Iterating the Pipeline
Experience shows that a challenging subject like collaborative identity and its manifestations on social platforms will not be studied in a sufficient way in a first attempt. Therefore, we expect an iterative process involving social scientists and computer scientists that will (hopefully) lead to answers to the research questions.

Workgroup 2: Shonan Model for Content Selection & Analysis

Gerhard Heyer, Kazufumi Fukuda, Fabian Schäfer, Cathleen Kantner, Maciej Piasecki, Chris Biemann, Hiroshi Yoshida, Stefan Jänicke

1. Introduction

In the NII Shonan meeting 132 on modelling cultural processes during March 10-14, 2019, a set of researchers discussed ways to better understand and model cultural processes as complex, self-regulating, media based communication dynamics. In order to foster the discussion, we set up three working groups each combining the different backgrounds and countries of participants to allow for an interdisciplinary and intercultural exchange about modeling cultural processes:

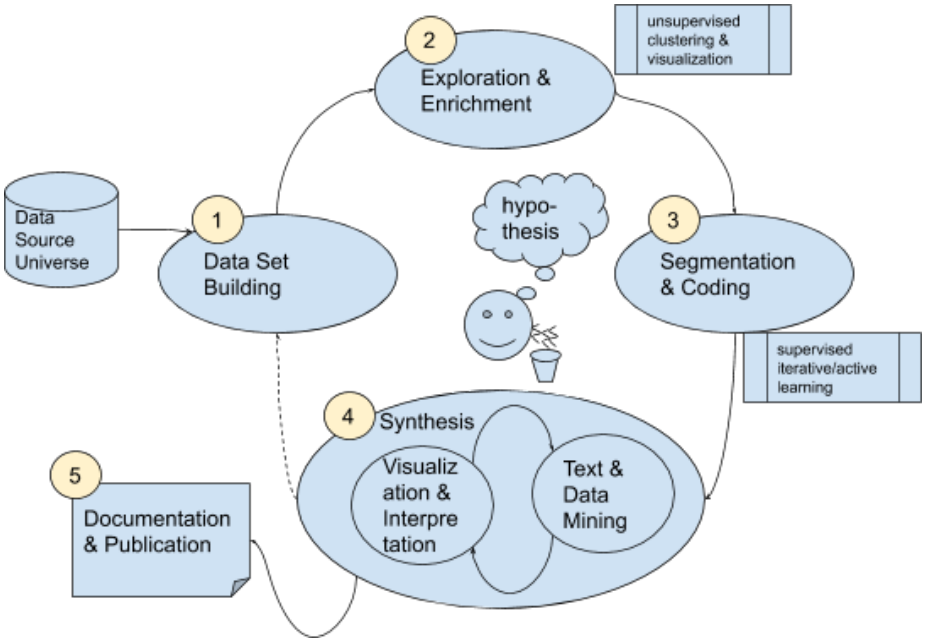
- (WG1) identity in the information age;
- (WG 2) shifting contents, shifting meanings across media and across time;
- (WG 3) constructions of culture.

The background of discussion in our working group 2 was the common practice of modelling data and processes by using meta-data on the one hand, and standard text and data mining on the other. It was agreed that the goal of the discussion was to find a unified approach that combines both practices of modelling contents, and how they change over time. From this discussion, a model for content selection, enrichment and analysis emerged that subsumes a range of use cases from the social sciences and the humanities, supported by computer science with machine learning and visualization. This model assumes a model of research as a hermeneutic circle. It breaks down datafied research into steps so as to allow us to develop guidelines for operationalization and automatization. While following established setups of science, its value lies in the explication of steps and its recommendation on how to use automation within intellectually driven research processes, as well as its wide applicability in many fields, including the digital humanities, the social sciences and investigative journalism.
The researcher with her evolving notion of a research project is placed at the center of this model. She could conceive of the research project as a question(s), hypothesis or a social concept/phenomenon to be studied. We assume that her conception will evolve as she progresses. We also assume an iterative process where she may return to earlier steps and revise decisions based on the evolution

of the project. The model is represented as a research cycle of 5 steps, illustrated and presented below. This is a simplification for purposes of interdisciplinary collaboration in the development of workflows, tools, shared vocabulary and use cases.
After defining the model and detailing the steps, playing attention to the potential of automation, we will describe a range of use cases that can be subsumed under this model.

2. The Shonan Model

Now the parts of the process, which is also depicted in figure (1) below, will be detailed.



To elaborate the concept and verify or falsify research hypotheses, the first step is the selection of the dataset (1) , which could consist of a corpus of text, media content or other primary data with associated metadata. Since any rough selection of source materials might produce unexpected results including unwanted by-catch, the next step is the Exploration and Enrichment (2) of the data. For this, we recommend the use of unsupervised clustering techniques coupled with a visualization to get an initial overview of the dataset and to quickly be able to remove outliers (e.g. with topic model visualizations or other fuzzy clustering techniques). This results in a refined, in-domain selection of materials, on which close reading and coding/annotation (3) is now performed to explicate the phenomena of interest with respect to the research question. This needs coffee. While the process of annotation/coding is manual and might involve the identification of passages and the assignment of hierarchically organized labels, it can be supported by supervised machine learning techniques that either propose labels based on previous assignments (iterative machine learning) or pick examples that might likely be assigned the same label (active learning). This

step corresponds to what is known as coding in the social sciences and annotation in linguistics and can be understood as a process that adds research project specific metadata on data - be it on entire data items such as documents as well as on parts such as passages, sentences, words, picture regions etc. These annotations/codes/subjective metadata form the basis for the subsequent synthesis step (4), where findings are aggregated, visualized and interpreted. This step feeds into the documentation of the results as well as of the process towards a scholarly publication (5). Further, following the hermeneutic cycle paradigm, this re-shapes the research question/hypothesis/concept - a step done intellectually by the researcher, triggering another iteration of the entire process, starting with an altered data selection, different exploration and enrichment, augmented annotation/coding and a more advanced synthesis. Note that the four steps do not have to be executed in sequence – it is always possible to revise previous steps on the basis of such intellectual insight.

2.1 Data Set Building (Corpus Gathering)

- Researcher starts with initial intuitions or a vague idea of the concepts or phenomena to be investigated. She gathers a collection of data with which to study the phenomena. This may involve scraping, or selection from larger collections, crawling the internet, or even creation of data. Generally, there is a universe of data, where the dataset is selected from on the basis of search queries or metadata ranges (bulk selection, not individual selection, to be able to scale to large datasets).

- We assume there is a theory driving the framing of the research question and hence building of the data set. This might be a naïve theory that will be reformulated in the research (which may then lead to a rebuilding/altering of the data set.)

2.2. Exploration and Metadata Enrichment

The computational task at this stage is unsupervised clustering. Topic modeling as one such method can be used to define subsets relevant to a research community. These subsets should be public and contain metadata commonly agreed upon. Metadata for such subcollections can be retrieved by search engines and aggregated by metadata harvesting.

- Researcher explores the data to understand it and begin to formulate hypotheses that can be operationalized. Exploring the dataset is also a way of exploring her own naïve formulation of the project. She is focused on better specification of the concepts/phenomena and the ways in which they can be recognised in the data.

  - Manual methods - close reading - browsing and looking into the data. Also, surveying the metadata. For example she might skim the titles, authors, and dates of the items gathered by some process or she might randomly read some items closely.
  - Automated methods, e.g. unsupervised clustering, possibly interactive and iterative, coupled with visualisation methods, for a targeted and quick bulk selection of relevant material.

  - Visualization aids at illustrating characteristic features of the data set. Those might be statistical overviews like (interactive) charts and contentual overviews like tag clouds. The purpose is to deliver the shape of the current data set in a visual form and to guide the scholar to feature dimensions with lacking or over-represented information. Visualizations can further be designed to support annotation.

- Enrichment of the data items (of any granularity) with metadata. She may be drawing on disciplinary research to add metadata to the items that will allow her to ask certain questions.
  - acquired from the original sources,
  - manually created following the comparison of concepts/phenomena against the collected data

- Note that this is where one type of shift in content takes place as metadata gets layered on content, explicating aspects relevant to the current research question, which might or might not be shared with other researchers as additional metadata.

2.3. Segmentation and Coding

The computational task at this stage is iterative supervised learning/classification. Specific concepts (as defined by the researcher) are being applied by way of annotation to a fine grained layer of text (such as words or sentences). Such classes can be considered to be metadata as well. The generation of annotated text can be supported by machine learning such as active learning or bootstrapping approaches. - It needs to be noted that the researcher can always interact and correct the annotations

- Now the researcher begins to analyze (in the sense of break into parts) and then code or annotate things in the data set that are of research interest.

- She makes decisions about the granularity of data items to be coded (annotated) by concepts significant to the phenomena under investigation.

- Preparation of the coding (annotation) guidelines
  - iterative process with several experimental phases aimed at achieving good inter-annotator agreement,
  - possibly supported by automatic tools, e.g. in a form of test training and annotation, Active Learning, Bootstrapping, Iterative supervised learning etc.
  - Semi-automatic annotation editors: passage identification and classification / coded annotation. Iterative annotation: suggest codes and spans during the process of linearly annotating a set of documents. Active learning annotation: use tools that suggests similar passages in automatically chosen order.

2.4. Synthesis

The goal of this step is to find patterns in the annotated data. It can be supported by a number of approaches and tools, such as statistical analyses, neural nets, or pattern based analyses. Examples are network analysis, time

series analysis, anomaly detection. Many of those tools are linked to specific visualizations and call for interactive fine tuning. Results of this stage may also generate metadata that can be used for further annotation of text. Results of the text and data mining stage need to be interpreted by the researcher. Visualization at this stage summarises data generated by the text and data mining stage.

- At this point, the researcher begins to synthesize an interpretation of the phenomena as represented by the annotated data. She will be using tools that propose views on the annotated data that may prompt insights into the phenomenon.
- Coded (annotated) data from the previous step are the basis for constructing (setting up or tuning the)
- Distant reading vs quantitative analysis vs computational analysis, e.g. by the Text and Data Mining methods
- Human analysis: Interpretation & Visualization
- Visualizations are designed to support hypothesis verification and generation. Distant reading visualizations show summaries of text/data mining results. Close readings allow inspection of individual data/text elements. A seamless transition between both perspectives is necessary to build trust in the underlying computational analysis methods: it is imperative that synthesized results allow accessing the underlying data to ensure full provenance.

2.5. Documentation

During the whole workflow, metadata have been generated semi-automatically for (1) the definition of subcorpora, (2) the fine-grained annotation of text, and (3) the results of the text and data mining analysis. All of these metadata can be used to evaluate the research workflow and adjust it to better fit the research question where necessary.

- Finally, the researcher wants to document the insights (interpretations) in ways that can be shared with others. She may be exporting passages (with citations), statistical results, visualizations, and so on for use in publications.
- She may also want to prepare a study dataset to be published with documentation for others to replicate her findings or for teaching.
- The enriched dataset can become a dataset gathered for another project thus closing the loop.
- She may want to document the processes of the previous research steps so as to try the same research flow on a different dataset.

It must be emphasized again that this process is not strictly sequential and does not follow a waterfall model. Rather, the researcher can always use intellectual revising to enter another loop, or to go back to previous steps in the current loop, in order to e.g. adjust the data set selection, revise the enrichment, alter the coding scheme etc. This is symbolized by the dashed line between steps 4 and 1 in the figure above.

3. Use Cases

Use Case 1: Humanities

1. Corpus Creation
   *Hathi Trust to study irony in Tudor drama

2. Exploration and enrichment
   Adding metadata about the plays from my knowledge of Tudor drama

3. Segmentation and Coding
   Searching for known ironic passages in known plays and coding them. Using tools to find similar passages. Training a machine to propose ironic passages. Eventually it might be possible for the machine to code the rest.

4. a. Text and Data Mining
   ...
   b. Visualization, Interpretation, Quotation
   Visualize distribution of passages over time, over authors, and over comedies/tragedies. Begin to form interpretation about Tudor irony.

5. Documentation
   Export the visualizations that make my point. Gather statistics and example quotes. Export a process visualization for the book. Export an "archive" that will be published online with the paper/book for others to use in recapitulating my results.

6. Feedback Loop to research question, theory

Use Case 2: Social Science

1. Corpus Creation
   Starting with a research question (rooted in a certain metatheoretical paradigm and an ongoing scientific controversy in our discipline)

   - Select relevant data sources in order to trace down the "social fact" you are interested in like "identity" or "power" (e.g. electronic text archives, ...)
   - Segmentation of texts, importing text in database, metadata, preprocessing
   - Cleaning the raw corpus (doublets, sampling errors / false positives)

2. Exploration and enrichment
   Getting an overview of the corpus (e.g. frequencies, collocations, topic models, first time series, ...) Adding notes or tags (new meta data) about those preliminary first findings

3. Segmentation and Coding
   Searching for highly interesting

   - Subcorpora
   - Text passages

Manual coding / annotation of a (random or layered random) sample of texts

- Coding interface would be helpful
- ti-marking function would be helpful (also in order to link it to learning algorithms in the future)
- Functionality for continuous intercoder-reliability (on different levels) would be helpful (compare e-Identity, RECON)
- For us also output functions for smaller subcorpora were helpful in order to process them further in commercial software tools like Atlas.ti or Provalis (not everything can / needs to be reprogrammed in the project specific tools)

4. a. Text and Data Mining
   Yes
   b. Visualization, Interpretation, Quotation
   Yes.

   - Nice visualization helps, however, certain fancy functions cannot be printed in scientific articles or books (obstacles: moving depictions, 3D, copyright issues for pictures ...)
   - Data output functions are helpful in order to allow for the statistical analysis of the data generated from the text data in common statistical analysis (e.g. descriptive statistics, regressions, ARIMA time series, ...) output not just of the graphs but also the relevant coefficients, tests of the preconditions that allow to use a certain statistics ....(necessary for documentation of the whole)
   - Partly statistical analysis gains value from combination with other data sources (surveys, event data, demographic or macro-economic)

5. Documentation

   - Export visualizations
   - Find the respective typical and / or exceptional quotations
   - Process visualisation is helpful (however, the whole research process / cycle usually is not just in one tool)
   - If done with SPSS or R in highly quantitative studies it becomes more an more common to publish also the syntax
   - Publication in scientific journals, papers books nice, if for both project partners (IT, SocSc) it results in innovative contributions to their resp. fields (here it is also good to have agreed in advance on a common publication strategy: e.g. in which types of journals which authors will be named first, ... how to cite each others earlier work...)
   - Of course, we write our social science text ourselfs :-), however for interdisciplinary projects it is important to have time for that. Software creation, data work etc. cannot last up to the end of the project duration. There must be in between results to work with.

6. Feedback Loop to research question, theory

    - Wrap up, what worked and what not
    - Usually a possible reframing of theory, research questions, or consideration of new corpora is not done within one project, but rather in the planning of the next projects building on recently made experiences

Use Case 3: Disaster Archive

1. Corpus Creation
   Collecting photographs and other various resources for the future use by local governments, schools and research institutions

2. Exploration and enrichment
   Analyze the type of collected materials collecting know-hows for metadata creation Create metadata for each collected material Subject headings, thesauri don't exist. Folksonomy? Need resources such as geographical names and their changes over time

3. Segmentation and Coding
   Keywords, location, temporal information analysis Photographic image analysis — seems difficult without contextual information

4. a. Text and Data Mining
   Yes Technologies required: Image analysis Topic detection Metadata aggregation
   b. Visualization, Interpretation, Quotation
   Metadata visualization to help access

5. Documentation
   Annotations for linking the resources with community memory Geographic/temporal annotations...

6. Feedback Loop to research question, theory

Use Case 4: Investigative Journalism (new/s/leak project) www.newsleak.io

1. Corpus Creation
   Background corpus given by a set of leaked documents. Subcorpora are selected by metadata and fulltext search.

2. Exploration and enrichment
   Tool shows network of named entities and keywords from sub-corpus. Users can iteratively refine sub-corpus by including/excluding selectors.

3. Segmentation and Coding
   Documents can be tagged, thus marked as belonging to a specific case under investigation. Entity labels can be altered, added, allowing for the annotation of relevant keywords.

4. a. Text and Data Mining
   The entire tool is an interactive data mining environment
   b. Visualization, Interpretation, Quotation
   .. which has an interactive visualization.

5. Documentation
   Views and selectors can be saved for later use. In investigative journalism, leaks are the sources for further investigations that typically happen outside of the data collection, so underpinning of results is not suo much of an issue in this use case.

6. Feedback Loop to research question, theory
   This use-case integrates the feedback loop very tightly. A larger instantiation of feedback loop would be to add more background data on the basis of findings, but this does not match the reality of document leaks, which are typically one-time events.

Use Case 5: Creating Subject Data for Video Games

1. Corpus Creation
   Gathering text and/or metadata from the source of the records. (e.g. Wikipedia, Wikidata, Mobygames or any type of references)

2. Exploration and enrichment
   Getting an overview of the data, Analysing the type of the components (chapter).

3. Segmentation and Coding
   Making structured data from the analysis.

4. a. Text and Data Mining
   Find the keywords (subjects/topics) for works of video game through the automatization (e.g. text mining, topic model)
   b. Visualization, Interpretation, Quotation

   - Interpretation of the result from the mining is needed for the creating useful data.
   - Some type of visualization (e.g. network or cluster) will support to the analysis.

5. Documentation

   - Publish the created data on the online catalogue
   - Write the text of guideline for explain the spec of the describing elements (items)
   - Published the research paper of this analysis (if we can)

6. Feedback Loop to research question, theory
   Collect user's response on online catalog. Evaluate/Critique published data for generating new data or research.

Seems to be another instantiation: Text Mining for Qualitative Data Analysis
CK: Gregor Wiedemann worked with Gerhard Heyer in Leipzig in the eHumanities Project "Postdemokratie und Neoliberalismus", constructing the Leipzig Corpus Miner