Archiving and Analysing Elections: How can Web Archiving, Digital Humanities and Political Science go together?

Tobias Beinert (Bavarian State Library)
Markus Eckl, Sebastian Gassner (University of Passau, Chair of Digital Humanities)
Florence Reiter (University of Passau, Jean Monnet Chair for European Politics)
Web Archiving at the Bavarian State Library

- Selective Harvesting since 2012 for Specialised Information Services and Bavarica with the Web Curator Tool and OpenWayback
- Permissions for harvesting, long-term preservation and access requested
- Approx. 1600 websites archived with several snapshots
- Manual and semi-automated quality control
- Access via BSB’s catalogue and the gateways of the Specialised Information Services

-> Need to explore better ways of exploring the content of web archives and strengthen relationship to research community
Project with the University of Passau

• Project Partners: Chair of Digital Humanities and Jean Monnet Chair for European Politics

• Use of methods and tools from the Digital Humanities for data sets of web archive collections -> Exploratory Study

• Case Study on the Bavarian state election (2018) and the European Elections (2019): How do political actors and parties frame the European Union throughout their election campaigns?

• Aim to publish successfully tested tools open source and research

• Improve collection management by data analysis

• Explore theoretical framework of web archiving as research source material (Definition of corpora and completeness)
Event Crawl Elections: First steps and challenges

• Defining the corpus
  • Many actors with few snapshots vs. few actors with many snapshots
• Redundancy of snapshots/data
• Including Social Media and News Websites
• Web Archiving vs. Data Scraping
• Use of Web Curator Tool 1.6, 1.7 beta and 2.0
• Use of webrecorder for social media
Processing the data

• Primary focus of analysis of text-based materials
• Evaluation of existing tools
• Extracting HTML and text cleaning with Python Scrips
• Building a Mongo database
Analysing the data with DH methods

- Link Mining to identify gaps in the collection
- Topic Modelling (Latent Dirichlet Allocation – LDA)
- Visualisation
- Network Analysis
- Word Networks
First results – Topic Modelling LDA

Blei 2012

---

**Topics**

- gene: 0.04
- dna: 0.02
- genetic: 0.01
- life: 0.02
- evolve: 0.01
- organism: 0.01
- brain: 0.04
- neuron: 0.02
- nerve: 0.01
- data: 0.02
- number: 0.02
- computer: 0.01

**Documents**

**Topic proportions and assignments**

---

**Seeking Life’s Bare (Genetic) Necessities**

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to keep it alive? Two researchers who work in the genome-mining business have taken different approaches to a critical question in basic biology. One is at the University of California, Santa Barbara, and the other at Cold Spring Harbor Laboratory. They've both side-stepped the 200-gene minimum estimate suggested by an international effort to sequence the first living organism.

In their recent analysis, the Santa Barbara researchers sequenced 221 genes of the bacterial species Mycoplasma genitalium and found that only 188 were necessary to keep the organism alive. The scientists suggest that the 188-gene core of life could be further reduced to just 144 genes. The analysis builds on the seminal paper by Lawrence Siegal and collaborators published in Nature in 1998.

---

*Image credit: Science Magazine*
### First results – Topic Modelling LDA

<table>
<thead>
<tr>
<th>Words</th>
<th>Topic Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.022&quot;partei&quot; + 0.009&quot;grüne&quot; + 0.009&quot;söder&quot; + 0.008&quot;politisch&quot; + 0.008&quot;wahl&quot; + 0.008&quot;wähler&quot; + 0.006&quot;prozent&quot; + 0.006&quot;mehr&quot; + 0.005&quot;freie&quot; + 0.005&quot;europawahl&quot;</td>
<td>Europe Election</td>
</tr>
<tr>
<td>0.021&quot;europa&quot; + 0.012&quot;europäischen&quot; + 0.011&quot;land&quot; + 0.009&quot;deutschland&quot; + 0.008&quot;europäische&quot; + 0.007&quot;frankreich&quot; + 0.007&quot;brüssel&quot; + 0.007&quot;präsident&quot; + 0.006&quot;regierung&quot; + 0.006&quot;deutsche&quot;</td>
<td>Europe, Gemany &amp; France</td>
</tr>
<tr>
<td>0.016&quot;flüchtling&quot; + 0.013&quot;begriff&quot; + 0.010&quot;deutschland&quot; + 0.008&quot;seehofer&quot; + 0.008&quot;land&quot; + 0.007&quot;mensch&quot; + 0.007&quot;sprache&quot; + 0.006&quot;politisch&quot; + 0.006&quot;asylbewerber&quot; + 0.005&quot;integration&quot;</td>
<td>Migration &amp; refugees</td>
</tr>
</tbody>
</table>
**First results – Topic Modelling LDA**

<table>
<thead>
<tr>
<th>Words</th>
<th>Topic Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.014*&quot;britisch&quot; + 0.011*&quot;großbritannien&quot; + 0.010*&quot;brexit&quot; +</td>
<td>Brexit</td>
</tr>
<tr>
<td>0.009*&quot;london&quot; + 0.008*&quot;brite&quot; + 0.005*&quot;parlament&quot; +</td>
<td></td>
</tr>
<tr>
<td>0.004*&quot;david&quot; + 0.004*&quot;premierministerin&quot; + 0.004*&quot;land&quot; +</td>
<td></td>
</tr>
<tr>
<td>0.004*&quot;theresa&quot;</td>
<td></td>
</tr>
<tr>
<td>0.019*&quot;facebook&quot; + 0.012*&quot;nutzer&quot; + 0.010*&quot;datum&quot; +</td>
<td>Facebook &amp; Internet</td>
</tr>
<tr>
<td>0.009*&quot;unternehmen&quot; + 0.008*&quot;netz&quot; + 0.008*&quot;internet&quot; +</td>
<td></td>
</tr>
<tr>
<td>0.006*&quot;mehr&quot; + 0.006*&quot;mensch&quot; + 0.005*&quot;inhalt&quot; +</td>
<td></td>
</tr>
<tr>
<td>0.005*&quot;plattform&quot;</td>
<td></td>
</tr>
<tr>
<td>0.010*&quot;landwirt&quot; + 0.010*&quot;bauer&quot; + 0.010*&quot;volksbegehren&quot; +</td>
<td>referendum bees</td>
</tr>
<tr>
<td>0.008*&quot;mehr&quot; + 0.008*&quot;bienen&quot; + 0.007*&quot;landwirtschaft&quot; +</td>
<td></td>
</tr>
<tr>
<td>0.006*&quot;naturschutz&quot; + 0.006*&quot;pflanze&quot; + 0.006*&quot;artenvielfalt&quot; +</td>
<td></td>
</tr>
<tr>
<td>0.005*&quot;insekt&quot;</td>
<td></td>
</tr>
</tbody>
</table>
Thank you!
Questions?

beinert@bsb-muenchen.de