Introduction to Digital Libraries

hussein suleman
uct cs honours 2004

Course Structure

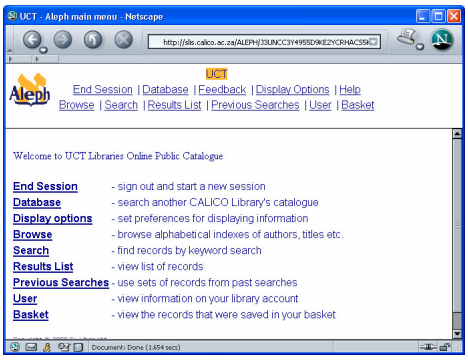
- 23+1 lectures
  - intro
  - 18 dl topics - lectures
  - 4 dl issues – discussions
- 3 programming assignments
  - DL package installation / configuration
  - OAI-based DL service
  - DL construction from components
- take-home final

Course Topics

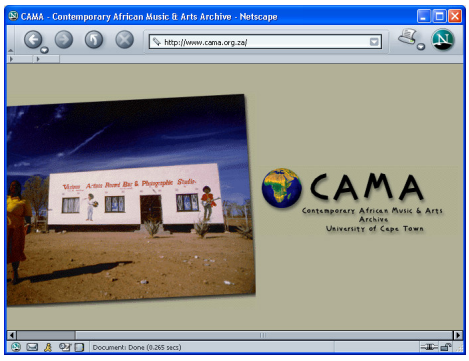
- definitions and examples
- data/service model
- overview of XML
- metadata, DC
- repositories, searching, software
- OAI, interoperability, architecture
- interfaces, portals
- IP/DRM, preservation, access, economics

Definitions and Examples

Example 1/15



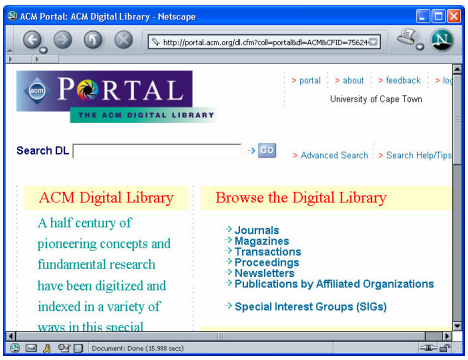
Example 2/15



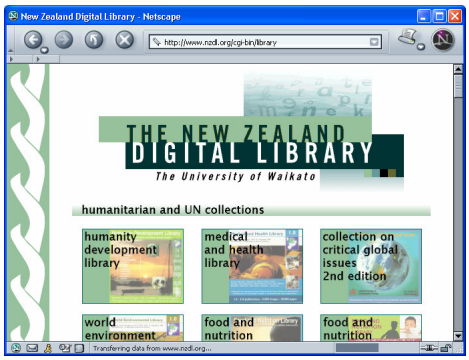
Example 3/15

Example 4/15

Example 5/15



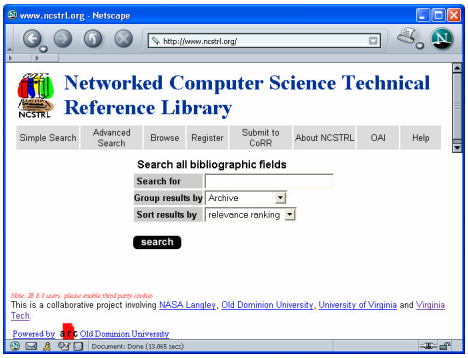
Example 6/15



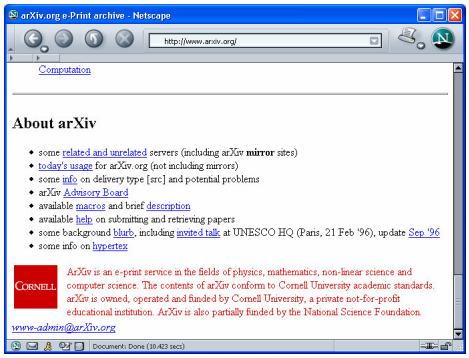
Example 7/15

Example 8/15

Example 9/15



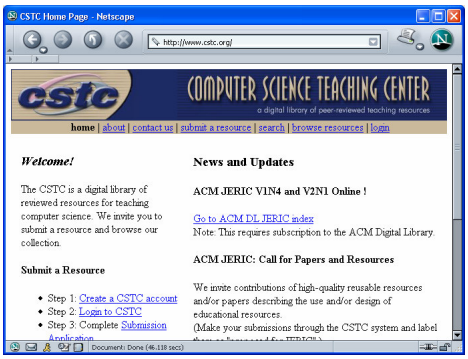
Example 10/15



Example 11/15

Example 12/15

Example 13/15



Example 14/15

Example 15/15

Definition 1/5

- "Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities."
  - Greenstein, Dan (2000) DLF Draft strategy and business plan. Available http://www.diglib.org/about/strategic.htm

Definition 2/5

- "Digital libraries are complex data/information/knowlege (hereafter information) systems that help: satisfy the information needs of users (societies), provide information services (scenarios), organize information in usable ways (structures), manage the location of information (spaces), and communicate information with users and their agents (streams)."
  - Fox, Edward A. (1999), DL Self-Study: definitions. Available http://ei.cs.vt.edu/~dlib/def.htm

Definition 3/5

- "Systems providing a community of users with coherent access to a large, organized repository of information and knowledge."
  - Lynch, Clifford and Hector Garcia-Molina (1995), "Interoperability, scaling, and the digital libraries research agenda: A report on the May 18-19, 1995 IITA Digital Libraries Workshop". Available http://diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html#2
- "The virtual or digital library is not an oxymoron- it is redundant. ... Since we did not bother to qualify our libraries by calling them clay libraries or papyrus scroll libraries, why now do we have to call them digital libraries?"
  - Braude, Robert and Samuel J. Wood (1999), "Virtual or actual: The term library is enough", Bulletin of the Medical Library Association, p. 87.

Definition 4/5

- A digital library is "a world of literature, history, photographs, movies and maps open, free of charge, to any curious mind that wants to meander through the electronic equivalent of library stacks."
  - Lipkin, Richard (1995), "The library that isn't there: Digital libraries transform books, photos, and videos into bits and bytes", Science News, Vol. 147, No. 22, pp. 344-346.

Definition 5/5

- "a focused 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."
  - Witten, Ian and David Bainbridge (2002), How to Build a Digital Library, Morgan Kaufman, p. 6.

So what is a Digital Library ?

- collections of digital objects
- palette of services
  - storage and preservation
  - access and use
- users
  - information seekers
  - information producers
  - information managers
- systems
  - network- and storage-based computer systems

Is the WWW a digital library ?

Variety of Perspectives

- Computer Science
  - technical issues
  - preference for automatic solutions e.g. Google
- Library Science
  - policies and organisational issues
  - preference for human-mediated solutions, e.g. library cataloguing
- Information Science
  - philosophical issues ?
- Physics, Chemistry, Medicine, Economics, etc.
  - practical issues – how can we leverage digital libraries to solve our information management problems?

CSC400 Perspectives

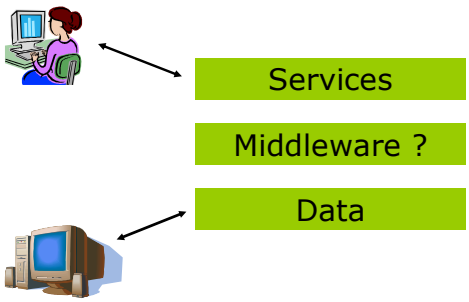
- Digital Libraries are a conceptual space within which we define technology and policy to organise information effectively to address the needs of users.
- Digital Libraries provide a framework within which to devise advanced mechanisms for information management on the WWW and beyond.

The Data and Services Model

DL as computer system

- Software package(s) to manage data (aka digital objects) and provide access to users, either locally or through a Web-based service.
- Software is used to
  - provide services to users
  - mediate data between layers
  - manage data storage and access
- Layers are not necessarily distinct!

The 3-tier DL model



Digital Object Types

Type	Example
Text	
Hypertext	
Image	
Video	
Audio	
3D Model	
Interactive Visualisation	
Software	

Common Features

- Can be created/destroyed
- Can be serialised and stored/retrieved electronically
- Can be transferred from one system to another (e.g.)
- Can be described
- Can be linked to (e.g.)

Examples of Services

- Google search
- Yahoo! directory
- Mailing lists
- Kalahari.net
- UCT Library catalogue
- -
- -
- -
- -

User management

- Authentication
  - Check users are who they claim to be.
- Authorisation
  - Check users are allowed to perform the tasks they are attempting.
- Maintain user information/profiles.

Searching

- Searching focuses on automatic/manual algorithms for indexing and querying.
- Indexing:
  - Transformation of information to support efficient discovery/retrieval.
- Querying:
  - Accessing transformed data to obtain results sorted in order of relevance, date, etc.
- a.k.a. Information Retrieval (IR)
- a.k.a. free-text databases
- Good example: Google
- Bad example: UCT website

Browsing

- Access subsets of data by categorical classification.
- Manual or automatic classification
- Single or multiple category membership
- Linear or hierarchical
- Is Searching = Browsing? Can searching be used as a surrogate for browsing?
- Example: Open Directory Project

Submission

- Add new digital objects to a DL.
- Content
  - digital objects
  - descriptions of objects
- Explicit submission vs. Harvesting vs. Crawling
  - Explicit submission = submission by local users
  - Harvesting = obtaining material from external sources
  - Crawling = finding material by automatically sifting through public collections e.g., WWW

Review

- Check submissions for appropriateness, quality, completeness, correctness, etc.
- Modes of review
  - Editorial review
  - Peer review
  - User review
- DL must support workflow for review processes.
- Security/Privacy issues must be addressed.
- Example: Online conference management

Annotation

- Add commentary or associated information to a digital object.
- Generalisation for reviews, ratings, discussions.
- May be stored as part of object or as separate objects.
- Link to objects and other annotations must be well-defined.
- Example: User feedback in online stores

Recommendation

- Suggest possibly relevant items based on past behaviour.
- Individual- or group-based recommendation
- a.k.a. Collaborative filtering (for groups)
- a.k.a. Selective Dissemination of Information (SDI) (for automatic push-services)
- Example: Amazon.com's recommended items

Middleware

- Databases and collections of data with standard/shared data formats
  - traditional approach
- APIs to access data, enabling use of different databases
  - current production systems
- Protocols to access data/services, enabling component-wise development of systems
  - current experimental systems

Why use 3-tier architecture ?