Libraries in a data-centered environment

Jakob Voß

Ticer Summer School, August 22th, 2012

Prolegomena

The importance of Data

The importance of Libraries

Summary

Appendices

Section 1

Prolegomena

So what about the Cloud?

- It‘s a hype
- It’s a buzzword (cloud = bullshit)[1]
- Better know exactely what is referred to by “cloud”

Which notion of cloud do libraries refer to?

[1] to impress and persuade, unconcerned with falsehoods (Frankfurt, 2005)

Three notions of the Cloud

Figure: Infrastructure as a Service (IaaS)

Figure: Platform as a Service (PaaS)

Figure: Software as a Service (SaaS)

Software as a Service (aka “web application”)

- Software that you don’t have to install or update.
- Software that hides some of its complexity.

Any software is inherently more complex then the task it automates. Don’t expect software to simplify anything!

Section 2

The importance of Data

Data vs. Applications

"Data matures like wine, applications like fish"
— James Governor

Data vs. *Applications*

- For immediate consumption
- Requirements and business logic change
  - Technical developments and trends
  - People's requirements change

Data vs. Applications

- Can be used in different context and times, if it is well done:
    - respect special properties of data
    - respect different notions of data

Special properties of data

Bits can freely be rearranged. Eventually, data

- can be copied
- can be modified

very efficiently, without any traces or differences between “original” and “copy”.

Special properties of data

Digital Collections and descriptions of data are data again.

Data challenges

This is where libraries are needed!

- Preservation
- Authenticity
- Provenance
- Identity

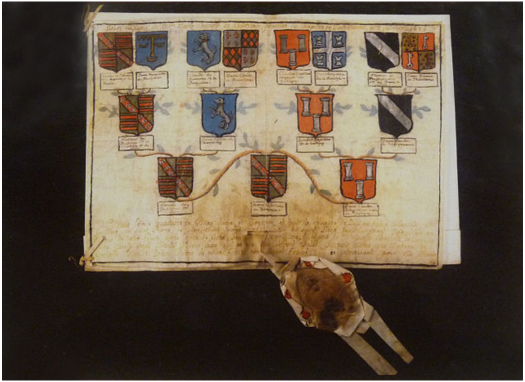
Data challenges: Preservation

- All data needs a carrier
- Unsolved problem in general, but established discipline

Data challenges: Authenticity

- Data modification leaves no traces
- Related to preservation but more about trust

Data challenges: Provenance



- Data copy leaves no traces
- Digital signatures and trust (again)

Data challenges: Identity

- A single bit changes the whole dataset
- Which modifications matter?

Notions of data

- It's also a hype
- It's not a buzzword (data $\neq$ bullshit)[2]
- Better know exactely what is referred to by "data"

Which notion of data do libraries refer to?

[2] Not primarily used to impress (but also used without care)

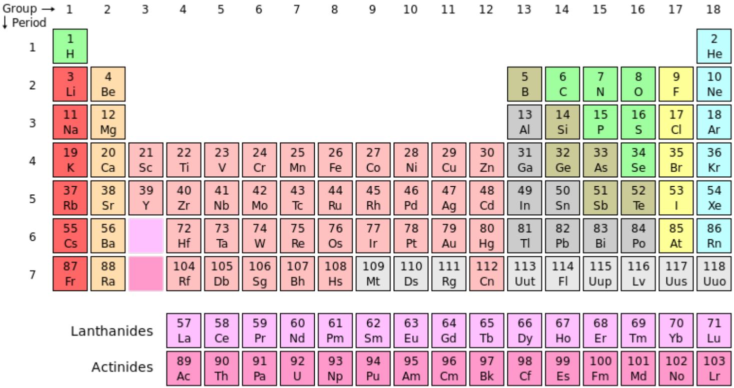
Three notions of data[3]

- Data as facts
- Data as subjective observations
- Data as communications

[3] As identified by Ballsun-Stanton (2012)

Data as facts

- Hard numbers, product of reproducible measurements, scientific facts
- Used to reveal (the real) world



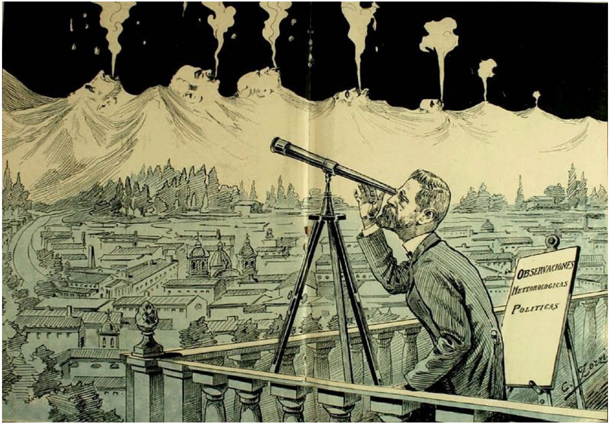
Group → ↓ Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba		72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra		104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo
Lanthanides				57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
Actinides				89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Data as facts in libraries

- Created by libraries
  - Holding counts
  - Patron information
  - Formal metadata
- Collected by libraries
  - research data

Data as subjective observations

- Product of recorded observations, sense-impressions that must be filtered
- Used to construct (our) reality



Data as subjective observations in libraries

- Created by libraries
  - Subject indexing
  - User studies
  - Analysis of publication trends
- Collected by libraries
  - research data

Data as communications

- Transferred or stored sign, a container of meaning in form of sequence of bits
- Used to describe (any) reality

Data as communications in libraries

digital objects, electronic resources, informational objects, electronic publications, . . . , **digital documents**

- Collected by libraries

Data as communications/documents in libraries

A **document** is not information but a recorded “evidence in support of a fact” (Briet, 1951),[4] which can be any possible statement. This notion somehow got lost in the history of library and information science / documentation science (Ørom 2007).

- **Advice:** Don’t mess with research data as facts or as observations but treat them as documents, like other (digital) publications!

[4] See Buckland (1997,1998) for an introduction.

The importance of data in libraries

- More and more communication and publication is digital.
- If libraries still care for documents, they have to care for data.

Does your library actually care for data or is activity around data outsourced to the "tech people"? Would you outsource any activity around books to the "book people"?

Section 3

The importance of Libraries

What does a library do?

A library collects, arranges, and makes available (published) documents (among other services) to meet user needs.

This should also apply to digital documents:

- **collect** data
- **arrange** data
- **make available** data

Collect data

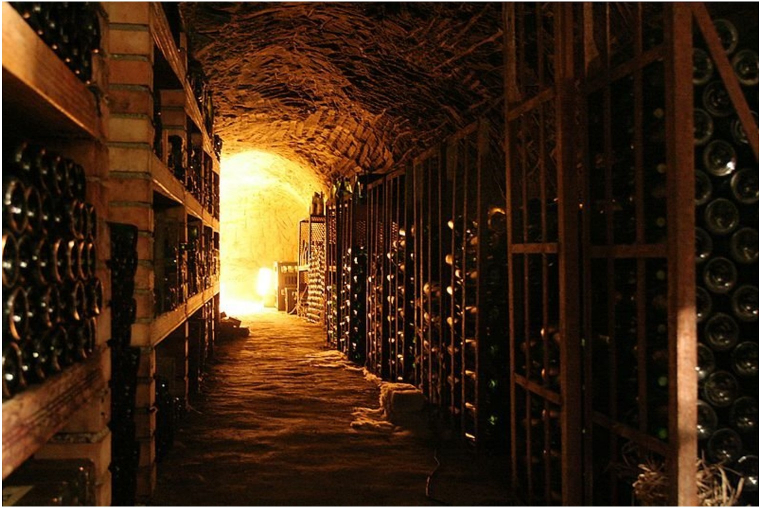


Figure: Data needs care

Figure: How many libraries store digital objects

The eResource fallacy

If libraries license eResources to be accessed from publisher sites, libraries are only temporary, intermediary retailers.

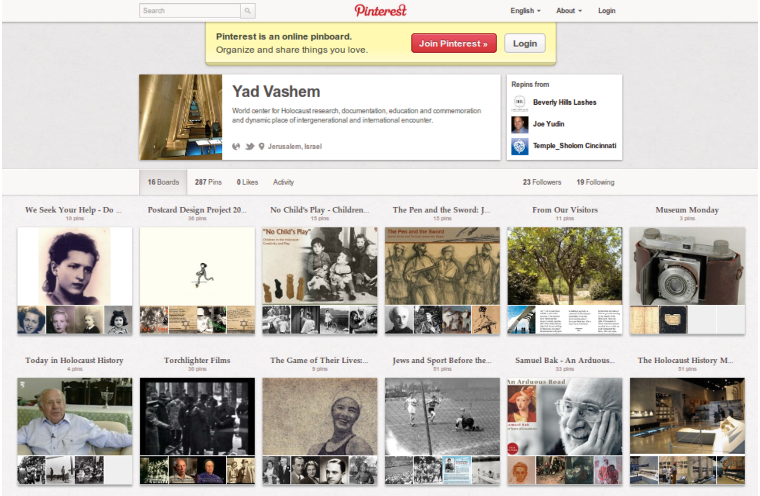
- **Advice:** Data that cannot be **copied and modifed** is lost. Actually collect and process digital documents (or you won't be in the document business anymore)

Why and how can libraries care for digital documents?

- Libraries are used to deal with documents (data as communications)
  - collect digital document
  - arrange digital documents
  - make available digital documents
- Libraries can respond to the **data challenges**, because of:
  - Trust
  - Neutrality
  - Persistence
- Ensure that documents can be used as data:
  - copying
  - modification

Example: Collections

- Create and present your digital collections
- in different forms on different platforms



Pinterest is an online pinboard.
Organize and share things you love.

Yad Vashem

World center for Holocaust research, documentation, education and commemoration and dynamic place of intergenerational and international encounter.

Jerusalem, Israel

Repins from

Beverly Hills Lashes

Joe Yudin

Temple_Sholom Cincinnati

16 Boards | 287 Pins | 0 Likes | Activity | 23 Followers | 19 Following

We Seek Your Help - Do ...
10 pins

Postcard Design Project 20...
36 pins

No Child's Play - Children...
15 pins

The Pen and the Sword: J...
15 pins

From Our Visitors
11 pins

Museum Monday
3 pins

Today in Holocaust History
4 pins

Torchlighter Films
30 pins

The Game of Their Lives:...
9 pins

Jews and Sport Before the...
51 pins

Samuel Bak - An Arduous...
33 pins

The Holocaust History M...
51 pins

<!–

flickr von YAHOO!

Sie sind nicht angemeldet Anmelden Hilfe

Example: Annotations

One must be able to

- annotate your documents
- link to (parts of) it
- to track changes when reusing it (revision control)

Prologus.

Valerij Maximi Romane urbis iuriſperitiſſimi: in librũ factorũ et dictorum memorabiliũ ad Tiberiũ ceſarem Prefatio incipit.

Vrbis Rome exterarumq; gentiũ facta ſimul ac dicta memoratu digna / que apud alios latius diffuſa ſũt / ut quambreuiter cognoſci poſſint: ab illuſtribus electa auctoribus / deligere cõſtitui. ut documenta ſumẽ uolentibus: longe inquiſicõnis labor abſit. Nec mihi cuncta cõplectẽdi cupido inceſſit. Quis eni omnis eui geſta modico voluminũ numero cõprehenderit? Aut quis compos mentis: domeſtice peregrineq; hiſtorie ſeriem felici ſuperiori ſtilo conditam: vel actẽciore cura: vel preſtanciori facundia: traditurũ ſe ſperauerit? Te igitur huic cepto penes quẽ hominũ deorũq; cõſenſus / maris ac terre regimen eſſe voluit certiſſima ſalus patrie ceſar: inuoco. cuius celeſti prudentia: virtutes de quibus dicturus ſum benigniſſime fouentur. vicia ſeueriſſime vindicãtur. Nã ſi priſci oratores ab ioue optimo maximo / bene orſi ſunt: ſi excellẽtiſſimi vates a numine aliquo principia traxerũt: mea paruitas / eo iuſtius ad fauorẽ tuũ decurrerit: quo cetera diuinitas opinione colligitur tua. pñti fide paterno auitoq; ſideri par videtur. Quorũ eximio fulgore: multũ cerimonijs nr̃is inclite alacritatis acceſſit. Reliquos eni deos accepimus: ceſares dedimus. Et quoniã iniciũ a cultu deorũ petere in animo eſt: de conditione eius ſũmatim diſſeram.

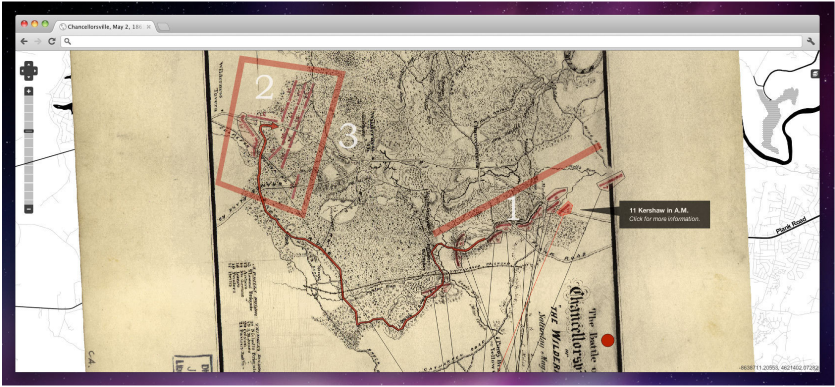
Deſinit prefatio.

Tituli primi libri.

¶ De religione. Ca. primũ. ¶ De neglecta religione. ca. ij. ¶ De ſimulata religione. c. iij. ¶ De auſpicijs. c. iiij. ¶ De

et hiſtorie

S.

 [5]

[5] Netline.org screenshot by David McClure with map tiles by Stamen Design, under CC BY 3.0, data by OpenStreetMap, under CC BY SA, maps from the Hotchkiss Map Collection at LoC

Section 4

Summary

The situation

- Software is inherently complex and becomes obsolete
- Data ages less rapidly if it can be used in **different context and time**

In the end all content will be digital – get used to it!

Care for data!

"Data that is loved tends to survive" — Kurt Bollacker

- respect special properties of data
  - copying and modification must be possible
    - Be able to store whereever you like
    - Be able to transform to any form you like
    - Be able to combine with any data you like
  - digital collections and metadata are also data
- respect different notions of data
  - focus on data as communications instead of digging into details of research data
- Provide solutions to the data challenges
  - preservation
  - authenticity
  - provenance
  - identity

Where to start

- Collect digital publications!
  - Start archiving public websites, blogs, mailing lists etc.
  - Create and manage data/document repositories (see yesterday's talks)
  - Invest in preservation
  - Exchange digital documents with other libraries and initiatives (see following talk by Herbert van de Sompel, LOCKSS...)
- Provide data as accessible as possible (Open Data)
- Publish your own digital publications
- Allow annotating and connecting with your digital documents
  - simple access will not be the primary role of libraries
  - Data needs to be usable in different context and times

Section 5

Appendices

References

- Ballsun-Stanton, Brian (2012): Asking About Data: Exploring Different Realities of Data via the Social Data Flow Network Methodology. PhD thesis
- Briet, Suzanne (1951): Qu'est-ce que la documentation? Editions documentaires, industrielles et techniques
- Buckland, Michael (1997): What is a "document"? In: Journal of the American Society of Information Science (JASIST) 48.9, pp. 804–809
- Buckland, Michael (1998): What is a "digital document"? In: Document Numérique 2.2, pp. 221– 230
- Frankfurt, Harry G. (2005). On Bullshit. Princeton University Press
- Ørom, Anders (2007): The concept of information versus the concept of document. In: Skare et al. (eds.) : Document (re)turn. Contributions from a research field in transition. pp. 53-72. Peter Lang

Image Credits and licenses

All images from Wikimedia Commons:

- construction.jpg CC-0 by Schweinepeterle (Rolf H.)
- cubicals.jpg CC-BY-SA by David R. Tribble
- mcdonalds.jpg CC-0 by Raysonho@Grid Engine
- wine.jpg CC-BY-SA by Rafael Garcia-Suarez
- fish.jpg CC-BY-SA by mahalie stackpole
- periodictable.png CC-0 by Cepheus
- droste.jpg CC-BY Zzubnik
- winecellar.jpg CC-BY-SA by Che (Petr Novák)
- cloud.jpg CC-0 by Sidik iz PTU
- regal.jpg CC-0 by Czarna Trucizna
- telescope.jpg CC-0 by C. Zorzi
- tree.jpg CC-BY-SA by Ji-Elle
- twins.jpg CC-0 by William Morris Agency
- earthquake.jpg CC-0 by Bert Cohen
- vermeer.jpg CC-BY-SA Kunstkenner2305
- clown.jpg CC-BY by Hamdan Zakaria

Got questions? Just Ask!

use of bibliographic data?
problem in your library?
ask libraries.stackexchange!
ask libraries.stackexchange!

http://libraries.stackexchange.com

Q&A about libraries and information science

This presentation as digital document

Source code and images of this presentation are available at `https://github.com/jakobib/ticer2012` to be copied and modified under CC-BY-SA license.

Original data from libraries

- Data not used as documents:
  - facts: library data (holdings, patrons, loans, formal metadata)
  - observations: subject metadata (descriptions)
  - communication: your publications
- Can be used in other context and times
  - Example: Authority files, connected via VIAF
  - Good use of data refers to you as source and authority
- **Advice:** Provide as much as possible your original data and documents. Just publish and care for this documents like any other acquisitions.