Libraries as 'epistemological agents’ in Digital Humanities research

Prof. Dr. Jan Christoph Meister
Universität Hamburg
Reviving the past and keeping up with the future – the libraries’ role in preserving and providing access to newspapers and news media
role of physical libraries may become even more important. They are not just a place to house resources, but one in which to create meaning from them.

Libraries of the 21st century provide a welcoming common space that encourages exploration, creation, and collaboration.

They bring together the best of the physical and digital to create learning hubs.

The Role of the Library in 21st-Century Scholarly Publishing (ctd.)

- librarians have the potential to **play a leading role** in moving forward with new models of scholarly communication
- research library is in a position to **provide both the platform** and
- many of **the skills needed** to enable the creation of new forms of scholarship and to disseminate the resulting content to a wide audience of users
Kate Wittenberg, Council on Library and Information Resources (2014)
The Role of the Library in 21st-Century Scholarly Publishing

- focus on users → on disseminating information and services that respond to users' needs in whatever form seems most appropriate to the content
- information literacy → are users actually creating new ways of evaluating content that are different from those with which publishers are comfortable?
- role of information professionals → rapid development and change in the current digital environment ... new models for academic publishing are needed

Reviving the past and keeping up with the future – the libraries’ role in preserving and providing access to newspapers and news media

- “libraries are also agents of media transformation by digitizing their historical holdings of, i.a., newspapers and journals as cultural heritage”
- “how can libraries cooperatively support preparation of the digital raw-material for historical research, text and data mining, other digital humanities or educational purposes?”

http://blogs.sub.uni-hamburg.de/ifla-newsmedia/ [20.04.2016]
Digital Humanities approaches in news and media studies

„Smart data“ approaches → indepth close reading of small & exemplary samples relies on interactive procedures:

- digitization & creation of a relevant digital corpus
- categorization & annotation of corpus elements via meta-data
- hypothesis driven iterative analysis/interpretation-cycles

„Big data“ approaches → bottum-up distant reading relies on automated procedures

- creation of large, indiscriminate collections
- automatic segmentation and structuring of collection items (parsing, normalisation, named entity recognition etc.)
- automatic statistical & ML-based exploration (pattern recognition, topic modelling etc.)
Small data/close reading vs. big data/distant reading: an example

... plus approx. another 104.999.990 documents related to our topic
Interactive close reading of the corpus

http://voyant-tools.org

http://catma.de
Topic Modelling: a *Distant Reading* method for algorithmic analyses of large corpora

- fully automatic: does not require annotated training data
- purely statistical: does not analyze linguistic structure, but random „bag of words“
- scalable: simple to run and robust to variation in language and data quality
- clustering based: uses contextual clues to group related words and to distinguish between uses of ambiguous words
- synthetic: reduces the dimensionality of a corpus to several hundred clusters or topics, represented as distributions over the full vocabulary

Topic Modelling analysis of the first ten documents in a collection of news articles on

http://mallet.cs.umass.edu/about.php [20.04.2016]
### Results 1

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>resources, technology, staff, traditional, management, access, development, social, communication, provide</td>
</tr>
<tr>
<td>2</td>
<td>space, digital, learning, books, physical, students, create, printed, school, collaboration</td>
</tr>
<tr>
<td>3</td>
<td>public, st, brecher, century, philanthropy, center, foundation, directors, arlington, importance</td>
</tr>
<tr>
<td>4</td>
<td>library, services, online, skills, technologies, librarian, technology, project, people, patrons</td>
</tr>
<tr>
<td>5</td>
<td>librarians, books, people, training, open, health, teaching, city, life, day</td>
</tr>
<tr>
<td>6</td>
<td>information, publishing, users, publishers, scholarly, work, models, based, scholars, process</td>
</tr>
<tr>
<td>7</td>
<td>content, community, role, research, digital, environment, make, form, changing, tools</td>
</tr>
<tr>
<td>8</td>
<td>librarians, important, reference, internet, change, search, ability, students, good, stakeholders</td>
</tr>
<tr>
<td>9</td>
<td>information, knowledge, access, services, user, century, st, users, centre, critical</td>
</tr>
<tr>
<td>10</td>
<td>library, libraries, share, computer, age, free, literacy, year, formats, individual</td>
</tr>
</tbody>
</table>

- remove English stopwords
- 10 topics
- 200 iterations (random “bags of words“)
- output: show first ten words per topic
Results 2

<table>
<thead>
<tr>
<th>Topic 14: access university important librarian databases researchers years recent consult catalogues print real abstracting titles business</th>
</tr>
</thead>
<tbody>
<tr>
<td>➔ Relationship to Universities</td>
</tr>
</tbody>
</table>

- remove English stopwords
- 25 topics
- 500 iterations (random „bags of words“)
- output: show first 15 words per topic
- possibility to assign interpretative labels to 15-word topics
Results 3

- remove English stopwords
- 25 topics
- 750 iterations (random „ bags of words“ )
- output: show first 15 words per topic
- possibility to assign interpretative labels to topics increases
…so: what is the libraries‘ role in this (and in other) Digital Humanities research?

„Topic models are powerful and scale to large data sets, but their ease of use can be deceptive.“

Matthew Jockers & David Mimno (2013)
large data sets: the good, the not-so good and the fake & useless

Deutsches Textarchiv: discipline- and genre independant collection of (digital full) texts in German 1600-1900, approx. 2400 items [4/2016] – excellent, high-quality collection of digital resources

Google Books: currently approx. 30 mio digitized volumes. A powerful resource, but with methodological short comings – lack of transparency, proprietary & copyright issues, cultural bias

Deutsche Digitale Bibliothek: portal linking to content provided by approx. 30.000 institutions - politically motivated, practically useless for DH research
Opportunities vs. risks of digital collection based research into news and media

- Large data sets held or made accessible by libraries, such as newspapers & newsmedia archives, invite the application of Digital Humanities methods
- Methods & approaches like Topic Modelling will over the next 5-10 years become *business as usual* and increasingly complement qualitative research
- Universities and research institutions without a dedicated DH department/center lack the expertise to protect against the risks of unreflected, simplistic „datafication“ research
→ Libraries as *Epistemological Agents* can provide *data management & analysis expertise*

**data management**
- aggregation & quality control of large digital corpora
- documentation of rules & procedures followed in corpus composition (reproducability of results!)

**data analysis**
- offer *best practice* oriented consultation to researchers
- provide, monitor and maintain a basic suite of robust „tried and tested“ analytical tools that can be reliably used by non-DH specialists (e.g., parsers, segmenters, topic modelling & sentiment analysis tools, visual and sound analysis tools, markup tools, visualizers, etc.)
- provide a **DH helpdesk** for the non-specialist user of DH applications
The library’s role as epistemological agent in DH research

Conceptual DH: theory, concepts, methods, tools development

the library’s DH helpdesk service

Curated digital collections

Non-curated digital collections (Google Books)

Dynamic ad-hoc collections (web crawling)

DH methods & procedures

Applied DH (quantitative/qualitative, statistical/hermeneutic, automatic/interactive, etc.)