Universität Konstanz

DBVIS

Digital Libraries

Jun.-Prof. Dr. Tobias Schreck

WS 2014/15

1. Introduction

Organization

- **Lecture**
  - Mondays, 11:45 - 13:15, G300

- **Exercises (biweekly blocked)**
- Fridays, 13:30 - 15:00, G227
  07.11., 21.11., 05.12., 19.12., 09.01., 23.01., 06.02.
  - Mix of exercise sheets and a practical assignment (implement a prototype, or present a paper)

- **Award 5 credit points**
  - Achieve at least 50% of exercise program scores
  - Written exam, dates:
    09.02.2015 (A702, 11.45-13.15) / 13.04.2015 (t.b.a.)
  - Bonus of 1/3 grade for achieving at least 80% of the exercise scores

Organization

- **All students need to register at the beginning of the semester**
  - Studis http://studis.uni-konstanz.de/
    - Deadline for first exam: 01.12.2014 - 31.01.2015
    - Deadline for second exam : 01.03.2015 - 23.03.2015
  - LSF http://www.studium.uni-konstanz.de/lehrveranstaltungen/
    - Deadline: 06.10.2014-16.11.2014

- **ILIAS Repository**
  - Lecture Slides
  - Exercise assignments, upload solutions
  - https://ilias.uni-konstanz.de/ilias/goto_ilias_uni_crs_374700.html
  - Apply for access using written registration sheet handed out in lecture

1. Introduction

- Learning Goals
  - Understand basic concepts, terminology and definitions of the subject
  - Understand the traditional role of libraries and expectations towards digital libraries
  - Get to know important areas of service for Digital Libraries; understand the various challenges of setting up and operating digital libraries
  - Preview of the lecture topics

1. Introduction

1.1 The Role of Libraries

1.2 History of Digital Libraries

1.3 Influential Ideas, Definitions, Models

1.4 The World Wide Web

1.5 Lecture Outline

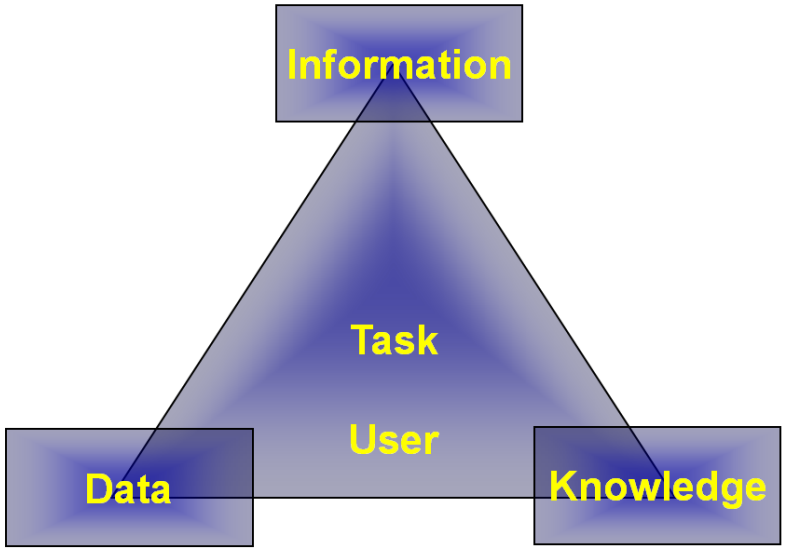
1. Introduction

1.1 The Role of Libraries
1.2 History of Digital Libraries
1.3 Influential Ideas, Definitions, Models
1.4 The World Wide Web
1.5 Lecture Outline

1.1 The Role of Libraries

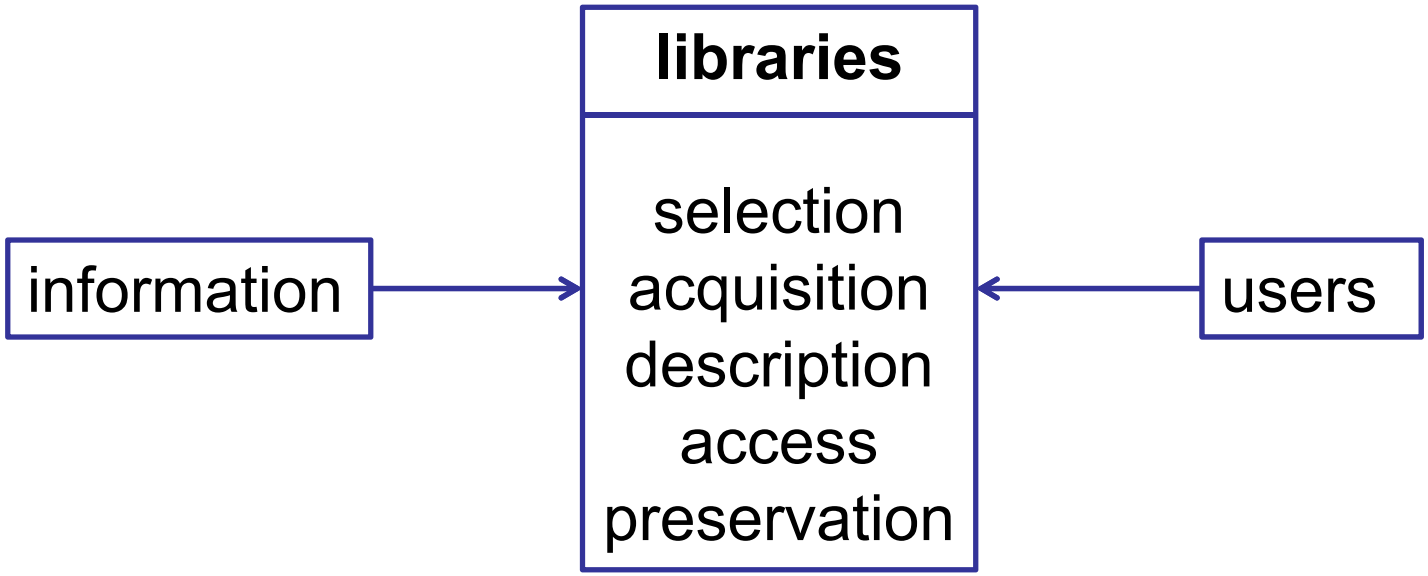
- **Information as a resource**
  - Scientific discovery process
  - Technology transfer
  - Communication across time and space
  - Libraries have historically been **the** institutions to organize information

- **Important trends**
  - Specialization in the scientific disciplines progressing (Leibniz)
  - Tremendous progress of IT
    - Digitization, storage, transmission of data
    - Automatic processing capabilities
  - The Internet



1.1 The Role of Libraries

Libraries act as mediators between information and users. They play an important role in information retrieval and knowledge transfer.



Private funding models exist, but many libraries are financed by the public as part of national information infrastructure programs.

1.1 The Role of (Paper) Libraries

Traditionally, information is stored in *paper form* (books, journals, etc.) and libraries have long experience in working with paper documents

- Selection – Definition of collections
- Acquisition – Physical objects
- Description – Catalogs
- Access – Shelves, Lending schemes
- Preservation – Controlled environment, Media care

Library of Congress in Washington D.C.

1.1 The Digital Age

Digital Information Processing

- Acquisition, production, storage
- Data integration, data mining
- → Large and increasing amounts of documents available

Data-intensive application domains

- Business
- Research
- Engineering

Digital Information Production

- Retrodigitization
- Authoring (borne digital)
  - Traditional (publisher, peer-review)
  - Social media (Blogs, Twitter, Forums, …)

Share of digital information

2000: 25%

2002: 50% (Begin *Digital Age*)

2007: 94% (300 Exabyte)

Estimated growth rates (1986-2007)

Storage: 23%

Network: 28%

Compute: 56%

Source: Science, according to [F&L 3/2011]

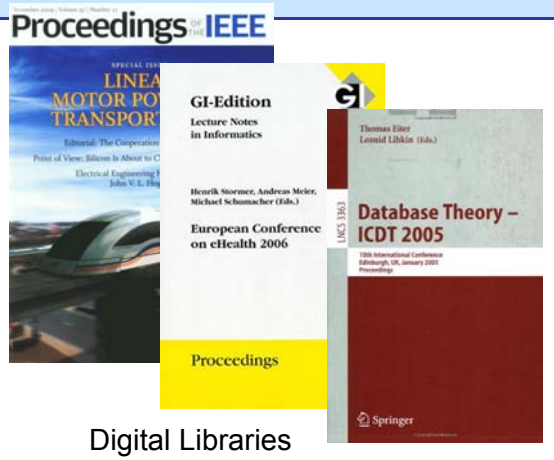
1. Forms of Digital Libraries and Documents

Textual Data Repositories

- Digital Libraries
- Web
- Social Media

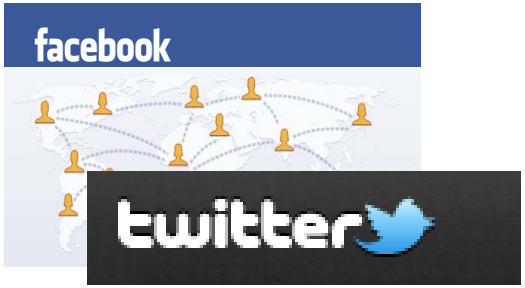
Non-textual Data Repositories

- Image repositories
- 3D Object repositories
- Data repositories



Digital Libraries

Customer Reviews (Amazon.com)



PROBADO3D Archive (http://www.probado.de/3d.html)

Sloan Digital Sky Survey (http://www.sdss.org/)

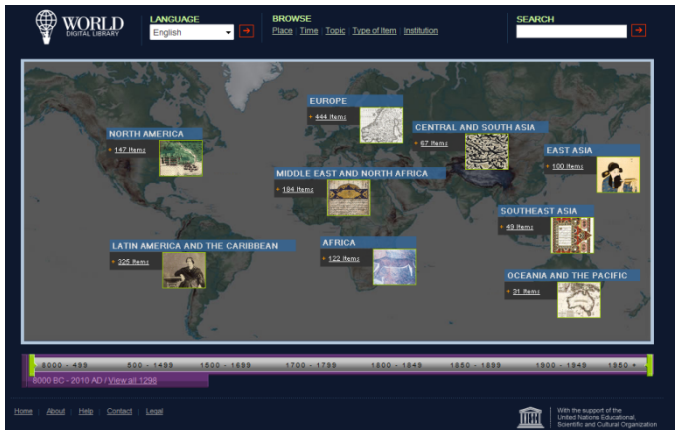
Victoria State Library Image Collection (http://www.slv.vic.gov.au/)

1.1 Why digital libraries?

Increasingly, *digital documents* arise (either digitally born, or from retro digitization efforts)

Important implications for role of libraries:

- Economic
  - Electronic access (who pays? restrict access?)
- Digitization
  - How to organize the digitization process?
  - Quality standards?
- Indexing
  - How to automatically index and search in digital documents?
- Long term preservation
  - How to provide accessibility? How to securely store data?
- Multimedia and non-textual documents



1.1 Why digital libraries?

Evolution of technology

- Computer technology
  - CPU and integrated chips
  - Random Access Memories – from KB to GB
  - External memories
    - Tapes, hard disks, floppy disks, Memory sticks, CDs, DVDs
- Communication technology (networks)
  - (Telephone) line speed
  - Point to point (leased lines)
  - Local Area Networks
  - Inter-networking (TCP/IP)

Result: more and more information are now stored online/in digital form

1.1 Where are we now?

In 1964, Arthur Samuel predicted that paper libraries would disappear, except for paper museums. But this did not happen, Why?

- Costs:
  - 1 Billion dollar to digitize 100 Mio Books; plus additional funds to compensate copyright holders (←→ Google scan project)
  - Institutional transformation
- User access
  - Many users still like paper books and even card catalogs
  - Difficulties to design good user interfaces
  - Affordance of computers by everyone
  - Reluctance to trade a system that works for one which may not work

Lesk: Transition from academics (early adopters) to eventually everyone who will be able to use it.

- unsuccessful example: Doomsday Project of BBC (see next slides)

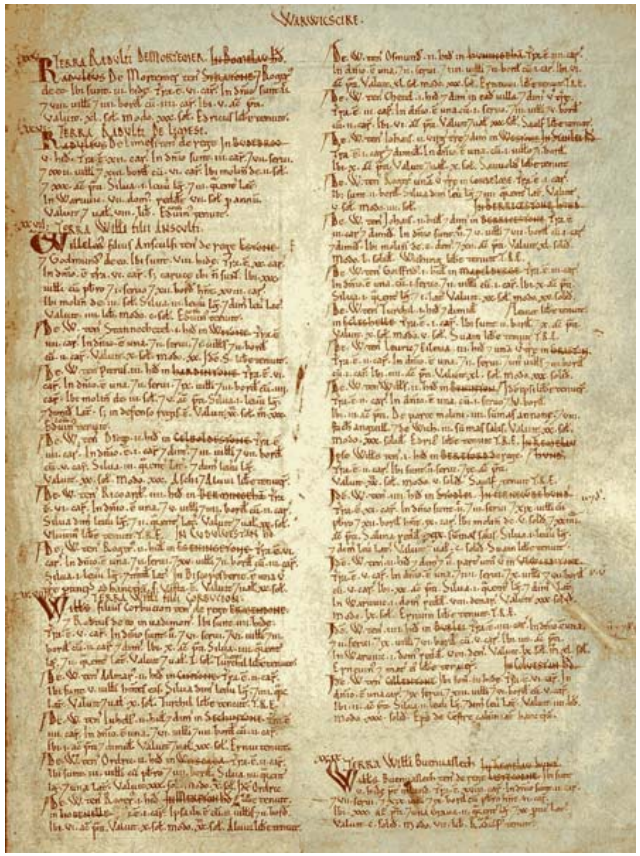
1.1 Example Problem of Traditional Libraries

Destruction of hundreds of thousands of scrolls in the library of Alexandria 48 BC
(xylograph; Source: http://www.spiegel.de/)

1.1 Example Problem of Traditional Libraries

Collapse of the Cologne City Archive on March 3, 2009
(Source: http://www.spiegel.de/)

1.1 Example Problem of Digital Library Project



Domesday Book (1086), commissioned by King William I of England

1.1 Example Problem of Digital Library Project

- **BBC Domesday Project (1986)**
  - Partnership between BBC, Acorn Computers Ltd., Philips, Logica
  - Compiled1984-1986
  - > 1 million participants surveyed
  - Stored on 2 laserdics (300 MB each) Images, video, digital data

1.1 Example Problem of Digital Library Project

Problems

- Choice of wrong medium
  - LD and computer system became obsolete during production
  - Data could not be read, software could not be executed
- Re-acquisition of data later, but copyright issues prevented distribution
- Significant efforts to recover and convert the material (emulator, extraction of disk images, conversion for web)

Domesday Preservation Group

.about. .people. .reports. .faq.

If you reached this website via the domesday1986.com domain unfortunatly the site is currently off-line as Adrian Pearce the producer of that web site sadly died on January 27th 2008. [Details]

In November 1984 the BBC announced the start of the BBC Domesday Project, to compile a new electronic Domesday Book in a bid to provide a comprehensive image of Britain in 1985. The completed project was timed for release in 1986 to celebrate the 900th anniversary of the 1086 Domesday book.

Despite the huge scale of this project, surprisingly it is now extremely difficult to access the project. Data recorded in old computer formats increasingly can no longer be accessed on new equipment while magnetic storage media and optical discs have physically decayed, ruining precious data. Fragile obsolete equipment is required to read the data and very little usable equipment still exists.

The BBC Domesday project is a high profile example of 'digital obsolescence' where technology has changed to the point that this once prestigious project has essentially disappeared less than a quarter of a century after it was completed.

The historical and cultural significance of the Domesday project makes it one of the most significant digital resources produced in recent history. Significant efforts must be made to ensure that this resource is not lost and is once again made accessible the people who contributed to it.

1. Introduction

1.1 The Role of Libraries
1.2 History of Digital Libraries
1.3 Influential Ideas, Definitions, Models
1.4 The World Wide Web
1.5 Lecture Outline

1.2 History of Digital Libraries

- How information is transmitted and used changes over time with technology and society.
  - E.g., music of Bach
    - Delivery on instruments → printed score sheets → LPs and CD/MP3.
    - Usage: spiritual → concerts → movie and shopping mall music
  - Forms of information transmission
    - Recitation (Homer)
    - Scripts, monasteries
    - 18th century: Increase in literacy in society and demand for printed books
    - Increase in electronic editing and consuming (Mobile devices, Ebook readers etc.)
  - Political notions of information access restriction (e.g., Chinese emperor Shih Huang Ti) or freedom of access (e.g., USA)
  - Change driven by technology and society; no determinism

- Lesk: Need for libraries and universities to actively shape process of information access

1.2 Comparison of Library Sizes

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

1.2 Growth of Held Volumes and Periodicals

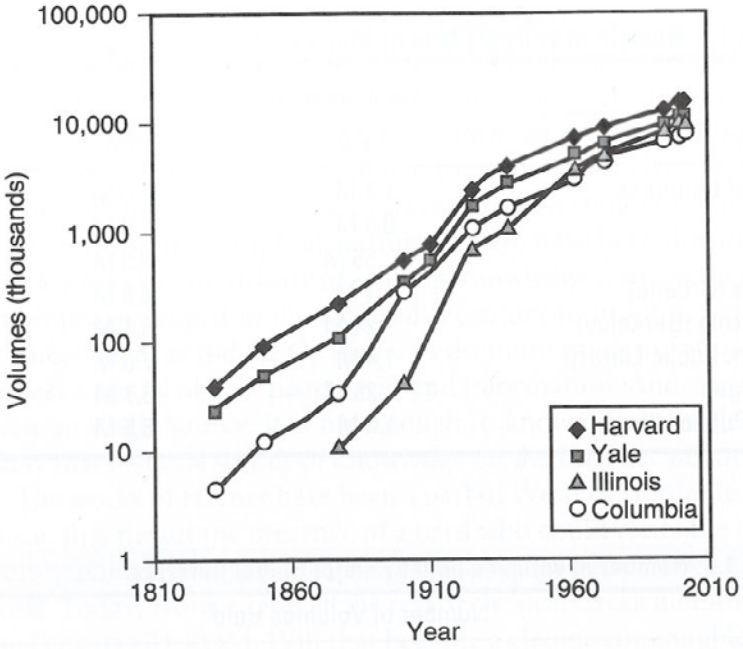


Figure 1.2 Rate of growth in university libraries for the past 100 years.

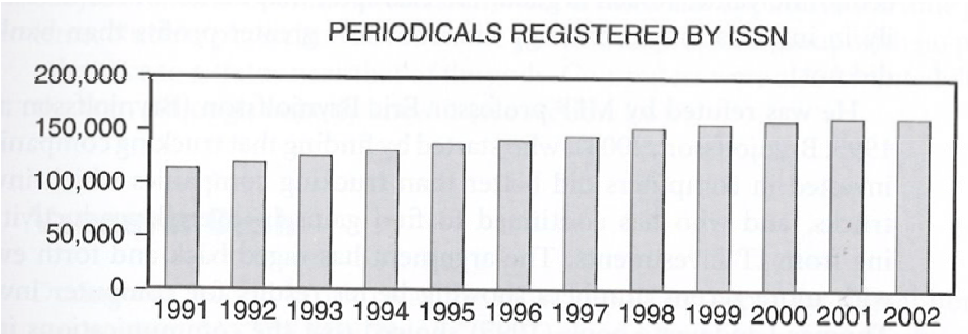


Figure 1.4 Recent rate of growth of periodicals.

1.2 Digital Storage Capacities

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

1.2 Value of Information

- What is the value of information on a macro level?
  - General belief: access to information is good for productivity
  - Universities spend about 4% of budget for libraries
  - But it is hard to measure the dependency between information supply and productivity
    - Largest libraries are in Washington, London, and Moscow; but largest economic growth is in Beijing, Bangalore and Seoul
    - Steven Roach (Morgan Stanely analysis) claims that US slowdown since mid 70s attributed not to oil crisis but to investment in IT (but claim is heavily contested by other economists).
  - Dot-com bubble burst (then: assumptions that 30% of productivity due to information) in 2000 de-emphasized the belief, but still it is held valid that information is beneficent factor.
  - Overall, it is unclear how to quantitatively measure the relation between information supply and economic growth

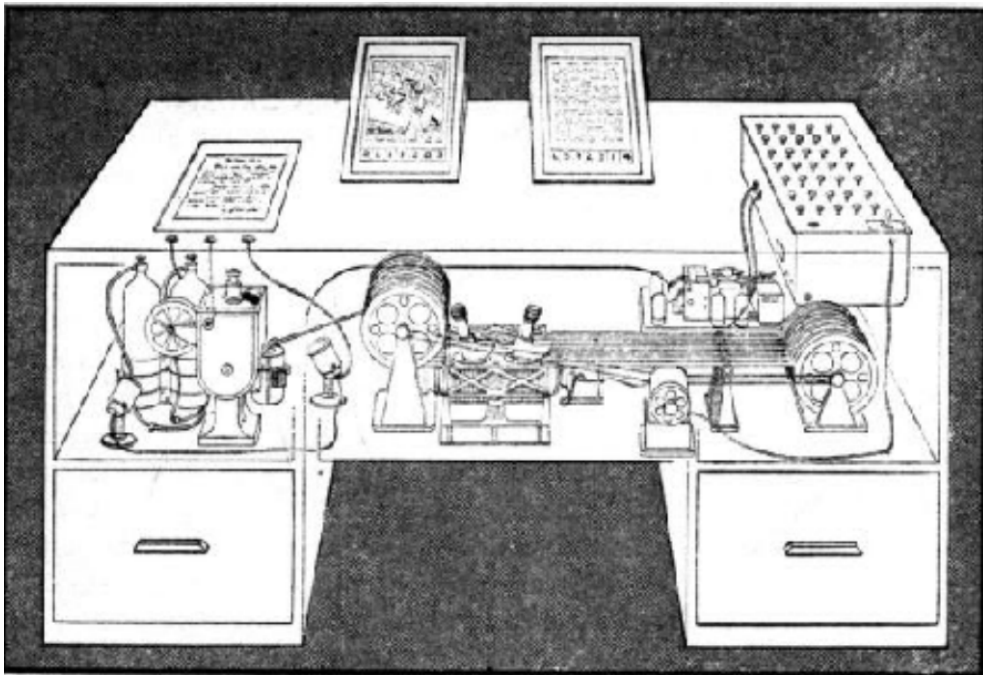
1. Introduction

1.1 The Role of Libraries
1.2 History of Digital Libraries
1.3 Influential Ideas, Definitions, Models
1.4 The World Wide Web
1.5 Lecture Outline

1.3 Influential Ideas, Definitions, Models

Vannevar Bush (1945)

- Head of US science during WW2, Prof. at MIT
- Use of “knowledge” and team work to advance science
- The Memex: mechanized private archive and library (microfilms)
- “trails” of information
  - associative links
- No “free text” search

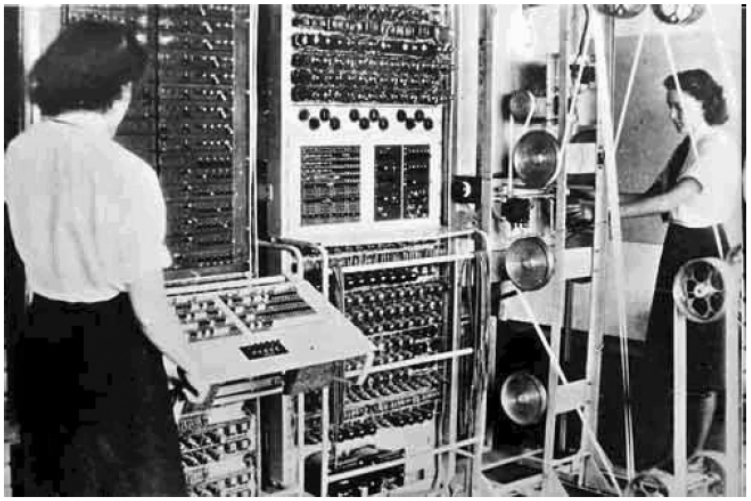


Memex: As visualized by Life Magazine in 1945

1.3 Influential Ideas

Warren Weaver (MIT, 1947)

- Proposed machine analysis of documents (then: translation).
- Father of *automatic approaches*
  - Automatic access, statistical methods
    (Information Retrieval)



Colossus computers (1940s)

Two schools of thought:

Manual and collaborative approach (Bush)

Statistical-automatic approach (Weaver)

1.3 Influential Ideas

JCR Licklider (Libraries of the future - 1965)

- Head of US Dept. of Defense, Information Processing & Technologies
- The book foresees the research and development needed to build a Digital Library
  - Time-sharing just beginning
  - "Big" memories around 32K
  - Networking "to be invented"
- Rather accurate overall view of what a DL could look like in 1995
  - Under-estimation of computing power
  - Over-estimation of progress in artificial intelligence and natural language processing

1.3 Definitions and Models

Gladney H.M, et. al. 1994

"A digital library service is an assemblage of digital computing, storage, and communications machinery together with the software needed to reproduce, emulate, and extend the services provided by conventional libraries based on paper and other material means of collecting, storing, cataloguing, finding, and disseminating information."

1.3 Definitions and Models

Borgman, 1996

"Digital Libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information…they are an extension and enhancements of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds, static or dynamic images) and exist in distributed networks"

- Borgman identified two major aspects
  - DL researchers from Computer Science focus on content for user communities and therefore emphasize the enabling technologies
  - Library professionals appear to emphasize DLs as services
- DLs require the skills from both sides

1.3 What is important?

- Site Neutrality
  - *Access-Anytime (24*7)*
  - *Anywhere (Office, Residence, Travel)*
  - *By Anyone*
- Open Access and Sharing of information
- Greater variety and granularity of information
- Up-to-date-ness
- New forms of rendering (New Genre)
- Integration of digital media into traditional collections
- Digital libraries are different in that they are designed to support the creation, maintenance, management, access to, and preservation of digital content

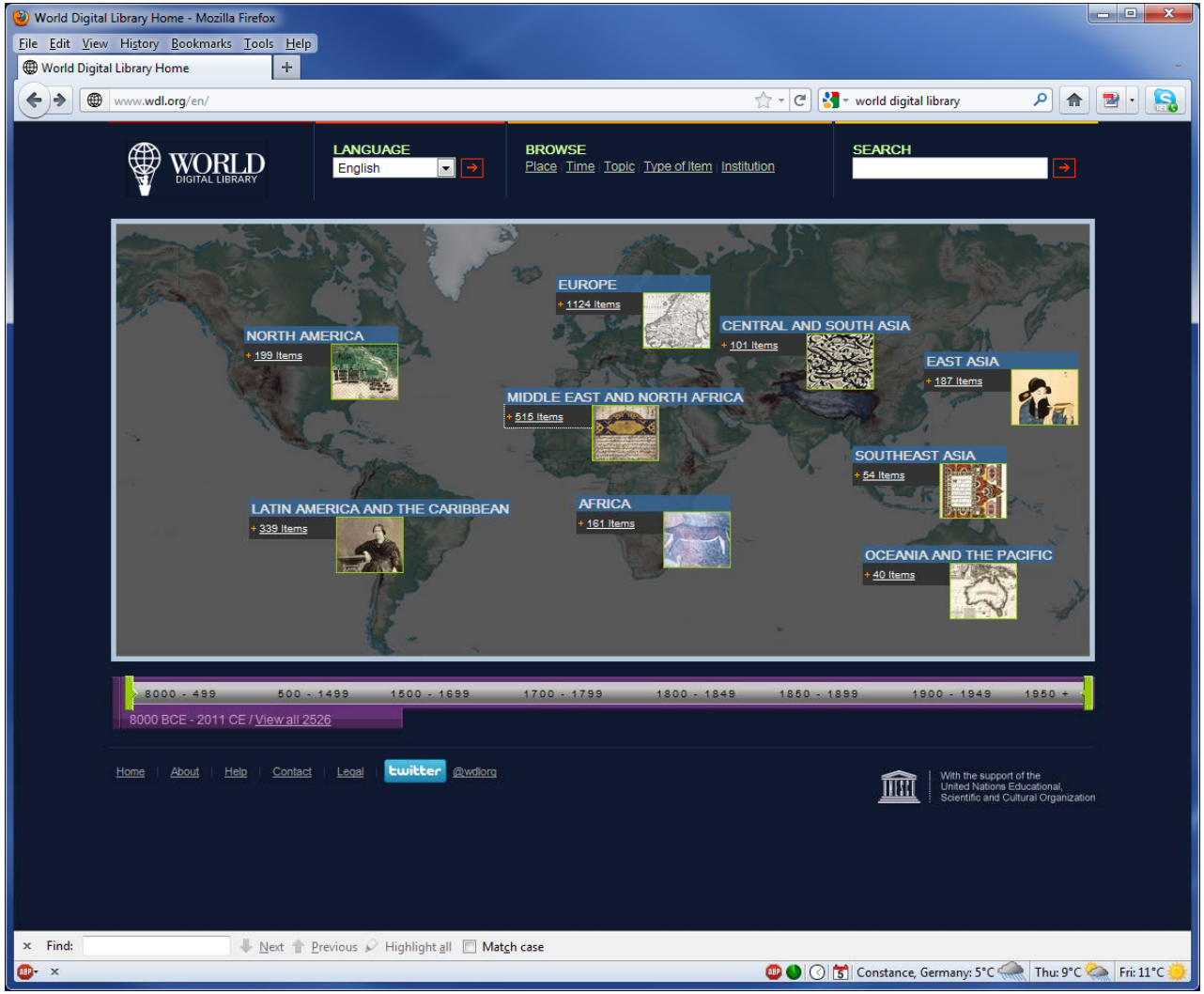
1.3 Digital Library Example Systems

Europeana

1.3 Digital Library Example Systems

Deutsche Digitale Bibliothek

1.3 Digital Library Example Systems



World Digital Library

1.3 Five Elements in Various Definitions of DL

- The digital library is not a single entity;
- The digital library requires technology to link the resources of many;
- The linkages between the many digital libraries and information services are transparent to the end users;
- Universal access to digital libraries and information services is a goal;
- Digital library collections are not limited to document surrogates: they extend to digital artefacts that cannot be represented or distributed in printed formats.

Association of Research Libraries (1995)

1.3 Traditional libraries vs. digital libraries

Challenges: how can the libraries fulfill their role as mediators between information and users in digital time?

- Selection – Definition of collections?
- Acquisition – Physical objects?
- Description – Catalogs?
- Access – Shelves?
- Preservation – Controlled environment?

1.3 What is a digital library?

- A Service? An Architecture?
- A set of Information Resources?
- A set of tools to locate, search, retrieve information?
- Possibly the tools to create such resources and services also fall within the purview of DLs?
- Digital face of traditional libraries ?
- Both digital collections and traditional ones included?

1.3 Where are we now? - DELOS

The DELOS Network of Excellence on Digital Libraries
EU Initiative

Define and conduct a joint program of activities in order to integrate and coordinate the on-going research activities of the major European research teams in the field of digital libraries for the purpose of developing the next generation digital library technologies.

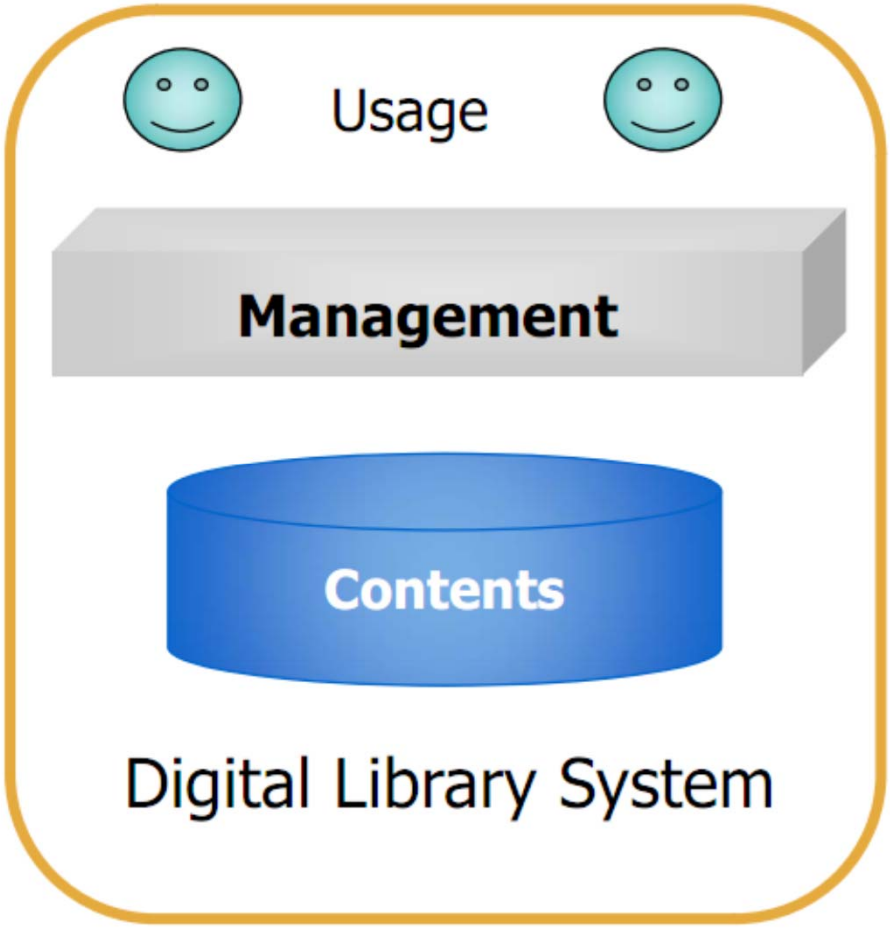
http://www.delos.info/

1.3 What a DL should be like – DELOS visions

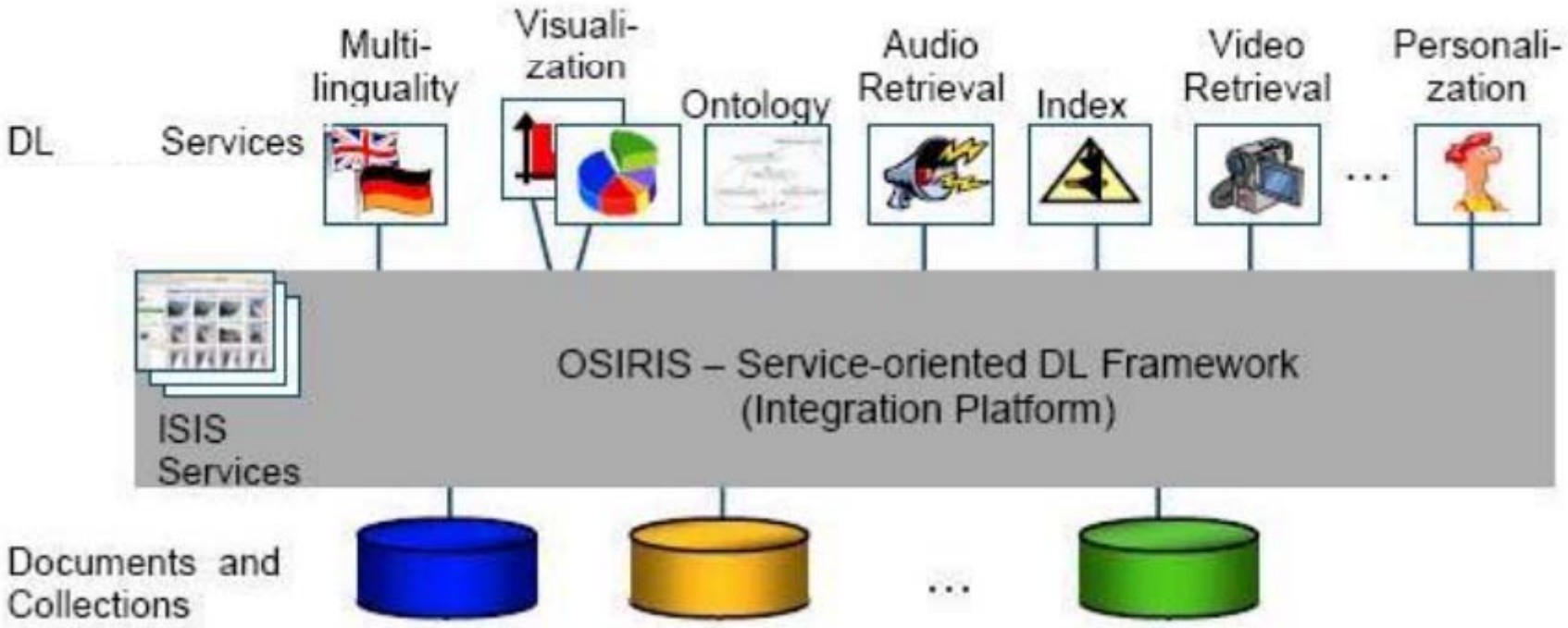
- Digital libraries should enable any citizen to access all human knowledge anytime and anywhere, in a friendly, multi-modal, efficient, and effective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices (about year 2000)

- The potential exists for digital libraries to become the universal knowledge repositories and communication conduits for the future, a common vehicle by which everyone will access, discuss, evaluate, and enhance information of all forms (about year 2005)

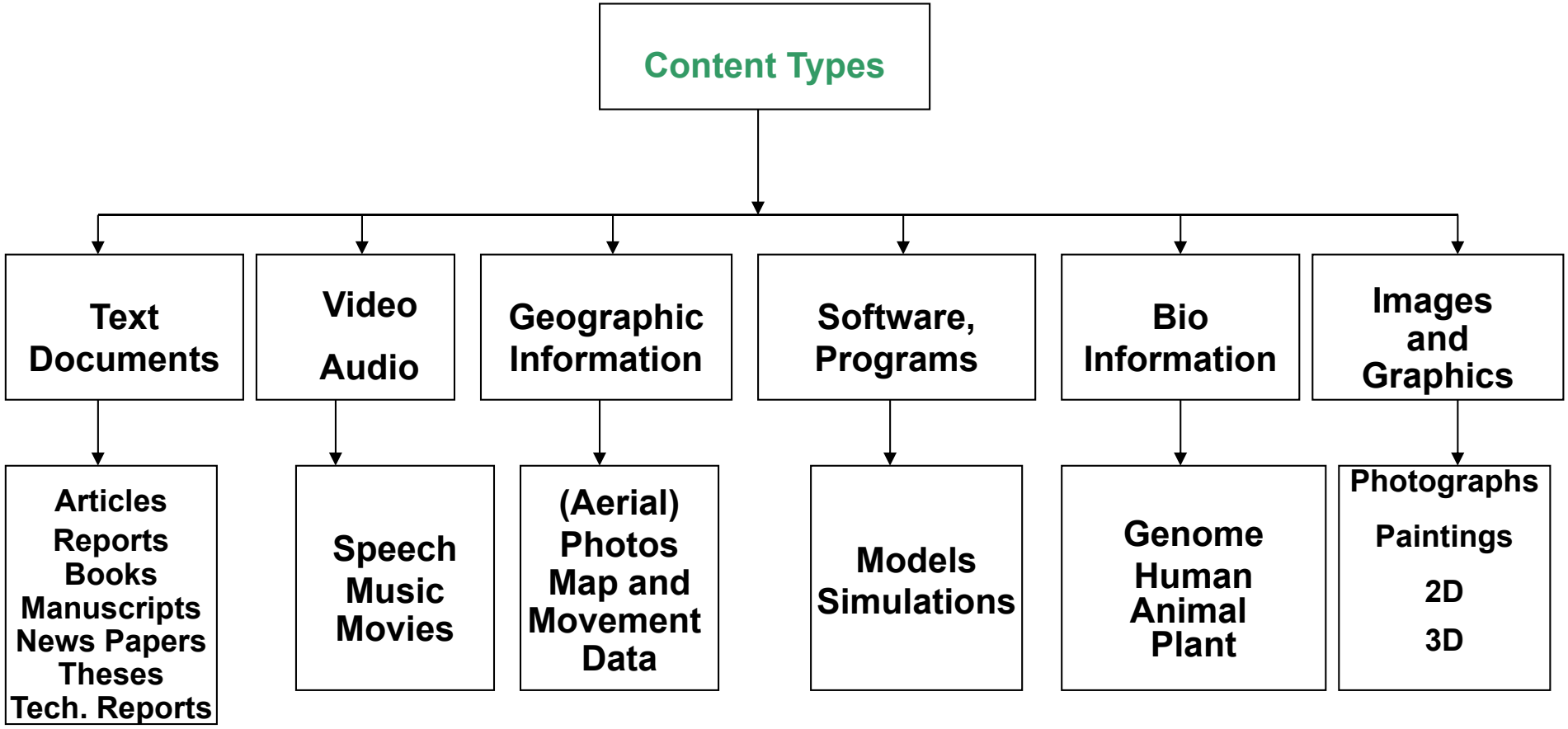
1.3 DELOS Conceptual Framework



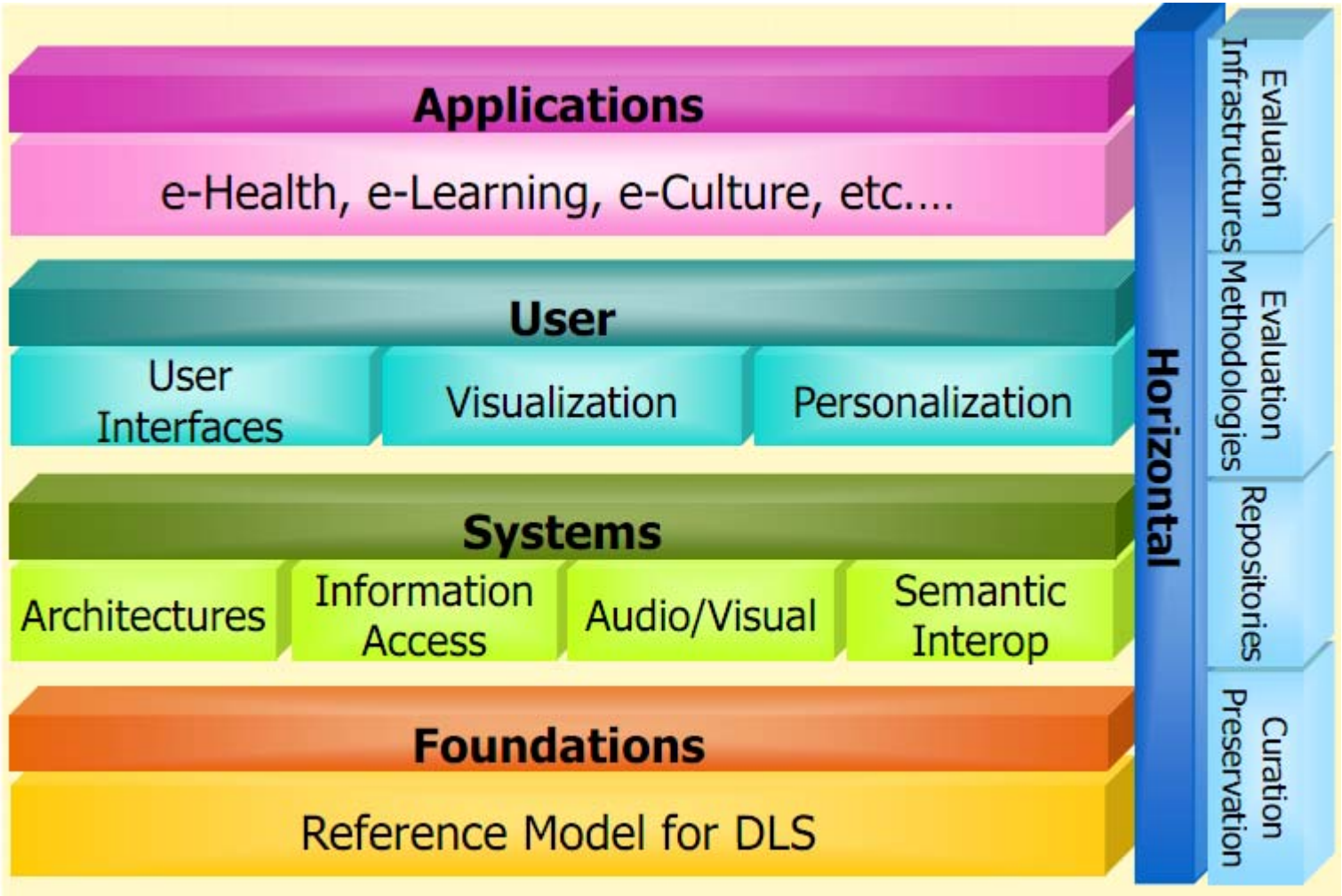
1.3 DELOS DL Management System



1.3 Digital library contents



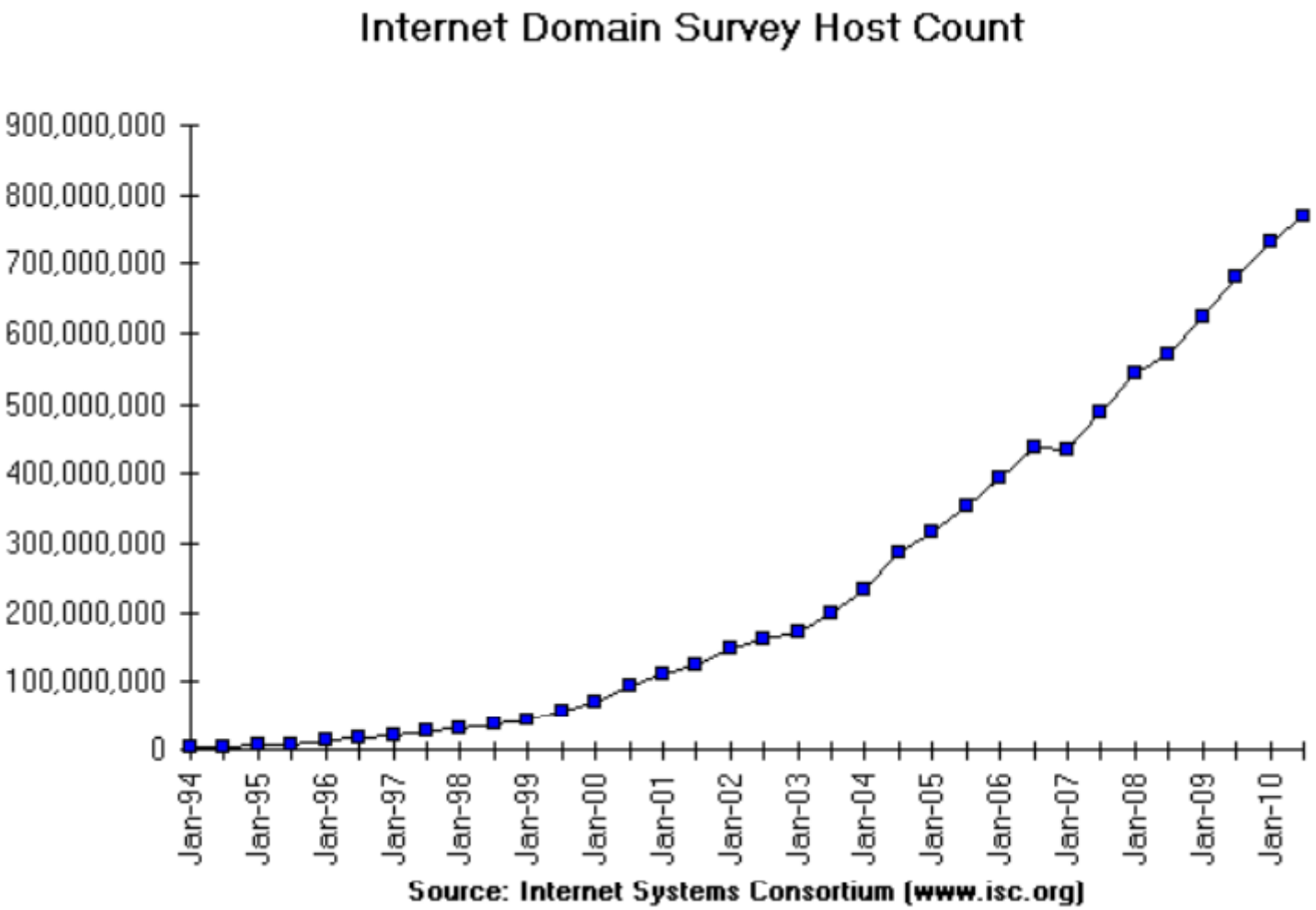
1.3 Digital Library Research



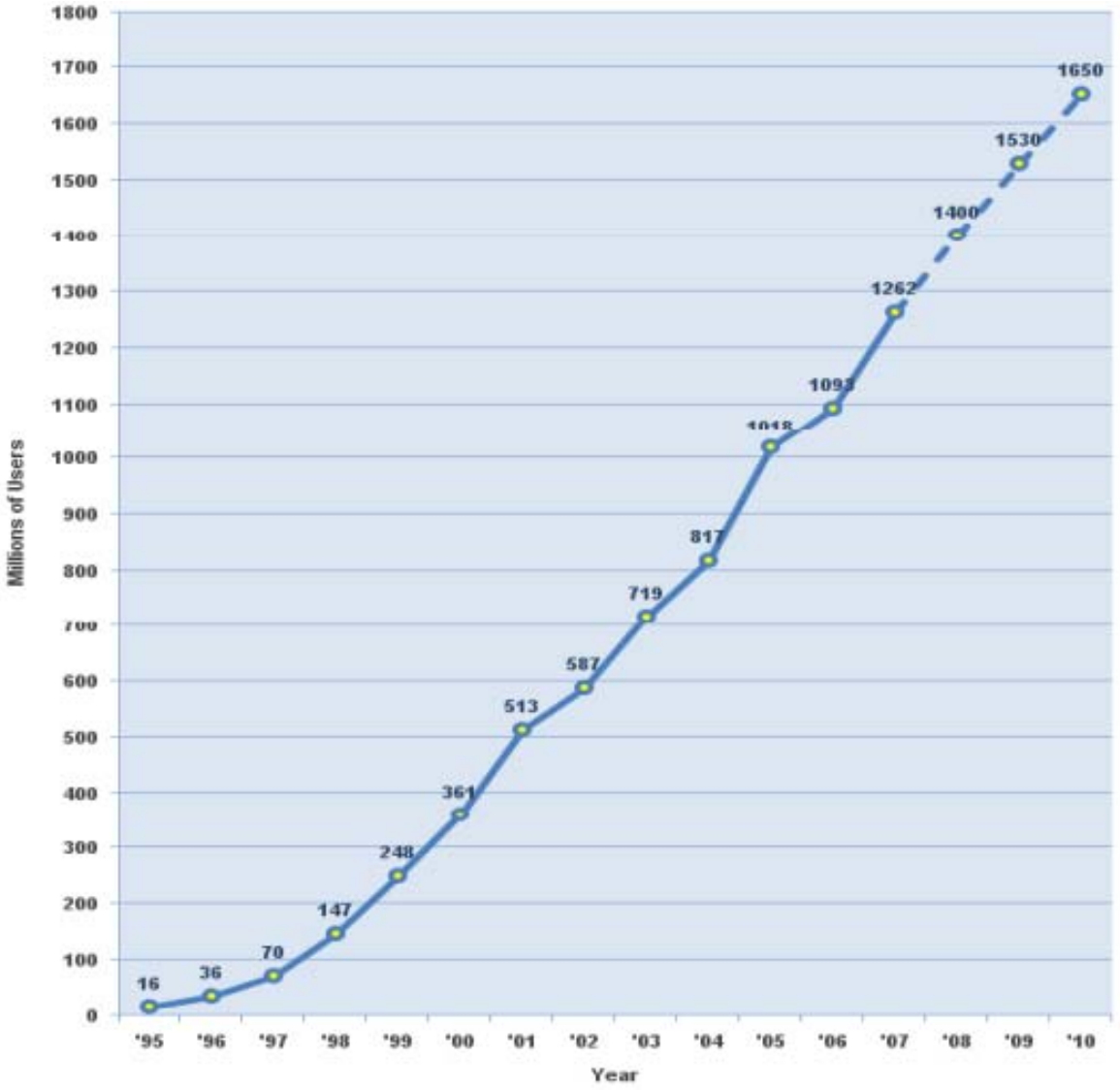
1. Introduction

1.1 The Role of Libraries
1.2 History of Digital Libraries
1.3 Influential Ideas, Definitions, Models
1.4 The World Wide Web
1.5 Lecture Outline

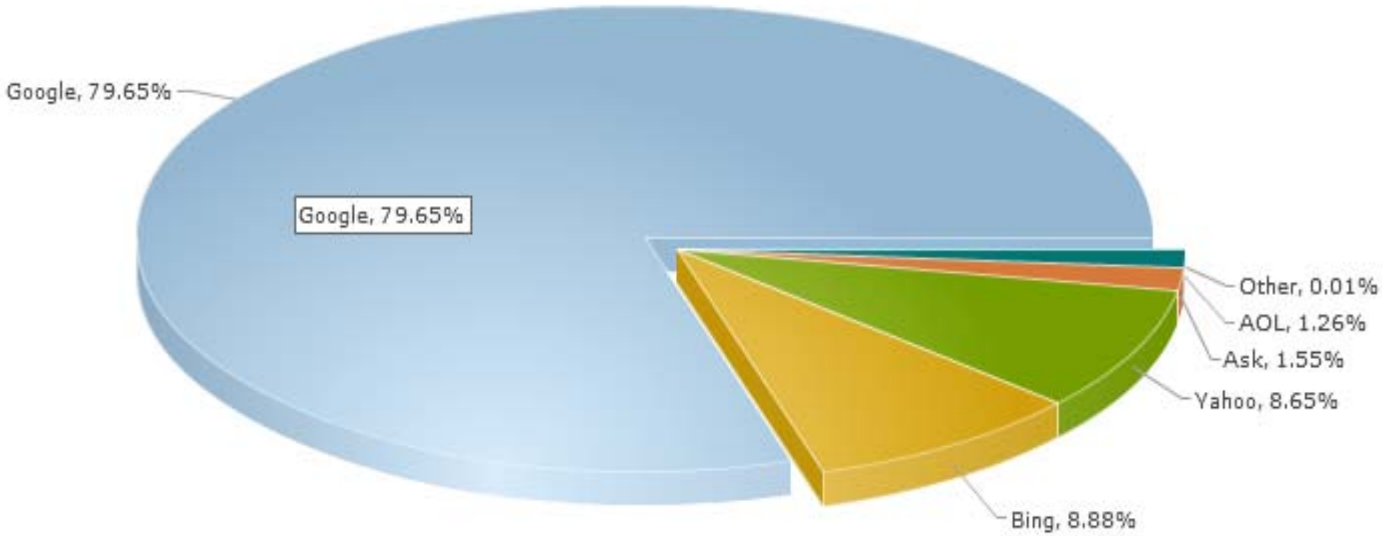
1.4 Growing number of hosts



1.4 Growing number of Internet users



1.4 Growing number of searches



Source: Statowl Internet Usage Statistics, http://www.statowl.com/

1.4 Growth of the World Wide Web

Is it good or bad?

- Rich information sources
- Fast information retrieval
- Convenient - you don't have to go to the library
- …

but …

- Information overload
- Wasteful - US: 300 pound of paper consumption per person/year
- How much time is lost "surfing the web"?
- How to find information when you need it?

Solutions? ***Organization*** **–** ***libraries***

1.5 (Planned) Lecture Timetable

Date	Lecture
20.10.2014	Introduction
27.10.2014	Text Documents
03.11.2014	Images of Pages
10.11.2014	Knowledge Representation
17.11.2014	Collections, Digitization
24.11.2014	Preservation
01.12.2014	Survey of Digital Libraries
08.12.2014	Multimedia Retrieval
15.12.2014	Research Data
22.12.2014	Visual Search and Exploration
12.01.2014	Personalization
19.01.2014	Evaluation
26.01.2014	Economic Aspects
02.02.2014	Current Research Topics
09.02.2015	Exam (take one)

Date	Exercise
07.11.2014	t.b.a.
21.11.2014	t.b.a.
05.12.2014	t.b.a.
19.12.2014	t.b.a.
09.01.2015	t.b.a.
23.01.2015	t.b.a.
06.02.2015	t.b.a.
All exercises are from 13:30-15:00 in G227	

1.5 Literature

- M. Lesk: *Understanding Digital Libraries*. Morgan Kaufmann; 2nd edition (2004).
- S. Rüger: *Multimedia Information Retrieval*. Morgan and Claypool Publishers (2010).
- White and Roth: *Exploratory Search - Beyond the Query-Response Paradigm*. Morgan & Claypool 2009.
- W. Arms: *Digital Libraries*. MIT Press (2001).
- A. Endres, D. Fellner: *Digitale Bibliotheken - Informatik-Lösungen für globale Wissensmärkte*. dpunkt.Verlag, 2000.
- I. Witten: *How to Build a Digital Library*. Morgan Kaufmann; 2nd edition (2009).
- R. Baeza-Yates, B. Ribeiro-Neto: *Modern Information Retrieval - The Concepts and Technology Behind Search*. Addison Wesley (2010).
- J. Zhang: *Visualization for Information Retrieval*. Springer, 2008.
- M. Hearst: *Search User Interfaces*. Cambridge, 2009.
- Hey, Tansley und Tolle: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.