Methods for User-Centered Design and Evaluation of Text Analysis Tools in a Digital History Project
Ben Heuwing, Thomas Mandl, Christa Womser-Hacker
Department of Information Science and Natural Language Processing, University of Hildesheim
Universitätsplatz 1, 31141 Hildesheim, Germany
[heuwing|mandl|womser]@uni-hildesheim.de

ABSTRACT
This paper reports on the user centered, formative evaluation of tools and the validation of models for the analysis of historical textbooks in the context of the digital history project Children and their World. The goal of the project is to create methods for computer-supported, interactive analysis that can be applied to a large corpus of historical textbooks on history and geography (~5000 volumes). A first version of a tool for text analysis has been created based on a user centered design process, including a contextual study on current work practices of historians, participative design workshops, and the prioritization of requirements with the project stakeholders. In addition, several generations of text models used in these tools have been iteratively evaluated by the historians in the project. In the context of a cooperative validation study, researchers have used these tools to examine existing hypotheses from the field. The method enables the validation of text models regarding established knowledge, provides additional insights into the requirements for tools and visualizations, and helps to strengthen the expert users’ trust in the tools. Based on the findings, this paper proposes four principles for the analysis of a corpus of historical texts. Moreover, the methods presented are discussed regarding the application of user centered design in the context of digital humanities projects.

Keywords
Digital Humanities, Digital History, User Centered Design, Interactive Text Analysis

INTRODUCTION
The last years have seen a rising interest in the area of humanities computing, or digital humanities. The analysis and interpretation of texts is certainly the most important research activity in the humanities which needs to be supported by digital humanities tools (Toms & O’Brien, 2008). In general, humanists expect tools that respect and gradually enhance their existing research practices (Gibbs & Owens, 2012). However, user groups in the humanities are highly diverse, ranging from sporadic humanist users, which use tools that are optimized for a specific purpose and a specific set of data, through humanist expert-users of highly configurable tools, to active humanist programmers making use of available programming libraries. In addition, scholars from different disciplines in the humanities have distinct goals when analyzing and interpreting texts, which results in differences in their information behavior (Palmer & Cragin, 2008). Because of this diversity of processes and motivations, the specific research interests of the intended users have to be carefully studied before the findings are translated into optimized designs. In the context of research tools for the humanities, user centered design will not only have to inform the appearance and the utility of an interface, but also the transparency of automated processing steps of text mining and the utility of the models created. General advice on user centered development in the context of digital humanities projects is presented by Warwick (2012). However, the application of these methods in digital humanities projects has rarely been reported on.

The project Children and their World presented in this paper1 confronts the problem of how the world was represented in textbooks in the period of the German Empire between 1871 and 1914. For most children in this period, the textbooks that they used in school were their most important or even their only source of knowledge about the world outside of their direct experience. These popular media therefore are of high value for the research on the history of knowledge. The books are expected to contribute dominantly to the children’s perception of other countries and continents and to their understanding of the role of Germany in the world. For example, the exploding economic globalization during the 19th century and the simultaneously increasing role of nation states can be expected to be represented in the textbooks. Consequently, research questions of the project include the investigation of assumptions about the increase of contents regarding the national history of Germany in textbooks during this time.

---

1 http://welt-der-kinder.gei.de/en/the-project/
and changes in the descriptions of colonies and emigration. In addition to changes over time, researchers want to compare textbooks for different school types, regions in Germany, and religious denominations.

The version of the corpus used in our experiments currently consists of 2935 textbooks with 605830 pages, and mostly covers textbooks on history and geography. The corpus (example in Figure 1) has been digitized and is publicly accessible. In this context, the primary user group consists of the researchers in the project, i.e. expert users which have general knowledge about digital humanities methods and can configure and use the tools developed in the project as well as existing digital humanities tools. However, a core set of the functionality will be made available publicly, so a secondary user group of more sporadic users of the tools is targeted as well, mostly from the specialization of textbook research.

In this context, it is of importance to understand existing processes, but to also provide a creative environment to include the participating historians as expert users into the processes of innovation and development. The project therefore adopted strategies from user centered design, including contextual user studies and participative workshops. In addition, during development a methodology is needed which helps to estimate the utility of the resulting tools from a domain perspective, because the corpus is prohibitively large for manual inspection and comparison. Therefore, an evaluation study of a first generation of tools has been conducted which systematically validates the text models for interactive analysis relative to established knowledge in the domain. In this study, analysis of existing hypotheses is conducted interactively and jointly by a domain expert and an expert for the tools provided. The idea for the study has been derived from existing research in the digital humanities (DiMaggio, Nag, & Blei, 2013; Evans, 2014) and has been formalized as a method which, in addition to the primary goal of validating modeling techniques, also helps to establish trust in the technologies provided. The method is also well suited for formative evaluation, because during the cooperation of the experts new approaches to analysis can be developed. In the process, existing problems and new ideas can be derived from the conversation of the experts, similar to the approach of co-discovery in user testing (e.g. Rieh, Yang, Yakel, & Markey, 2010). The study focuses on the validation of text models and the formative evaluation of the tools. In addition, the user centered process used to create the tools in the project is presented and discussed as an example of how to incorporate user centered design into projects in the digital humanities.

RELATED WORK
We will give an overview on existing approaches to user research in the digital humanities and on methods for text analysis in the area of digital history, focusing on topic models and methods for their evaluation.

User Studies in the Digital Humanities
While a technical focus still appears to be predominant in many digital humanities projects, the application of user centered design methods has been promoted for the creation of new tools (Gibbs & Owens, 2012; Warwick, 2012). The variety of users of digital tools in the humanities is made visible through the broad conception of such tools, ranging from software libraries to be used in scripting environments, to configurable, general-purpose text-mining software, and to tools and interfaces created and offered specifically to explore an existing data-set. In the field of historical research, researchers have been found to expect digital tools to improve their existing research processes and to support the methods they already employ (Gibbs & Owens, 2012). If new tools require the introduction of new methods and procedures, these should therefore be consistent with established approaches. The information behavior of historians emphasizes the collection of contextual information about sources and their assessment regarding a current research interest (Rhee, 2012). A tendency to distrust results of text-mining has been noted in the discipline (Gibbs & Owens, 2012).

Regarding the analysis of sources, researchers in the humanities may choose to select, to interlink and to

---

2 http://gei-digital.gei.de - Strötgen (2014) provides further information about the corpus

3 For an example cf. a collection of programming tutorials for historians - http://programminghistorian.org

4 E.g. Voyant, a collection of tools for text-analysis - http://voyant-tools.org

5 E.g. the interactive maps of the American Panorama project - http://dsl.richmond.edu/panorama
combine contents for further analysis (Collins & Jubb, 2012). In a survey on the use of tools, researchers in the area of digital humanities rate those tools that enable analysis by comparing texts, creating concordances (textual context), and analysing distributions of words or phrases as most useful to their research practices (Toms & O’Brien, 2008). However, basic tools for text manipulation to be used from the command line (“UNIX tools”) are also popular among the respondents. This may indicate a perceived inflexibility of available, more specialised tools.

For projects in the digital humanities, productive interdisciplinary communication constitutes an important factor of success (Collins & Jubb, 2012; Lin, 2012). In a project, a high level of collaboration between humanists and technologists is necessary to create tools which can enhance existing research interests, e.g. when existing resources are visualised or interlinked in new ways. This communication across disciplines can benefit from methods of participatory design. For example, participative design workshops with different user groups have been conducted define requirements for the collection and analysis of archival materials (Boukhelifa, Giannisakis, Dimara, Willett, & Fekete, 2015), emphasizing the need for personal collections of materials, and the possibility of annotating texts with private notes.

Support for Text Analysis in Historical Research
An impressive collection of tools has been developed to support text analysis in the humanities, based on technologies from the areas of information retrieval, natural language processing, corpus linguistics, and text-mining. For the computer-supported study of documents in the digital humanities, the extraction of facts from the texts („Facts from Texts“ - Thaller, 2012) can be distinguished from the analysis of the ways these text represent knowledge and sentiments about the world („Text as Text“ - Thaller, 2012), the latter including studies of historical semantics, historical literary studies, and the study of the history of science, or, in the case of our project, more generally the history of the representation of knowledge. However, text analysis of historical documents presents a number of challenges, primarily the quality of optical character recognition (OCR) and the general absence of conventions for spelling (Piotrowski, 2012). In many cases, no resources and tools exist which can be used with historical varieties of a language.

Term-Based Analysis
A basic analysis of term frequencies can provide important insights. For example, the variation of frequencies of relevant terms in titles of Victorian books can help to analyze the change of religious orientations in a society over time (Gibbs & Cohen, 2011). In this case, bigrams of words are extracted from the texts to examine the characteristic usage of specific terms over time. With the goal of applying tools from corpus linguistics to historical research questions, a methodology has been proposed which involves the creation of term groups by domain experts, the analysis of these terms in context to find additional relevant terms for a concept, followed by the time-based and comparative analysis of their frequencies and of their co-occurrences to analyze their usage and accompanying opinions (Baker et al., 2008). In this approach, the repeated analysis of examples displayed as concordances is intended to integrate quantitative and qualitative approaches.

Identification of Semantic Entities and Relations
In addition to the analysis of term frequencies, the identification and disambiguation of named entities such as person names, place names, or events, can be helpful to analyze the contents of texts. For example, Cybulski & Vossen (2011) report a recall rate of 49% and a precision of 57% for extracted historical events. A prominent example in historical research presents an analysis of the spatial focus of a local American newspaper over time, based on gazetteers and lexical resources (Blevins, 2014). Here, the identification of mentions of locations and places clearly does not try to extract “chunks of information” (Thaller, 2012), but to analyze the development of the discourse about these places. Additional ideas exist which can help to model discourses on historical events (Nuessli & Kaplan, 2014), and, on a more general level, on historical trends (Allen & Chu, 2014).

Statistical Models of Contents with Topic Modeling
A number of statistical approaches from the area of text mining can be applied to analyze text collections on a more abstract level, including unsupervised document clustering and topic modeling as well as supervised approaches such as document classification. Topic modeling is a prominent approach to unsupervised text-mining in the digital humanities. In most studies, topic modeling is implemented using LDA (latent dirichlet allocation - Blei, Ng, & Jordan, 2003), which analyzes the contents of documents based on the assumption that those terms which co-occur frequently with each other constitute a semantically meaningful topic. The modeling process results in a list of topics, each one represented as a distribution over the vocabulary of a corpus. In practice, however, only the most important terms attributed to the topic are taken into consideration when interpreting a topic. The distribution of terms for a topic can be used to infer its relevance for a document, resulting in a distribution over all topics for each document in the corpus.

In order to generate topic models which are suitable for a specific purpose, in most cases several generations of models have to be created to optimize the parameters for modeling (DiMaggio et al., 2013). The most important parameter is the number of topics to be created, which has to be specified in advance. In addition, the removal of stopwords, the selection of specific word forms, the documents or relevant text snippets used, and the definition of the delimitation of a document (e.g. sentences, paragraphs, pages or complete books) may influence the
result of the modeling process (cf. e.g. Newman & Block, 2006).

The interpretation of the topic models based on the terms and their weights in the topics often constitutes the first step of an analysis for historical research. If possible, topics are then labeled manually (e.g. Newman & Block, 2006; Evans, 2014). A possible next step in the aggregation of the results of topic-inference for all documents of a period to analyze topics over time, e.g. to explore trends in historical newspapers (Newman & Block, 2006; Yang, Torget, & Mihalcea, 2011; Eisenstein, Sun, & Klein, 2014).

Because the goal of our project is not the deduction and verification of factual information but the analysis of broad topical trends and opinions, topic modeling has been introduced as the primary method to analyze the text corpus in our project. As an unsupervised approach to text modeling, it is also expected to be less susceptible to the influence of preconceived opinions of researchers.

**Evaluation of topic models**

Evaluation of unsupervised methods for text mining, such as clustering and topic-modeling, constitutes a problem as there is no training data available for automatic comparison while at the same time most text corpora are too large to be inspected manually. Existing approaches to evaluate topic modeling mainly focus on the automatic evaluation of models. The performance of models on held-out parts of a text-corpus can be assessed automatically (Mimno, Wallach, Talley, Leenders, & McCallum, 2011), while semantic coherence of topics in a model can be assessed based on measures of term-similarity from external textual resources (e.g. Newman, Lau, Grieser, & Baldwin, 2010; Ramirez, Brena, Magatti, & Stella, 2012)

However, automatic evaluation measures do not necessarily conform to human judgements. Visualizations of topic-term matrices can be used together with an optimization of the order of topics and terms to manually assess and compare topics (Chuang, Manning, & Heer, 2012), including a comparison with user-generated topics to assess domain relevance (Chuang, Gupta, Manning, & Heer, 2013). Additional methods have been proposed to elicit human judgements about topic coherence (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009): Users are given the task to identify a random word injected into the top-terms of a topic by human judges (word intrusion), or a randomly chosen topic assignment for a document in a group of topics assigned by a topic model (topic intrusion). These approaches help to derive more objective and consistent judgements than having participants assess topics directly.

However, in the context of research projects a topic model should be understood as a “lense” (DiMaggio et al., 2013) that provides a distinct view for a specific research question. Therefore, its utility can only be assessed in relation to the research problem which it has been optimized for. In this perspective, evaluation can take place based on an exploration of the model by a domain expert, who links interesting aspects of the model and of the distributions of topics over time to the knowledge he or she has about a corpus, as well as about external events which might have influenced the contents of a corpus (e.g. Yang et al., 2011). In contrast, domain knowledge can also be stated beforehand as a set of hypotheses about the topics and their distributions in the corpus and over time (e.g. DiMaggio et al., 2013; Evans, 2014).

**RESEARCH GOALS & METHODS**

In this project, we apply methods of user centered design to design and to evaluate tools which support existing processes and to devise new and innovative approaches for the interactive analysis of historical texts. In the initial phase of the project, we needed to …

(1.) … learn about existing processes in contextual interviews
(2.) … apply the results to the project in design workshops
(3.) … optimize early statistical text models with formative evaluation,
(4.) … prioritize requirements together with the intended users of the tools.
(5.) … validate resulting text models and evaluate the tools regarding requirements of interaction and visualization

This paper focuses on a cooperative validation study of the resulting prototype for text analysis (5.), based on topic models and faceted analysis. Because statistical topic modeling has been selected as the main method of text-mining for the project, it is important to validate the models from the point of view of users and with realistic tasks. While the coherence of topics can also be seen as an important criteria of the optimization process, in this context it cannot be evaluated with regard to external resources for text-similarity, because what constitutes a coherent topics is determined by the corpus. Domain relevance, i.e. relevance in regard to the research questions of the project, is evaluated in direct assessments by the users. In addition, there is the need to explore the corpus regarding new and unexpected contents. However, as a prerequisite for this, the trust of the researchers into the models used for text-mining has to be established first (Chuang, Ramage, Manning, & Heer, 2012). In addition, the results of the cooperative validation study offer insights into requirements at the level of text modeling and also helps to start a discussion about exploratory and confirmatory strategies of information analysis. The requirements and ideas for solutions resulting from this process will be presented according to four higher-level principles.

**PARTICIPATIVE DESIGN PROCESS**

First, we give an overview of the design process which led to the development tools and text models in the project, which incorporates ideas from user centered and from participative design.
Contextual Interviews

Contextual observations have been recommended for early requirements analysis in the digital humanities (Warwick, 2012). A contextual interview study with five participants was conducted to analyze existing information behavior and research processes of the wider group of users who specialize in textbook research. The interviews provide detailed insights into existing research processes of historians in a specialized field, the analysis of educational media. The study focuses on the structuring of results as part of the research process (“sensemaking” - Pirolli & Card, 2005). Conducting the interviews in the context of the working-place of the participant made it possible to analyze information structures in documents and tools used by the participants. Results show that during the initial phases of a project, researchers collect references to available sources. For a more detailed analysis, example documents are analyzed in order to find interesting categories. At this stage, quantitative overviews of sources and their contents can support the definition of sub-topics and research goals. The analysis of sources is often interwoven with the writing process.

Workshops

The results of these interviews have been discussed internally with the stakeholders of the project. In these workshops, the two historians working in the project represent the user group. During the first workshop, participants discuss the concepts identified in the interviews and organize them as an affinity diagram. Discussing these concepts from several perspectives helps to establish a common vocabulary among project members. Groups of concepts that emerge included the dimensions of analysis accessible via metadata compared to aspects of automated text annotation, aspects of the content in the textbooks, and the interpretation of these contents. In a following workshop, project members discuss the process and activities of analysis. Among other findings, participants state that, while being important to create an overview at the beginning of a project, quantitative analysis of sources would also be used for more detailed analysis later in the research process.

Prototyping

Early in the project, a prototype which enables search and analysis of the complete corpus was made available to the team (Figure 2). The prototype was continuously improved with new models and features for analysis. It is based on Apache Solr as a retrieval engine and generates visualizations based on the results of the integrated statistics functions.

At the time of the study, the prototype provided the possibility to formulate keyword searches and to use a wide range of dimensions of metadata as faceted filters, including year, region, place of publication, the number of editions, the sub-collection, a general subject area, the type of

6 http://lucene.apache.org/solr
7 Using D3.js, a framework for data visualizations
school, type of book, the religious denomination and gender of students targeted. For the analysis of aggregates, a histogram of result-counts by year of publication is created for each result set. In addition, the histogram displays the distribution of the average intensity of a currently selected topic over time. Other dimensions of metadata can be selected to create statistics on result counts and topic-intensity as cross tables. The intensity of all topics in a topic model for the current document selection can be displayed as a list, in addition presenting its distribution along a chosen facet as a heatmap. The topic models that are offered in the interface consist of 50, 100, and 200 topics, with an additional topic model emphasizing sentiment-bearing words. In addition to this data-driven, functional prototype, a set of wireframes displays additional functionality for analysis as an artifact on which to base the discussion of further requirements.

Prioritization of Goals
Based on the concepts and activities from the studies, a set of goals are formulated to discuss the priorities in the project. The formulated goals represent higher level system requirements, similar to epics in agile development processes (Patton, 2014), and consist of a title and a description of the goal from a user’s point of view. More comprehensive but longer usage scenarios, which have been proposed earlier, had not elicited much discussion. This is possibly a result of participants perceiving scenarios as prescriptive solutions to be accepted or rejected as a whole.

A first selection of goals is presented to the domain experts during a subsequent workshop. During a discussion of these goals, missing goals are added and unclear goals are rephrased or split into separate goals. The domain experts then sort the goals according to their priority for the project. Among the highest ranked functionality are the need to (1.) find representations of topics relevant for specific research questions, to (2.) analyze the distribution of topics across contexts while (3.) controlling for effects of the composition of the corpus during analysis, and to (4.) find relevant documents for specific topics.

Optimization of Topic Models
The primary method intended for accessing and analyzing documents and sub-corpora in the project are topic models based on LDA (Latent Dirichlet Allocation - Blei et al., 2003). In parallel to requirements elicitation and the design of the user interface, members of the project also assessed several generations of such models to optimize the parameters used for creation, for example the number of topics, the selection of word forms, the definition of pages or sentences as documents for processing, and the correction of errors resulting from the OCR-process. The domain experts contributed to this process, rating the coherence of topics and the comprehensiveness of topic models relative to the range of research questions defined for the project. Participants first tried to find names for a topic based on the ten most relevant terms for the topics, and then indicated words they would expect in the topic, words missing from the topic, as well as errors in the words, e.g. because of problems with text recognition or tokenization.

Providing names for the topics required considerable efforts of the domain experts, but also provided valuable insights into the coherence of topics and the expectation of the users. Researchers in the project appear to expect topics to be coherent in a historical sense (note that most of the textbooks in the corpus cover the subject of history), e.g. grouping together terms and names belonging to one historical era, instead of creating topics which group terms e.g. with a focus on government and administration or on war in general. As the domain experts in many cases had to resort to the documents assigned to the topics to arrive to an interpretation of a topic, the quality of the document assignment was added to the criteria for evaluation. In addition, a method for semi-automated evaluation of document retrieval has been created to enable faster iterations when generating topic models (cf. Schnoer & Gurevych, 2015).

COOPERATIVE VALIDATION OF MODELS
For the topics, tools, and result-visualizations created, it is important to validate how meaningful the representation of the topic model appears from the point of view of domain experts. Two historians from the project analyse the corpus regarding several hypotheses from historical research because its size prevents manual inspection. For these hypotheses, results from historical research already exist. From a valid modeling approach, we would expect the majority of hypotheses to be confirmed, even in the case that not all of them are correct. This situated approach to the validation of topic models has already been used to validate topic models from the point of view of domain experts (e.g. DiMaggio et al., 2013; Evans, 2014). In contrast, we see the inability of rejecting or validating a hypothesis as an indicator of the validity of the tools and models provided.

Method & Participants
For this study, historians from the project together with an information scientist selected four hypotheses from previous scholarly work (relevant themes highlighted):

1. Geography books in the context of migration from Germany have a higher focus on German colonies compared to history books.
2. In textbooks for primary schools, German colonies are covered to a lesser degree.
3. An increase of contents covering the recent history of Germany and Prussia and a decrease of contents about ancient history can be expected starting from 1890 because of a new orientation in Germany's official educational policy.
4. The German navy is covered to a similar degree in history and geography books.

Each hypothesis predicts thematic differences for subcategories of German textbooks or changes over time. However, thematic definitions are on different levels of abstraction. The scope and focus of these hypotheses was chosen to be representative for the project. As a result, this study can also help to evaluate the usefulness of the approach. However, in this regard, insights are limited to confirmatory tasks as opposed to exploratory analysis. During the study, the participants try to examine the hypotheses using the topics from the topic models. In addition, they are asked to formulate term-based queries. The results of these queries provide additional indicators of the validity of the topic modeling approach.

In the prototypical implementation, not all functionality is available from the interface. Because of this, the analysis is conducted in teams, with one historian as the domain expert who is assisted by an information expert familiar with the system. In addition, external tools are employed for the collection and visualization of results. This procedure enables the analysis of the conversation between participants, similar to a co-discovery situation in user testing (e.g. Rieh et al., 2010). Consequently, the participants receive assistance during the tasks and therefore the results are not representative for the evaluation of the usability of the interface. Nevertheless, missing functionality, missing affordances for existing functionality, and visualizations that are difficult to interpret can easily be identified from the comments and the reasoning process of the participants.

After each task, the domain experts give their personal judgment about whether or not it has been possible to confirm or to disconfirm the hypothesis, and by which means they have achieved this. In addition, they rate the reliability of the results as well as their expectation of the transparency of the results, both on a scale from 1 (strongly disagree) to (strongly agree).

**Results**

For the majority of hypotheses, historians are able to confirm the hypothesis (Table 1). This is an indicator of the validity of the tools provided. However, the topic-models in some cases lack expressiveness, as thematic aspects of the research questions could not always be formulated as queries. The last hypothesis is judged as disconfirmed by one participant. Even though this contradicts the validity of the research approach, the finding was mainly based on term-queries generated by the historian, indicating that the representation in the models might still be valid.

In the case of hypothesis 3, one participant remains undecided while the other sees the hypothesis as only partly confirmed. The latter is due to the evidence found for a trend to start later than predicted in the hypothesis, in the year 1900, which is an alternative that can also be expected based on existing theories in the discipline. The outcome therefore constitutes an original contribution to the field, as it may help to decide the controversy about which event triggered a change in the contents of textbooks. This case is an example of a more interpretative and exploratory form of analysis. As part of this strategy, participants try to find explanations for trends and pattern they encountered during the analysis and interpret documents in-depth to gather reliable evidence and to formulate new hypotheses. However, for an exploratory analysis, there is a potential risk that results might be biased, e.g. if only positive results are emphasized during this process, or if research questions are formulated according to the results found. Therefore, in this context different sources of evidence should be combined to support or disprove a thesis, e.g. from term queries, a topic-based analysis, and from the interpretation of selected documents.

While not completely distinct in practice, a second strategy gives priority to the early formulation of the results that are needed to confirm or to disconfirm a hypothesis. The analysis then proceeds using the intended methods, focusing on those aspect of the results relevant to the research question. In this case, analysis focuses on aggregates while documents in most cases are only scanned to evaluate their relevance to the current research question. For both strategies, the models in many cases do not offer topics of a suitable level of granularity to match the topics formulated in the hypotheses. As a result, participants sometimes switch the level of analysis, trying to analyze a topic using sub-concepts as examples or trying to assert a hypothesis on a higher, more general level. In both cases, if the distribution within a topic is uneven across categories of interest, the results of an analysis may be misleading. In order to mitigate this problem, topics are combined in order to display generalized quantitative results for either of several topics or result counts for documents with high values for all of the topics for more specific results.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Hypothesis</th>
<th>Confirmation Topics</th>
<th>Confirmation Terms</th>
<th>Reliability</th>
<th>Transparency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>partly confirmed</td>
<td>undecided</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>confirmed</td>
<td>confirmed</td>
<td>3.5</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>confirmed</td>
<td>confirmed</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>undecided</td>
<td>undecided</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>partly confirmed</td>
<td>undecided</td>
<td>4</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>undecided</td>
<td>disconfirmed</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1: Results for the hypotheses in the validation study for each participant, their judgment whether a
initial exploration of the corpus, interesting selections and variables can be selected for a more detailed analysis. For closer analysis, interesting measures can be compared across contexts, while the influence of biases across categories has to be controlled. At the same time, historians will collect documents for further interpretation and to present results.

Entry Points for Analysis
Exploratory analysis is important to formulate hypotheses and to find measures relevant to a research question. As exploratory analysis is an iterative process, recommendations of interesting terms, categories, or topics, should be presented to users based on the current result sets. Another possibility is to provide multivariate visualizations to identify outlier and categories of interest. As the specificity of topics in the models provided has been found to not be suitable for all tasks, it should also be made easier to find relevant topics in larger topic models and to combine them for a custom analysis, creating more specific or more general, user-defined topics.

Comparative Analysis across Multiple Contexts
A particularity of text analysis for historical research is the attention given to contextualization of sources on several levels and for many dimensions, e.g. compared to the analysis of literature. This behavior confirms existing findings about the information behavior of historians (Rhee, 2012). In the project, one of these contextual layers is already represented as categorical metadata. A comparison view for two sub-collections can enable more detailed analysis and will help to find distinctive categories, topics and terms for each sub-collection in relation to the other. Surrounding contents can be seen as another level of context for historians, as e.g. in hypothesis 1, where the analysis focuses on the question whether two themes (migration and colonies) co-occur in different categories. It is therefore important to analyze the change of correlations of topics and terms over time.

Aggregating evidence from multiple methods
Findings that are based on both the results of user generated term-queries and on those of an analysis based on topics are perceived as the most reliable ones in the validation study. Therefore, it should be possible to collect and validate results with different methods of analysis, e.g. term-frequencies, frequencies of multi-word expressions, or document clustering. This strategy has been described as “[...] examining the objects of study from as many reasonable and original perspectives as possible to develop convincing interpretations.” (Sinclair, Ruecker, & Radzikowska, 2013). The extensive background knowledge contributed by the domain experts represents an additional layer of analysis, e.g. about the role of specific authors or publishers in a larger discourse, which could be represented as user-generated metadata. As could already be seen from users’ behavior during the formative evaluation of models, we expect a combined analysis of aggregated information and the inspection of representative documents for interpretation at all stages of an analysis. In fact, from the point of view of acceptance of new methods in the discipline, the discovery of interesting sources prove to be the more worthwhile activity for historians.

Detect and Manage Bias in the Collection
The solutions described in the previous sections provide support for the identification of differences and correlations within the contents of the corpus. However, the corpus is not evenly distributed across all categories, which may introduce unforeseen biases into any quantitative analysis. Tools should therefore help to identify confounding variables by visualizing potential correlations of metadata-dimensions in the corpus or make tests of independence available for a given analysis to check other possible causes of a result. For example, this could help to identify if a rapid shift in the distribution of contents can be attributed to a single author which is overrepresented in a sub-collection. It should also be possible for users to check for diverging meanings of a topic in a sub-collection, based e.g. on local term-assignments for a topic or by displaying multi-word expressions. This is especially important as the corpus is very large and heterogeneous (Schmidt, 2013). In addition, thresholds for recommendations of interesting aspects will have to be selected carefully, and result visualizations need to include indicators of variance and outliers.

DISCUSSION OF METHODS & OUTLOOK
In the context of this project, participative and cooperative methods have been found to be useful to establish interdisciplinary communication and to create ideas for new tools. In addition, new insights from the point of view of historical research could be created through an exploratory strategy of information analysis during the validation study. This is a promising signal regarding the potential of the tools that have been created.

While the validation method presented in this paper does not provide easily quantifiable measures, it helps to establish trust in the methods provided, and to elicit additional insights into requirements. However, the tasks defined in the form of hypotheses are only representative of confirmative research questions. Consequently, there is still a need to evaluate tools and models regarding their utility for exploratory information analysis. Methods for evaluation could be based on those for exploratory search (White & Roth, 2009). The interface will also have to be evaluated in usability tests. Especially for a more general group of users with less experience in text analysis.

The experiences in the project provide important insights, which we expect to be relevant for the application of user centred design methods in similar projects in the maturing field of digital humanities. The requirements for text analysis in the project appear to be equally influenced by the established processes and expectations in the discipline, and by the highly specific, but continuously evolving research questions of the project. While it was possible to
observe important aspects of the general processes of information behavior in research projects within the discipline, e.g. through the contextual interviews conducted, surprisingly it resulted to be more difficult to create a common understanding of the specific research problems of the project. This may be due to the expectation of the researchers that the direct application of existing technology can enable more objective and unbiased scientific discoveries because the tools are less influenced by the research interests. However, direct transfer of technologies without adaptation may not bring satisfying improvements for the specific research interests in the project.

We therefore argue that user centred design practices are necessary to consider existing, implicitly or explicitly defined research practices when creating tools for the humanities. In addition, methods from participative design can help to moderate the process of interdisciplinary communication. However, artefacts created during the design process should enable experimentation, for example by incorporating state of the art search technology as a first step to access the corpus. In a similar manner, existing tools, when applied to the data used in the project, can serve as prototypes to elicit specific requirements and to enhance research questions.

ACKNOWLEDGEMENTS
This research was made possible by funding of the Leibniz Association under grant number SAW-2014-GEl-2.

REFERENCES


