Web Archive Research and the role of (digital) academic libraries

Thomas Risse
Goethe University Frankfurt
University Library J. C. Senckenberg

RESAW@Porto2018, 13. September 2018
Web Archives in Research

Different Research Fields
- Digital Humanities
  - Historians (long term)
  - Social Science (more short term)
  - Cultural anthropology
  - ...
- Law
- Computer Science and related areas
  - Information Retrieval
  - Temporal Analytics
  - Natural Language Processing
  - ...

Different Requirements
- Selection strategy, Time frames
- Provenance, Validation
- Persistence
- Analysis, Visualizations
- ...

A Typical Point of Departure

Web Content might be interesting for my Research

That’s too much data. I need some tools.

Please make a Web collection

Is the data complete? Where can I store it? Can I publish the data? How can I cite my data? ...

Web Crawler (HTTrack, Web Scraping Tools)

Private Web Data Collection
Sources of Web Content used in Research

Oh, there are other sources!!!

- **Archive.org**
  - Broad in scope
  - URL and fulltext access

- **ArchiveIT**
  - Collections curated by libraries
  - URL and fulltext (inside collections)

- **National Libraries**
  - Broad in scope, different selection strategies
  - Limited access with some exceptions (e.g. Portugal)

- **Organizational Archives, e.g. political parties**
  - Focused collections
  - Limited access due to copyright restrictions
Doing Research with Web Archives

"I'm new to Web Archives. Where should I start?"

Typical Questions
• Where to find relevant content?
• What can be done with the content?
• How to handle the amount of data?
• Which tools can be used?
• What are regulatory limitations?
• …

But there is more to know when working with Web Archives
• WAs differ from traditional libraries and archives
• WAs are never complete or fully coherent
• There are different versions of a page
• WAs are not a full copy of the Web (or Web Site)
• Different WAs have different selection strategies

Scholars need to get a good understanding of the sources, content properties and tools
→ Currently it’s their own task
Academic Libraries could Help

- Many researchers are used to work with (local) academic libraries
- Libraries have a long tradition and experience in working with content of various kinds
- Academic libraries are already offering various courses and services e.g. reference management, research data management, digital humanities
- Provide direct support in finding literature
- Libraries might have specialists in various areas
  - Special collection focus since many years
  - Knowledgeable librarians who curate the collections
How does it help with Web Archives?

Guiding the Users
• Basic knowledge on Web Archives is independent of the research field
• Academic Libraries could offer courses on Web Archive usage
  – General Introduction
  – Major tools
• Publish general guidelines
• Faculties provide complementary specific knowledge

In collaboration with the computing center
• Offer tools and services
Providing Access to Web Archive Content

From the researchers point of view
• Academic libraries would be a logical source/starting point to access Web archives
• Similar to publications and research data

Some Major Challenges
• Resources
  – Academic libraries have limited (or no) resources for doing Web archiving
    – Staff
    – Computation & Storage Resources
• Content Selection
  – What should be collected?
  – What are the researchers needs?
• Regulation
Regulation: Copyright and Privacy

- Regulations differ among the countries and states
- Typically Academic Libraries are not allowed to archive Web content and to provide access
  - There is no regulation that enables this function
  - Germany: Some state libraries which are also academic libraries are allowed
- National libraries are allowed to harvest
  - Harvest national domains with different selection strategies
  - Provide limited access
    - Maybe 70 years after the death of author of a page
      But huge amounts of orphan works
    - Maybe if the content is marked as open access
    - Other countries are more open: Portugal, US
  - Access is broad, fulltext search is not satisfying for focused research questions
  - Curated collections only exist for major topics (s. ArchiveIT)
  - Specialists in research areas are missing
Let’s dream a bit …

… of the nearly ideal library and research world
• Academic libraries would be allowed to harvest and archive the Web
• And could provide unlimited access

Web Harvesting
• What should a library harvest?
  – Everything? → Impossible
  – Selected Topics
    – On request? → Would only address research questions on current Web content
    – Topical areas of the Library → Would be a nice complement
• Access to content
  – Search among all collections might not be necessary
  – Fulltext search and browsing within collections should be provided
  – Any advanced search is welcome

Resources
• Libraries have only limited personal and monetary resources
• Most innovative services are financed by 3rd parties
• Sustainability is not guaranteed but necessary for Web archiving
… but it was only a nice dream

… the Regulation will not change
… the Resources will not increase
What do we have?

Academic Libraries
• Could offer courses on Web Archive usage
• Have specialists in certain areas
• No mandate to harvest or to provide Web archive content

National libraries / Archive.org
• Allowed to collect and to provide access
• Have the resources and experience for harvesting and Web archiving
• Technology for retrieval
• No specialists in specific research fields

→ Let’s join forces to provide academic related access
Providing a Window to Web Archives

The real legal problem of access: The storage location of the data

Countries with strict regulations
- It is forbidden to deliver archived content that is physically stored in this country

But
- It is allowed to store URLs to archived content in any country
- It is allowed to show users the archived URLs
- It is allowed to open a frame in the browser that loads the archived URL

Approach:
- Technology: Micro Archives + Viewer
- Organizational Structure at the library
Micro Archives consists of:

- **Crawl Specification**: URLs (to the Web or a Web Archive)
- **Meta-Information**
  - Labels and Comments for URLs
  - Semantic descriptions
- **Type**

**MiCrawler**

- Creates an instance of the Micro Archive
- Provides an unique identifier (e.g. DOI, URN)

Organizational Aspect
e.g. Specialized Information Services

- German Research Infrastructure Activities funded by German Research Association
- Build upon special collections and existing knowledge in libraries
- Provide researchers a comprehensive view on their topic
  - Aggregates library catalogues from all over Germany/International
  - Provide free access to licensed content (collections, databases)
  - Provide community services, e.g. blogs, researcher directories, tool boxes
- For example: University Library Frankfurt
  - Comparative literature, Linguistics, German Studies, African Studies, Jewish Studies, Performing Arts, Biodiversity
- Idea: Semi Curated Web Archive Windows
Semi Curated Web Archive Windows

Portal

- Curates Micro Archive
- Micro Archive
- Specification
- Catalog
- Storage
- MiCrawler (User Machine)
- Local Instance
- Archive.org
- Internet Archive (Direct Access)
<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Author</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Die Abtheilung der israelitischen Ritualgegenstände im städtischen historischen Museum zu Frankfurt am Main</td>
<td>Freimann, Aron</td>
<td>1900</td>
</tr>
<tr>
<td>2</td>
<td>Die Abtheilung der israelitischen Ritualgegenstände im städtischen historischen Museum zu Frankfurt am Main. [Druck des Frankfurter Intelligenz-Blattes]</td>
<td>Freimann, Aron</td>
<td>2012</td>
</tr>
<tr>
<td>3</td>
<td>Annalen der hebraischen Druckerei in Wilhermsdorf</td>
<td>Freimann, Aron</td>
<td>1903</td>
</tr>
<tr>
<td>4</td>
<td>Annalen der hebraischen Druckerei in Wilhermsdorf [Elektronische Ressource]</td>
<td>Freimann, Aron</td>
<td>2010</td>
</tr>
<tr>
<td>5</td>
<td>Annalen der hebraischen Druckerei in Wilhermsdorf</td>
<td>Freimann, Aron</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>Count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>undetermined</td>
<td>54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish Studies: Cultural History</td>
<td>38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jüdische Studien: regional - Germany</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish Studies: Religion</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish Studies: General</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jewish Studies: Literature</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>more ...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Search: freimann, aaron

#### Showing 1 - 25 of 153 for search: 'freimann, aaron'

<table>
<thead>
<tr>
<th>Year</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Aron Freimann and the Development of Jewish Bibliography in ... - Brill</td>
</tr>
</tbody>
</table>


- [der Festschrift für Aron Freimann - Jüdisches Museum Berlin](https://jmb.de/object/jmb-object-587722.jsessionid)


- [festschrift fuer aron freimann von marx - ZVAB](https://www.zvab.com/buch-suchen/titel/...aron-freimann/...marx/)

### Language

- German: 118
- Hebrew: 51
- English: 10
- Yiddish: 5
- Italian: 4
- Undetermined: 4
An Example

2001 Broad Archive (e.g. Internet Archive)

Direct Access in an iframe

Local Metadata
Conclusions

Academic Libraries should play an more important role
• Close contact to the users
• Knowledge about selected areas

Experienced in training users in using library services
• Offer training courses on Web archive usage

Resources and legal regulations are hindering
• Harvesting by libraries overall not realistic
• Could provide windows into large scale Web archives
• Curation of content for specific areas are possible
  – Requires more IT support to ease the task
• Requires more collaboration among libraries and Web archives