Back to our Data – Experimenting with NoSQL Technologies for Humanities Research
EHRI facts

• 2010 – 2014 (48 months)

• Supported by the 7th Framework Programme of the European Union

• Funding scheme: Integrated Infrastructure Initiative I3

• 20 partners (research institutions, libraries, archives, museums and memorial sites)

• 3 disciplines: history, archival science and digital humanities

• Coordinator: NIOD Institute for War, Holocaust and Genocide Studies, dr. Conny Kristel
Why EHRI?

- Fragmentation and dispersal of archival sources
  - Geographical scope Holocaust
  - Attempts to destroy the evidence
  - Migration of Holocaust survivors
  - Multiplicity documentation projects after the war

- Internationalization Holocaust research
  - Holocaust in Eastern Europe
  - New levels of collaborative research
Aims and Objectives

- Integrate Holocaust material and research
- Initiate new levels of collaborative research
- Enable historiographical progress (transnational and comparative research)
20 partners in 13 countries
The Adler case
Archives are natural graphs or at least trees
Research Datafication

- Latest buzz word in big data to describe the collecting of everything
- Datafication is not digitisation, especially in the context of humanities research
  - In our context there is already a huge difference between an archival description and a research object for Humanities
- Datafication emphasizes reuse for analysis and integration from different sources
The joined research work

- Identifying
  - Collections and Users

- Integrating
  - Standards and Thesaurus for a Virtual Observatory for Holocaust Research

- Analysing
  - Connecting Holocaust Research Objects
  - Archive Graph
Author: Dennis Nilsson.
# Eastern European Jewish Holocaust Victims

<table>
<thead>
<tr>
<th></th>
<th>Total pop. on the eve of WWII</th>
<th>Jewish pop. on the eve of WWII</th>
<th>% of total population</th>
<th>Estimated number of murdered</th>
</tr>
</thead>
<tbody>
<tr>
<td>USSR</td>
<td>168,000,000</td>
<td>3,020,000</td>
<td>1.8</td>
<td>1,050,000</td>
</tr>
<tr>
<td>Poland</td>
<td>32,040,000</td>
<td>3,300,000</td>
<td>10.3</td>
<td>2,950,000</td>
</tr>
<tr>
<td>Estonia</td>
<td>1,125,000</td>
<td>4,500</td>
<td>0.4</td>
<td>1,750</td>
</tr>
<tr>
<td>Latvia</td>
<td>1,830,000</td>
<td>91,500</td>
<td>5</td>
<td>70,750</td>
</tr>
<tr>
<td>Lithuania</td>
<td>2,302,000</td>
<td>168,000</td>
<td>7.3</td>
<td>141,500</td>
</tr>
</tbody>
</table>
EHRI’s focus on Eastern Europe

- Case study on Ukraine resulted in identifying about 1,000 collections, bringing the players in the field in contact with each other & starting up cooperations

- Similar workshops regarding Belarus, Lithuania and Poland

- Conference of archivists and researchers in Cracow, Poland in 2014 to discuss the state of Holocaust documentation in Eastern Europe
WP15 workshop on Ukraine on 19 July 2012 (IfZ)

- Case study on Ukraine resulted in identifying about 1,000 collections, bringing the players in the field in contact with each other & starting up cooperations
Holocaust victims

Total - Greece: ca. 65,000
From pre-war 77,000
Identification & Investigation

- Identify institutions that hold Holocaust-related material
- Identify Holocaust-related collections within these institutions (or in aggregators)
- Approx. 1500 institutions identified so far
- Holdings are often non-digital
42 countries on the investigation map

Approx. 1500 institutions identified so far
Integrating Archives

'Find all information about prisoners arriving in Therezin from the Netherlands in 1944 and the research on this‘

• WP 17: Enhance Standards and make Research Objects
• WP 18: Integrate Thesaurus
Data Integration

Problems with traditional approach

- A relational database excels at storing predictable data structures, and answering queries like average, count, maximum etc.
- These queries are not at the core of the research environment of EHRI
- The metadata is hierarchical & relational databases are not great at managing hierarchies

A graph database is used for the integrated collection store
• Stores data as edges & nodes with properties on both
• Supports browsing: Provides the ability to navigate through the (meta) data by traversing the edges
• Allows for integration of the controlled vocabulary graphs from Thesaurus
• Easily amendable with heterogenous metadata
• Not another triple store – Linked Data is part of the solution but not the solution
Heterogeneous metadata

User generated content

Thesaurus

Authorities

Collection Metadata
Heterogeneous metadata – a collection
Copies and multiple descriptions of same material
Instead of a Conclusion

Advantages

- Concentrates on Data
- Historiographers researchers walk the graph
- Powerful framework for future analysis
- Easily social: Annotations at the core of metadata enrichment as well as the VRE

Disadvantages

- Still hard work – not something for the light-hearted
- To deal with documents another tech is also needed

Instead of a Conclusion
Discussion of advantages and disadvantages

• Advantages
  – Concentrates on Data
  – Historiographers researchers walk the graph
  – Powerful framework for future analysis
  – Easily social: Annotations at the core of metadata enrichment as well as the VRE

• Disadvantages
  – Still hard work – not something for the light-hearted
  – To deal with documents another tech is also needed
Prototype

http://ehritest.dans.knaw.nl
Publications

• **General**
  – Tobias Blanke, Veerle van den Daalen, Michal Frankl, Conny Kristel, Reto Speck, The past and the future of Holocaust research: from disparate sources to an integrated European Holocaust Research Infrastructure, Festschrift Goettingen State and University Library, Forthcoming 2013

• **User requirements**

• **OCR/Information Extraction**
  – Tobias Blanke, Michael Bryant, Reto Speck, Conny Kristel, Information Extraction on Noisy Texts for Historical Research, Digital Humanities 2012 (Hamburg 2012)

• **Virtual Research Environment**
  – Mark Hedges, Tobias Blanke, Michael Bryant, Conny Kristel, Creating general-purpose virtual environments for (digital) archival research, Digital Humanities 2011 (Stanford, California)
Stay informed ...

Visit EHRI at
www.ehri-project.eu

Contact EHRI at
www.ehri-project.eu/contact-us

and receive EHRI Newsletters and mailings