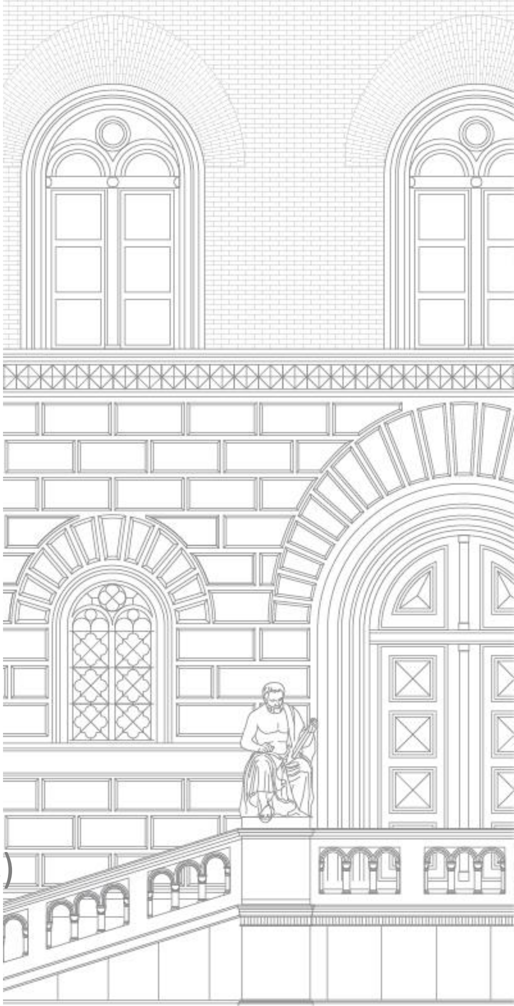
BSB Bayerische StaatsBibliothek
Information in erster Linie

UNIVERSITY OF PASSAU

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 Digitial 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

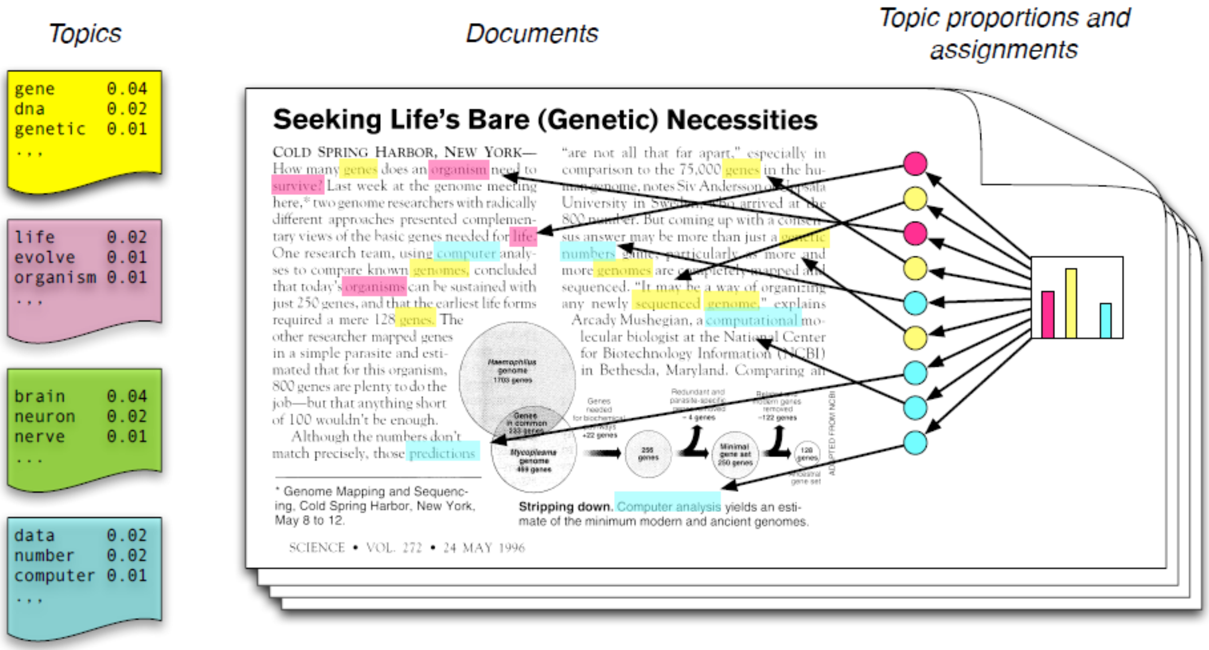
Processing the data

- 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


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
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
255 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Blei 2012

First results – Topic Modelling LDA

Words	Topic Name
0.022*"partei" + 0.009*"grüne" + 0.009*"söder" + 0.008*"politisch" + 0.008*"wahl" + 0.008*"wähler" + 0.006*"prozent" + 0.006*"mehr" + 0.005*"freie" + 0.005*"europawahl"	Europe Election
0.021*"europa" + 0.012*"europäischen" + 0.011*"land" + 0.009*"deutschland" + 0.008*"europäische" + 0.007*"frankreich" + 0.007*"brüssel" + 0.007*"präsident" + 0.006*"regierung" + 0.006*"deutsche"	Europe, Gemany & France
0.016*"flüchtling" + 0.013*"begriff" + 0.010*"deutschland" + 0.008*"seehofer" + 0.008*"land" + 0.007*"mensch" + 0.007*"sprache" + 0.006*"politisch" + 0.006*"asylbewerber" + 0.005*"integration"	Migration & refugees

First results – Topic Modelling LDA

Words	Topic Name
0.014*"britisch" + 0.011*"großbritannien" + 0.010*"brexit" + 0.009*"london" + 0.008*"brite" + 0.005*"parlament" + 0.004*"david" + 0.004*"premierministerin" + 0.004*"land" + 0.004*"theresa"	Brexit
0.019*"facebook" + 0.012*"nutzer" + 0.010*"datum" + 0.009*"unternehmen" + 0.008*"netz" + 0.008*"internet" + 0.006*"mehr" + 0.006*"mensch" + 0.005*"inhalt" + 0.005*"plattform"	Facebook & Internet
0.010*"landwirt" + 0.010*"bauer" + 0.010*"volksbegehren" + 0.008*"mehr" + 0.008*"bienen" + 0.007*"landwirtschaft" + 0.006*"naturschutz" + 0.006*"pflanze" + 0.006*"artenvielfalt" + 0.005*"insekt"	referendum bees

Thank you!
Questions?

beinert@bsb-muenchen.de