University Libraries – between Service Providers and Research Institutions

Wolf-Tilo Balke
Institut für Informationssysteme
Technische Universität Braunschweig
http://www.ifis.cs.tu-bs.de

36th Annual IATUL Conference 2015 – Hannover, Germany
The Royal Library of Alexandria

- Founded around 300 BC
- New idea: a central library to contain the collective knowledge of mankind
- Collection of about 500,000 scrolls including scientific content
Indexing and Access

• Meta data
  • Author, ISBN, …
  • Keywords
  • Classification

• Semantics?
Now digital...

• Online Public Access Catalog (OPAC)
  – Dynix OPAC (1983)
Still digital…

- GetInfo (TIB, 2015)
Scientific Information Services

• German Research Foundation (DFG) recently introduced a grant program for Scientific Information Services
  – Replaced the cooperative system of Special Subject Collections, dating back to 1949
  – Maintaining a special subject collection differs substantially from developing specialized information services
    • No common standards
    • Has to examine the needs of specialist communities
    • Needs to research and develop a suitable range of services
• Well,… maybe Google & Co?
  – Strong computer science competence
  – The size of the digital universe in 2007 was about $2,25 \times 10^{21}$ bits
    (281 Exabytes or 281 billion GB)

  – What about the indexing quality?
Semantic Indexing

• Heterogeneous formats
  – Classic publications
  – Primary research data
  – Simulations & models
  – Audiovisual content
  – …

• Major problems
  – Automatic indexing?
  – Search and access paths?
Scientific community cultures

- Example: chemistry
  - Structures as first class citizens
- CAS: manual indexing
  - Licenses about 30,000 USD
  - 10 million substances in 1990
  - 22 million substances in 2000
  - 50 million substances in 2009

Interfaces / retrieval services

...for every discipline?
...for every type of use?
...for every person?

C₆H₆ - Benzene

...and 217 others...
• Needed at least at level of **user groups**
  – Interdisciplinary needs
  – Different levels of detail

  – What is the Higgs boson? Article in Scientific American
    • Easy to understand answer (2 paragraphs, Northeastern University)
    • Some foundations (6 paragraphs, Santa Cruz Institute for Particle Physics)
    • Elaborate answer needing some expertise (9 paragraphs, Fermi National Accelerator Laboratory)
Information Extraction

• Typical automated approaches for mining semantics
  – Named entity recognition & extraction
  – Entity reconciliation and disambiguation
  – Structuring/classification of information
  – Generation of ontologies
  – ...

• Cooperation needed between
  – Information management science
  – Computer science
  – Domain
• Careful with **machine learning** – powerful, yet complex

  – Example: mathematics

  • Annotation with Mathematics Subject Classification (MSC)
  • Using support vector machines (SVM)
  • Confusion matrix for classes does not really convince…
• Where to extract suitable metadata for indexing?
  – Corpus-centric
    • Word-level analysis of publications
    • Analysis of similar publications
    • Analysis of cited/citing publications
  – Include external information
    • Exploit social annotations
    • Monitor usage types
• Gathering metadata from Web 2.0 information seems easier than it actually is
  – Quality of metadata vs. quality of content
  – Example: how to deal with scientific misconduct?

• Valuable research …
  • …is heavily cited?
  • …is often downloaded?
  • …has a lot of followers or like?

[Nature 478, 26-28 (2011)]
Again... Quality?

- Altmetrics / bibliometrics / scientometric
  - How valuable or influential is a publication or a researcher?
  - h-index, g-index, …
  - Citation counts
  - Co-author graph
  - Centrality, …
Again... Quality?

- Information generated on the Web 2.0
  - Reliable, trustworthy, honest?
  - How easy can it be manipulated?
  - What impact has the Web 2.0 on science, anyway?

- Collective intelligence...
  ...oder dictatorship of few?
Conclusions

• Library information services need more semantics and active research collaborations
  – How to extract a corpus’ full potential?
    • Adapt to needs of domain and user community
  – How can external information be included?
    • Deep investigations of meta data extraction, quality, robustness, and applicability
  – The holy grail of personalized knowledge spaces is still a distant hope…
Thanks for Your Attention!

balke@ifis.cs.tu-bs.de