Digital Libraries: Introduction

RAFFAELLA BERNARDI

UNIVERSITÀ DI TRENTO

P.ZZA VENEZIA, ROOM: 2.05, E-MAIL: BERNARDI@DISI.UNITN.IT

Contents

1. Course Structure

Programme First part about (Digital) Libraries; second part on Language Technologies and their application to DL; third part about new trends: seminars.

Text Books Michael Lesk 2004; Ian Witten et al. 2003 (on Libraries). Belew 2000, and Manning 2009 (on IR).

Exam

- Paper presentations in the third part of the course.
- Written exam on the first two parts.

http://disi.unitn.it/~bernardi/Courses/DL/11-12.html

2. Library

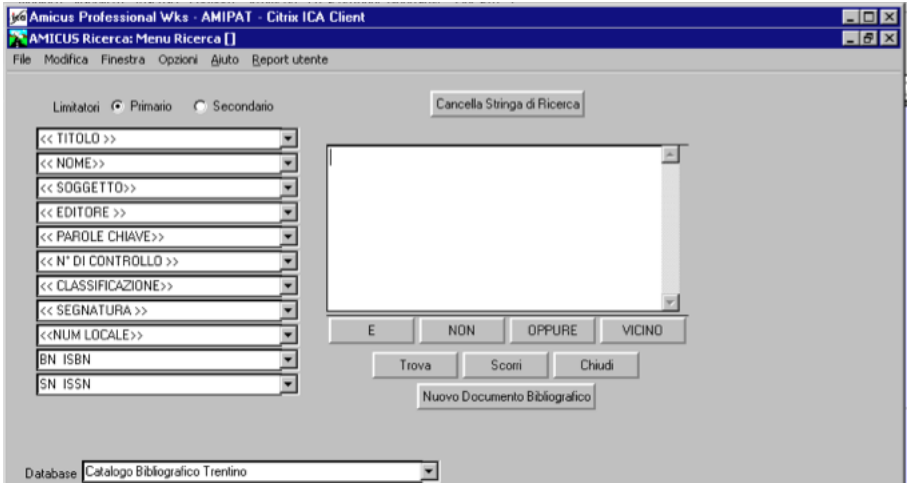
When you think of a Library, what do you think of?

When you think of a Digital Library, what do you think of?

3. CBT (Catalogo Bibliografico Trentino)

CBT: Integrates catalogues of Biblioteca universitaria, Biblioteche storiche di conservazione, Biblioteche pubbliche di base, Biblioteche specialistiche. (in 2004: 1.195.000)

3.1. CBT: Catalogue Interface

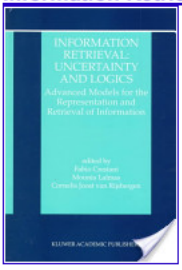


3.2. Google: Search interface and results

information retrieval ✕ Search

About 860,000 results (0.82 seconds) Advanced search

▸ **Information Retrieval**



Fabio Crestani, Mounia Lalmas ... - 1998 - 323 pages - Preview
Information Retrieval: Uncertainty and Logics contains a collection of exciting papers proposing, developing and implementing logical IR models. This book is appropriate for use as a text for a graduate-level course on Information Retrieval or Database Systems, and as a reference for researchers and practitioners in industry.
books.google.com - More editions

More from: Wikipedia, School of Computing Science
Buy: Amazon.com $285.00, Barnes & Noble $285.00, Borders $285.00, Buy.com $285.00, More »

And you can view the whole document, rank the results by relevance, get external info about the topic of the book. It corrects users' spelling mistakes, suggests for other keywords etc.

3.3. Goal of the course

Get a feeling of the technology behind CBT and the technology behind Google Books.

To this end, we need to look at:

- Evolution of Library and Evolution of Computer Science (to see how one has effect on the other)
- How and which information about records are stored.
- Information Retrieval (to retrieve the document)
- Language Technologies (to process the document)

4. Library Evolution

- Medieval Cathedrals monks kept libraries and copied books for each other by hand. There was an organized book trade.
- Printed books replaced manuscripts, the organization stayed the same.
- 18th century: changes in literacy, but the technology stayed the same.
- 19th century: greater technology advances

New role of libraries.

4.1. Writing time and page cost

- A compositor setting type in the 18th century was expected to set 3/4 words per minute.
- A Linotype operator could set about 10 words per minute.
- A modern keyboarder can do 50 words per minute.

As a consequence, London Daily Journal of March 4, 1728 cost 3 half-pennies for two dies of one sheet, while the Times of 1905 cost 1 penny for 24 pages.

Table 1.1 Number of volumes held by major US libraries.

	Volumes Held		
Institution	1910	1995	2002
Library of Congress	1.8 M	23.0 M	26.0 M
Harvard	0.8 M	12.9 M	14.9 M
Yale	.55 M	9.5 M	10.9 M
U. Illinois (Urbana)	.1 M	8.5 M	9.9 M
U. California (Berkeley)	.24 M	8.1 M	9.4 M
New York Public Library	1.4 M	7.0 M	11.5 M
U. Michigan	.25 M	6.7 M	7.6 M
Boston Public Library	1.0 M	6.5 M	7.5 M

Table 1.2 Number of volumes held by major global libraries.

	Number of Volumes Held				
Institution	Earlier	1910	1996	2002	Former name, if any
British Library	240 K (1837)	2 M	15 M	18 M	British Museum Library
Cambridge Univ.	330 (1473)	500 K	3.5 M	7 M	N/A
Bodleian (Oxford)	2 K (1602)	800 K	4.8 M	6 M	N/A
Bibliothèque Nationale de France	250 K (1800)	3 M	11 M	12 M	Bibliothèque Nationale
National Diet Library	N/A	500 K	4.1 M	8 M	Imperial Cabinet Library
Biblioteca Alexandrina	533 K (48BC)			240 K	Library of Alexandria

5. Computer Technology Evolution

Electronic computers and digital storage are the key inventions that made possible digital libraries.

- Volume: the first computers filled in one room. Today much more powerful machines fit on a chip
- Time: since 1960, there has been about 1000-fold speed increase.
- Storage: in the '60, less that one long text; today huge volumes of texts and videos.

Table 1.3 Memory sizes.

Unit	Exponent	Amount	Example
Byte	1	1 byte	One keystroke on a typewriter
		6 bytes	One word
		100 bytes	One sentence
Kilobyte	3	1000 bytes	Half a printed page; a tiny sketch
		10,000 bytes	One second of recorded speech; a small picture
		30,000 bytes	A scanned, compressed book page
		100,000 bytes	A medium-size, compressed color picture
		500,000 bytes	A novel (e.g., *Pride and Prejudice*)
Megabyte	6	1,000,000 bytes	A large novel (e.g., *Moby Dick*)
		5,000,000 bytes	The Bible
		10,000,000 bytes	A Mozart symphony, MP3-compressed
		20,000,000 bytes	A scanned book
		50,000,000 bytes	A 2-hour radio program
		500,000,000 bytes	A CD-ROM; the *Oxford English Dictionary*
Gigabyte	9	1,000,000,000 bytes	A shelf of scanned paper; or a section of bookstacks, keyed
		100,000,000,000 bytes	A current disk drive size
Terabyte	12	1,000,000,000,000 bytes	A million-volume library
		20 terabytes	The Library of Congress, as text
Petabyte	15	1000 terabytes	Very large scientific databases
		9 petabytes	Total storage at San Diego Supercomputer Center
Exabyte	18	A million terabytes	
		20 exabytes	About the total amount of information in the world
		5 exabytes	World disk production, 2001
		25 exabytes	World tape production, 2001

5.1. Storage Prices

What matters for storage are size, price and durability.

- 1956: 4.5 megabytes – cost 40K dollars.
- 2003: 300 gigabyte disk drive, 1 dollar per gigabyte (=300K books, can fit on 3.5-inch disk drive – few hundred dollars)

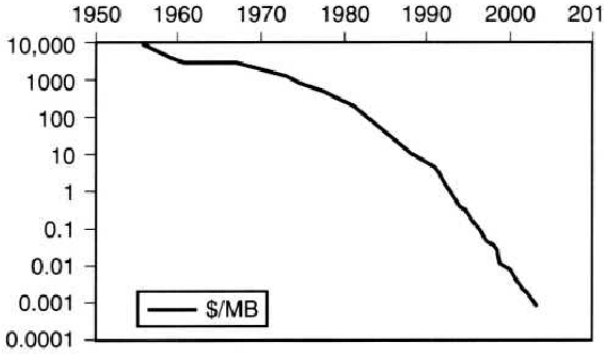


Figure 1.8 Changes in the price of memory.

5.2. Summing up: Libraries

Technology has made enormous advances in processor, input and output, storage devices and connectivity.

- In 1962 a good university research computer was an IBM 7090, which cost above 3 million dollar, a student to program it could be hired for 1.50 dollar per hour, so that the machine cost the equivalent of 2 million hours or the equivalent of 1000 years of work.
- Today we would buy a better machine for 600 dollars and pay an undergraduate 10 dollars per hour, so that the machine would cost the equivalent of 60 hours or less than two weeks of work.

Given the cost changes, libraries can now use computer for many functions, so long as they save even a little time or provide a little help, and we no longer are technically limited in building systems.

6. DL Evolution

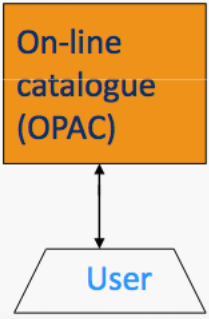
Digital Libraries do exist today?

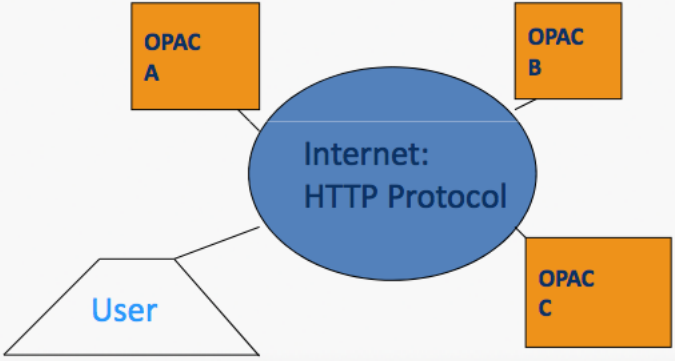
- Are they a transformation of “traditional library”?
- Are they an evolution of data bases?
- Are they (a subset of) the Web?

DLs are the intersection of different disciplines/technologies.

A “theory” of Digital Libraries has not been developed yet.

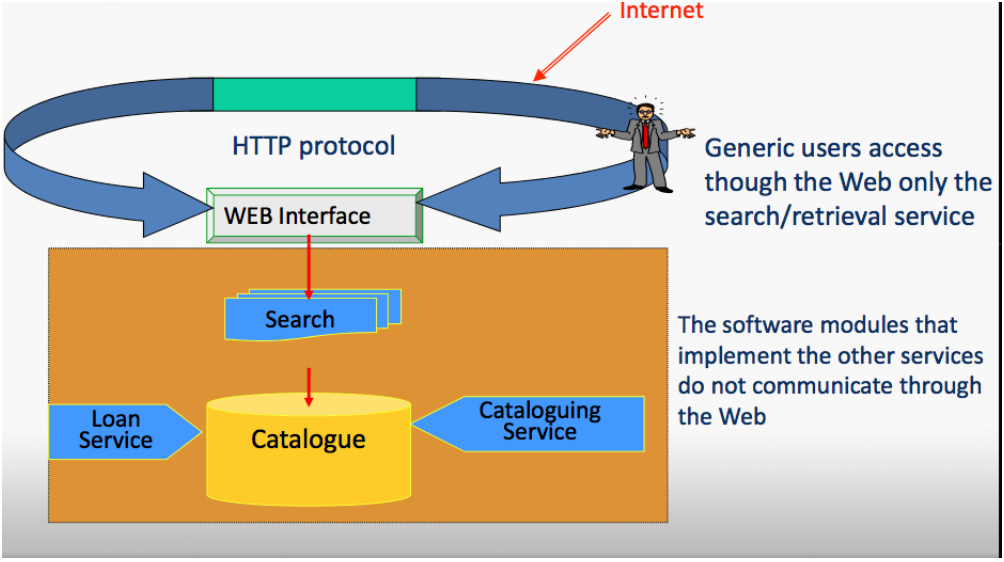
from a “direct” communication
(on-line) ....





.... to a communication via
WEB: HTTP protocol

6.1. Web Access to OPACs



6.2. Disadvantages

- Each catalogue is accessible through its own user interface
- User interfaces differ for:
  - access points
  - names of the access points
  - language
  - graphics
  - ...
- Users must be familiar with many interfaces
- No cross-searches are possible

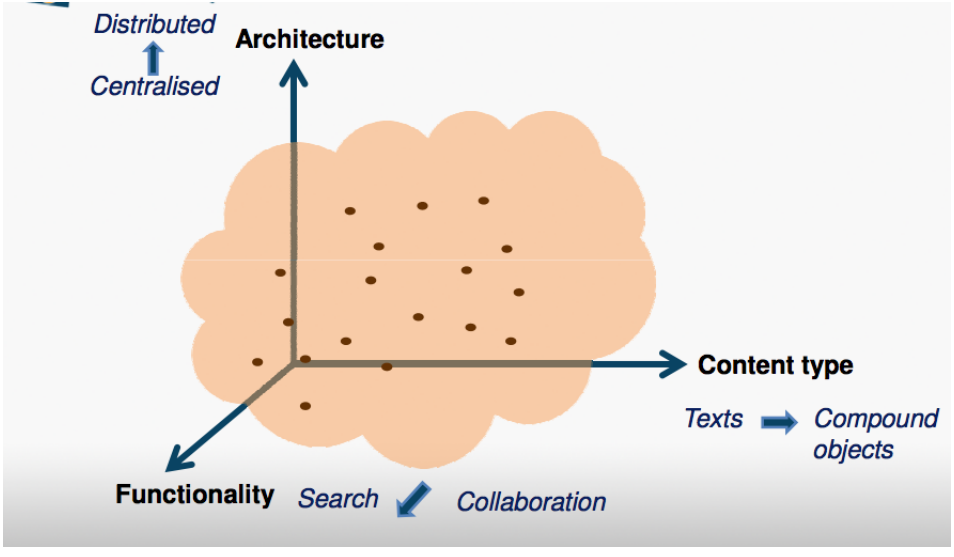
6.3. Digital Libraries comes to life

Phase I: 1994-1998 Digital Libraries Initiatives. Funded by National Science Foundation (NSF), Department of Defense Advanced Research Project Agency (DARPA); and National Aeronautics and Space Administration (NASA).

Objective "The focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks – all in user friendly ways."

Definition (1998) "An institution which performs and/or support (at least) the function of a library in the context of distributed, networked collections of information objects in digital form."

6.4. DL evolution axes



6.5. Content type: From Text to Compound Objects

New document types empower novel forms of communications and remote collaboration among the members of a community of interest:

1. Text objects
2. Multilingual objects
3. Multimedia objects
4. Annotated object
5. Compound objects

Multilingual Documents in different languages can be maintained in the same DL.

Multimedia Audio-video DL (mainly News)

Annotated Comment, Rating, Description; on the whole doc or on its parts; authored by different people; public or restricted.

Compound Objects Journals (dif. articles), Video (dif. sequences), tutorials (diff. lectures, moreover: videos, plus demos, plus related doc.) ....

6.6. Functionality: From Search to Collaboration

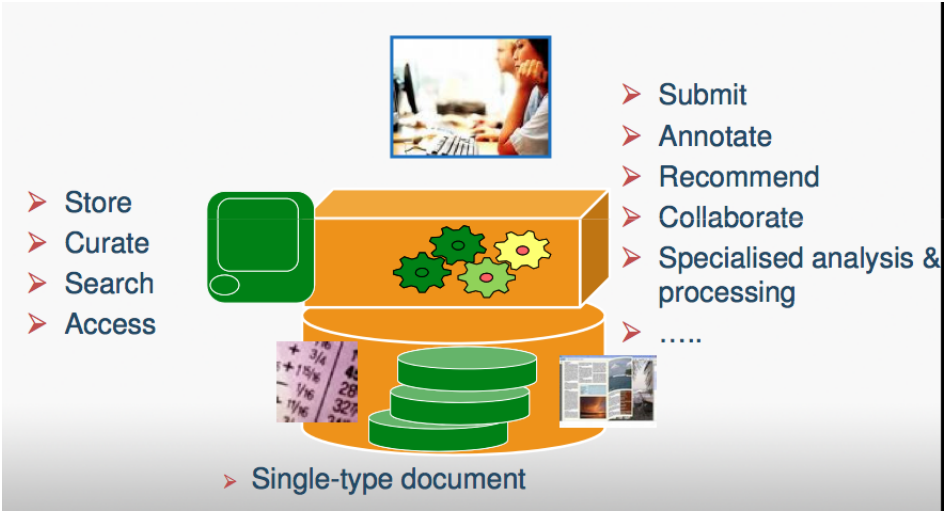
New document types impose a re-thinking of the "traditional" library services: Submission, Description, Search, Dissemination, ...

E.g. Submission of Video it must be possible to structure the video into meaningful parts (sequences, scenes, frames) ... and describe the video and its parts separately

Search Free text search, fielded search, monolingual and cross lingual, similarity search, search on annotations, ...

Communication New services can be included in a DL to improve its potential usage: Reccommenders, Peer-reviewing supporting services, social networking, ...

6.7. Architecture: From Centralised to Distributed to Federated

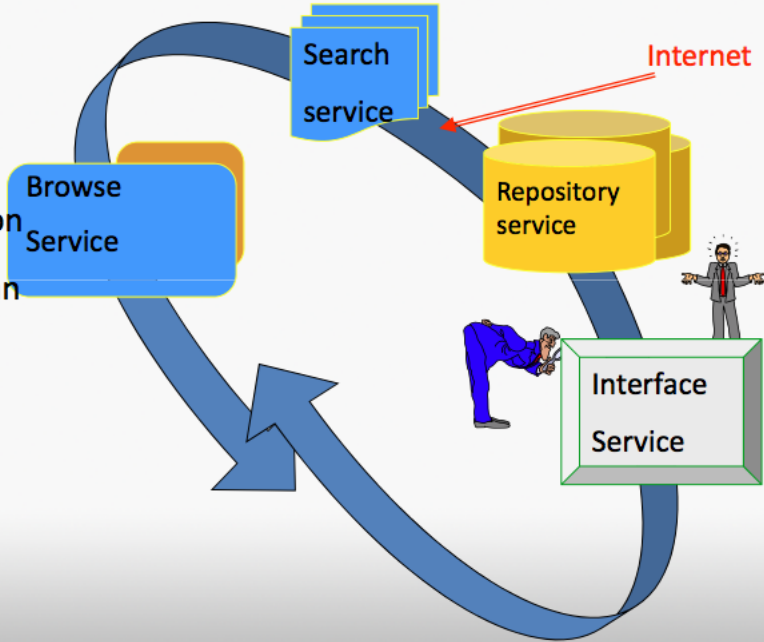


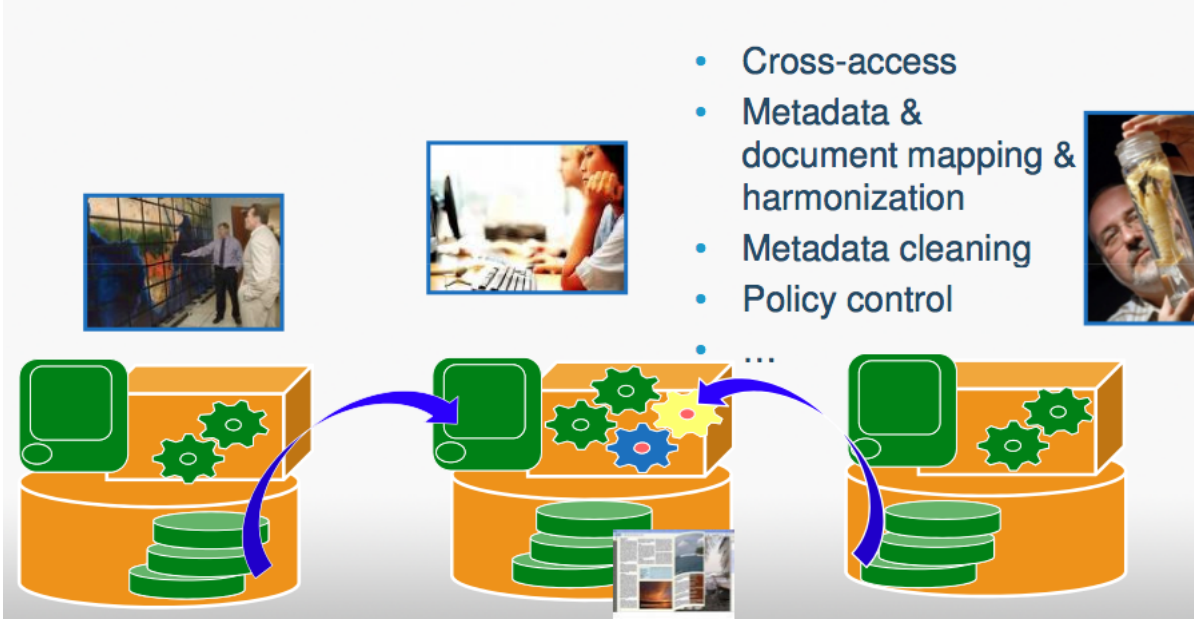
Distributed systems

The services are distributed on the Internet.

They communicate through an established protocol.

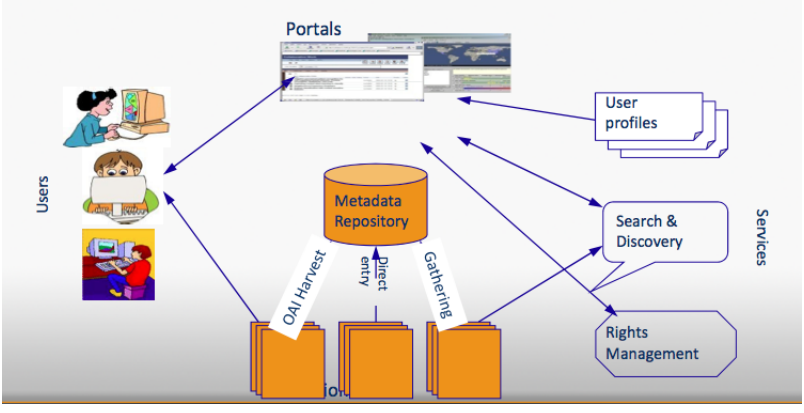
Users access the system through a Web interface



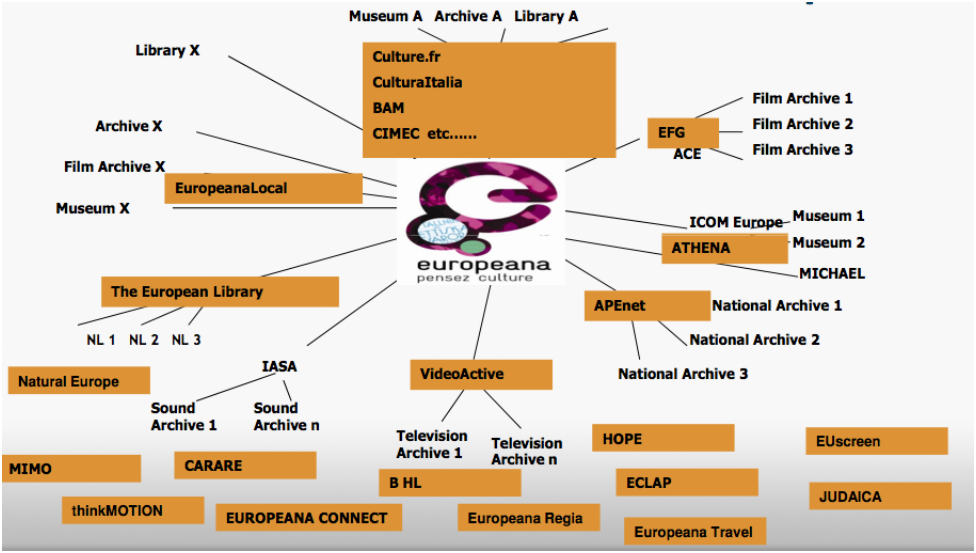

Cross-access
Metadata & document mapping & harmonization
Metadata cleaning
Policy control
…

6.8. A Federated Library: NSDL

US Nations' online library for education and research in Science, Technology, Engineering, Mathematics



6.9. Large Federation: Europeana



7. What is a DL?

Given definitions:

- Digital Libraries are organized *collections* of digital information. They combine the structure and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible. [Lesk, 1997]
- Digital Library is an *institution* responsible for providing at least the functionality of a traditional library in the context of distributed and networked *collections* of information objects. [Belkin, 1997]

- Digital libraries are a set of *electronic resources* and associated technical capabilities for creating, searching and using information; DL are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium. DL are constructed, collected and organized by (and for) a *community of users*, and their functional capabilities support the information needs and uses of that community. Researchers view Digital Libraries as content collected on behalf of user communities, while practising librarians view Digital Libraries as *institutions* or services. [Bormann, 1999]

- Digital Libraries are tools to serve research, scholarship and education; tools to access information; tools to provide services primarily to *individual users*. [Lesk, 1999]

- Digital Libraries are *organization* 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* [Soergel, 2002]

- A digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored *locally*, or accessed remotely via computer *networks*. A digital library is a type of *information retrieval* system. [Wikipedia]

7.1. DL: revision

(i) the Internet is The Digital Library;

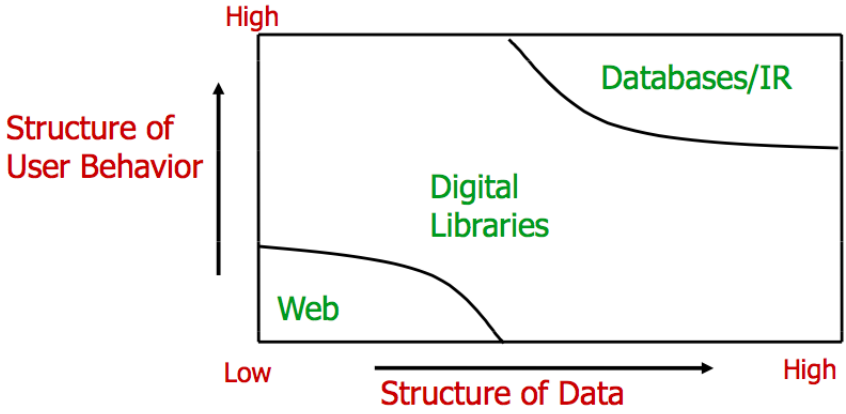
(ii) at some point there will be a single Digital Library or a single-window view of Digital Library collections;

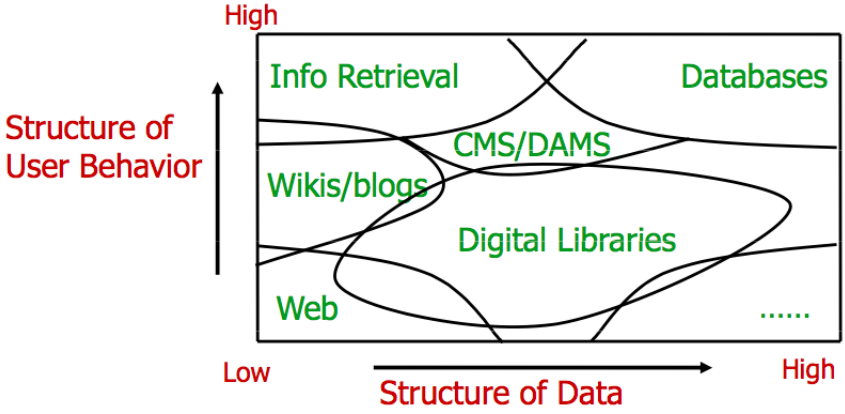
(iii) Digital Libraries are means to provide more equitable access to content from anywhere at any time; and

(iv) Digital Libraries are cheaper instruments than physical libraries

"None of the above claims is true. Hence, Digital Libraries impose *reinvention of the role of librarians and library models*." [Kuny, 1996]

7.2. DL: where does it sit?




High
Structure of
User Behavior
Info Retrieval
Databases
CMS/DAMS
Wikis/blogs
Digital Libraries
Web
.....
Low
High
Structure of Data

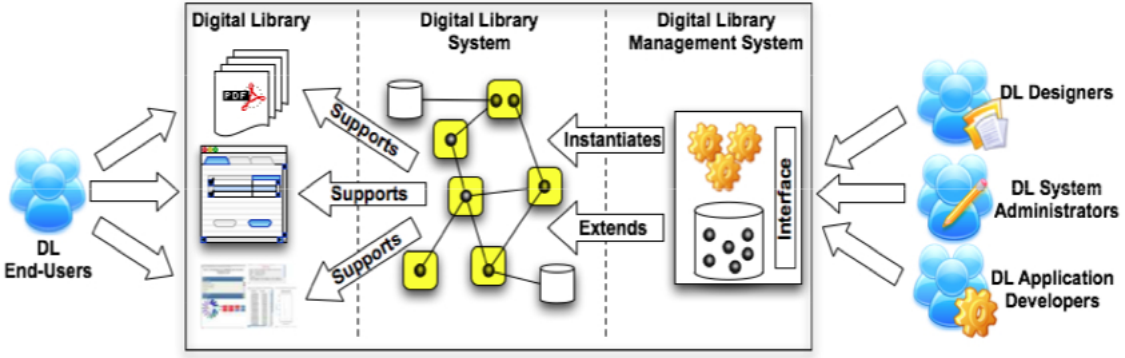
8. DL: Three entities

Digital Library An *organization*, which might be virtual, that comprehensively collects, manages, and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content of measurable quality and according to codified policies.

DL System A *software* system that is based on a defined (possibly distributed) architecture and provides all functionality required by a particular DL. Users interact with a DL through the corresponding DL System.

DL Management System A generic software system that provides the appropriate *software infrastructure* both (i) to produce and *administer* a DL System, incorporating the suite of functionality considered foundational for DL and (ii) to integrate additional software offering more refined, specialized, or advanced functionality.

8.1. DL actors


Digital Library
Digital Library System
Digital Library Management System
DL End-Users
Supports
Supports
Supports
Instantiates
Extends
Interface
DL Designers
DL System Administrators
DL Application Developers

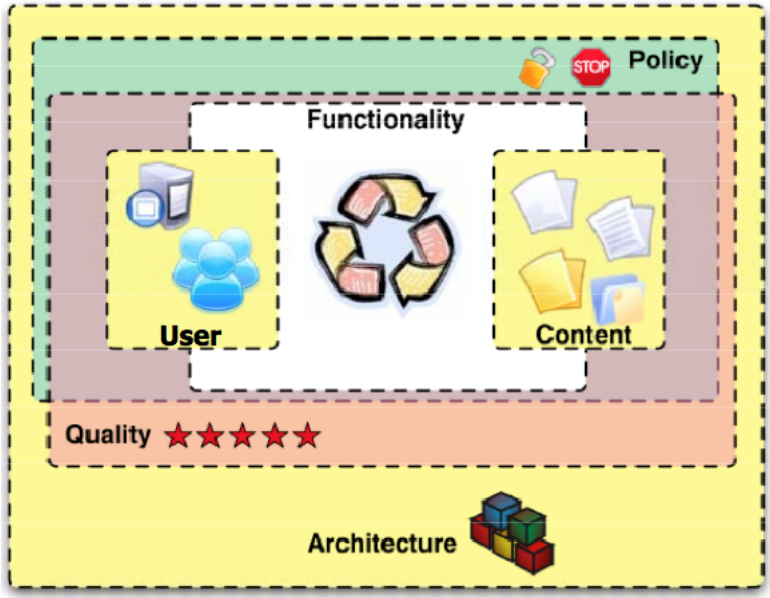
End Users They exploit the DL functionality for providing, consuming, managing the DL Content as well as some of its other constituents. They perceive the DL as a *stateful entity that serves their functional needs*. The behaviour and output of DL depend on its state at the time a particular part of its functionality is activated.

Digital Librarian They exploit their *knowledge of the semantic of the application domain* to define, customize, and maintain the DL so that it is aligned with the information and functional needs of its end-users. To perform this, they interact with the DLMS providing *functional and content configuration parameters*. The values of this parameters, which can be modified during the DL lifetime, configure the specific DL perceived by the end-users because they determine the particular DLS instance serving the DL.

System Librarian They select the *software components* necessary to create the DLS needed to serve the required DL (as specified by the Digital Librarian) and decide where and how to deploy them. They interact with the DLMS by providing *architectural configuration parameters*, s.a. the selected software components, the hosting nodes, and the components allocation. The value of the architectural configuration parameters can be changed over the DL lifetime. Any DL change of these parameters may result in the provision of different DL functionality and/or different quality.

DL Application Developers They *develop the software components* of DLMSs and DLSs, implementing necessary functionality.

8.2. DL objectives



Content the data and information that the DL handles and makes available to its users. Content is an umbrella that Library collects, manages, and delivers. It encompasses the diverse range of information objects, including such resources as objects, annotations, and metadata.

User various actors (whether human or machine) entitled to interact with DL. DL connect actors with information and support them in their ability to consume and make creative use of it to generate new information. User is an umbrella concept including all notions related to the representation and management of actor entities within a Digital Library. It encompasses such elements as the rights that actors have within the system and the profiles of the actors with characteristics that personalize the systems behaviour or represent these actors in collaborations.

Functionality the services that a DL offers to its different users, whether classes of users or individual users. While the general expectation is that DLs will be rich in capabilities and services, the bare minimum of functions would include such aspects as new information object registration, search, and browse. Beyond that, the system seeks to manage the functions of the DL to ensure that the functions reflect the particular needs of the dls community of users and/or the specific requirements relating to the Content it contains.

Policy set (or sets) of conditions, the terms and *regulations governing interaction between the DL and users*, whether virtual or real. Examples of policies include acceptable user behaviour, *digital rights management*, *privacy* and confidentiality, charges to users, and collection delivery.

Quality the parameters that can be used to characterize and evaluate the content and behaviour of a DL. Quality can be associated not only with each class of content or functionality but also with specific information objects or services. Some of these parameters are *objective* in nature and can be automatically measured, whereas others are *subjective* in nature and can only be measured through *user evaluations*.

Architecture the DLs entity and represents a mapping of the functionality and content offered by a DL onto hardware and software components.

9. The Internet, the Web, Google and Libraries

- Internet, electronic network between computers
- 1992: Berners-Lee – WWW (World Wide Web)
- 1997: Andreesen – interface
- First web crawlers at University of Washington. Today they make full text indexes and provide search services for free. High speed is achieved by keeping the index in RAM memory and by using enormous number of machine searching in parallel (Google quit saying how many machines it owned after the number passed 20K)
- Google gained acceptance by figuring out how to show good documents on top of the retrieved document list. The method was invited by Larry Page and Sergei Brin while graduate students in the Digital Library group at Stanford.

10. Conclusion

Frontal lectures by me.

Seminars leaded by students: we will assign a paper to be presented by one student.