Digital Libraries
Summarization

Andreas Rain

April 7, 2013

Contents

1 Introduction

Libraries are medators between information and users.

There are traditionally 5 basic tasks that have to be done by libraries:

- Selection
  "Definition of Collections"
- Acquisition
  "Physical Objects"
- Description
  "Catalogs"
- Access
  "Shelves, Lending schemes"
- Preservation
  "Controlled environment, media care"

1.1 Why digital Libraries?

- Faster network speeds
- more storage available
- easier accessability

Problems:

- How to provide access? Should there be payment of any kind?
- How to organize digitization of documents?
- Should there be quality standards? If so, how can the quality be measured?
- Indexing, Searching..
- How can the data securely be stored and preserved?

1.2 Memex

Designed by Vannevar Bush in 1945. He invented a machine (which couldn't be built at the time) which is similar to the computer today. It should give the user the capability of searching through your personal library via "hyperlinks" and textsearch. The books are printed on microfilms which can all be read by the machine.

1.3 Similarties between influential ideas

- Availability
- Open access, Sharing
- Greater variety, granularity of information

- Up-to-date-ness
- New forms of rendering
- Integration of digital media into traditional collections
  e.g. Adding digital pictures as references to a book

Five elements that are being shared throughout various definitions of DL:

- Not a single entity
- Linkage technology for resources
- Transparency for linkages
- Universal access to information
- Digital documents that can not be represented in a printed form

1.4 What should a digital library be like?

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).

1.5 Content types of digital libraries

- Text
- Video/Audio
- Geo Information
- Software, Source code
- Bio Information
- Images, Multi-Dimensional Graphics

1.6 Growth of the WWW

Good: A lot of information sources, fast retrieval and conveniency ...

Bad: Too much information, very redundant, searching the web can be an exhaustful task..

2 Text documents

2.1 Text documents in digital libraries

Two basic methodologies:

The ability of the computer to manipulate images: Scanning, Formatting, Displaying, Indexing..

The ability of the computer to manipulate text: Formatting, Searching, Language processing, document conversion...

2.2 Markup, Page description and Style sheets

There are two different aspects a text document should be divided into:

Structure: Helps identifying parts of the text document that are emphasized, finding footnotes, literature and so on. Often a markup spec. such as SGML or HTML helps to define such a structure.

Appearance: When structured, different parts of the document can be displayed in a different manner. Therefore you will need a stylesheet that is compiled for the text document and changes the appearance.

Page description languages such as PDF, PS and Tex take care of both, structure and appearance.

2.2.1 Markup: SGML

SGML is defined by ISO and comes in handy when being interpreted by different instances, such as people and computers.

SGML structure:
Content: Data about the subject

e.g. John Smith
+1-123-456-7890

Information about the content:

e.g. name
phone no.

The structure is hierarichal and modifiers can furthermore define the object (e.g. the phone no. is actually a fax no.).

Example structure:

```
<name> John smith</name>
<phone> +1-123-456-7890</phone>
```

Nested elements create a hierarichal structure (like xml does).

2.2.2 Markup: HTML

Should be well known by now..

2.2.3 Markup: MARC

Machine readable catalog: Was sed by the Library of Congress to make catalog cards readable by machines and was fragmented by different national versions. There is a universial version called "UNIMARC" which is the standard for exchanging information using this format.

2.2.4 Style sheets

Style sheets such as CSS and XSL have been developed to change the appearance of a document. Whereas CSS only defines output styles for given elements, XSL is able to create a whole new output file.

CSS (dis-) advantages:

Advantages:

- Easier maintanance
- Different appearance for different purposes (Desktop, Mobile, Print, ...)
- Separate storing of style definitions
- Can be applied on other sites aswell, easy to maintain a corporate design

Disadvantages:

- Browser compatibility...
  that means browser has to be able to interpret the defined styles. A lot of browsers use their own style definitions that will only work in their browsers. Espacially for sophisticated styles, such as gradients, rounded corners of blocks and so on browsers tend to use there own implementation and do not support the recently defined standards that came with CSS 3.0.

2.3 Searching text

The biggest advantage for digital text documents is that text based searches can be performed on the documents. A lot of problems come with text-based search such as:

- What voc. to use?
- How to handle queries
- Search algorithm of choice?
- Rank items? How?
- and many more...

2.3.1 Linear searching

Allows you to: Search for exact matches and of course regular expressions.

Standard linear search:
You just move the text that has to be matched along the searchstring until it matches.

Boyer-Moore algorithm This algorithm is more efficient than the standard linear search, since it takes advantage of situations where the string can't be matched within the next characters. It goes through the string in reverse instead of forward.

The three main ideas are:

1. Examine the searchstring P from right to left.

2. "bad character shift rule", avoids repeating unsuccessful comparisons

3. "good suffix shift rule", aligns only matching pattern characters against target characters already successfully matched

For a more detailed view, see the lecture slides with example graphics to this algorithm.

2.3.2 Inverted Files

Index the words in the docment by their positions and later on sort the index database. For more than one document see the example below:

Document	Text
1	Gold silver truck
2	Shipment of gold damaged in a fire
3	Delivery of silver arrived in a silver truck
4	Shipment of gold arrived in a truck

Number	Term	Times; Documents
1	a	3; 2,3,4
2	arrived	2; 3,4
...	...	...

Binary search for inverted files: If the listof indexes is already sorted, you can perform a binary search which allow you to search in $O(\log_2(n))$.
Two-stage linear scan: At first you find the approx. position of the word, using a shorter list (e.g. a list that only contains every 100th word) and based on this you scan the larger list.

Advantages: Fast searching, used by sophisticated searching engines such as the google search enging...
Disadvantages: No querying, generation of the list required, update process is difficult, storage ...

2.3.3 Hash tables

For every word, a hash is being computed based on a hash function and this information then is stored. A good hash function is not easily acquired, as seen in the lecture slides. But there are ways to get a "Perfect hash".

Advantages: can be faster than binary search
Disadvantages: Hash table construction and sorting has to be performed before-hand aswell, need for a good hash function, adding values is easy, removing one needs collision checking...

2.3.4 Data structures

A good data structure can improve linear searches of text documents (hash tables are data structures aswell..).

One mentioned in the lecture are "Tries":

This structure allows you to store many words using less space, since you can branch after letters that are in common. E.g. "be - ll, be - ar, b - id" ... . So basically every letter creates a new branch which gives a maximum of 26 branches for each node if you only consider a to z in a lower case.

2.4 Language processing

Most of linguistics is about syntax, but the problems with information retrieval derive mostly from semantics, which is why linguistics are only of little use to this problem. But phrase detection works quite well...

Two types of retrieval errors

- Synonyms are a common error factor
- Ambigous words (rock- music, stone)? How to distinguish? Semantics ...

Thesauri

- gives a label for each idea and then a list of equivalent terms -> Knowledge represenation!

3 Knowledge Representation

3.1 Introduction

A common problem is, that searches return either nothing or too much. This is why knowledge representation attacks this issue, by representing knowledge in a better way so that searches return better results.

3.2 Cassification and Onthologies

Definitions

Classifcation A classification is a structure that organizes concepts into a hierarchy, possibly in a scheme of facetes

Onthology An onthology is a specification of conceptualiation and consists of Concepts (human, animal, food ...), Instances, Properties, Relations and Rules.

3.2.1 Arstoteles Onthology

In "Science of Being" aristoteles proposed that everything in the world can be categorized using the following 10 categories: Substance, Quantity, Quality, Relation, Place, Time, Position, State, Action, Passion.

This concept describes a certain being at a certain time and point in space. There may be more than one instance for the same substance, for different positions, different times and actions.. and so on.

3.2.2 Library Classification Systems

Most libraries created their own classification systems, like

The Library of Congress

Before 1812 they had only 18 different classes in which their documents could be put into. After 1814 Jefferson proposed a newer classification system that e.g. took apart the class "Medicine, surgery, chemistry" and created distinct classes for each and even more.

The British Museum Library

Before 1808 the system of this library rather described the source than the content of books themselves. This is why some classes even were libraries. After 1808 the system reflected the content of the books and a lot of classes have been created.

End of 19th century
Big classification systems like the Dewey system (1876), New Library of Congress System (1898) and in Europe the Universial Decimal Classification have been invented. They created Classes that could be split into subclasses. E.g. "Scienece" was a super class and could be split into different scientific subjects.

3.3 Indexing: words and thesauri

A thesaurus is a structure that manages the complexities of terminology and provides conceptual relationships, ideally through an embedded classification/onthology. (Basically it can find semantic links).

3.3.1 Finding the exact meaning of a word

Geoffrey Sampson probabilistic approach using statistical methods to determine what a word is.

Garside 1987 pair of words at a time.

Ken Church 1988 trigrams, better results.

Lesk 1986 count overlaps between the definitions of different sense of nearby words.

3.4 Metadata & Semantic Web

What is metadata?
Common: Data about data, describing the object.

What's it good for?
Better description, information about the data, historical information, properties summarization, standard description.

There are proprietary metadata definitions like for JPEG, Probado 3D, PDF, .DOC and so on. But there are also standardized metadata definitions.

3.4.1 Dublin Core

is a set of name/value pairs that describe online resources. There are 15 elements in the original definition:

Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title and Type.

Since the original definition in 1995 many more elements have been added.

The elements can be devided into three categories such as Content elements (Coverage, Description, ...), Intellectual property elements (Contributor, Publisher, Creator...) and Instantiation elements (Date, Format, ..).

3.4.2 Semantic Web

is an attempt to provide information that is held by databases using the web. Programmers can access special data files that are built upon a special data scheme and therefore are able to interpret them. A controlled vocabulary helps to do this.

3.5 Vector models

A document can be represented by a vector of terms: Words, Phrases. Correlations between these vectors can be a sign of similarities of the documents.

Values tht can be used: Boolean (Absence, Presence), Term frequency, Document frequency, and TF-IDF (tf * inverse document frequency).

3.5.1 Vector representation

Weight terms high if they are frequent in relevant documents and also infrequent in the collection as a whole. So you can assign a weight to each term and each phrase.

3.5.2 Standard IR Eval.

Precision: $\frac{relevant}{retrieved}$
Recall: $\frac{relevant}{relevantincollection}$

3.5.3 E-Measure

The E-Measure combines precision and recall into one number using

$$E = 1 - \frac{b^2PR + PR}{b^2P + R}.$$

4 Images of pages

4.1 Scanning

Different types: Usual scanners, Microfilm scanners, 3D Scanners..

Advantages: Very economic and the common way to make analog material machine-readable.

4.1.1 Quality

300 DPI is good to converse digital text that has been printed into an image again. But for archival purposes a much higher resolution is needed.

Anti-aliasing is used to improve the result of the pictures. Pixels are treated as areas and over a wider range some grey-scales are being added so it looks smoother.

4.1.2 Advantages / Disadvantages of Images of Pages

Advantage: Familiarity and easy to create
Disadvantage: Big size, not easily adaptable, copying is not easy..

4.2 Conversion

OCR is used to converse scanned images into plain text. The technologies have improved over the past but are still far from perfect. They can now be over 99% accurate which is a considerably nice result.

4.2.1 Image formats

Lossless compression: file size reduced without the loss of quality.
Lossy compression: taking advantage of limitations of the human eye, information that couldn't be seen anyway is taken away.

4.3 Indexing images of pages

- Index description text of the image
- Index by other criteria
- Index using OCR

4.4 Shared text image/system

This topic is about pages that consist of more elements e.g. text, tables, equations, figures and schemes.

CORE (chemical online retrieval experiment): methods to solve the above issue have been developed.

- deskewing pages
- equations and table in the database
- figures are captured using the caption word ”figure”
- bit density plot - separate text from illustrations

5 Collections, Digtization and Preservation

5.1 Lirary Collections

Traditional libraries: Usually printed documents, volume is important.
Digital libraries: Quick delivery, View-on-demand.

5.1.1 Quality control

Traditional: Accuracy, timeliness, references, methodology, novelty, reproducability. Typically provided by publishing house. Peer-Review. Also: Bibliographic measures: Impact factor, citation statistics.

Digital documents: More difficult due to lower costs. Version control is hard to implement.

5.2 Preservation issues

5.2.1 Traditional material

Issues: Paper bleaches out, gets instable in its structure, the same counts for books. Vinyl gets scratche, CDs aswell.

Cheap solution: Microfilm. It has a long durability, is inexpensive to produce and even less expensive to copy. But: the medium is far from user friendly.

5.2.2 Digital material

Advantages

- lossless copying
- inexpensive copying
- redundantly distributable

Disadvantages

- Not preservable for a long time
- complex distribution has to be developed in order to preserver digital media
- formats can expire and may not be readable in a forseeable future
- a lot of material, which to preserve? quality, peer-review, citations ...

5.2.3 Arm's three material categories

Obviously valuable: This means it has to be preserved and indexed
Obviously worthless: Discard
Unsure (90%): Has to be checked by a curator

5.3 Digitization of Collections

Types of items: Originals, Prints/Copies, High volume - low risk (text books)..

5.3.1 Collections in a library

Collects, organizes, serves and preserves. Is user oriented, wants to provide information to users. Avoids the multiple digitization of high volumen - low risk items.

5.3.2 Collections in an archive

Collects, sorts and preserves. Doesn't have the need to provide the information, just to preserve it. Therefore collections of documents have to be systematically collected. In most cases the originals are being preserved.

5.3.3 Collections in a museum

Collects, exhibits, preserves and makes research possible over items that are either original or duplicates thereof.

5.3.4 Digitization workflow

1. Plan -technical goals, content goals, financial goals
2. Preparation - access to material, transport/insurance, identification, scanning
3. Digitization/Hard-/Software - scanning, storage, postprocessing/indexing
4. Parameters - Resolution, Colors, Postprocessing options..
5. Quality Control - OCR, signatures, image quality check
6. Storage / Meta Data - Lossy/Lossless

5.4 Sharing & Distribution

Traditionally storage, budget and user-req. trade-off..

Accessing models: Open access, Loan systems, Magazines, Fair usage exceptions,..-

6 Survey of Digital Library Systems

Motivation for building a DL:

- Web
- Low barriers
- Political competition

6.1 Text-oriented DLs

Project Guttenberg: Digitizing copy-right free books, community-based, more than 36000 books, export formats vary (HTML, Kindle, ePub, ASCII).

Gallia: Scanned 1.5M docs., Focus on history, culture, french literature...

Making of US & A: 19th century material, Education, Psychology, American history, Sociology, Religion, Science... 10000 books, 50000 articles...

JSTOR: 1000 complete journals, 12M Pages, about humanities, social sciences and others

Summary: Many prjects, large variety, not easy to estimate what has been done...

6.2 Image-oriented DLs

State library of victoria: 100k pictures from 650k picture collection digitized so far, history of victoria, paintings drawings, prints...

American Memory Project: Millions of pictures scanned, Record of American history..

ARTstor: >1M images, Art, Architecture.., Community-based

Implications: Manual indexing of images is expensive, Object recognition difficult, many domain-specific solutions...

6.3 Audio-oriented DLs

US national gallery of the spoken word: Speeches, Radio recordings and so on..

Animal sound archive: 120000 bioacstical recordings, most of material is on tape, digitization is an ongoing process...

Variations project: Audio, video, score images, score nations, >10000 recordings

Probado: Score sheet msic digitization, extracting from audio cds, content-based indexing, synchronization

6.4 Video-based Projects

Vanderbilt news archive: News broadcasts from ABC, NBC, CBS ..., Major news events, Digitization ongoing, access by sbscription..

Challenges: indesing videos manually is expensive, autmatic processes include image analysis, NLP (speeh recognition) and face recognition and matching.

6.5 3D Objects

Probado 3D: Collects 3D objects, manages metadata, provides information and objects...

6.6 Scientific Data

Example DLs: Sloan digital sky survey, PANGAEA, Protein Data Bank and many more..

Probelms: Too much data, Search is difficult...

7 Multimedia retrieval

Library Workflow Model

1. Acquire

2. Markup
3. Index
4. Server
5. Archive

Key challenges: Understanding document type, subject domain and user req.

Understanding of subject domain and user reqs.: Which material to acquire (Same as with collections...)? What is relevant? How to index? Querying? Presentation?

7.1 Content-Based Search

- Identify/measure properties useful to describe content
- Aggregate measures to create a descriptor (feature vector, graph)
- Define distance measure
- Searching modality: Query-by-example

7.1.1 Modeling of the problem "Jelly bear"

Image as function: $\overrightarrow{c}(n,m) = \begin{pmatrix} r(n,m) \\ g(n,m) \\ b(n,m) \end{pmatrix}, r,g,b \in \{0,...,255\}$

Descriptor based on average color: $\overrightarrow{d}_f = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \overrightarrow{c}(n,m) \cdot \frac{1}{N \cdot M}$

Distance: $Q(i,j) = |\overrightarrow{d}_f^{[i]} - \overrightarrow{d}_f^{[j]}|$

7.1.2 Wealth of Descriptors and Feature Selection

- Images: color, texture, shape...
- Video: Objects, Movements, Text, Speech,...
- Graphs: Topological features..
- Chemical Compounds: Atom contm types, functional properties

To find a good feature selection, experimenting with different features has proven to have good results.

7.1.3 Generic architecture for content-based retrieval

See lecture slide 18 in chapter 7.

7.2 Distance Measures

Metric properties:

1. $d(x, y) >= 0 (non - negativity)$
2. $d(x, y) = 0 \equiv x = y (definitivity)$
3. $d(x, y) = d(y, x) (symmetry)$
4. $d(x, y) <= d(x, z) + d(z, y) (triangleinequality)$

7.3 Query modalities

7.3.1 Text-based

A query is being done using a text search. The result then can be of any kind of media. There has to be a mapping and the software has to find the results for the given query.

7.3.2 Example-based

Lazy learning: nearest neighbors are found on runtime while executing the query. No learning process needed, the software learns on the fly.

7.3.3 Sketch-based

To do this more easily digital objects are being converted into sketches. Then a distance can be measured more easily. For a deatiled example see ls. 31 / ch. 7.

This highly depends on the sketching skills of the user.

7.4 Explorative search

In this approach a browsing mechanic has to be provided to the user. The user doesn't know what he wants to find until he found it. Therefore he should be able to browse the data in a nice manner.

8 Research Data

What is Research data? *Defined by factual records, use as source for scientific research, commonly accepted in scientific community, systematic, partial representation of the subject being investigated, digital, computer-readable format, marginal cost of distribution.*

8.1 Fourth Paradigm

- Large scientific projects allocate large parts for infrastructure
- Need for Laboratory
- Need for scalable data analysis tools

- self-describing and schematized data formats
- interoperability of data and literature

8.2 Publishing data?

Data can be published either synchronous or asynchronous to to the article is being published with.

Asynchronous: The author can send the data to a data archive whilst a journal peer-reviews the article. They can be published distinct from one another. ls. 11 / ch. 8

Synchronous: Again, the data sets go to a data archive and the article to a journal. But the editor and reviewer from the journal are in contact with the data curator from the data archive. Both are being published together later on (if accepted by both instances ofc.). ls. 12 / c. 8

8.2.1 OECD Guidlines

Purpose: Improve efficiency, effectiveness of global science system.

Principles (1)

- Openness - low cost, internationally available, easy to use
- Flexibility
- Transparency - how has the data been created? reproducable?
- Legal conformity - national security, privace, trade secrets ...

Principles (2)

- Protection of intellectual property
- Formal responsibility
- Professinalism
- Interoperability - Internation documentation standards have to be met
- Quality - Data curation
- Security - Integrity, Completeness, Security of research data has to be guaranteed.

Principles (3)

- Efficiency
- Accountability
- Sustainability

8.3 Example data repos

Oak Ridge National Laboratory Distributed Active Archive Center, PANGAEA, NASA's Earth Observing System ... See ls.

8.4 Data Cite Consortium

Founded in 2009 with the Goals to provide access to scientific research data in the internet and increase acceptance of research data as legitimate, citable contribution to scientific records.

The datacite consortium has 19 members and datasets are being registered using the DOI concept. Additionally they provide services for Metadata searches and an OAI Interface.

When registering a dataset, metadata about the object has to be provided. Mandatory metadata are: Identifier, Creator, Title, Publisher and the PublicationYear.

8.5 Visual search in research data

VisInfo project: Content-based Visualization of time series data. This project allows for an explorative search on time series data.

Backend workflow:

1. Scientific primary data repo.
2. Import data
3. Time series database
4. Preprocessing
5. Preprocessed database
6. Feature extraction
7. Time series descriptors
8. Clustering

The users then can browse through the visualized time series data and analyze it, search in it and so on.

8.5.1 Bivariate data

Regressional Feature Vector: The idea is to to perform regression tests on the data and pick the one that fits the data best.

9 User search and exploration

Main retrieval cases:

- Structured data: Databases
- Text collections
- Non-Textual and multimedia data

Classic approach for a user search: **Look-up based**

- Information needs to be well-defined
- User has to know the domain
- Query definition
- Often just a one time action
- Visal or command line

Disadvantages:

- Not how humans usually interact with information
- Non-dynamic search process
- Does not consider context of information need and use

Alternatives:

- Interactive approaches
- Explorative search systems
- Query-less systems

9.1 Typical user tasks

Querying: specify require content, precise formualtion, understanding query language and domain aswell..

Browsing: Navigation, Discovery, May be followed by queries

Analysis: Identify groups in data, relationships, correlation, evaluation and data mining..

9.2 Search models

Why?: Support human search behaviour / describe it, search interface design foundation, many models for many purposes..

9.2.1 Curiosity and browsing

By berlyne in 1960: Analyzed human nature of curiosity and rooted the model on it.

Two types: Specific curiosity (you know what you'sre looking for), and diverse curiosity (you want to browse, you don't know what you're looking for..).

9.2.2 Information Foraging Theory

Foraging (”nahrungssuchend” o.a. ”wissbegierig”), refine the query result selt by querying on it again.

9.2.3 Berry Picking Model

Information items are picked once at a time. Each item gives new perspective and influences the next picking step.

9.2.4 Sensemaking and Situational Awareness

People want to connect facts (people, places, events) to anticipate developments and act on that: Recognize the knowledge gap, create a mental model, search for information, analyze and synthesize information...

9.2.5 Information seeking model

See ls. 19 / ch. 9..

9.2.6 Information Visualization Reference Model

See ls. 20 / ch. 9..

9.3 Explorative search

If the search goal is ill-defined an explorative search can be of use.

Useful features that should be provided:

9.3.1 Querying and rapid query refinement

Refinement of keywords and auto-completion (see google search), Suggestions and of course immediate search results for a fast search. This gives the illusion of an interactive search.

9.3.2 Offer facets and metadata based on result filtering

Filter the results by facets (e.g. metadata fields), give the possibility to cluster them and provide an overview (classes, categories, onthologies)..

9.3.3 Leverage search context

Provide a highlighting of relevant objects so the query can be refined base on the relevant items

9.3.4 Offer visualization to support insight and decision making

Provide visual help and make use of the human visual processing power.

9.3.5 Support learning and understanding

Delivery meaning-ful results, e.g. explanations of why the result is in the resultset. Provide novice users with search trails of expert users and only show an overview of documents to novice users.

9.3.6 Facilitate collaboration

Perspectives for many users, pool expertise. Provide a collaborative search interface to the user.

9.3.7 Offer histories, workspaces and progress updates

Explorative search may take more than one session. Revisit previous sessions (Structured history, search progression, workspaces).

Show the space of the remaining data so the user can have an overview.

9.4 Overview and Navigation

9.4.1 Browsing

Challenges for the visual interface:

- Intuitive and effective navigation
- Show importan aspects
- Structured/Sorted list of documents

9.4.2 Application examples

See the ls. 47-53 / ch. 9