collections and associated functions to get, store, and delete digital objects
  - A set of metadata catalogs for all the collections in the repository
  - A set of services, including at least indexing, searching and browsing
  - A society

Document Representations

- Digital objects
  - Identifier
  - Metadata
  - Content
- Representations

Identifiers

- Unique object identifier – within a DL
- Unique identifier independent of location
  - URN/URI
  - Digital Object Identifier (DOI)
    - Designed for managing intellectual property
    - Implements the Handle system for resolving persistent identifier to current location

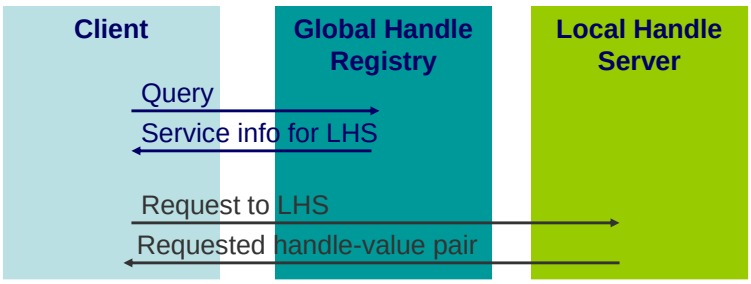
The Handle System®

- The Handle System (http://www.handle.net)
  - Protocols, namespace and infrastructure
  - Associates a persistent name (handle) for a digital resource with information needed to locate and access that resource
  - Developed, managed by Corporation for National Research Initiatives (CNRI)
- Syntax

  <Handle> ::= <Prefix> "/" <Handle Local Name>
  - Prefix is naming authority; guaranteed globally unique
  - Suffix is local name

The Handle System®

- Two level hierarchical model
  - Global Handle Registry
  - Local Handle Services
    - may have any number of handle servers



Digital Object Identifier (DOI)

- DOI (http://www.doi.org)
  - Uses The Handle System for name resolution
- DOI adds:
  - Numbering syntax
  - Organization for assuring persistence
  - Consistent rules for naming and numbering and an API for interoperability
  - Data model with data dictionary to facilitate mapping between metadata schemes

Digital Object Identifier (DOI)

- Can be assigned to various manifestations and abstractions of intellectual property
  - A song
  - A performance
  - A recording
  - Score
  - Lyrics
- Can apply at any level of granularity
  - A journal, an issue, an article, a table within an article

Metadata: example standards

- Machine Readable Cataloging (MARC) records
  - Detailed bibliographic information
  - Designed to catalog physical library holdings
- Dublin Core Metadata
  - International standard for metadata elements for describing resources on the Web
  - Fifteen “core” elements

Dublin Core Metadata

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type

Dublin Core with Qualifiers

Element	Qualifier	Scope Note
contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors
contributor	advisor	Use primarily for thesis advisor
contributor	author	
contributor	editor	
contributor	illustrator	
contributor	other	
coverage	spatial	Spatial characteristics of content
coverage	temporal	Temporal characteristics of content
creator		Do not use; for harvested metadata only
date		Use qualified form if possible
date	accessioned	Date DSpace takes possession of item
date	available	Date or date range item became available to the public
date	copyright	Date of copyright
date	created	Date of creation or manufacture of intellectual content if different from date.issued
date	issued	Date of publication or distribution.
date	submitted	Recommend for theses and dissertations
identifier		Catch-all for unambiguous identifiers not defined by qualified form; use identifier.other for a known identifier common to a local collection instead of unqualified form
identifier	citation	Bibliographic citation for works that have been published as a part of a larger work, e.g. journal articles, book chapters
identifier	govdoc	Government document number
identifier	isbn	International Standard Book Number

from http://dspace.org/technology/metadata.html (this only shows part of the table)

Document Representation

- Metadata
- Structure
  - Chapter, section, paragraph
- Presentation
  - Layout, appearance
- Format (document type)
  - HTML, PDF, Multimedia formats, etc.
- Content

Structure and presentation

- Standard Generalized Markup Language (SGML)
  - Metalanguage for defining markup languages
  - Used for publishing and sharing machine-readable documents
  - Provides syntax for tagging document structure and metadata
  - Handles both layout and structure, but they are separated
  - XML is a much-simplified subset of SGML

Structure and presentation

- Text Encoding Initiative (TEI)
  - Subset of SGML tags and modular DTD
  - Defines a standard for encoding texts used by researchers, especially in humanities (application independent)
  - Defines several hundred elements
  - TEI-Lite is a basic “core” subset (~130 elements)

Plain text example

CHAPTER 38

READER, I married him. A quiet wedding we had: he and I, the parson and clerk, were alone present. When we got back from church, I went into the kitchen of the manor-house, where Mary was cooking the dinner, and John cleaning the knives, and I said –
'Mary, I have been married to Mr Rochester this morning.' The housekeeper and her husband were of that decent, phlegmatic order of people, to whom one may at any time safely communicate a remarkable piece of news without incurring the danger of having one's ears pierced by some shrill ejaculation and subsequently stunned by a torrent of wordy wonderment. ...

From: http://www.tei-c.org/Lite/U5-eg.html

TEI-Lite example

```
<pb n='474'/>
<div1 type="chapter" n='38'>

<p>Reader, I married him. A quiet
wedding we had: he and I, the parson
and clerk, were alone present. When we
got back from church, I went into the
kitchen of the manor-house, where Mary
was cooking the dinner, and John
cleaning the knives, and I said
&mdash;</p>
<p><q>Mary, I have been married to Mr
Rochester this morning.</q> The
housekeeper and her husband were of
that decent, phlegmatic order of people,
to whom one may at any time safely
communicate a remarkable piece of
news without incurring the danger of
having one's ears pierced by some shrill
ejaculation and subsequently stunned by
a torrent of wordy wonderment. ... </p>
```

Finding Documents

- Searching traditional libraries
  - Card catalog or Online Public Access Catalog (OPAC): search on
    - Title
    - Author
    - Subject
  - Use classification code to find shelf location
    - Dewey Decimal
    - Library of Congress

Finding Documents

- Searching digital libraries
  - Metadata
  - Full text
- Searching the Web
  - Full text + page rank? + anchor text? + ???
  - Limit search by characteristics
    - URL
    - Links
    - Site

What about distributed collections?

- Distribution
  - Who has what?
  - How do you get it?
  - How do you merge results from multiple sources?
- Scaling
  - How do you do this efficiently?
- Basic approaches
  - Distributed search – metasearch on request
  - Search a repository of metadata already harvested from distributed sources

Z39.50

- ANSI/NISO Z39.50 – standard for interoperable searching of library catalogues
  - Establishing, managing a connection
  - Communicating a query
  - Returning results
  - Facility to find out what databases are available and what access points are provided (EXPLAIN)
- Supports distributed searching in client/server environment

OAI-PMH

- Open Archives Initiative Protocol for Metadata Harvesting
  http://www.openarchives.org//OAI/2.0/openarchivesprotocol.htm
- Interoperability framework
- Two classes of participants
  - Data providers: expose metadata using OAI-PMH
  - Service providers: use harvested metadata to provide value-added services

OAI-PMH

- Harvester: application that issues OAI-PMH requests to collect metadata from repositories
  - Requests expressed as HTTP requests
  - Results returned as HTTP responses
    - Must be XML documents that validate against a particular XML Schema
    - Contain metadata records

D-Space

- Open source DL system http://www.dspace.org/
  - Repository for multiple types of digital objects
    - Designed for university research output
  - Allows easy submission of documents which are then stored and indexed
  - Supports creation, indexing, searching of metadata
    - Uses a qualified version of Dublin core
  - Supports OAI-PMH as a data provider

D-Space

- A D-Space service is composed of multiple communities (departments, institutions)
  - A community may have multiple collections
- Provides digital preservation services
  - Bit preservation
    - Exact bit-level copy across changes in physical storage media
  - Functional preservation
    - Preserves ability to use material in the same way as formats evolve (via migration or emulation)
    - Not all formats supported

Greenstone

- Software to create and manage a digital library
  - Open source: http://www.greenstone.org
- Supports OAI-PMH as data provider or harvester
- Functions include:
  - Build collections
    - Plug-ins for document types (PDF, HTML, email, etc)
  - Search and browse
    - Full text search at multiple levels of granularity
      - e.g. title, section heading, section
    - Browse based on various metadata types
  - Administer library
    - Apply access controls
    - Log user activity
    - Update collections

Fedora

- Open source software for digital object repository http://www.fedora.info/
- Runs as a service within a web server
- Supports delivery of dynamic content via web services
- Supports OAI-PMH

Fedora

- Object model:
  - A digital object aggregates metadata with one or more versions (formats) of a document, or multiple text and multimedia objects that comprise a document
  - Access representations via different URLs
- Underlying components are data streams
  - URL can be local surrogate for remote content
  - URL can access virtual representations, rendered by associating a data stream with a service operation (e.g. converting from one image format to another)
- Metadata can express relationships (e.g. collection membership or article *in* journal)

E-Prints

- Software to facilitate self-archiving of scholarly work http://www.eprints.org
- Institution runs an E-print Archive that is publicly available through a web server
- Depositor fills in metadata fields, attaches full-text document
  - Can include preprints, postprints, links to publisher's proprietary version
- Archive is OAI-PMH compliant
- Facilitates open access to research output

Citation analysis

- CiteSeer
  - http://citeseer.ist.psu.edu/citeseer.html
- Google scholar
  - http://scholar.google.com

CiteSeer

- Digital library and search engine
  - For scientific literature
  - Primarily computer and information science
  - http://citeseer.ist.psu.edu/citeseer.html
- Automated citation indexing and linking
  - Displays citation statistics, context of citation
- Uses web crawling and submission to find papers
  - Extracts metadata
  - Provides links to full-text, when available

Some Digital Library Examples

Will look at examples in class

- ACM Digital Library
  - http://www.acm.org/dl
- Networked Digital Library of Theses and Dissertations (NDLTD)
  - http://www.ndltd.org
- Networked Computer Science Technical Reference Library (NCSTRL)
  - http://www.ncstrl.org
- National Science Digital Library (NSDL)
  - http://nsdl.org

- Next: Making Information Findable