Brauchen die Digital Humanities eine eigene Methodologie?
Übersicht

- Aspekte der Operationalisierung geistes- und sozialwissenschaftlicher Fragestellungen
- Beispiel „ePol – Postdemokratie und Neoliberalismus“
- Zusammenfassung
### Aspects of operationalization

#### DH project character

<table>
<thead>
<tr>
<th>DATA MANAGEMENT</th>
<th>DATA PROCESSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>data modeling:</td>
<td>requirements: iterated process of (1) devising research needs, (2) approaches to fulfil them (3) assessment of results</td>
</tr>
<tr>
<td>defining entities (E) and their relations (R)</td>
<td></td>
</tr>
<tr>
<td>service provision / translation of ER-models into research infrastructures:</td>
<td></td>
</tr>
<tr>
<td>● DBMS + GUI</td>
<td></td>
</tr>
<tr>
<td>● MAXQDA et al.</td>
<td></td>
</tr>
<tr>
<td>● ...</td>
<td></td>
</tr>
</tbody>
</table>

---

#### Roles

- **humanities**
- **computer science**
Aspects of operationalization

**WRONG** *(technology-driven)*

- We use *(incidentally)* available / ad-hoc compiled text collections
- We apply generic *(black box)* tools (well) known to us
- *(Afterwards)* We try to establish a meaningful connection between our initial research question and the automatically generated output data

**RIGHT** *(requirements-driven)*

- We select / compile a text collection by *carefully specified criteria* beforehand
- We apply procedures *best matching* our data / research interest. If not available, we develop them based on *requirements* we identify systematically.
- We use our text collection as basis for a *profound validation* of our initial research question
Aspects of operationalization

How can we support the process of „operationalization“?

For social sciences

• *Apply / adapt established methodology of qualitative and quantitative research*
  - But, use (large) written text corpora instead of survey data

• *Understanding: Identifying / extracting meaningful entities*
  - *Complement qualitative research based on texts* (e.g. MAXQDA)

• *Explaining / Quantitative paradigm: measuring (causal) relations*
  - *Hypothesis development + validation* by empirical testing on text collections

• *A reconciliation of qualitative and quantitative methods?*
Aspects of operationalization

How can we support the process of „operationalization“?

For social sciences and humanities in general

- Apply requirements engineering and modeling as part of a software engineering process
  - Identify domain model and workflows
  - Identify key functions and define their implementation

- Requires „mutual understanding“ / „common language“ of humanities scholars and computer scientists
Example: *ePol* – Tracking down economization

- Joint research project in the eHumanities; cooperation with Prof. Gary Schaal, department for political theory (HSU)
- Period: June 2012 – May 2015
- 2 x 1,5 academic employees, + student assistants
- licences for longitudinal retro-digitized German newspaper corpora (Zeit, FAZ, Süddeutsche, TAZ)
  → 3.5 Million documents
Research question:

In the debate on post-democracy political theory claims that contemporary western democracies tend to justify politics in a neoliberal manner. This manner, characterized by means of economization, has become dominant over the past decades.

→ ePol-projects: empirical evaluation of this hypothesis for Germany
Research issues from the computer science perspective

1) Operationalization of political science research questions for computer-assisted text analyses
2) Development of a technical research infrastructure for large scale text corpora adaptable to heterogenous research interests
3) Development of specifically adapted analyses procedures
4) Procedures to evaluate computer generated results
After each subgoal:
- evaluation $E_n$ of intermediate results
- adaption of requirement specification and functional specification
Requirements Engineering

**subgoal 1: identification of relevant documents**
- retrieval of documents relevant regarding aspects of neoliberal economization and arguments
  - manual search
  - computer-aided search (retrieval with reference dictionaries)
  - extraction of reference dictionaries → „Wörterbuch des Neoliberalismus“
- document management
- retrieval for manual full text search and machine composed queries

**subgoal 2: identification of argument structures**
- locating argumentative structures in thematically separated text collections
  - topic detection
  - manual annotation of arguments
  - retrieval of text passages similar to annotations
- category tools
- annotation tools
- classification and pattern recognition

**subgoal 3: analysis/empirical validation**
- empirical evaluation of retrieved text parts
- qualitative and quantitative description of resulting categories
- validation of automatically generated results
- quality measure
- refinement of training data
- visualization of results
- data export
- quality measure computation
- graphical user interface
- data export interfaces: CSV, R, ...

**After each subgoal:**
- evaluation $E_n$ of intermediate results
- adaption of **requirement specification** and **functional specification**
Requirements Engineering

subgoal 1: identification of relevant documents
- retrieval of documents relevant regarding aspects of neoliberal economization and arguments
  - manual search
  - computer-aided search (retrieval with reference dictionaries)
  - extraction of reference dictionaries → "Wörterbuch des Neoliberalismus"
- document management
  - retrieval for manual full text search and machine composed queries

subgoal 2: identification of argument structures
- locating argumentative structures in thematically separated text collections
  - automatic detection
  - manual annotation
- category tools
  - classification and pattern recognition

subgoal 3: analysis/empirical validation
- empirical evaluation of retrieved text parts
  - qualitative and quantitative description of resulting categories
  - validation of automatically generated results
  - quality measure
- refinement of training data
- visualization of results
- data export interfaces: CSV, R, ...

After each subgoal:
- evaluation $E_n$ of intermediate results
- adaption of requirement specification and functional specification
subgoal 1: identification of relevant documents
- retrieval of documents relevant regarding aspects of neoliberal economization and arguments
- manual search
- computer-aided search (retrieval with reference dictionaries)
- extraction of reference dictionaries → „Wörterbuch des Neoliberalismus“
- document management
- retrieval for manual full text search and
- machine composed queries

subgoal 2: identification of argument structures
- locating argumentative structures in thematically separated collections
- topic detection
- manual annotation of arguments
- retrieval of text passages similar to annotations
- category tools
- annotation tools
- classification and pattern recognition
- validation of automatically generated results
- quality measure
- refinement of training data
- visualization of results
- data export
- empirical evaluation of retrieved text parts
- qualitative and quantitative description of resulting categories
- quality measure computation
- graphical user interface
- data export interfaces: CSV, R, …

Tasks
- Preprocessing: data cleaning, meta data unification, tokenization, Tagging, Chunking, NER
- data storage + retrieval system
  - Mongo DB
  - Apache Solr
- web frontend / user interface
  - full text search
  - dokument / collection management
  - user system
  - result visualization/storage/export
- generic tools
  - frequency extraction
  - topic detection
  - cooccurrence extraction
- specific tools
  - term extraction via topic models / reference corpora
  - document retrieval for topic and parlance (via contextualization of query terms)

After each subgoal:
- evaluation of intermediate results
- adaption of requirement specification and functional specification
Requirements ePol-project

subgoal 1: identification of relevant documents
- retrieval of documents relevant regarding aspects of neoliberal economization and arguments
- manual search
- computer-aided search (retrieval with reference dictionaries)
- extraction of reference dictionaries → "Wörterbuch des Neoliberalismus"
- document management
- retrieval for manual full text search and machine composed queries

subgoal 2: identification of argument structures
- locating argumentative structures in thematically separated collections
- topic detection
- manual annotation of arguments
- retrieval of text passages similar to annotations
- category tools
- annotation tools
- classification and pattern recognition

subgoal 3: analysis/empirical validation
- empirical evaluation of retrieved text parts

Tasks
- Preprocessing: data cleaning, meta data unification, tokenization, Tagging, Chunking, NER
- data storage + retrieval system
  - MongoDB
  - Apache Solr
- web frontend / user interface
  - full text search
  - dokument / collection management
  - user system
  - result visualization/storage/export
- generic tools
  - frequency extraction
  - topic detection
  - cooccurrence extraction
- specific tools
  - term extraction via topic models / reference corpora
  - document retrieval for topic and parlance (via contextualization of query terms)

After each subgoal:
- evaluation of intermediate results
- adaption of requirement specification and functional specification
User interface for text analyses (Zeitungstexte)

Leipzig Corpus Miner: A web interface for
- corpus exploration
- document collection management
- control of analysis processes
- visualization of results

http://www.epol-projekt.de/
Conclusion

● What we need
  - modeling processes for systematic analysis of requirements from both perspectives (humanities and computer science)
  → open issue for future research

● What we recommend
  - adapted procedures / individual solutions rather than generic tools
  - iterated refinement of requirement specifications as well as feature specification during the overall analysis process
  - strong emphasis on evaluation / validation of computer-generated results