Welt der Kinder
Children and their World

User Centered Development for the DH – Experiences from the project

Ben Heuwing, Christa Womser-Hacker, Thomas Mandl
Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim
heuwing|womser|mandl@uni-hildesheim.de

Children and their World

Assumption: Text books reflect contemporary world interpretation of children:

General school attendance started

Textbooks as main information source for young adults

No travelling for most people

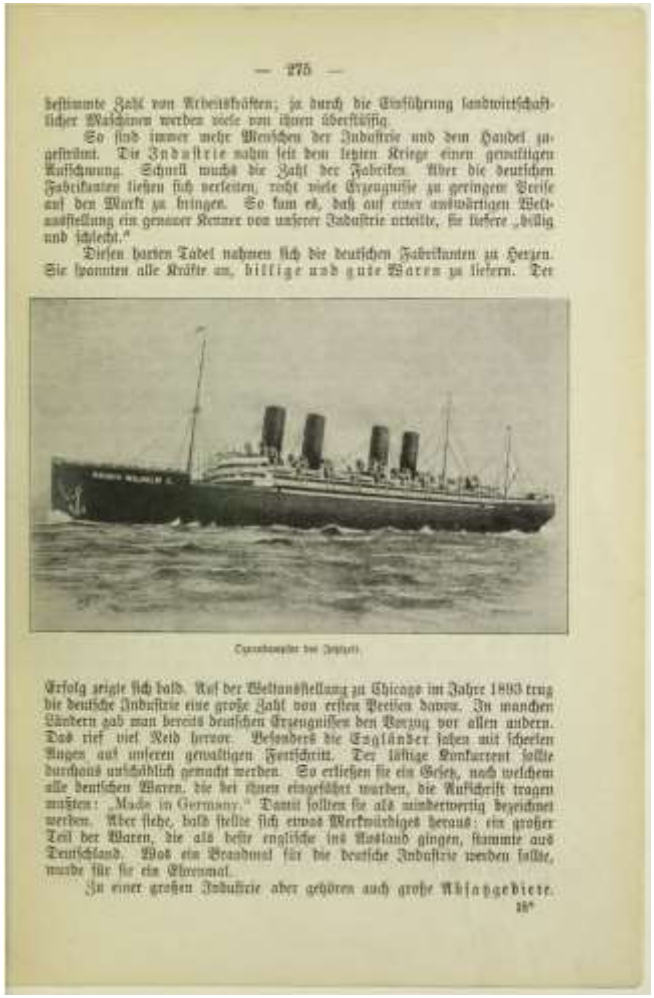
Historians want to explore German text books from the period between *1850* and *1918*

Accelerated production of knowledge

Simultaneous processes of Globalization & Nationalization

Combination

of established hermeneutic methodology with innovative DH-methods



— 275 —

bestimmte Zahl von Arbeitskräften; ja durch die Einführung landwirtschaftlicher Maschinen werden viele von ihnen überflüssig.

So sind immer mehr Menschen der Industrie und dem Handel zugeströmt. Die Industrie nahm seit dem letzten Kriege einen gewaltigen Aufschwung. Schnell wuchs die Zahl der Fabriken. Aber die deutschen Fabrikanten ließen sich verleiten, recht viele Erzeugnisse zu geringem Preise auf den Markt zu bringen. So kam es, daß auf einer auswärtigen Weltausstellung ein genauer Kenner von unserer Industrie urteilte, sie liefere „billig und schlecht."

Diesen harten Tadel nahmen sich die deutschen Fabrikanten zu Herzen. Sie spannten alle Kräfte an, billige und gute Waren zu liefern. Der

Cyrenhampfer der Imperator.

Erfolg zeigte sich bald. Auf der Weltausstellung zu Chicago im Jahre 1893 trug die deutsche Industrie eine große Zahl von ersten Preisen davon. In manchen Ländern gab man bereits deutschen Erzeugnissen den Vorzug vor allen andern. Das rief viel Neid hervor. Besonders die Engländer sahen mit scheelen Augen auf unseren gewaltigen Fortschritt. Der lästige Konkurrent sollte durchaus unschädlich gemacht werden. So erließen sie ein Gesetz, nach welchem alle deutschen Waren, die bei ihnen eingeführt wurden, die Aufschrift tragen mußten: „Made in Germany." Damit sollten sie als minderwertig bezeichnet werden. Aber siehe, bald stellte sich etwas Merkwürdiges heraus: ein großer Teil der Waren, die als beste englische ins Ausland gingen, stammte aus Deutschland. Was ein Brandmal für die deutsche Industrie werden sollte, wurde für sie ein Ehrenmal.

Zu einer großen Industrie aber gehören auch große Absatzgebiete.

18*

Funded by the Leibniz Association

Georg Eckert Institute for International Textbook Research

Prof. Dr. Simone Lässig, Dr. Robert Strötgen, Dr. Andreas Weiß,Maik Fiedler

Ubiquitous Knowledge Processing Lab (UKP) TU Darmstadt, and German Institute for International Educational Research (DIPF)

Prof. Dr. Iryna Gurevych, Dr. Richard Eckart de Castilho, Carsten Schnober

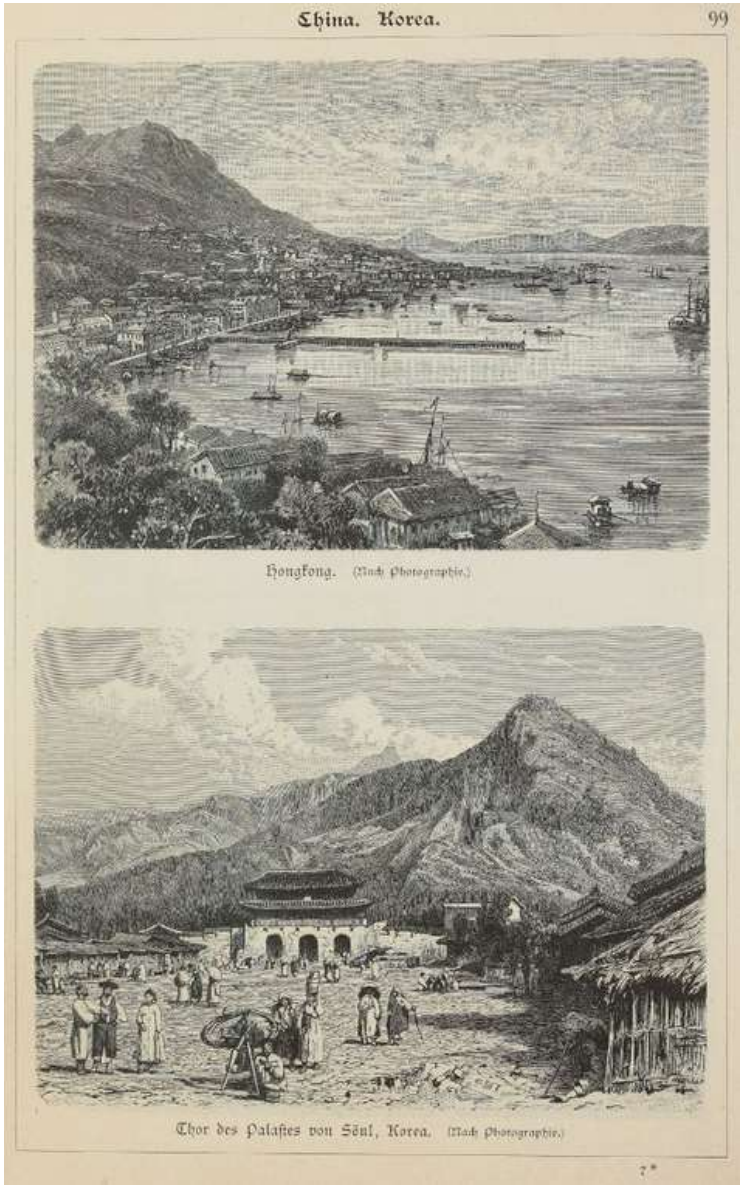
Institute for Information Science and Language Technology in Hildesheim (IWIST)

Prof. Dr. Christa Womser-Hacker, Prof. Dr. Thomas Mandl, Ben Heuwing

Other Partners

Institute of Popular Culture Studies, University of Zurich
Bavarian State Library
Göttingen Centre for Digital Humanities
Universitätsbibliothek Technische Universität Braunschweig

Research Questions: Example



Perceptions of alterity and race in the context of colonialism

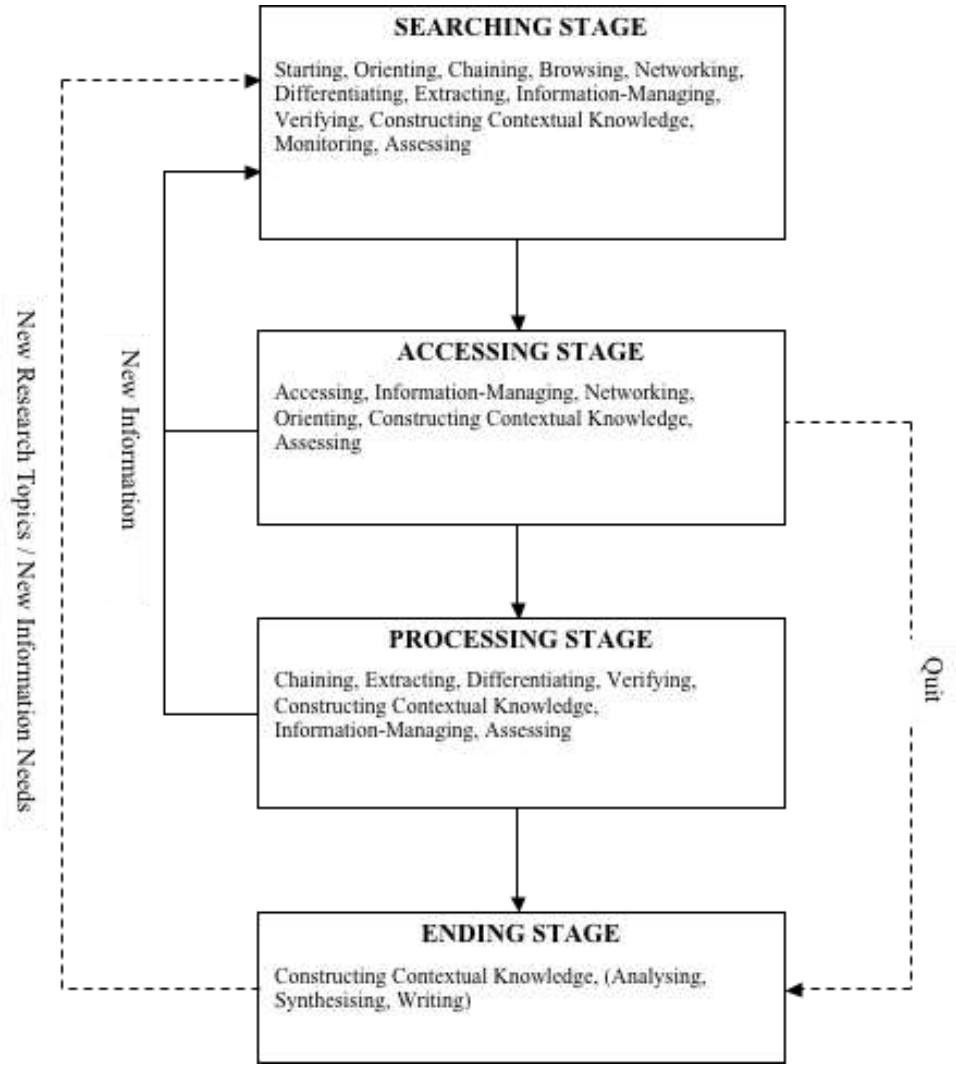
- How are foreign countries described?
- How is the role of Germany in the World presented?
- What is the role of citizens of foreign countries in Germany?
- How widespread were notions of race that were later pursued by national socialism?

Information Seeking Behavior in the DH

Requirements Analysis in the DH

Interactive Analysis of Large Text Corpora

Research Interests of Information Science



Rhee, Hea Lim (2012): Modelling historians' information-seeking behaviour with an interdisciplinary and comparative approach

Information Seeking of Historians

Starting

Orienting

Chaining

Browsing

Differentiating

Verifying

Extracting

Accessing

Monitoring

Networking

Information-managing

Constructing contextual knowledge

Assessing

Analysis? Corpus Level? Distant Reading?

Corpus

Data: ~6,000 digitized textbooks comprising ~900,000 pages

(~3,500 books, ~600,000 pages available yet)

Cannot be penetrated by hermeneutic methods alone

OCR

OCR-Errors: Estimated accuracy rate: 90-95% (characters)

Spelling Variations in the original document and across corpus

Metadata

Author, year & place of publication, publisher

Porject internal classification of: *type of school, religious denomination (catholic/protestant), grade*

Subcorpus of childrens' books ~500

to be added

Pre-processing

OCR post-processing

Goal: identify erroneous words, replace with correct spelling

Rule-based: static mapping for recurring OCR mistakes

Statistical approaches: learn recurrent errors

Pre-processing for Retrieval and Topic Modeling:

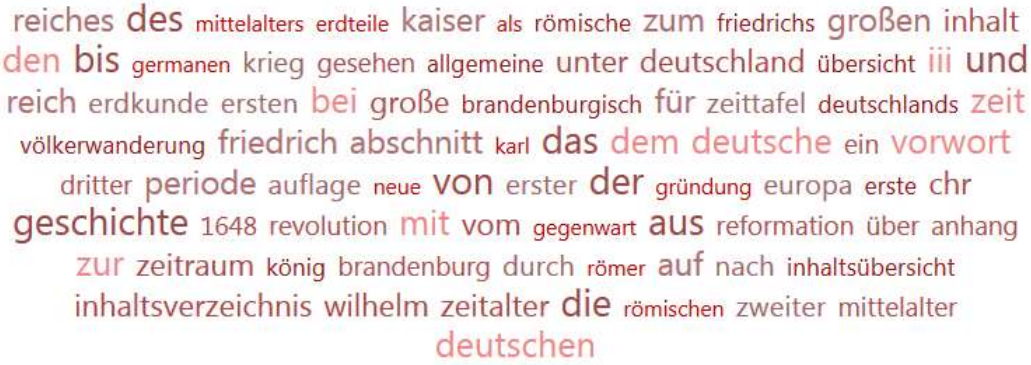
Stop-word filtering, lemmatization

Filtering: capitalized words, sentiment words, named entities

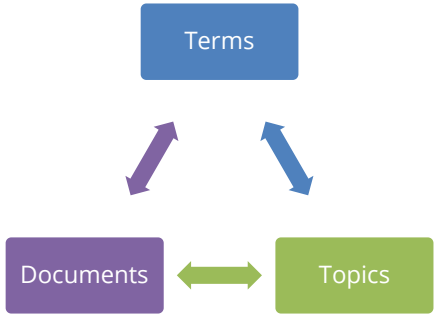
Open Problems

Normalization of orthography?

Changes in meaning?



Topic Modeling



Generative Model based on co-occurences of terms in documents

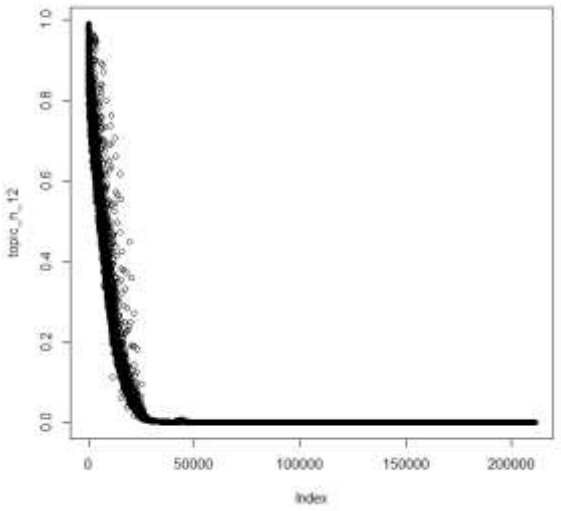
Latent Dirichlet Allocation (LDA) – Blei 2003

A topic comprises all visible vocabulary words with weights assigned

Intuition: words with largest weights are representative for a topic *t*

Application:

1. Generate *k* topics → *k* topic weights for each word in vocabulary
2. Each document is a mixture of topics
   For an (unseen) document, infer the proportion of each topic in that document (based on the model)



Topic number k: 50, 100, 150, 200, ...

Larger *k* increases topic specificity: more intuitive topics, but more redundancy and less clear inference

Sample Topics
Top 10 Words

Capital letters (200 topics):

Sieg, Eroberung, Zug, Tod, Herrschaft, Krieg, Iii, Kampf, Unterwerfung, Aufstand

Alexander, Perser, Babylon, Cyrus, Syrien, Israel, Euphrat, Aegypten, Asien, Darius

Reich, deutsch, Reiche, deutsche, Grenze, groß, Mark, Nation, Herrscher, Nachfolger

Named entities (200 topics):

Athen, Athener, Sparta, Athens, Griechenland, griechischen, Salamis, athenischen, Xerxes, Aristides

Joseph, Isaak, Abraham, Juden, Jakob, Gott, Kanaan, Israel, Abrahams, Gottes

Westfalen, Bielefeld, Minden, Paderborn, Dortmund, Herford, Arnsberg, Soest, Hagen, Hamm

Sentiment words (50 topics).

Recht, Schuld, hart, streng, Klage, verbieten, bestrafen, schwer, Weise, schuldig

Tod, Erbe, Besitz, Anspruch, Streit, Macht, Recht, anerkennen, Vertrag, Konstanz

führen, Macht, erklären, mächtig, unterstützen, unabhängig, vollständig, glänzend, Unterstützung, kräftig

Large Corpus

Rich Metadata

Assessments on Topics

Manual Annotations (?)

Lexical resources: Sentiment lexica (?)

Disambiguation of Historic Place Names, Person Names (x)

External Ontologies (DNB)

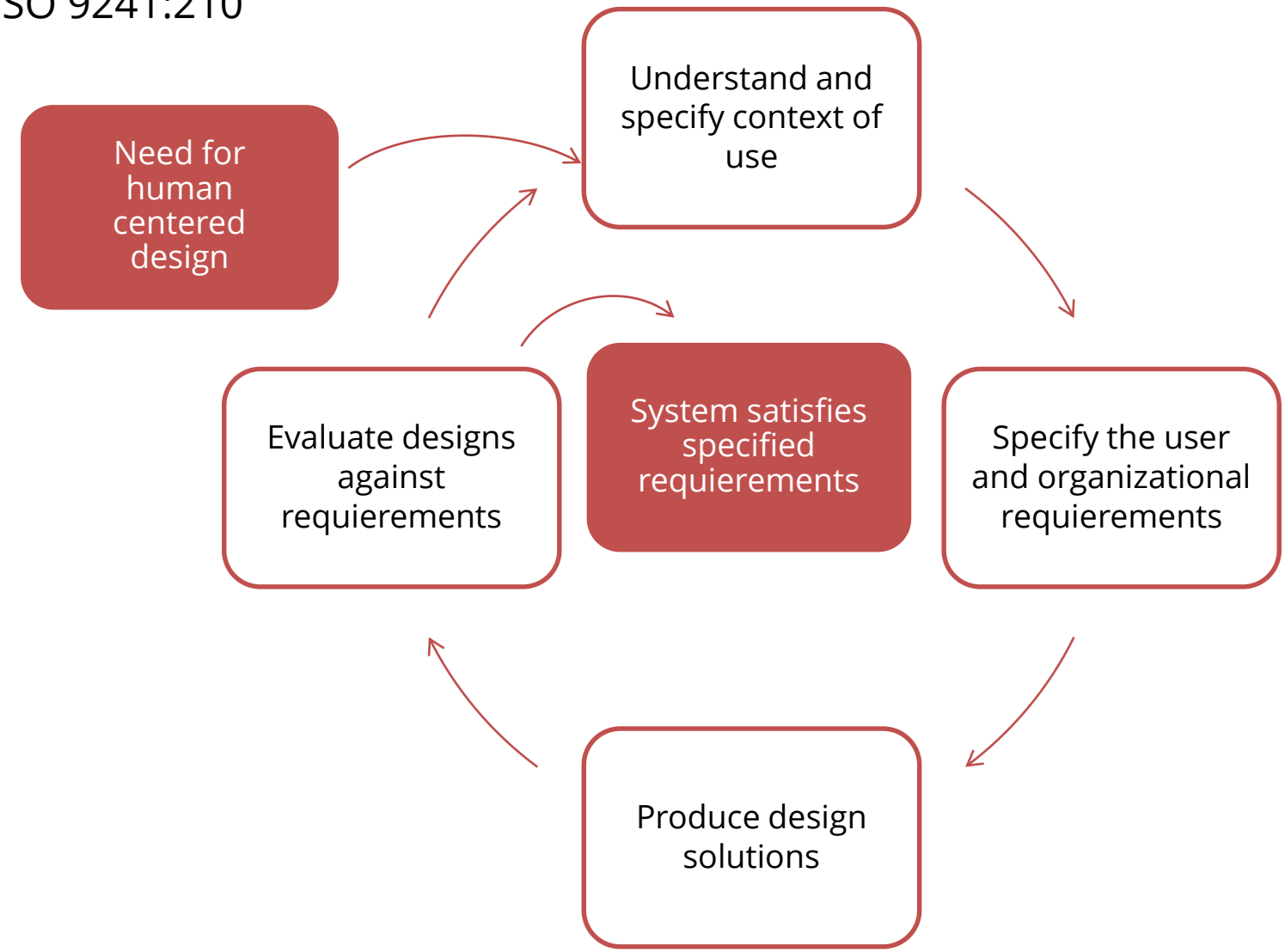
Summary: Resources

How can these resources be employed for a given research questions?

Co-Development of research methodology and technology

Human-centred design process + Participation

ISO 9241:210



Understand and specify context of use

Understand and specify context of use

Interviews

Contextual Interviews with 5 Historians

Including two from project WdK

Analysis of information representations used for research

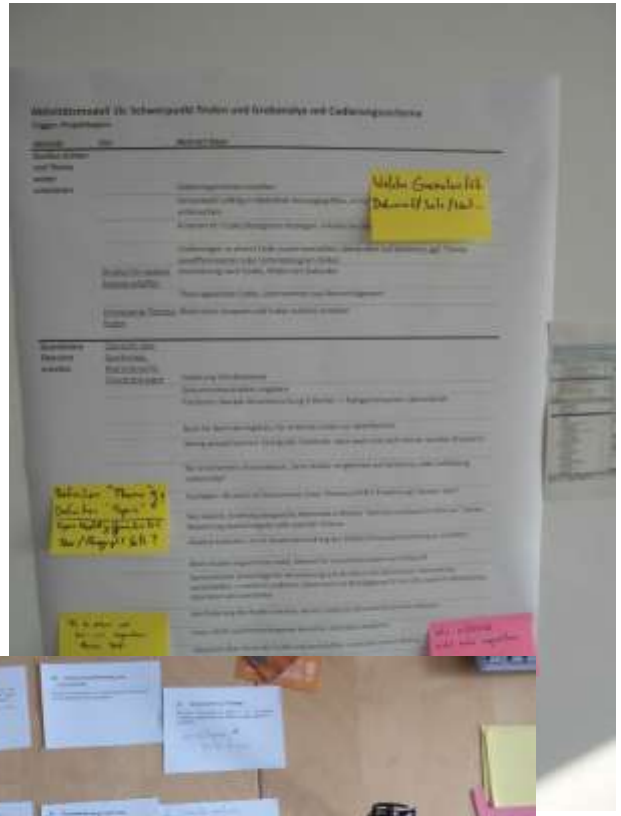
Internal Workshop to discussion results

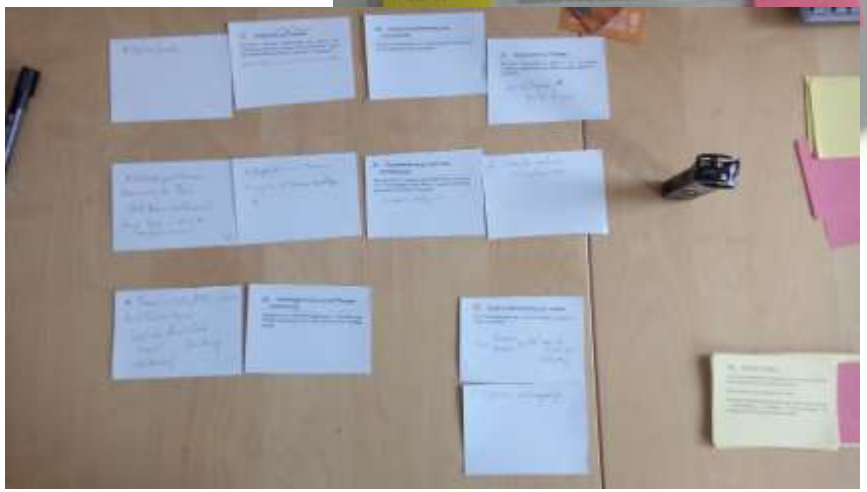
Prioritization of requirements

Finding common terminology and understanding of goals

Participation:

Cooperation and Co-Design





Specify user requirements

Proposed Workflow for analysis

1. **Discover Dimensions of Analysis**
   - Navigate through corpus
   - Find changes, pattern and trends for sub-collections
   - Create document sets and select topics for in depth analysis
   - Find relevant documents
2. **Quantitative comparisons across dimensions**
   - Find differences between sub-sets (terms, topics)
   - Direct comparisons: topic intensity, topic correlations
   - Subjunctive interface
3. **Qualitative Analysis & Comparison**
   - Identify, collect and compare relevant examples from texts
4. **Evaluate and correlate results**
   - Be aware of ones own preconceptions
   - Be aware of skewed distributions across dimensions in the collection

Entry Points for Analysis & Exploration *Serendipity*

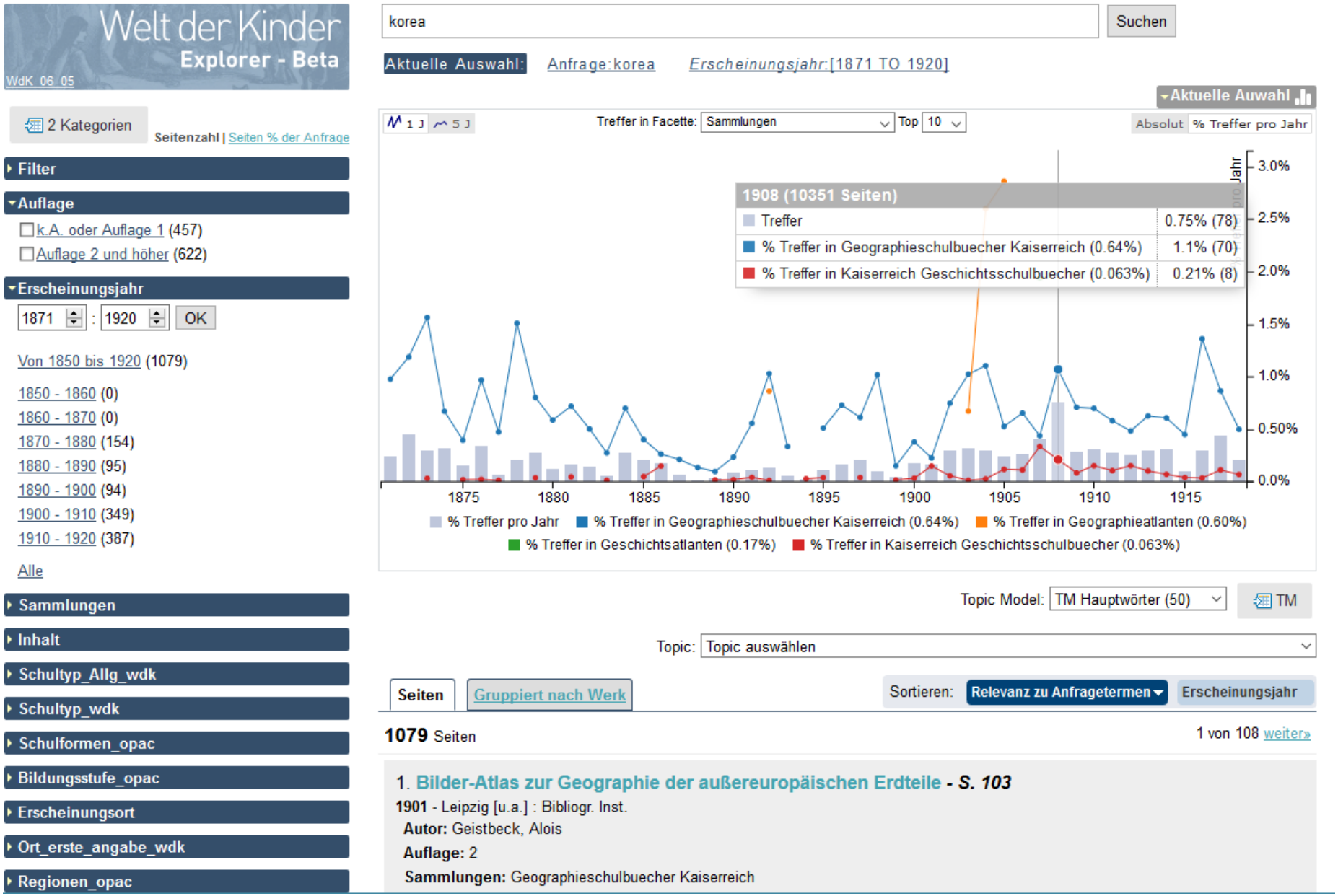
Analysis across Multiple Contexts

Evidence from Multiple Perspectives

Detect and Manage Bias in the Collection

Create Design Solutions

Working Prototype



Evaluate designs against requirements

Evaluating Topic Models

Assessment of Domain Experts (Historians)

Coherence of Topics

Value for research Questions

Dokumente zu den Topics: http://wdk.uk	Relevanz für Fragestellung			Überraschung	
Topic	*Sehr relevant*	*relevant*	*Nicht relevant*	*Ist das Topic unerwartet?*	*Kommentare*
T16: [Gebirge, Alpen, Klima, Höhe, Teil, Flüsse, Ebene, Norden, Süden, Boden]	☐	☐	☑	☐	
T15: [Heinrich, Kaiser, Otto, Friedrich, Papst, Italien, Deutschland, Sachsen,	☐	☐	☑	☐	

Task-Based Validation: Models & Interface

Cooperative, task-based exploration

Applying tools to real tasks

4 Hypotheses which are results of previous historical research

Extending Tools

Ad-hoc analysis in Excel et al.

Interdisciplinary learnings

Coverage of the recent, national history of Germany and Prussia increases

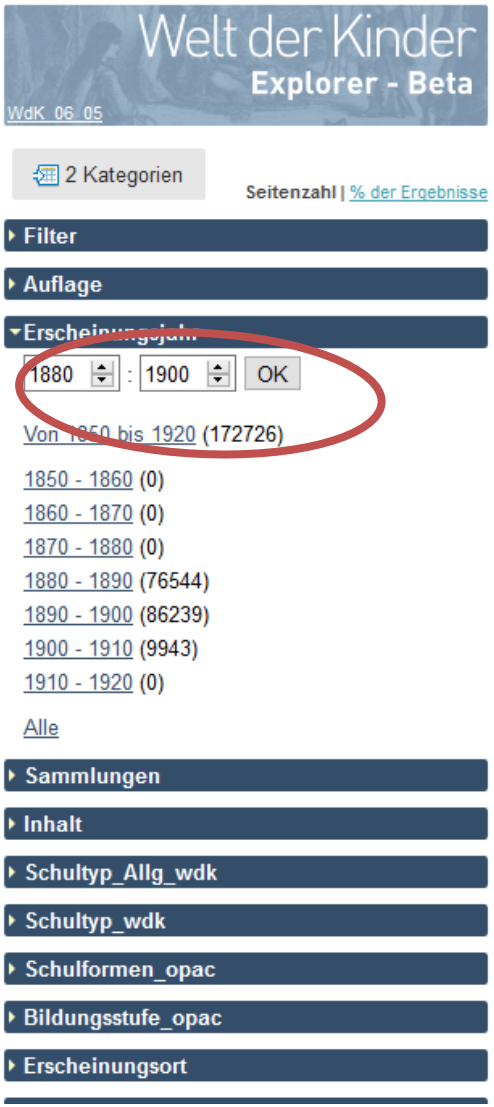
because of a new orientation in Germany's official educational policy in the year 1890,

accompanied by a decrease of contents about ancient history

Example - Hypothesis

Teilnehmer 1

Festlegen: Untersuchungszeitraum 1880-1900, Topics



#	Topic	Mittel
	fuß]	
38	T37: [luther, lehre, dichter, schrift, kirche, werk, wittenberg, universität, lied, gedicht]	0,97%
17	T16: [handel, industrie, ackerbau, gewerbe, viehzucht, stadt, arbeiter, fabrik, arbeit, verkehr]	0,89%
31	T30: [friedrich, wilhelm, könig, kaiser, iii, karl, kurfürst, prinz, johann, preußen]	0,87%
32	T31: [wald, baum, tier, pferd, boden, feld, wiese, mensch, pflanze, rind]	0,86%
49	T48: [rom, römer, italien, rmer, krieg, stadt, spanien, volk, gallien, hannibal]	0,82%
2	T1: [gott, tempel, mensch, gtter, priester, zeus, held, erde, opfer, himmel]	0,78%
39	T38: [napoleon, schlacht, franzose, heer, sieg, juni, paris, preußen, mai, leipzig]	0,76%
47	T46: [stadt, hauptstadt, straße, eisenbahn, dorf, einw., handel, verkehr, einwohner, berlin]	0,71%
1	T0: [general, armee, heer, mann, truppe, schlacht, feind, franzose, festung, preußen]	0,66%
21	T20: [athen, athener, griechenland, sparta, stadt, spartaner, flotte, perser, insel, theben]	0,61%
35	T34: [insel, meer, küste, halbinsel, ozean, afrika, kap, land, europa, teil]	0,61%
23	T22: [papst, kirche, rom, kaiser, bischof, heinrich, gregor, erzbischof, papste, vii]	0,60%

Participant 1
Topics for 1880-1900

172726 Pages

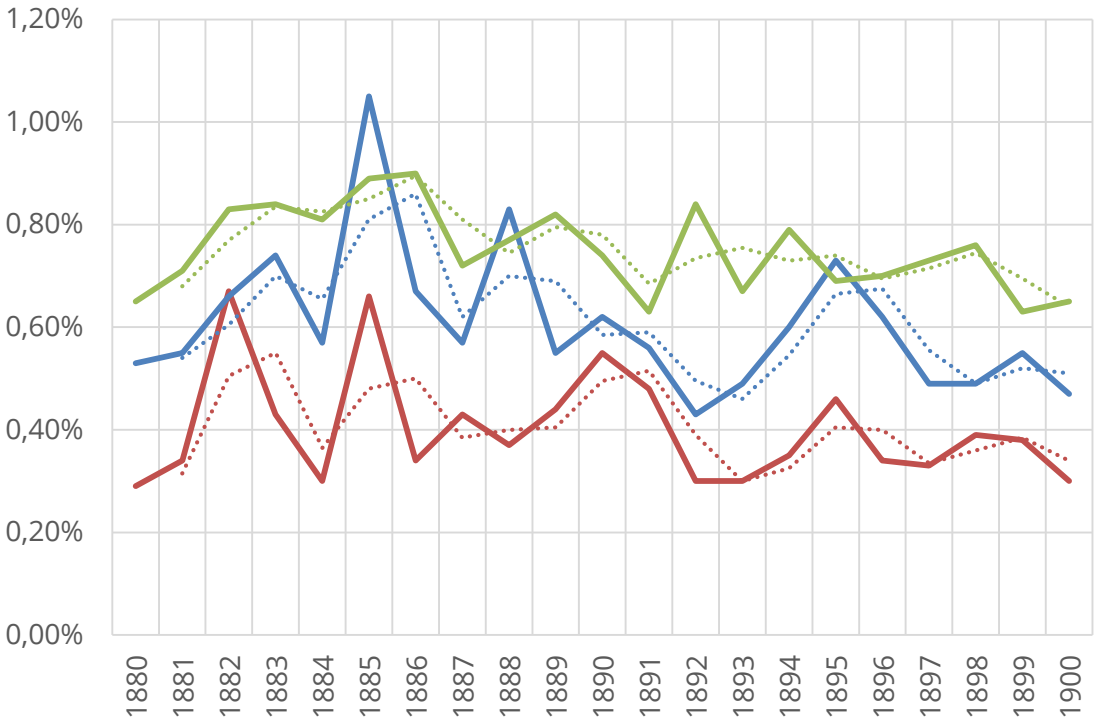
Antike:

T20: [athen, athener, griechenland, sparta, stadt, spartaner, flotte, perser, insel, theben]

T5: [rom, cäsar, senat, pompejus, jahr, sulla, antonius, marius, csar, provinz]

Französische Revolution & Befreiungskriege

T38: napleon, schlacht, franzose, heer, sieg, juni, paris, preußen, mai, leipzig] Topicfilter

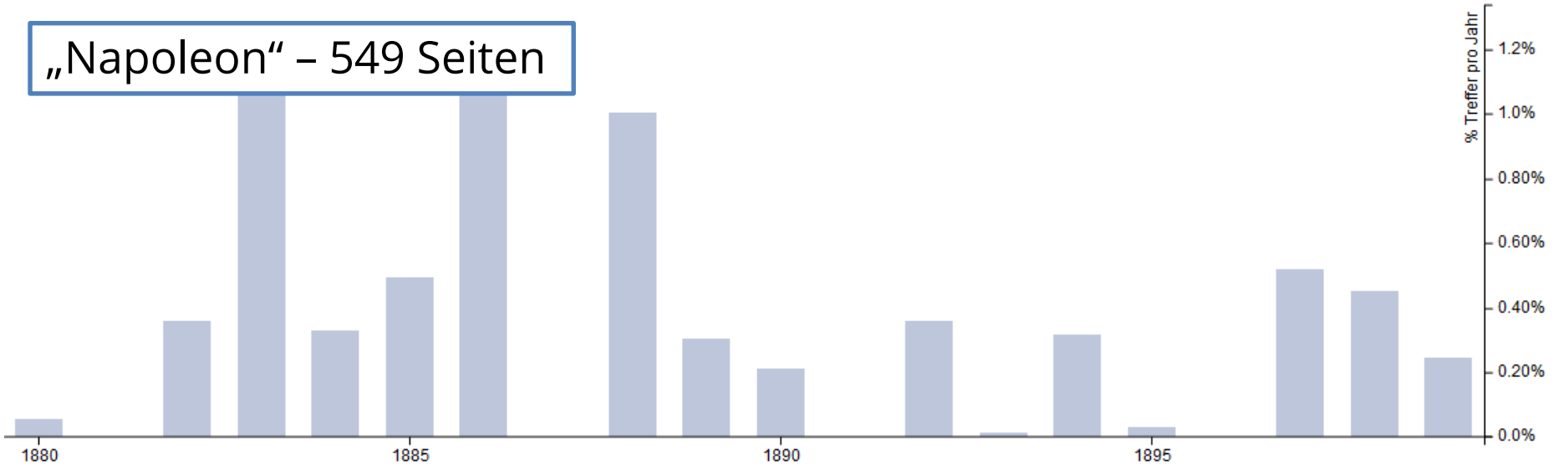


Participant 1

Term-Search on Textbooks for „Gymnasium“

„Antike“ – 76 Seiten

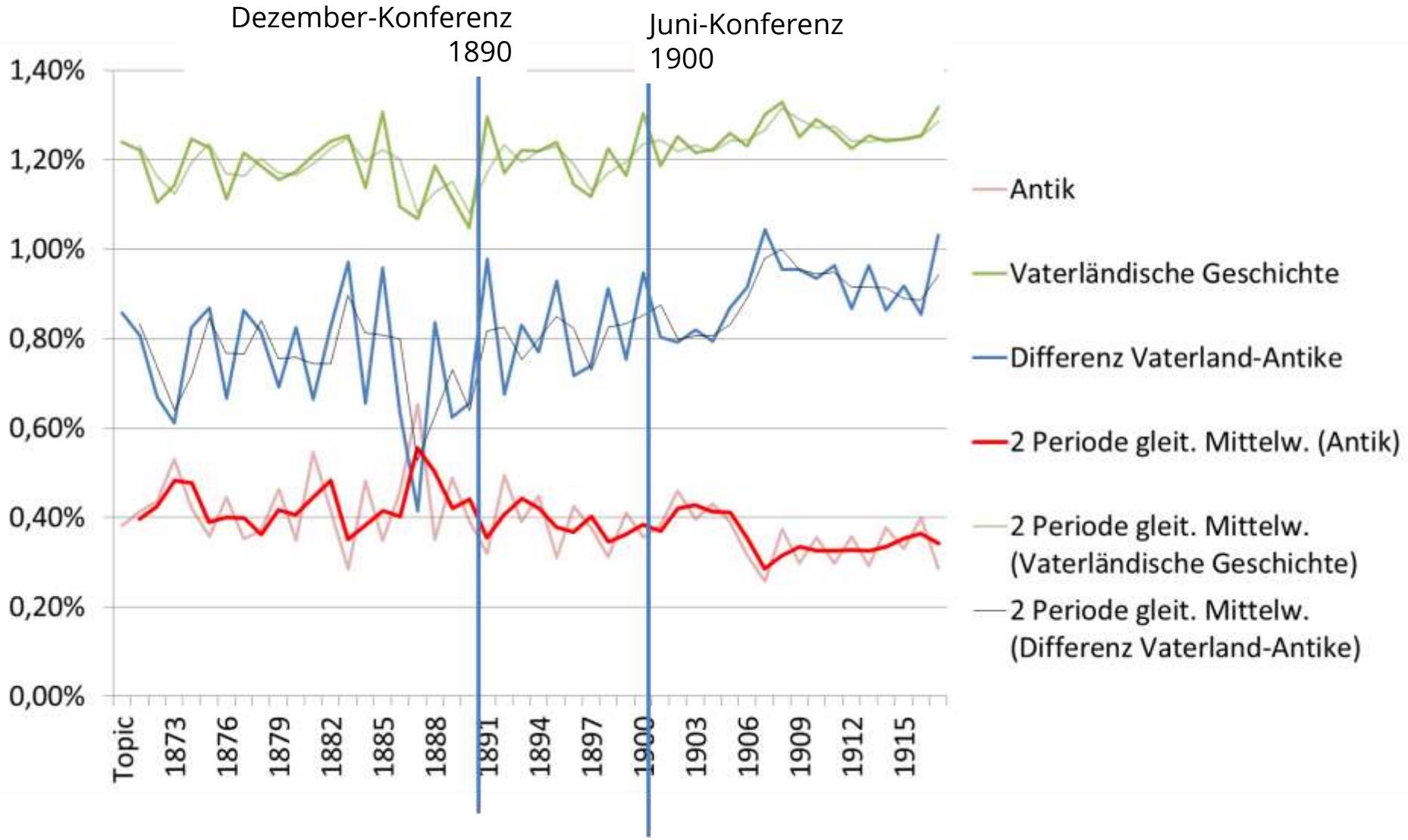
„Napoleon“ – 549 Seiten



Participant 2

Combination of Topics

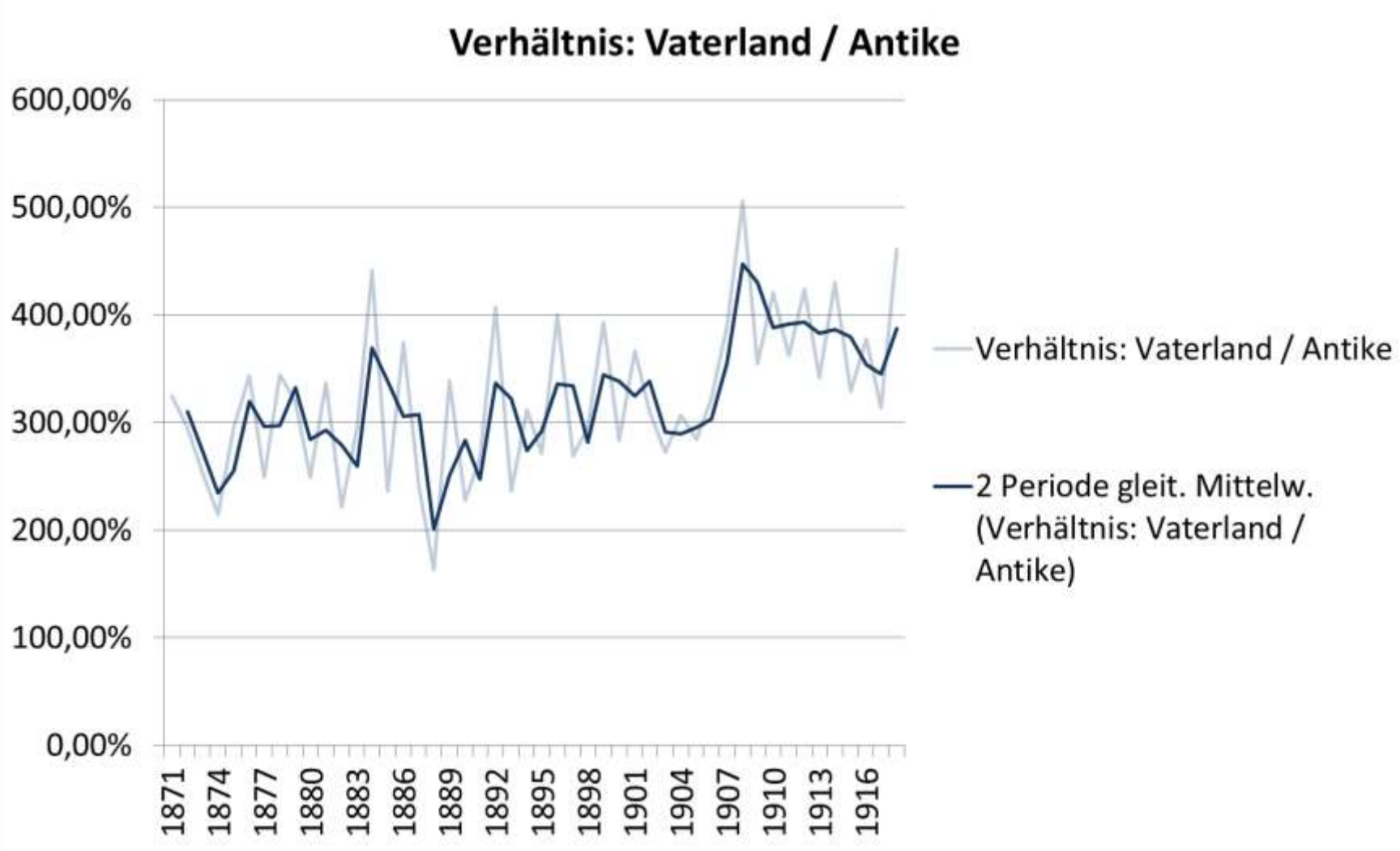
Untersuchung des Verhältnisses für größeren Zeitraum



Teilnehmer 2

Kombination von Topics

Untersuchung des Verhältnisses für größeren Zeitraum



Gesamtergebnis der Validierungsstudie

These	Tn	Beurteilung	Grundlage	Belast-barkeit	Transparenz
1: Kolonien - Erdkunde/ Geschichte	1	Teilweise bestätigt	Topics	4	5
	2	bestätigt	Topics & Terme	3,5	4
2: Kolonien - Volksschulen / Höhere Schulen	2	bestätigt	Topics & Terme	4	4
3: Vaterländische Geschichte / Antike	1	Nicht eindeutig		3	4
	2	bestätigt (in der Tendenz)	Topics	4	3,5
4: Kriegsflotte Erdkunde / Geschichte	2	widerlegt	Terme	3	2

Results of the Validation Study

Tools and models appear to be valid

Importance of Transparence and Validity

Demonstrate validity using different tools and methods

Level of analysis

Topics in many cases are too specific or to general

→ Implement Topic-Combination as a tool

Planned Steps of Evaluation

User Studies, based on Word-Intrusion, Topic Intrusion

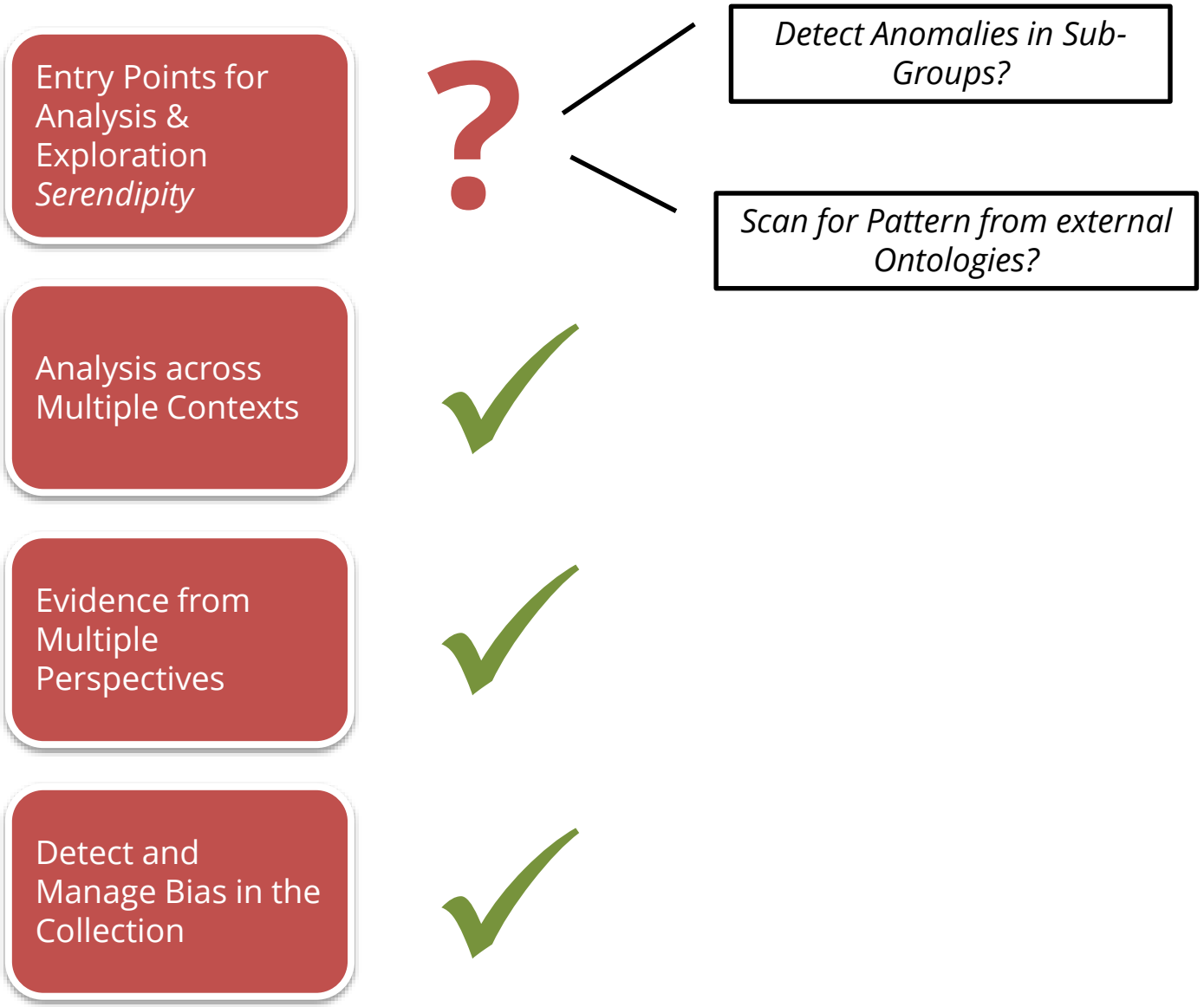
- Evaluate new approach to text modeling in comparison to prototype (Chang et al. 2009)

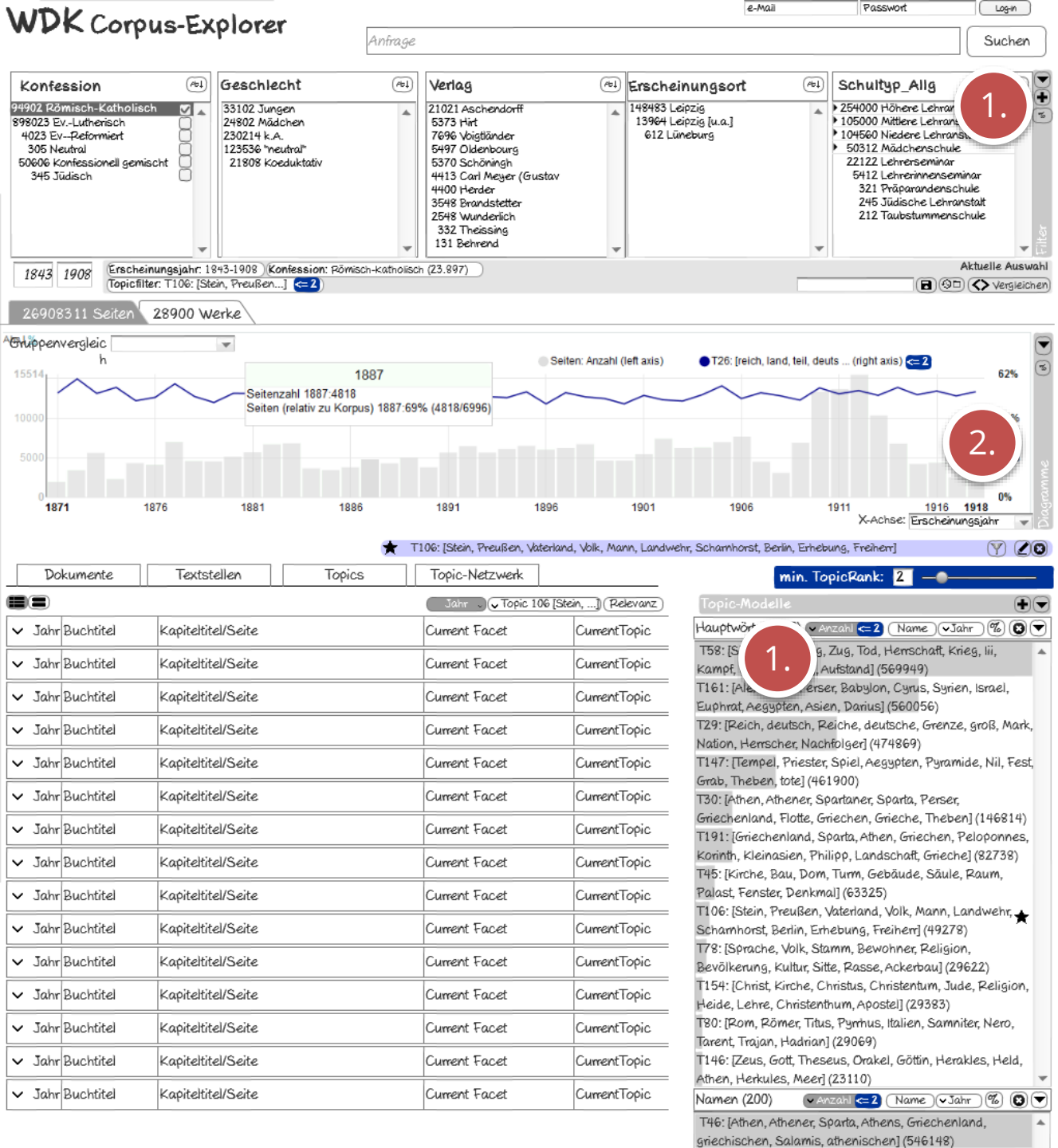
IR-Evaluation Studies

- TopWords from Topics (tf*idf) compared to ranking by Topic-Intensity derived from models

Additional interactive Evaluation Studies

- Focus on Usability
- Focus on subjective measures of Recall & Precision from Interactive Information Retrieval


Entry Points for Analysis & Exploration *Serendipity*
?
Detect Anomalies in Sub-Groups?
Scan for Pattern from external Ontologies?
Analysis across Multiple Contexts
Evidence from Multiple Perspectives
Detect and Manage Bias in the Collection



1. Facets & Dynamic Topic-Facets

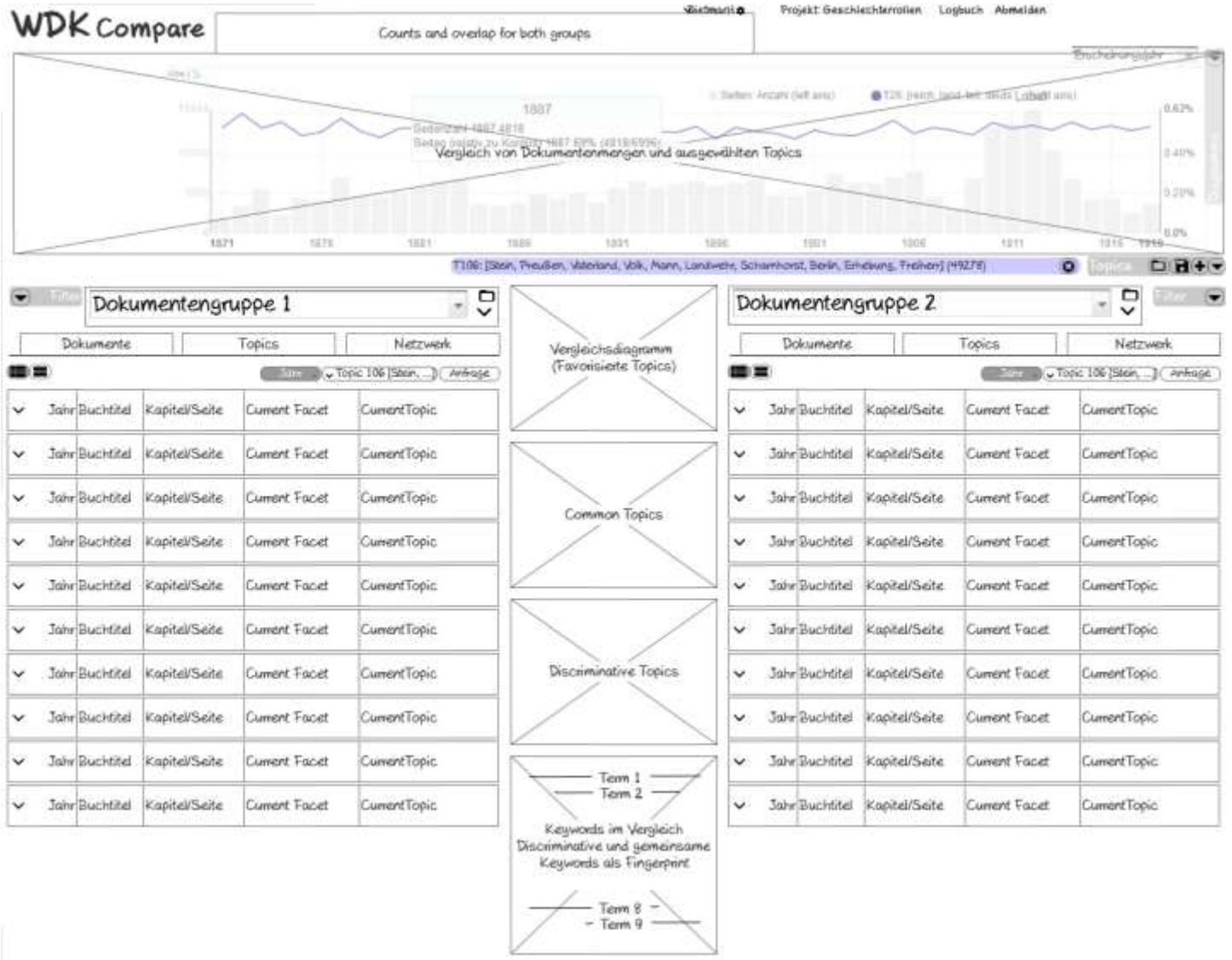
Relative to Corpus: Most relevant Dimensions in a given selection

2. Visualising & Comparing Results

Trends across dimensions

Compare Sub-Collections

With relative metrics and measures

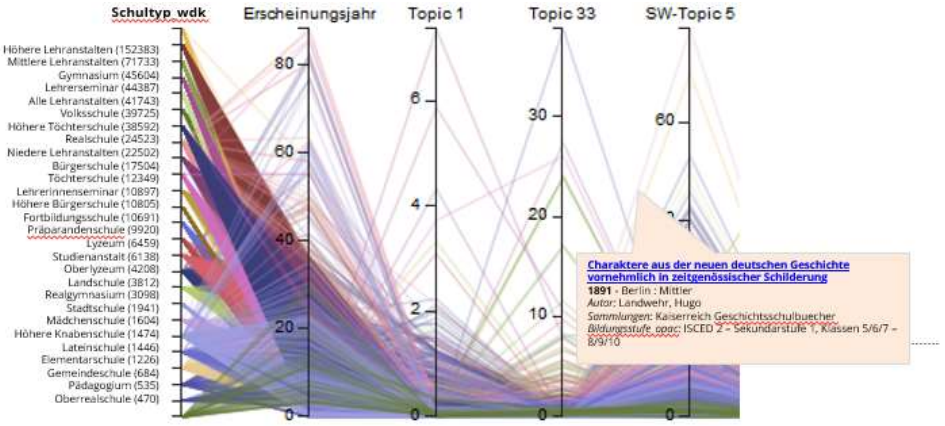


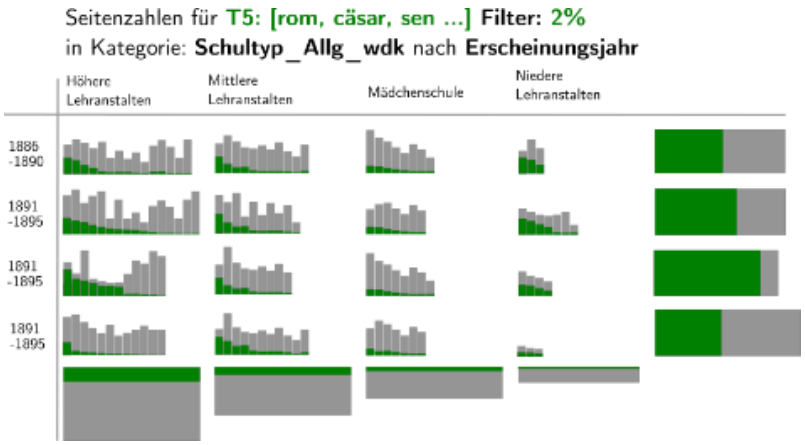
Visualizations grouped by work

Multivariat Analysis

Include context information

Control for biases in the corpus across two variables

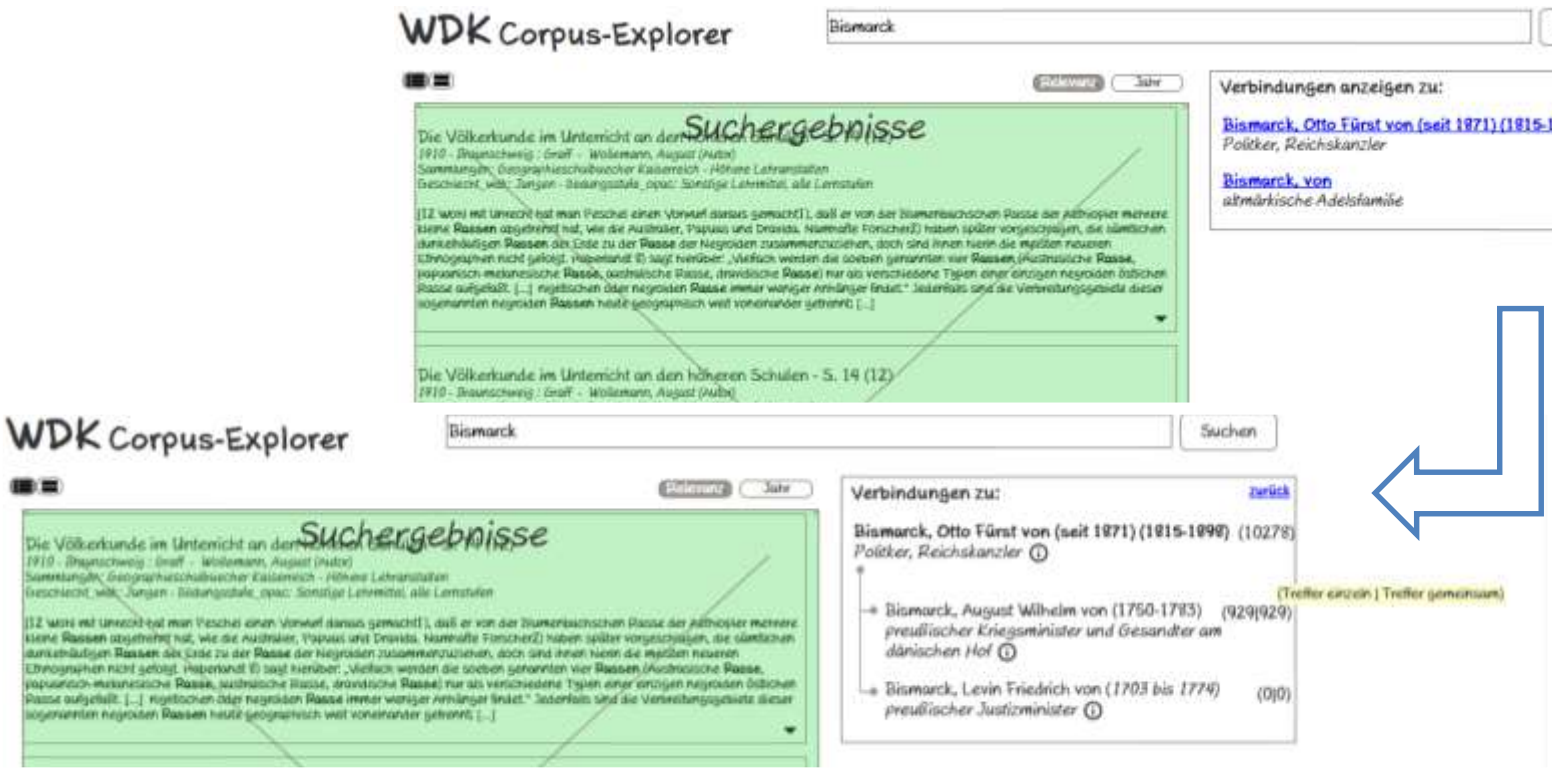




Incorporate External Knowledge Sources

Related Persons as Query-Suggestions

Analysis of narratives based on history



Open questions...

Document ranking for topic models?

Relevant document with multiple top topics do not appear on top → not visible

More specialized modeling techniques necessary?

e.g. Interactive topic modeling? Dynamic and/or structural topic models?

Interpretation and trust in models and visualizations more important than perplexity/novelty (Chuang et al. 2012)

Subjectivity and evaluative Language

Targeting topics, named entities, ...? Ideologies?

Statistical significance ...

... of differences in document sets?

... of correlations of topics in document sets?

Infrastructure needed for DH (Digital History) based on large corpora

Robust Correction of OCR

Robust Text-Normalization

cf. Efforts of BBAW

Robust Disambiguation of Entities (Events, Persons, Places, Dates) and Linking to external Knowledge Sources

Integration of manual Annotations at scale

Literature

Blei, D. M. ; Ng, A. Y. ; Jordan, M. I. (2003): Latent dirichlet allocation. In: *Journal of Machine Learning Research* Bd. 3, S. 993–1022

Chang, J. ; Boyd-Graber, J. L. ; Gerrish, S. ; Wang, C. ; Blei, D. M. (2009): Reading Tea Leaves: How Humans Interpret Topic Models. In: *NIPS*. vol. 22, pp. 288–296

Chuang, J. ; Ramage, D. ; Manning, C. ; Heer, J. (2012): Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *CHI '12*. New York, NY, USA: ACM, S. 443–452

Freund, L.; Toms, E. G; Waterhouse, J. (2005): Modeling the information behaviour of software engineers using a work-task framework. In: *Proceedings of the American Society for Information Science and Technology* Bd. 42, Nr. 1

Rhee, Hea Lim (2012): Modelling historians' information-seeking behaviour with an interdisciplinary and comparative approach. In: *Information Research* Bd. 17, Nr. 4 - http://InformationR.net/ir/17-4/paper544.html

Womser-Hacker, Christa (2013): Information Seeking Behaviour (ISB). In: Kuhlen, Rainer; Semar, Wolfgang; Strauch, Dietmar (Hrsg.) *Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis*. 6., völlig neu gefasste Ausgabe. De Gruyter Saur