Close and Distant Reading Visualizations for the Comparative Analysis of Digital Humanities Data

Der Fakultät für Mathematik und Informatik
der Universität Leipzig

eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt

von Dipl.-Inf. Stefan Jänicke

geboren am 17. März 1982 in Oschatz

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Gerik Scheuermann (Universität Leipzig)
2. Prof. Dr. Hans Hagen (Technische Universität Kaiserslautern)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 06.07.2016 mit dem Gesamtprädikat *summa cum laude*.

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbstständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

.....
(Ort, Datum)

.....
(Unterschrift)

Biography

Stefan Jänicke graduated in Computer Science at Leipzig University, Germany, in 2009. Over the last years, he has gained experience in developing information visualization techniques in close collaborations to humanities scholars in a number of digital humanities projects. From 2009 to 2013, he worked as a research assistant at the Göttingen State and University Library in the projects *europeana-connect* and *HeyneDigital*. From 2011 to 2012, he also worked for Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS in the project *Deutsche Digitale Bibliothek*. Since 2012, he works as a research assistant in the Image and Signal Processing Group at Leipzig University, where he was involved in the digital humanities projects *eAQUA*, *eTRACES* and *eXChange*. Besides contributions for visualization journals and conferences, he has had the opportunity to also present his research at the annual digital humanities conference in the past three years. His main research topic is information visualization, especially the development of visualization techniques to support digital humanities and text analysis research is of peculiar interest. The developed visualizations, he presents on his website http://www.vizcovery.org.

Acknowledgments

Many people contributed in order to fill the following pages. First, I thank my supervisor Gerik Scheuermann for his support in the years of preparing this dissertation and for giving me the opportunity to delve into the fascinating digital humanities field. I thank Christian Heine for teaching me how to write research papers in our joint work on my first three published articles. I further thank Ralf Stockmann, Christian Mahnke, Mustafa Dogan and Kristine Voigt with whom I worked on various digital humanities projects at the University of Göttingen. Concerning an instructive year working for the Fraunhofer IAIS, I thank Sebastian Bothe, Karl-Heinz Sylla and Vera Hernandez Ernst who have a great share in making GeoTemCo a successful visualization. I thank Marco Büchler for the freedom to discover and work on open visualization challenges in the digital humanities field during the eTRACES project. I am indebted to the (digital) humanities scholars David Joseph Wrisley, Annette Geßner, Greta Franzini, Judith Blumenstein, Michaela Rücker, Eva Wöckener Gade, Michael Cade-Stewart, Charlotte Schubert and Josef Focht for sharing their thoughts, for applying our visualizations, for suggesting improvements, and for their collaborative work on research articles. I thank Muhammad Faisal Cheema, Martin Reckziegel and Thomas Efer for inspiring discussions on information visualization and on solutions for digital humanities research problems. I also thank all of my former and current colleagues who are not mentioned by name for numerous conversations on computer science and other topics of interest. For their bureaucratic support, I like to thank Karin Wenzel, Petra Gamrath and Renate Henkeler.

The final *Thank you!* goes to my parents Reinhardt and Iris, my sister Silvana, my grandmother Anita, and especially, to my wife Christin and our daughter Anna.

Contents

The greatest value of a picture is when it forces us to notice what we never expected to see.

John Tukey

1 Motivation

Traditionally, humanities scholars carrying out research on a specific or on multiple literary work(s) are interested in the analysis of related texts or text passages. But the digital age has opened possibilities for scholars to enhance their traditional workflows. Enabled by digitization projects, humanities scholars can nowadays reach a large number of digitized texts through web portals such as Google Books[1] or Internet Archive.[2] Digital editions exist also for ancient texts; notable examples are PHI Latin Texts[3] and the Perseus Digital Library.[4]

This shift from reading a single book "on paper" to the possibility of browsing many digital texts is one of the origins and principal pillars of the digital humanities domain, which helps developing solutions to handle vast amounts of cultural heritage data – text being the main data type. In contrast to the traditional methods, the digital humanities allow to pose new research questions on cultural heritage datasets. Some of these questions can be answered with existent algorithms and tools provided by the computer science domain, but for other humanities questions scholars need to formulate new methods in collaboration with computer scientists.

Developed in the late 1980s [Hoc04], the digital humanities primarily focused on

[1] https://books.google.com/
[2] http://www.archive.org
[3] http://latin.packhum.org/
[4] Ed. Gregory R. Crane. Tufts University. http://www.perseus.tufts.edu

designing standards to represent cultural heritage data such as the Text Encoding Initiative (TEI)[5] for texts, and to aggregate, digitize and deliver data. In the last years, visualization techniques have gained more and more importance when it comes to analyzing data. For example, Saito introduced her 2010 digital humanities conference paper [SOI10] with: "In recent years, people have tended to be overwhelmed by a vast amount of information in various contexts. Therefore, arguments about 'Information Visualization' as a method to make information easy to comprehend are more than understandable." A major impulse for this trend was given by Franco Moretti. In 2005, he published the book "Graphs, Maps, Trees" [Mor05], in which he proposes so-called *distant reading* approaches for textual data that steer the traditional way of approaching literature towards a completely new direction. Instead of reading texts in the traditional way – so-called *close reading* –, he invites to count, to graph and to map them. In other words, to visualize them.

This dissertation presents novel close and distant reading visualization techniques for hitherto unsolved problems. Appropriate visualization techniques have been applied to support basic tasks, e.g., visualizing geospatial metadata to analyze the geographical distribution of cultural heritage data items or using tag clouds to illustrate textual statistics of a historical corpus. In contrast, this dissertation focuses on developing information visualization and visual analytics methods that support investigating research questions that require the comparative analysis of various digital humanities datasets. We first take a look at the state-of-the-art of existing close and distant reading visualizations that have been developed to support humanities scholars working with literary texts. We thereby provide a taxonomy of visualization methods applied to show various aspects of the underlying digital humanities data. We point out open challenges and we present our visualizations designed to support humanities scholars in comparatively analyzing historical datasets. In short, we present (1) *GeoTemCo* for the comparative visualization of geospatial-temporal data, (2) the two tag cloud designs *TagPies* and *TagSpheres* that comparatively visualize faceted textual summaries, (3) *TextReuseGrid* and *TextReuseBrowser* to explore re-used text passages among the texts of a corpus, (4) *TRAViz* for the visualization of textual variation between multiple text editions, and (5) the visual analytics system *MusikerProfiling* to detect similar musicians to a given musician of interest. Finally, we summarize our and the collaboration experiences of other visualization researchers to emphasize the

[5] eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. 2.8.0. 2015-04-06. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

ingredients required for a successful project in the digital humanities, and we take a look at future challenges in that research field.

Overview of Publications

This dissertation is based on the following publications by the author:

Chapters 2 & 8:
- On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges [JFCS15]
- Visual Text Analysis in Digital Humanities [JFCS16]

Chapter 3:
- Comparative Visualization Of Geospatial-Temporal Data [JHSS12]
- GeoTemCo: Comparative Visualization of Geospatial-Temporal Data with Clutter Removal Based on Dynamic Delaunay Triangulations [JHS13]
- Visualizing Uncertainty: How to Use the Fuzzy Data of 550 Medieval Texts? [JW13]
- Utilizing GeoTemCo for Visualizing Environmental Data [JS14]

Chapter 4:
- Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies [JBR+16]
- TagSpheres: Visualizing Hierarchical Relations in Tag Clouds [JS16]

Chapter 5:
- Visualizations for Text Re-use [JGBS14b]
- Designing Close and Distant Reading Visualizations for Text Re-use [JEBS15]

Chapter 6:
- 5 Design Rules for Visualizing Text Variant Graphs [JGBS14a]
- Improving the Layout for Text Variant Graphs [JBS14]
- TRAViz: A Visualization for Variant Graphs [JGF+15]
- A Distant Reading Visualization for Variant Graphs [JG15]

Chapter 7:
- Interactive Visual Profiling of Musicians [JFS16]

Remark

Although this dissertation is the work of only one author the pronoun “we” is used. One reason is that the presented works were carried out in collaboration with other researchers, another reason is the typical writing style most readers are familiar with.

Any job very well done that has been carried out by a person who is fully dedicated is always a source of inspiration.

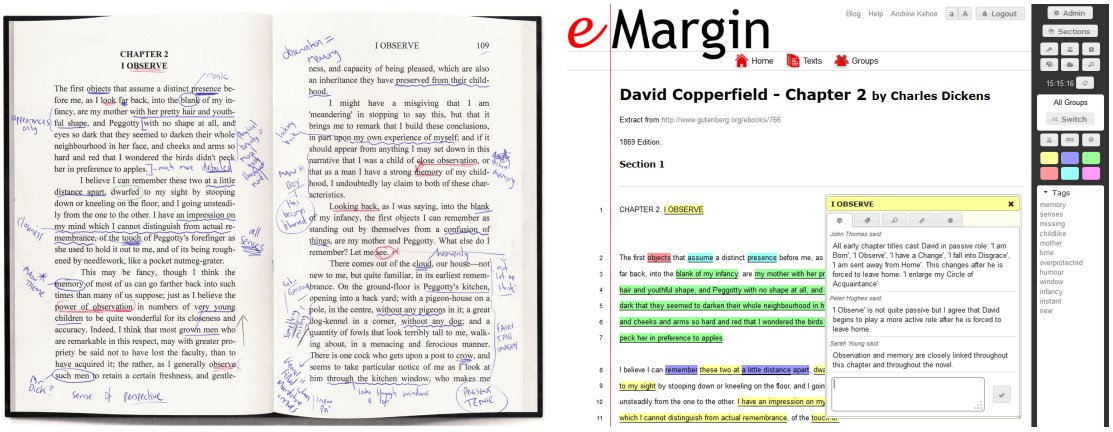
Carlos Ghosn

2 Close and Distant Reading Visualizations

While the close reading of a text is a traditional method that has its roots in antiquity when Aristotle close read the works of Plato [McC15], distant reading is a rather novel idea that was introduced by Franco Moretti at the beginning of the 21th century. In contrast to Moretti, Jockers uses the terms *micro-* and *macroanalysis* instead of close and distant reading [Joc13]. Inspired by micro- and macroeconomics, he focuses on quantitative literary text analysis using statistical methods. As the related works we observed solely provide visualization techniques, we decided to use the traditional, more common terms *close* and *distant reading*, but we also considered related methods using different terminologies. This chapter introduces close and distant reading techniques and draws a line from the digital humanities to information visualization by combining both techniques.

2.1 Close Reading

The close reading of a text became a fundamental method in literary criticism in the 20th century [Haw00]. Nancy Boyles [Boy13] defines it as follows: "Essentially, close reading means reading to uncover layers of meaning that lead to deep comprehension." In other words, close reading is the thorough interpretation of a text passage



(a) Traditional close reading. **(b)** Digital close reading with *eMargin*.

Figure 2.1: Examples of close reading of the second chapter of Charles Dickens' *David Copperfield* (Figures reproduced with permission from Kehoe et al. [KG13]).

by determining central themes and analyzing their development. In particular, close reading includes the analysis of [Jas01]:

- individuals, events, and ideas, their development and interaction,
- used words and phrases,
- text structure and style, and
- argument patterns.

The result of a traditional close reading approach is shown in Figure 2.1a. In this example, the scholar used various methods to annotate various features of the source text, e.g., the usage of different colors (blue, red, green) and underlining styles (straight or wavy lines, circles). Furthermore, numerous thoughts are written next to the corresponding sentences. Although most humanities scholars are trained in this traditional approach of close reading, today's large availability of digitized texts and of digital editions through web portals like Google Books or Project Gutenberg,[1] opens up new possibilities for close reading, and especially for sustainable and collaborative annotation. Figure 2.1b shows a straightforward approach of visualizing various scholars' annotations of a digital edition [KG13] within the web-based environment

[1] http://www.gutenberg.org/

eMargin.[2] There, colors are used to highlight different text features, and a pop-up window lists the comments of collaborating scholars. In Section 2.4 we outline different approaches to support close reading by visualizing supplementary human- or computer-generated information.

2.2 Distant Reading

While close reading retains the ability to read the source text without dissolving its structure, distant reading does the exact opposite. It aims to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s). Moretti [Mor13] describes distant reading as "a little pact with the devil: we know how to read texts, now let's learn how not to read them." In 2005, he introduces his idea of distant reading [Mor05] with three examples using:

- *graphs* to analyze genre change of historical novels,
- *maps* to illustrate geographical aspects of novels, and
- *trees* to classify different types of detective stories.

Although the proposed methods and the intention of distant reading are controversial in the humanities [GH11a, Mar12, CRS+14], many works in the digital humanities domain are based upon Moretti's idea. Figure 2.2 shows Posavec's Literary Organism [Pos07], a distant reading of Jack Kerouac's *On the Road* in the form of a tree. While a non-interactive infographic, Posavec's approach perfectly illustrates the idea behind Moretti's distant reading, as it turns away from the traditional close reading by providing an abstract view of a literary text. The branching structure represents the ordered hierarchy of content objects from chapters down to words, and themes are drawn with different colors. In Section 2.5, we present a list of different distant reading techniques developed for a wide range of research questions in the digital humanities.

2.3 Combining Close and Distant Reading

Many methods include close reading as well as interfaces that provide distant reading visualizations, which allow to interactively drill down to specific portions of the data.

[2] http://eMargin.bcu.ac.uk/

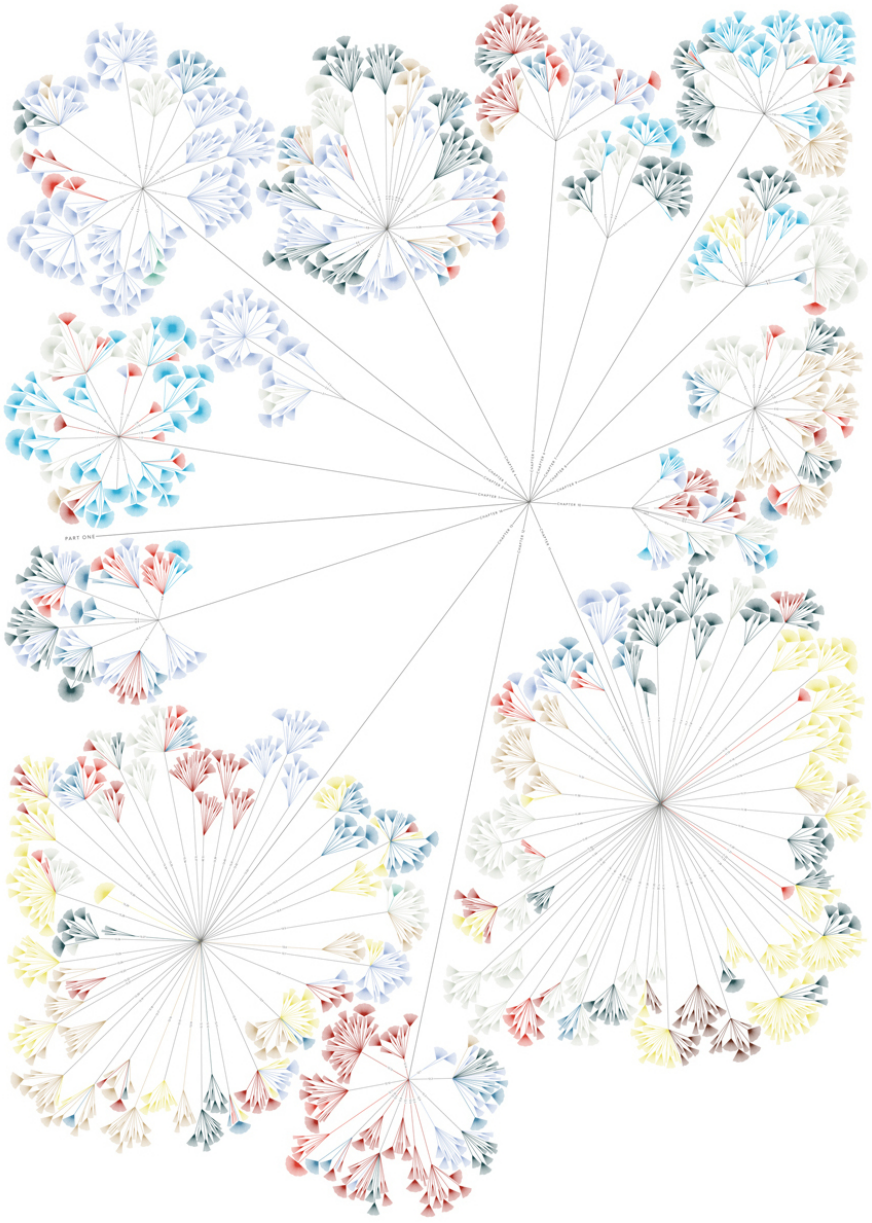


Figure 2.2: Distant reading example shows the structure of and the themes in Jack Kerouac's *On the Road* (Figure reproduced with permission from Posavec [Pos07]).

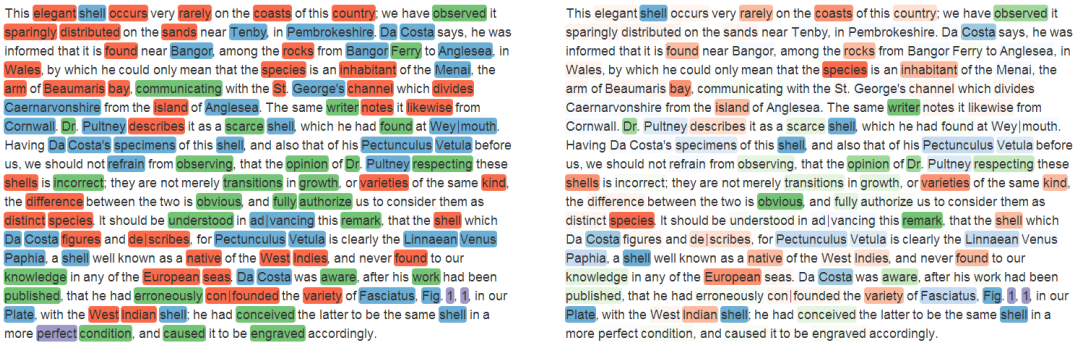
This suggests that direct access to the source texts is important for humanities scholars when working with visualizations. For example, Bradley [Bra12] asks whether it is "possible to develop a visualization technique that does not destroy the original text in the process." Similarly, Beals [Bea14] asks: "In an age where distant reading is possible, is close reading dead?" Coles et al. argue that distant reading visualizations cannot replace close reading, but they can direct the reader to sections that may deserve further investigation [CL13].

When distant reading views are interactively used to switch to close reading views, Ben Shneiderman's Information Seeking Mantra "Overview first, zoom and filter, details-on-demand" [Shn96] is accomplished. It follows that an important task for the development of visualizations is to provide an overview of the data that highlights potentially interesting patterns. A drill down on these patterns for further exploration is the bridge between distant and close reading. In Section 2.6, we take a look at three approaches to combine close and distant reading techniques.

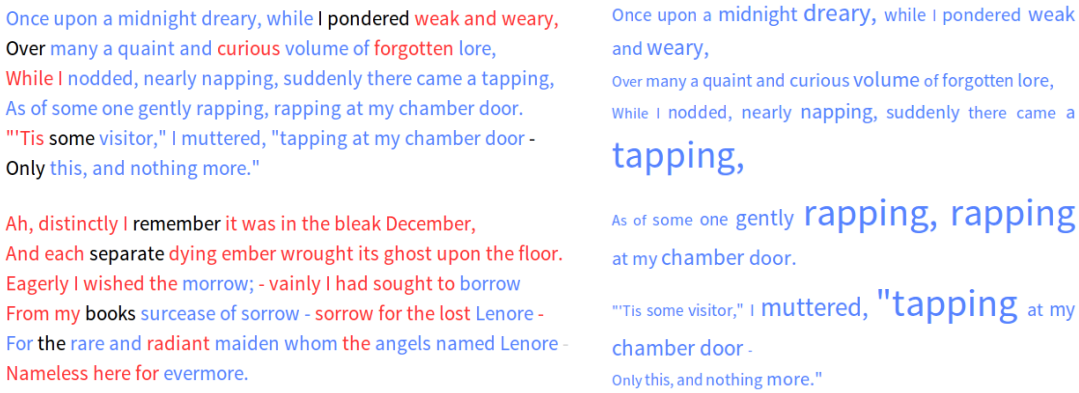
2.4 Close Reading Techniques

A visualization that allows to close read a text requires retaining the structure of the text in order to facilitate a smooth analysis. With additional information in the form of manual annotations or of automatically processed features of textual entities or relationships among them, a plain text can be transformed into a comprehensive knowledge source. Some visualizations provide only plain close reading views without additional information, others attend to the matter of enhancing the close reading capabilities of the humanities scholars. To visualize such additional information for a great variety of purposes, the researchers made use of the techniques listed below.

Color is the visual attribute most often used to display the features of textual entities and it is applied in different ways. In most cases, a colored background is used to express various types of information about a single word or an entire phrase (Figure 2.1b). The tool Serendip [AKV+14] varies the transparency of background colors to encode the importance of individual words (Figure 2.3a). Font color is also frequently used for this purpose [WMN+14], an example is given in Figure 2.3b (left). Colored circumcircles (Figure 2.5c) around words are used only once [MLCM16]. When displaying digital editions of literary texts, insertions are underlined. This might be the reason that this metaphor of underlining words is also rarely used to enhance close reading [CWG11]. Overall, coloring is a suitable method to express a



(a) Colored backgrounds and backgrounds with varying transparency (Figure provided by Alexander et al. and based on [AKV+14]).



(b) PRISM uses color to highlight the classification of words and font size to encode the number of annotations (Figures under CC BY 3.0 license based on [WMN+14]).

Figure 2.3: Color usage for close reading.

great variety of textual features. Among other purposes, coloring is used to highlight the automated or manual classification of words or phrases [KJW+14, WMN+14], to mark common words [JRS+09, Mur11] and aligned text segments [RPSF15, ZNMS15] in parallel texts, or to visualize various sound patterns in poems [CTA+13, Ben14].

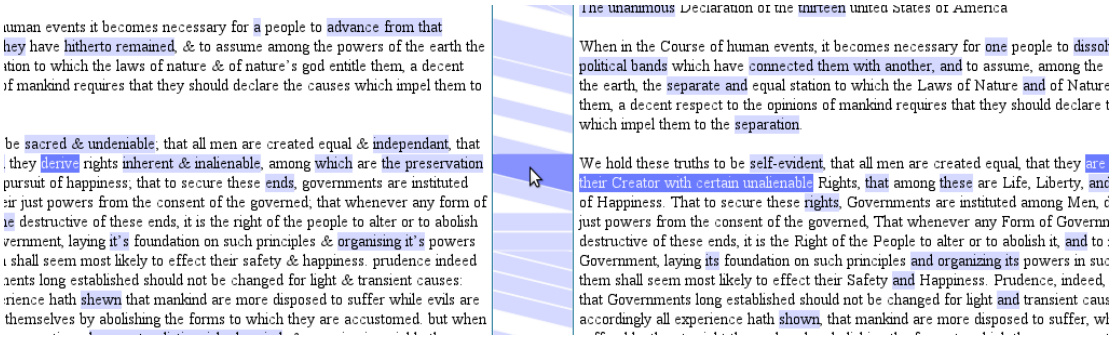
Font size is another method of visualizing features of textual entities. Adopted from tag cloud design [VWF09], this metaphor serves best to highlight the significance or weight of a textual entity in relation to the given text or corpus. Within the web-based tool PRISM [WMN+14], users collaboratively group the words of literary texts into different categories. The collected statistics are used to display the number of annotations of each word by variable font size (Figure 2.3b, right). In [CWG11],

varying font size is used to visualize the importance of text passages according to the user's preferences.

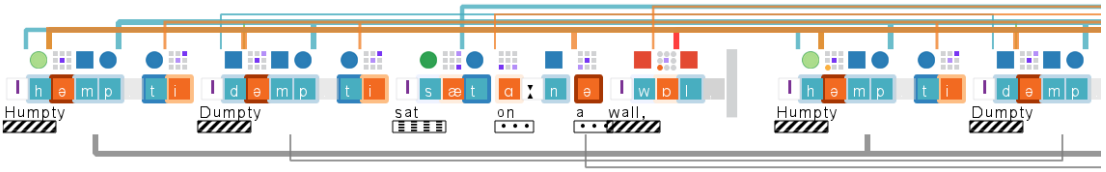
Glyphs attached to individual textual entities are convenient techniques to visualize abstract annotations that are hardly expressible with plain coloring or varying font size. Most examples we found enhance the close reading of poems. For the visualization of a poem's hermeneutic structure, Piez deploys glyphs in the form of rectangular and circular maps [Pie10, Pie13]. An example is given in Figure 2.4. In [ARLC+13], phonetic units are drawn atop each word using color to classify phonetic types (Figure 2.5b). Additionally, pictograms illustrate phonetic features. The Myopia Poetry Visualization tool uses rectangular blocks to visualize poetic feet and the spoken length of syllables [CGM+12]. Goffin explores the placement and design of so-called word-scale visualizations, which are small glyphs enriching the base text with additional information [GWFI14]. For example, the background color of words contained in digital copies speaks for OCR certainty. Furthermore, small interactive bar charts show variants of observed words.

Figure 2.4: Close reading example with glyphs illustrating the hermeneutic structure of a poem (Figure provided by Piez and based on [Pie10]).

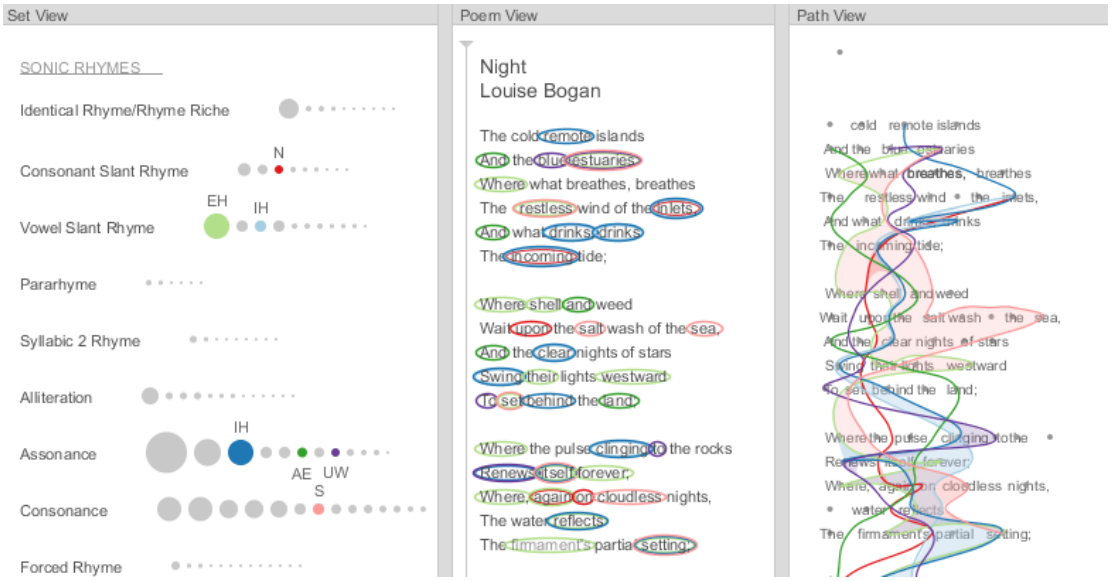
Connections aid to illustrate the structure among textual entities. One usage of connections in close reading is to highlight subsequent words in a variant graph to track variation among text editions [BGHE10]. Other approaches juxtapose the texts



(a) Juxta commons supports close reading of two parallel texts by connecting related text passages (Figure provided by Wheeles et al. and based on [WJ13b]).



(b) Poem Viewer uses glyphs to encode phonetic units, and connections show phonetic and semantic relationships (Figure reproduced with permission from Abdul-Rahman et al. [ARLC+13]).



(c) Circumcircles and connections highlight assonance, consonance, and slant rhyme sets (Figure produced with Poemage[3] based on [MLCM16]).

Figure 2.5: Using connections for close reading.

of various editions and visually link related text passages [WJ13b, HKTK14], as instantiated in Figure 2.5a. Connections can also be used to visualize sentence structure [KZ14] or phonetic and semantic relations in poems [ARLC+13], like shown in Figure 2.5b. A similar application is Poemage,[3] where paths are drawn between a poem's words sharing the same tones (see Path View in Figure 2.5c) to support the analysis of occurring sonic patterns [MLCM16].

2.5 Distant Reading Techniques

A visualization that displays summarized information of the given text corpus facilitates distant reading. The process of transforming such information into complex representations can be based upon a large variety of data dimensions, e.g., various types of metadata of textual entities, automatically processed or manually retrieved relationships between textual elements, or quantitative and qualitative statistics about unstructured textual contents. We grouped various visualizations that provide a rather abstract distant reading view of a given text corpus into six categories.

Heat maps or block matrices are often used to highlight textual patterns. Thereby, a heat map may reflect structural elements of a text [JRS+09, VCPK09, KJW+14] or the structure of an entire corpus [CDP+07, Mur11, BGHJ+14]. In such scenarios, the coloring of rectangular blocks helps to analyze the distribution of specific textual patterns [CWG11, MH13, AGZH15, JKH+15]. Another example is the usage of heat maps to show relationships among various texts in a corpus. The similarity for each tuple of texts within the corpus can be determined by counting similar text passages, and the result can be visualized as a heat map [GCL+13, FKT14], e.g., to highlight the similarity between Shakespearean plays [RRRG05]. Heat maps are also applied to visualize similarities or differences among text editions [PMMR15], or to highlight re-used passages between the texts of a corpus [RARC+15, ZNMS15]. For the analysis of potentially plagiarized texts, so called Difflines reveal structural differences between several suspicious text fragments and their alleged originals in a Focus+Context view [RPSF15], an example is shown in Figure 2.6. Alexander et al. propose two matrix representations [AKV+14]. The RankViewer illustrates the ranking of words belonging to topics and the CorpusViewer shows relations to certain topics for each document of a corpus. Heat maps are also used in [MSR+15] to display "high-level summaries" of topic modeling results. Fingerprinting techniques,

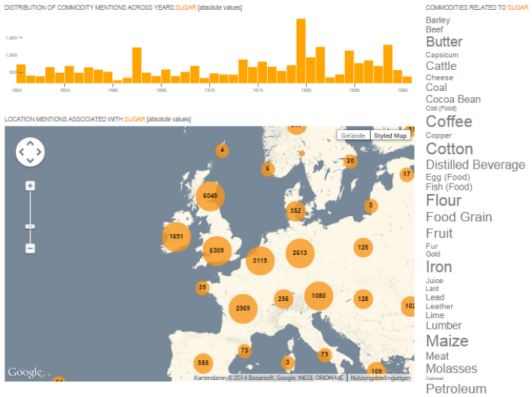
[3] http://www.sci.utah.edu/ ~ nmccurdy/Poemage/

Figure 2.6: Heat map highlighting potential plagiarized text passages in a PhD thesis (Figure reproduced with permission from Riemann et al. [RPSF15]).

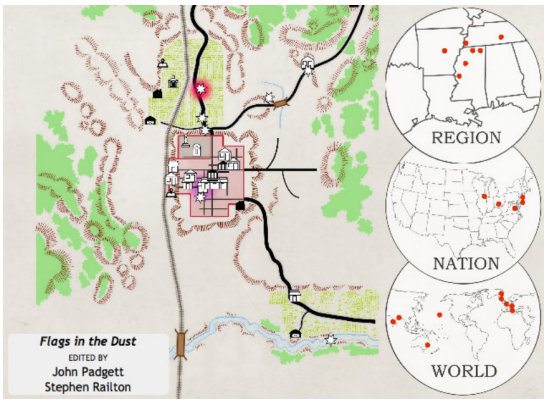
as introduced in [KO07], visualize characteristic textual features of literary works, or can be further used to reveal interpersonal relationships between characters in prose literature [OKK13]. Finally, heat maps are used to visualize the similarity [CTA+13] or the flow [FS11, Ben14] of sound in poems.

Tag clouds are intuitive visualizations to encode the number of word occurrences within a single section, a whole document or an entire text corpus by using variable font size [VCPK09, FKT14, GTAHS15]. By applying significance measures, the visualization can be limited to displaying only characteristic tags [ESK14, KJW+14, HAC+15] (examples can be seen in Figure 2.7a and Figure 2.11a). Tag clouds can also summarize the major tags for certain time periods [CLT+11, Bea12, CLWW14] or topics inherent in a text corpus [BJ14, JOL+15, MSR+15]. The usage of tag clouds to explore the classification of speculative fiction anthologies [HFM16] is shown in Figure 2.8b. Beaven also uses tag clouds to illustrate collocational relationships of a single word [Bea08] and to compare the collocates between two words [Bea11]. In some of the mentioned works, tag coloring is used to express additional information such as the temporal evolution of a word's significance or the classification of tags.

Maps are widely used to display the geospatial information contained in a text. Many approaches project the placenames mentioned in a text or in an entire corpus onto maps. With the help of contemporary (e.g., GeoNames[4]) and historical gazetteers (e.g., Pleiades[5]), the extracted placenames can be enriched with geographical coordinates, and their visualization on a map supports the analysis of the (fictional) geographic space described in the source text(s). Some approaches use thematic [DFM+08, ÓML14] or density maps [GH11b, GDMF+14, BB15b] for this purpose, but the usage of glyphs in the form of circles is more frequent [Tra09, DWS+12, HAC+15, Wil15] as it simplifies the interaction with individual plotted places (e.g., see Figure 2.7a). Two works that focus on mapping the geographical knowledge of ancient Greek authors draw connections between glyphs to illustrate travel routes [EJ14] or to highlight the strength of the relationship between placenames, which is reflected by the number of co-occurrences [BPBI10]. In contrast to the previous works, the geospatial metadata associated with individual corpus texts (text creation place) can be used for mapping [MBL+06]. The visualization of Faulkner's fictional *Yoknapatawpha County* includes various means of geographic mapping [DNCM14]: on the one hand, the imagined geography and, on the other, real placenames displayed on the geographic levels *region*, *nation* and *world* (Figure 2.7b).



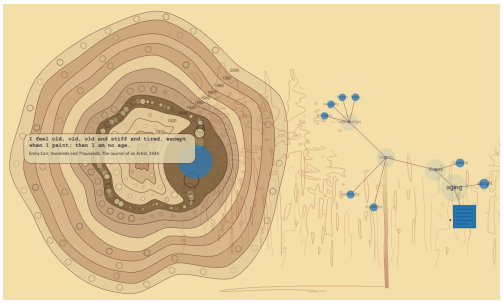
(a) Linked views for the exploration of commodity trading (Figure reproduced with permission from Hinrichs et al. [HAC+15]).



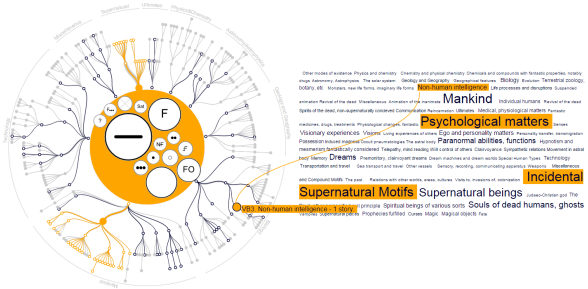
(b) Fictional map of *Yoknapatawpha County* and related places (Figure reproduced with permission from Dye et al. [DNCM14]).

Figure 2.7: Maps supporting distant reading.

[4] http://www.geonames.org/
[5] http://pleiades.stoa.org/



(a) A combination of an abstract timeline view and a tree in EMDialog (Figure provided by Hinrichs based on [HSC08]).

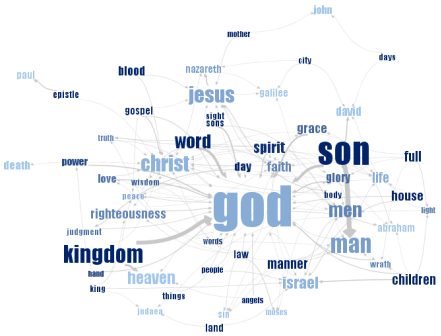


(b) Exploring the tree structure and representative tags of a novel classification scheme (Figure provided by Hinrichs based on [HFM16]).

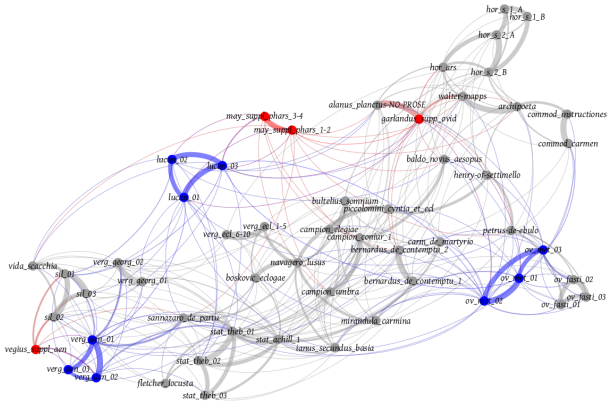
Figure 2.8: Linked views for text collection analysis.

Timelines are appropriate techniques to visualize historical text corpora carrying various types of temporal information. One approach is the use of the text's metadata [HFM16]. An example is the exploration of events in news articles [ESK14] (see Figure 2.11a). Sometimes, the temporal information about events reported in a text needs to be extracted in order to visualize (fictional) calendars [DNCM14, GDMF+14, ÓML14, HAC+15], e.g., like shown in Figure 2.7a. For the exploration of placenames in Herodotus' *Histories*, a timeline is used to show where certain placenames occur in the text [BPBI10]. A somewhat abstract timeline view is shown in [HSC08]. Here, a so-called tree cut section, whereby each ring represents a decade, visualizes statements from and about Emily Carr's life and work (Figure 2.8a). Streamgraphs are popular techniques that produce aesthetic visualizations and allow to track the evolution of themes over time [BW08], thus generating enhanced versions of the timeline metaphor. Such visualizations are often based on newspaper sources [CLT+11, KBK11, DWS+12, CLWW14] or political text archives [Kau15, Poi15] to support the analysis of contemporary topic changes. Using a research paper pool [ARR+12], the changing importance of research topics can be explored. Streamgraphs may also be used to visualize storylines or to illustrate plot evolution and changing locations in literary texts [LWW+13]. Based upon Hollywood screenplays, the tool ScripThreads visualizes action lines of movie characters [HPR14].

Graphs are the most often applied method to visualize certain structural features of a text corpus. A common usage is the visualization of relationships between the texts (represented as nodes) of a corpus in the form of a tree [HFM16] as shown



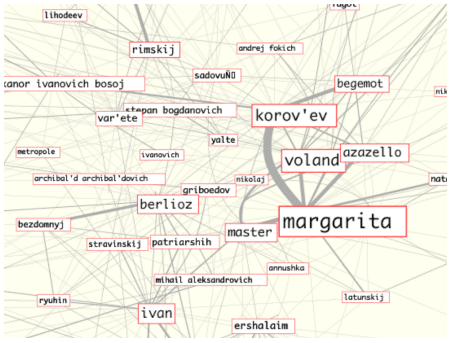
(a) Phrase net for the pattern '[word] of [word]' in the New Testament (Figure based on [vHWV09] and reproduced with permission from Smith [Ope09]).



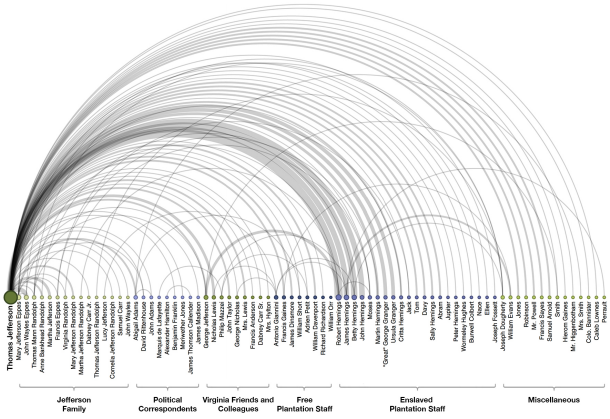
(b) Network of 55 poetic texts. Imitated texts are marked in blue, sequels are marked in red (Figure reproduced with permission from Eder [Ede14]).

Figure 2.9: Graphs supporting distant reading.

in Figure 2.8b, or a network [WH11]. Proximity can be used to express the similarity of these texts [EX10, MRMK15], e.g., based upon stylistic closeness [Joc12, CEJ+14]. For instance, Eder visualizes a network of poetic texts based on stylistic features [Ede14], thereby highlighting connected editions and sequels (Figure 2.9b). Graphs can also be used to illustrate topics (as nodes) and proximities of topic models applied to text corpora [Kau15]. In other works, nodes represent the words of a corpus and links are drawn to reflect semantic relationships [AGL+07, RFH14, OGH15] or co-occurrences [VCPK09, KKL+11, WJ13a, BB15b]. In [HAHB15], a graph visualizes cross references in historic encyclopaedia by linking related entries. Further applications are the visualization of scene changes and character movements in Shakespearen plays [RRRG05], as well as the display of conceptual [Arm14], contextual [HSC08] or multilingual [GZ12] information. Phrase nets [vHWV09] connect textual entities that appear in the form of a user-specified relation (syntactic or lexical), e.g. phrases like '[word] of [word]' within the Bible's New Testament, as shown in Figure 2.9a. All aforementioned works apply force-directed algorithms for the placement of nodes. Radial graphs can be used to unveil the relationships of words within poems [MFM13], or, again, to highlight the similarity among texts and, in this case, as nodes radially grouped by authors [Wol13]. The Word Tree [WV08], also used in [MH13], visualizes sentences sharing the same beginning in the form of a tree. Finally, we found a method that visualizes plain event trigraphs extracted through



(a) Excerpt from the social network in Mikhail Bulgakov's *Master and Margarita* (Figure provided by Maciej Ceglowski and reproduced with permission from Coburn [Cob05]).



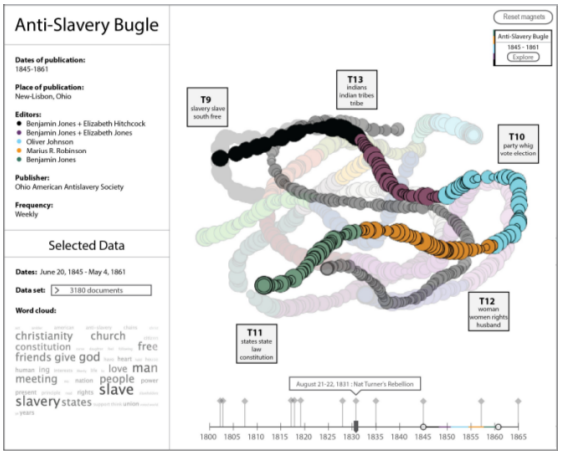
(b) Thomas Jefferson's social relationships (Figure reproduced with permission from Klein [Kle12]).

Figure 2.10: Social networks supporting distant reading.

phrase mining algorithms and thus providing metaphors to display uncertain information [MLSU13]. Social networks are graphs visualizing the relationships between people. Such representations are widely applied in the digital humanities to illustrate the relationships between characters in literary texts [CSV08, Tót13, BB15a, TFK15]. In these graphs, the size of a node can be used to encode the frequency of a character name in the text [BHW11, Pet14], the thickness of an edge [Cob05] (see Figure 2.10a) or the proximity of the nodes [Poi15] can serve to reflect the strength of a relationship, and edge style can be used to classify the type of relationship [KOTM13]. As per the aforementioned works, Kochtchi uses a force-based graph layout to visualize social networks automatically extracted from newspaper articles [KLB14]. In contrast, radial layouts and parallel coordinates are used in [Boo13]. For the visualization of Thomas Jefferson's social relationships (Figure 2.10b), the nodes placed on a horizontal axis are connected with arcs [Kle12]. Riche proposed a layout for Euler diagrams, which can also be utilized to visualize relationships between characters extracted from Shakespearean texts [RD10]. Finally, GeneaQuilts smartly visualizes large genealogies extracted from literary texts such as the Bible [BDF+10].

Miscellaneous methods also produce beneficial results for certain research questions. For the analysis of word statistics, tree maps can be used to illustrate the occurrences of adjectives in fairy tales [WJ13a]. For the exploratory thematic analysis of historical newspaper archives [ESK14], an application of the dust-and-magnet

metaphor [YMSJ05] yielded useful results (Figure 2.11a). Another topical analysis technique uses a landscape metaphor to visualize the topology-based clustering of articles taken from the New York Times Corpus [OST+10] (Figure 2.11b). Many techniques also attend to the matter of supporting the analysis of rather small text collections or individual works. Sankey diagrams can be used to compare the categories of words contained in two books [HCC14], and to highlight plagiarized text passages when juxtaposing a PhD thesis to potential sources [RPSF15]. In [GCL+13], a parallel coordinates and a dot plot view, which is used for filtering purposes, visualize the similarity of parallel text sections. For the analysis of repetitions in Gertrude Stein's *The Making of the Americans* [CDP+07], parallel coordinates visualize the frequency of phrases across sections, and TextArc [Pal02] is used to explore the repetition of individual words. Some miscellaneous methods were also developed to provide rather abstract text views. For the visual representation of phonetic patterns, a so called Path View is introduced in [MLCM16]. The tool PlotVis allows users to model and interact with XML-encoded literary narratives in 3D [PBD14]. A further complex tool named "Simulated Environment for Theatre (SET)" supports the story flow simulation of theatrical plays [RSDCD+13]. It consists of various 2D interfaces illustrating the "line of action" and a 3D interface populated by character avatars.



(a) Combination of a dust-and-magnet visualization, a timeline and a tag cloud for browsing historical newspapers (Figure reproduced with permission from Eisenstein et al. [ESK14]).

(b) Topological landscape visualizes thematic clusters in the New York Times corpus (Figure provided by Oesterling based on [OST+10]).

Figure 2.11: Miscellaneous distant reading methods.

Many digital humanities visualizations support either the close or the distant reading of texts. Still, an important feature for literary scholars when working with distant reading visualizations is direct access to source texts or, in other words, close reading. Some works also support close *and* distant reading, and combine both techniques most often in the form of coordinated views [BWKK00]. According to the given research task, there are three different methodologies for this combination.

Bottom-up methods focus primarily on close reading, and distant reading visualizations are generated in dependency on the scholar's input during close reading. In [GCL+13], the user selects a desired text passage in Shakespeare's *Othello*, which is shown in various German translations. Distant reading visualizations are processed (parallel coordinates view, dot plot view, heat map) based on that selection. In [Mur11], the literary scholar selects a certain phrase during the close reading process. Next, that phrase is searched within the text corpus and the phrase's distribution is shown in the form of a heat map. A similar approach is applied when annotating literary texts [AGZH15]. Places related to Edinburgh are marked, and a linked heat map that displays the distribution of all annotations is accordingly updated. In [OGH15], the user explores automatically tagged named entities of scientific papers in close reading mode. After editing, a graph reflecting contained entities and relationships among them is generated. Another bottom-up approach supports the semi-automatic alignment of early new high German text variants [MRMK15]. A graph displaying the similarities between text editions is updated as annotations are collected in close reading sessions.

Top-down & bottom-up approaches taken within one visualization entity allow for switching between close and distant reading while taking into account manipulations of the preceding view. In [JRS+09] and [PMMR15], the user can switch between heat map (distant reading) and text view (close reading). A side-by-side navigation between source text (close reading) and distant reading graphs showing the relationships among textual entities are illustrated in [WV08] and [RFH14]. Here, textual entities can be selected in both the graph and the text, triggering mutual updates. A typical use case for the combination of top-down and bottom-up behaviors are visual analytics methods. The VarifocalReader [KJW+14] hierarchically visualizes a document with the help of distant views (heat map, tag clouds) and close reading techniques (use of color, digital copy), thus supporting hierarchical navigation. In close reading mode,

automatically acquired classifications of textual entities can be manually modified, which subsequently affects distant views. The same applies to social networks automatically extracted from newspaper articles [KLB14]. The user browses the graph, opens close reading views associated with individual nodes and annotates the source text, which, again, affects the distant view and is used for classifier training. WordSeer [MH13] allows for a multifaceted perusal of a text corpus. For selected textual entities, several close and distant reading views can be used to browse the corresponding source texts. Within the close reading views, the user can group words into classes, which can then be used as a starting point for text corpus analysis. For the analysis of sound in poems, the tool Poemage dynamically links the provided close reading visualization (poem view) to the distant reading visualization (path view) to support the exploration of occurring phonetic patterns [MLCM16].

Top-down strategies are mostly applied when combining close and distant reading visualizations. Such methods implement the Information Seeking Mantra in its original meaning. Initially, a distant view on the textual data is shown, the user can often manipulate the visualization by means of filtering and zooming, and finally retrieve details-on-demand by clicking on a potentially interesting data item. In some cases, the texts are simply shown at the end of the information seeking pipeline, e.g., in [HSC08, DWS+12, RSDCD+13, Wil15, HFM16]. Observed words or text patterns are often highlighted in the close reading view by way of coloring (e.g., [VCPK09, Wol13, AKV+14, HPR14, HAC+15]). Various colors can thereby illustrate word categories [CDP+07], e.g., toponym types in the Herodotus Timemap [BPBI10] or topological cluster information [OST+10]. In some systems, close reading is related to the preceding distant reading very closely. In [BGHJ+14], the connection between close and distant reading is achieved by zooming. The distant view, a heat map, highlights certain patterns, and zooming allows the close reading of individual passages. Similarly, the navigation between distant plagiarism overviews and the close reading of plagiarized passages is organized in [RPSF15] and [ZNMS15]. The CorpusSeparator presented in [CWG11] is a distant view used to generate a weighted tag list (dependent on corpus statistics). Based upon these weights, the close reading view of a text (illustrated with Shakespeare's *A Midsummer Night's Dream*) is manipulated by coloring and sizing lines.

2.7 Open Challenges

As shown in this chapter, the application of approved visualization techniques was beneficial to support the investigation of various research tasks. Only few presented works propose methods that facilitate a comparative analysis of digital humanities data, most often to comparatively visualize automatically determined topics or clusters [OST+10, AKV+14, ESK14]. We worked together with humanities scholars on research questions that also require comparative views onto certain data facets, but appropriate comparative techniques did not exist and a straightforward usage of basic visualizations was not possible. Below, we list the research problems addressed in this dissertation – outlined in the next five chapters – for which we designed visualizations that take advantage of the aforementioned close and distant reading techniques.

Comparative Visualization of Geospatial-temporal Data. Visualizing geospatial as well as temporal metadata are popular methods to provide an intuitive visual access to large cultural heritage datasets. The Trading Consequences project, which investigates commodity trading in the past centuries, provides linked views for exploration purposes [HAC+15]. The geospatial information is shown on a map with multiple shapes representing one or multiple data items, and a time chart illustrates the temporal change inherent in the underlying dataset (see Figure 2.7a). Similarly to other projects, the geospatial and the temporal information are both aggregated, so that research questions regarding the comparison of data aspects – in this case the geospatial-temporal comparison of different commodities – cannot be posed. In order to support investigating this type of research question, Chapter 3 presents a distant reading visualization, consistent of interactive linked views (*map* and *timeline*).

Using Tag Clouds to Analyze Faceted Textual Summaries. Tag clouds are distant reading techniques often used in the digital humanities to visualize textual summaries. Figure 2.11a shows a system that embeds a tag cloud displaying frequent tags in a newspaper article collection. Although a clustering into topics was determined, the tag cloud still shows a summary for the entire dataset. Within one of our digital humanities projects, scholars also had multifaceted textual summaries at hand, and they wanted to analyze occurring similarities and differences. In Chapter 4, we present two designs that attend to the matter of comparatively visualizing these multiple textual summaries in a *tag cloud*. As close reading is very important for the collaborating

humanities scholars in these scenarios, we implemented a top-down strategy to enable the access to the underlying text passages.

Visualization of Text Re-use. The analysis of re-used patterns among different texts is a very interesting task in digital humanities [RARC+15, ZNMS15]. Usually, the visualizations are tailored for specific usage scenarios (e.g., [RPSF15]), but solutions for the visual analysis of Text Re-use occurrences in an entire, arbitrary text collection are not present. Chapter 5 introduces a *heat map* visualization that juxtaposes all texts of a corpus in order to provide hints about the amount of re-used text passages and the Text Re-use pattern types between each text tuple. In addition, a close reading interface facilitates the inspection of these Text Re-use patterns by drawing *connections* between related passages and by using *color* to illustrate similarity.

Visualization of Textual Variation. Many system exist that juxtapose (only) two editions of a text, and highlight similar text patterns with connections. But for the close analysis of similarities and differences among these patterns, which can be modeled in the form of a so called Variant Graph, appropriate visualizations does not exist. The Word Tree [WV08], which visualizes only sentences sharing the same beginning, cannot be applied to Variant Graphs. Chapter 6 describes a layout and a design for Variant Graphs in order to support the close analysis of textual variation using *font size* to express frequency and *connections* between subsequent terms. An additional *heat map* visualization allows for posing distant reading research questions on various editions of entire, hierarchically structured texts, and it uses a top-down strategy to drill-down to individual Variant Graphs.

Interactive Visual Profiling of Musicians. Computationally analyzing similarities among texts is a typical research task in digital humanities [RARC+15, ZNMS15]. Discovering similarities in historical groups of persons is a rather untypical task, which is usually done in a traditional fashion by referring to print media. In Chapter 7, we computationally analyze similarities among musicians. Therefore, we present a visual analytics system that implements a profiling for musicians similar to a musician of interest for the first time digitally. Thereby, the provided visualizations – a *map*, a social network *graph* and a so called *ColumnExplorer* – comparatively show the characteristics of different musicians, which were extracted and collected from several text sources including print media.

Space and time are the framework within which the mind is constrained to construct its experience of reality.

Immanuel Kant

3 Comparative Visualization of Geospatial-Temporal Data

Although the amount and types of data available through public Web resources is seemingly endless, finding information is still largely performed by text queries. Popular search engines rank the typically huge amounts of query results based on relevance and popularity. When too many irrelevant items remain, the user is required to restate the query by adding or replacing search terms. Unfortunately, repeated refinement can lead to frustration. An alternative is to spatialize the data and allow the user to refine queries using mouse-based navigation. While the data amount is increasing, data also became more structured. Many platforms provide rich data sources annotated with geospatial and temporal metadata. This metadata can be used to provide a contextual overview of the data in many forms: topical, geospatial and temporal being some of the most popular. Users already familiar with searching in geographic environments like Google maps can find results presented directly on a map rather than in a list, emphasizing the geospatial aspect. A tool providing more contextual overview and filter capabilities allows for synergetic effects.

This section presents the design of the web-based application *GeoTemCo* that enables the synergetic exploration of datasets (e.g., search results for topical queries) in a geospatial and temporal context. It employs a map view for the geospatial context,

a time view for the temporal context, and a detail view to inspect data items individually. *GeoTemCo* is designed to visualize data items of multiple datasets, which allows for comparing spatial distributions and temporal trends. All views are linked and provide filtering mechanisms, so that the visualized data can be analyzed dynamically according to the desires of the user. Our design is based on some related published systems that each allow only a subset of our requirements: comparing multiple datasets, refining temporal context, and scaling to a large number of items, both computationally and visually. We achieve the latter by zoom-dependent aggregation of data item locations into non-overlapping circles, thus avoiding visual clutter in the map view.

Although *GeoTemCo* was not particularly designed for humanities data, it became a valuable tool in a number of digital humanities projects, e.g., Europeana,[1] DARIAH-DE,[2] or Heyne Digital.[3] The reason might be that the investigation of many research questions in that field requires the comparative analysis of data facets, especially the comparison of geospatial data is a common task. For illustration purposes, we will outline representative digital humanities usage scenarios. In addition, we will demonstrate *GeoTemCo*'s applicability to other research domains.

3.1 Related Work

There has been done a lot of research on the usage of visual interfaces to interact with large data resources. In the following, we will list closely related works in different categories.

Visual Data Exploration Tukey illustrates the benefit of integrating interactive visual tools into the data exploration process to expedite the undirected search for structures and trends [Tuk77]. This so called Visual Data Exploration process follows the Information Seeking Mantra [Shn96], which means, the first presentation of the information is an overview over the whole data, followed by the search for and drill down on interesting patterns, exploiting capabilities of the human visual system. The search for a specific information, which starts with a vague imagination and leads to understanding in a stepwise process is called exploratory search [Mar06]. Much research in the field of Visual Data Exploration was dedicated to the analy-

[1] http://labs.europeana.eu/apps/europeana4d
[2] https://de.dariah.eu/geobrowser
[3] http://heyne-digital.de/

sis of data with geospatial and temporal metadata. An overview of tools, principles, challenges, and the concept of the analysis of geotemporal data are given by Andrienko and Andrienko [AA05, AA06]. Further overviews of the related fields of thematic cartography and geovisualization are given by, e.g. Dent [Den99] and Slocum et al. [SMKH09].

Multiple Views Baldonado [BWKK00] proposed guidelines for the usage of multiple views in visualization to increase the user's ability to receive deeper insights if the data is shown under multiple aspects. Causal relationships as well as unforeseen connections can be found easier. Among many related implementations that exploit synergetic effects from using linked views for geospatial-temporal data, we take a look at the most related ones in the following. The Web application *VisGets* [DCCW08] employs four linked views: a location view, showing the result items as small glyphs on a map, a time view, showing histograms of results for year, month, and day resolution, a tag view, showing most-frequent words in a size proportional to their importance, and a results view, showing small textual or image thumbnails of results arranged as a table. VisGets offers query refinement in space (selecting a glyph), time (selecting a year, month, or day), and by topic (selecting a tag). Each refinement affects the presentation in all other views. VisGets also provides many mechanisms to interact with the results of a single query, but a comparison of different queries as possible with *GeoTemCo* is not directly supported. Because no aggregation of closely positioned glyphs takes place, visual clutter often ensues. The time view supports only query refinement for months within the same year, or days within the same month. The provided time resolutions make it also inconvenient to work with datasets spanning centuries or just hours and minutes. *GeoVISTA CrimeViz* [RRF+10] enables comparison and analysis of different crime incident types. The map view shows individual circles only when zoomed in. In overviews, the incidents of all types are aggregated into hexagonal bins, thereby disabling comparison. The display of the bins using transparency interacts with the map: in dense areas the map is occluded and the underlying map can bias the transparency perception. In the CrimeViz time view, the numbers of incidents per time period of different types are stacked. It can show the total trend well, but is insufficient for comparing the different incident types [CM84a]. The supported linked interactions are asymmetric: a selection in the time view (year, month, week, no time ranges) filters items on the map but not vice versa. CrimeViz's data source offers eight different incident types, of which only the

three smallest (with a maximum total of 637 incidents per year) are used. The other five types, ranging from 2,600 to 9,300 incidents per year each, are omitted. The problem with many data items in a map view becomes apparent in the visualization of the Iraq conflict incidents by *The Guardian* [Rog10]. It does not make use of the different casualty types present in the data source or time information, and suffers from "red-dot fever"[4]: glyphs overlap in the map to an extent where the spatial distribution can no longer be determined reliably (see Figure 3.7a). To support large datasets, our tool thus automatically aggregates data items in the map to non-overlapping circles. It becomes indispensable when presenting glyphs of different visual properties, since overlaps could bias the perceived distributions.

Spatial Binning We tailored a binning algorithm to cluster overlapping circles appropriately. The increasing data sizes to be shown in scatter plots forced the development of binning strategies. Data-independent, top-down approaches like rectangular binning [Wil05] and hexagonal binning [CLN86] split the plane into rectangular or hexagonal bins. The number of items per bin can be reflected in different ways, e.g., a bin coloring with associated colors from a predefined color map. Furthermore, specific shapes (e.g., circles, hexagons) reflecting the bin count with size can be placed in the bin area. A special case for shapes are so called sunflower plots [CM84b] that reflect the number of items in one bin of the scatter plot with a sunflower glyph, which has a specific number of petals for a specific number of points. These top-down binning strategies are widely used in geoapplications [AA05]. However, Novotny [Nov04] remarked that the result can be misleading since cluster centers might be split into distinct bins. He proposes K-means [Llo82] as a bottom-up binning approach, but it requires an appropriate selection of the number of clusters. General clustering algorithms cannot make assumptions on the dimensionality of the data and can therefore be slow. We propose to use a fast data-driven clustering employing a dynamic 2D Delaunay triangulation algorithm. Delaunay triangulations are typically used in geographic information systems to model terrain [SMKH09], but to the best of our knowledge, we are the first to employ it as a clustering algorithm for clutter removal of glyphs.

[4] http://mappinghacks.com/2006/04/07/web-map-api-roundup/

3.2 *GeoTemCo* Design

GeoTemCo's design is inspired by Marian Dörk's VisGets [DCCW08]. It also consists of a map view showing the position of query results, a time view showing the distribution of results in a time span, and a detail view showing textual contents or thumbnails of data items arranged in a table. Our system differs in that we allow comparison of multiple datasets (e.g., results from classical keyword or term-based topical searches), show tag clouds for selected data on demand in the map, provide more flexible selections in the time domain, and render glyphs in the map avoiding visual clutter. Internally, we support any simple statistical graph, with the time view being a special case, and a designer can add more views based on the data to show. Figure 3.8a gives an example view composition.

We chose colors to mark the different datasets because of their effective use to discriminate categorical data [Ber83]. An alternative is to use small multiples similar to the system *LISTA-Viz* [HK10], but we found the use of small multiples makes comparison of scattered data with irregular spatial distribution difficult. Furthermore, colors can serve as the visual link between the different views. As the number of colors that can be easily distinguished by humans is limited [War13], we restricted our method to four datasets to ensure both a good distinction from the map, as well as allowing colors to mix in the time view. We presume map colors to be mostly dark, cold, or unsaturated, and select very light colors for deselected and very saturated colors for selected circles to ensure that the thematic overlay pops out in comparison to the base map. The four base colors used for the four datasets are red, blue, green, and yellow – to accommodate color impaired users while preserving red for single type datasets.

Each view provides native navigation and selection that results in updates of the other views. The map and the time views provide simple zoom and pan. Because of the way we aggregate data in these views an animation between zoom levels or time resolution changes is not performed. We reflect selections by marking table entries and the corresponding fraction of map glyphs and time graphs with saturated versions of the datasets' base colors. When selections are performed by a mouse drag gesture, the impact of releasing the mouse at this point is immediately reflected in the other views. Selections can be modified by dragging shapes or clicking on table entries in the detail view.

3.2.1 Map View

The dominant view in our visualization in terms of screen space is the *map view*. It is a thematic map [SMKH09] comprising a base map and a thematic overlay. The base map can be a contemporary map, and for humanities data our system provides 23 different historical maps showing political borders from 2000 BC to 1994 AD.[5] Overlaying data over historic maps can benefit applications in the humanities (see e.g. Tsipidis et al. [TKK11] for archaeological data, or the HESTIA project [BPBI10], which investigates the differences between imagined geographic distances and real distances in ancient Mediterranean space). Because a dataset may span a time range for which multiple maps are available and there is no concept of "average political border" we show the map closest to the median time stamp occurring in the dataset as default and allow the user to switch maps.

For the overlay, we chose a proportional glyph map over isopleth and choropleth maps due to the scattered nature of our data. We can use neither dasymetric nor dot maps, as these require ancillary information and a cartographer to apply this information correctly. The spatial distribution could also be shown via heat map, but to preserve legibility of the base map and to show multiple datasets, color mixing would ensue, against which Ware [War13] argues. Also, humans are more accurate judging areas than they are judging color tones, making areas a better candidate for quantitative values [CM84a]. Using glyphs also allows to group glyphs and to make every data item individually accessible for interaction.

We disallow the glyphs to overlap in order to avoid visual clutter. While drawing rules such as "always let the smaller overlap the larger" can reduce the risk of occluding small glyphs, this is only a solution if glyphs do not differ in their other visual attributes, like shape or color. Instead, we merge circles based on their size, distances, and the current scale in an iterative process. The overlap removal algorithm, which can be directly attached for single input sets ($m = 1$), is described in the next section. In the case of multiple ($m = 2, 3, 4$) input sets, we compose multiple circles $c_1, \ldots, c_m$ into a more complex glyph – a *circle group* – whose bounding circle b will be used for aggregation. We chose them over pie charts, because these improve comparisons of data at the same point at the expense of comparison of the global distribution. The grouping process is illustrated in Figure 3.2d-f. Initially we place the centers of m prototype circles on the vertices of a regular m-polygon. The prototype circles' ra-

[5] provided by Thinkquest, see http://library.thinkquest.org/C006628/

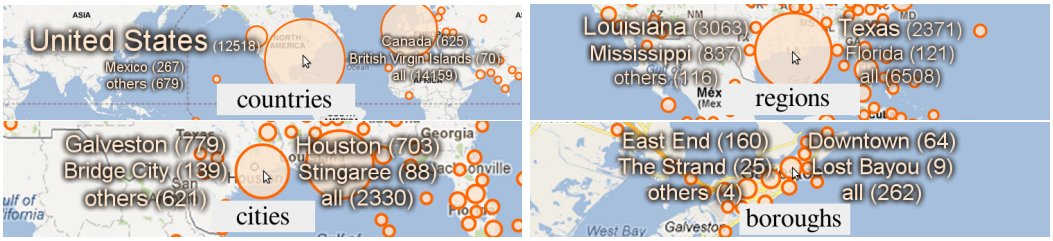


Figure 3.1: Place name tag clouds with different levels of detail

dius is set equal to the largest circle to group. Then we move the m circles using their correct radius from their polygon vertex closest to the polygon center without leaving the prototype. Finally we construct the bounding circle by moving it from the polygon center towards the center of the largest circle until its boundary touches two circles. Although this wastes some space, it is quick to compute and allows the map to be seen through. To ensure the map legibility underneath large circles, we draw all circles semi-transparent.

Selections in the map view can be specified by clicking on items or by drawing circles, rectangles, polygons, or clicking on an administrative region (e.g. country), which then selects the items in the surrounding polygon. Each of the map's circles is associated with a details-on-demand tag cloud showing the most frequent place names in a font scaled proportional to their frequency. The cloud provides a preview of how a glyph arising from agglomeration would split if zoomed in. If the data offers different levels of detail for a place, we choose the label dependent on the current zoom level. We distinguish between country, region, city, and borough level. We replace missing levels by the next coarser or the next finer level. Figure 3.1 gives example place name tag clouds. In the comparison setting we use lines to link tags between datasets, as can be seen in Figure 3.4a.

3.2.2 Time View

Using the terminology of Harris [Har99], our time view is a segmented area graph using time for the x-axis. A segmented area graph is a line chart where the area under the line is filled. $T_1, \ldots, T_n$ partition the interval $T = [t_{min}, t_{max}]$ of the given dataset into intervals of regular duration: either seconds, minutes, hours, days, weeks, months, quarters, years, or decades. It is chosen to maximize the number of intervals without

exceeding 400. Short units typically arise from dynamic data sources and large units arise from historical data. The resolution unit changes automatically when the user zooms inside the time view.

Whereas the x-axis of the time view is directly defined by $T_1, \ldots, T_n$, the y-direction shows the number of data items that fall in each interval using binning. For multiple datasets in the comparative setting we use multiple bins for each time interval. In the final visualization the bins' sizes are shown as overlapping segmented area graphs rather than bar charts because the former is better suited to direct comparison of the groups' time distribution. We shade the area under each line using a semi-transparent version of that dataset's color, ensuring that all curves are visible and also hinting at the area that would have been present in the bar chart. Through the use of blending, the limitation to four colors as well as the stacked area graphs' shape mitigates ambiguities. The user can switch to logarithmic scale if datasets to be compared have largely different totals.

The time view allows both the clicking on one bin and the selection of a time range using a mouse drag gesture. The selected time range can be animated to analyze time-dependent geospatial changes. Finally, a time selection can be used for filtering purposes, which triggers updates in all views of the system.

3.2.3 Detail View

This view is the only one that does not include any aggregation. The data items of multiple datasets are shown in multiple tab windows. For each data item that match the current filtering, we display the provided textual information that is organized in table format. For browsing purposes, each of the table columns can be sorted alphabetically. A column can hold each information in HTML markup, including linked image thumbnails. In the detail view, a current selection can be refined before filtering by removing or adding individual data items.

3.3 Overlap Removal Algorithm

In the following let k denote the number of supported scales (i.e. magnifications). The scale doubles with each level l ($1 \leq l \leq k$). Furthermore, let N denote the number of points P, $p_i = (0.5 + lon_i/360°, 0.5 + lat_i/180°)$ in a normalized space. Each circle i represents n_i points. We define a minimum radius r_{min} dependent on the average font size of common Web mapping services' labels (e.g. Google Maps, Bing

Maps) so that circles have salience no smaller than labels. The maximum radius was found empirically as $r_{max} = 4\log_2(N+1)$. A circle's area A_i is a linear interpolation between the corresponding minimum and maximum circle areas A_{min} and A_{max} based on n_i:

$$A_i = A_{min} + \frac{n_i - 1}{N - 1}(A_{max} - A_{min}).$$

Because neighboring circles are most likely to overlap and we want to merge close circles before far circles, we use a dynamic Delaunay triangulation as supporting data structure, allowing us to quickly find and merge overlapping circles.

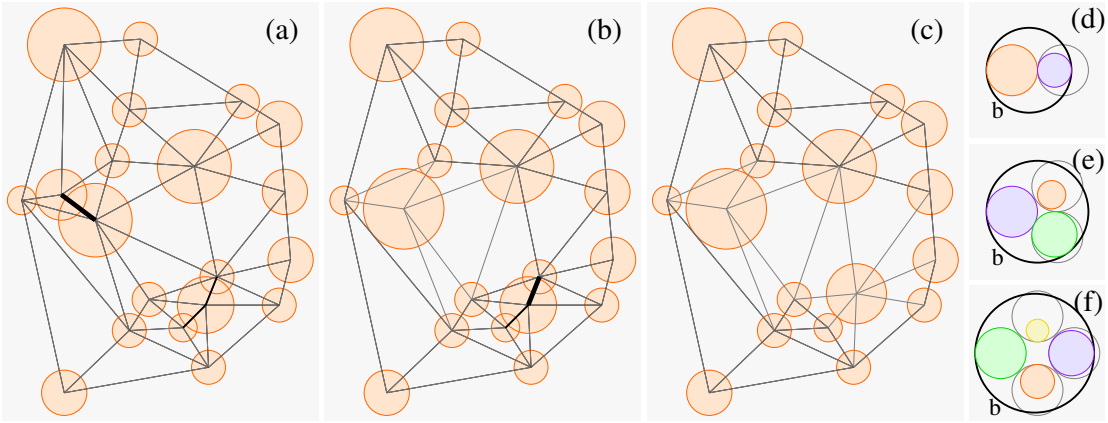


Figure 3.2: Aggregation of single and multiple items. The next occlusion to be resolved is marked by line thickness.

The algorithm is detailed in Algorithm 1 and illustrated in Figure 3.2a-c. It initially creates a dynamic Delaunay triangulation of the given points in the normalized space merging duplicates. After initialization, all edges of the triangulation are inserted into a priority queue with a priority ψ depending on the amount of overlap. ψ relates the distances after transformation from normalized space to screen space and the radii of the circles in pixel. The factor containing l in the formula for ψ assumes that the map image for level 1 has dimensions 256×256. For simplicity and speed we use the ratio of the minimum distance desired plus $\varepsilon \geq 0$ and the real distance as priority. Circles overlap or are too close when $\psi > 1$. After constructing the priority queue, we repeatedly find the overlap of highest priority and remove it by merging the circles, which affects both the Delaunay triangulation and the priority queue, and finish when

Algorithm 1 OverlapRemoval(P)

$D \leftarrow$ empty Delaunay triangulation
for $i = 1$ to $|P|$ **do**
  if $p_i \in D$ **then**
    $p_j \leftarrow$ duplicate of p_i in D
    $n_j \leftarrow n_j + 1$
  else
    Insert(p_i,D)
    $n_i \leftarrow 1$
  end if
end for
for $l = k$ to 1 **do**
  $Q \leftarrow$ empty priority queue
  for all edges $\{p_i, p_j\}$ in D **do**
    $\psi \leftarrow (\varepsilon + r_i + r_j)/(2^{7+l}||p_i - p_j||)$
    if $\psi > 1$ **then**
      Insert($\{p_i, p_j\}$,ψ,Q)
    end if
  end for
  while Q not empty **do**
    $\{p_i, p_j\} \leftarrow$ highest priority element of Q
    $n_{ij} \leftarrow n_i + n_j$
    $p_{ij} \leftarrow \frac{n_i}{n_{ij}} p_i + \frac{n_j}{n_{ij}} p_j$
    Delete(p_i,D)
    Delete(p_j,D)
    Insert(p_{ij},D)
    Update(Q)
  end while
end for

there are no more overlaps. Then we proceed with the next scale, which implicitly halves point distances but not radii in the formula for ψ.

In our implementation we use the algorithm proposed by Kao et al. [KMS91]. It is a randomized incremental algorithm that constructs the initial triangulation in $O(N \log N)$ expected time, and can find a duplicate to a given point in $O(\log N)$. As the triangulation contains $O(N)$ edges, construction of the priority queue using a simple heap requires $O(N \log N)$ time. Each merging of circles requires three elementary operations on the triangulation, each with $O(\log N)$ expected time, and an expected constant number of updates to the priority queue, each with $O(\log N)$ time. As building the priority queue and merging circles takes place for each scale we can give the trivial upper bound of our algorithm as $O(kN \log N)$.

3.4 Usage Scenarios

GeoTemCo[6] – the implementation of the above outlined design – works completely within the client's browser, performing visualization, filtering and interaction, and allows to load data locally or communicates to a server for dynamic data sources. *GeoTemCo* supports both KML and JSON exchange formats. Since the overall support for JavaScript as well as its browser performance has increased in the last years, we decided to implement *GeoTemCo* completely in JavaScript. We adapted two Open-Source JavaScript libraries for implementing our views: OpenLayers[7] is used to provide both the thematic layer and the base map from different Web mapping services, e.g., Google Maps, Open Street Map or historic maps hosted on a GeoServer[8] instance. For our time view, we extended the Simile Widgets Timeplot.[9]

The design was iteratively developed and improved during three projects. The prior purpose of *GeoTemCo* in the two digital humanities projects Europeana[10] and Deutsche Digitale Bibliothek[11] was the comparative visualization of digital library contents. A further purpose in Deutsche Digitale Bibliothek is the occlusion free visualization of institutions belonging to the project;[12] a Europeana use case can be

[6] http://geotemco.vizcovery.org/
[7] http://openlayers.org/
[8] http://geoserver.org/
[9] http://www.simile-widgets.org/timeplot/
[10] http://www.europeana.eu/
[11] https://www.deutsche-digitale-bibliothek.de/
[12] https://www.deutsche-digitale-bibliothek.de/about-us/institutions#map

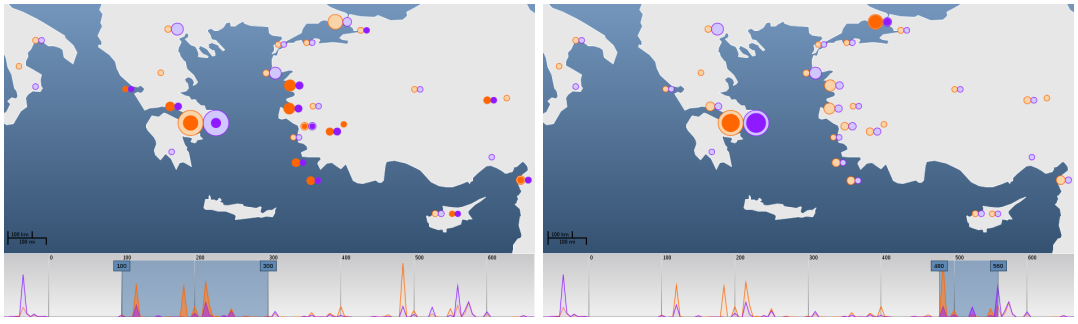
found in [JHSS12]. In the biodiversity project BioVeL,[13] *GeoTemCo* was used to analyze the migration of marine organisms; various examples are presented in [JS14]. Further usage scenarios are outlined below.

3.4.1 Digital Humanities Projects

GeoTemCo turned out to be a valuable tool for the digital humanities. We first take a look at representative use cases from related projects.

I. eAQUA

During the *eAQUA* project [BHG08, HBBS11], humanities scholars used *GeoTemCo* to explore the geographic distribution and propagation of words extracted from Latin and Ancient Greek texts. The provided metadata spans the eastern Mediterranean region with Greece, Italy, Turkey, and North Egypt and a time range from 8 BC to 10AD. The resulting data divisions are shown in a suitable historic context by automatically displaying adequate historic maps. An example is given in Figure 3.3, which shows occurrences of the words *Plato* and *Aristotle* on the map of 400 AD. In the first period – Middle Platonism (1st to 3rd century) – we see a widely spread distribution of *Plato* and *Aristotle* (Figure 3.3a) in the Greek-Ionic region. The second period – Neoplatonism (4th to 6/7th century) – shows a lot of occurrences in the territories of Athens and Constantinople (Figure 3.3b). This indicates a correlated movement of both topics from rural regions to metropolises.



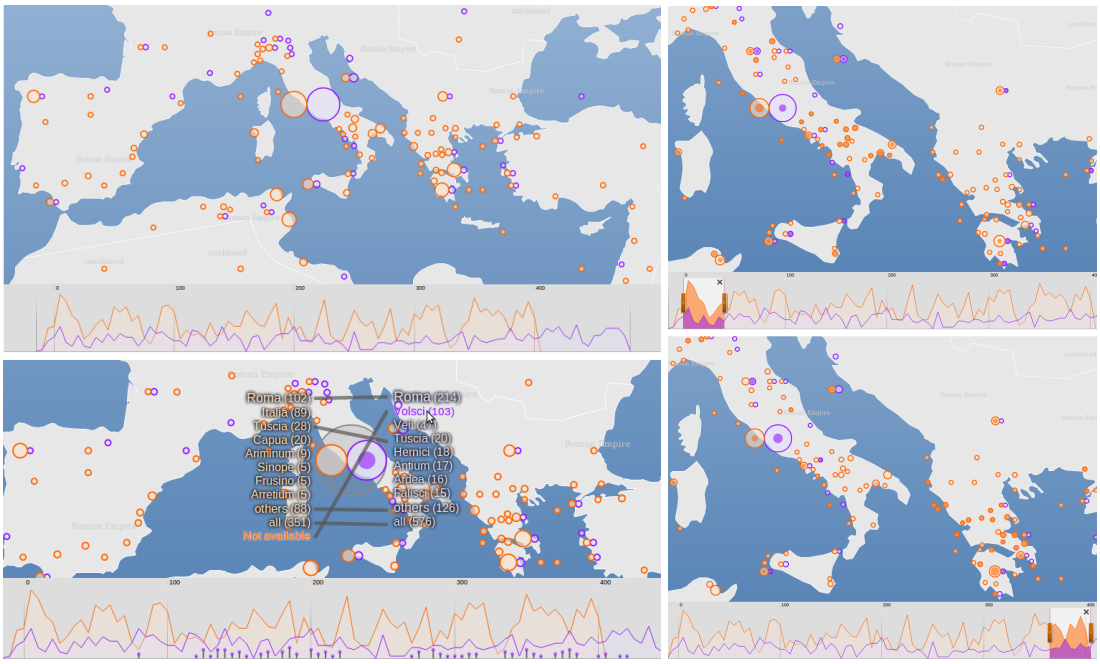
(a) Occurrences during Middle Platonism **(b)** Occurrences during Neoplatonism

Figure 3.3: Occurrences of the words *Plato* (red) and *Aristotle* (blue) in Ancient Greek texts.

[13] https://www.biovel.eu/

II. Books from Google Ancient Places

Google Ancient Places [BIBK11] is concerned with the analysis of books on history. *GapVis*[14] is the corresponding visualization where historic places mentioned in books are plotted onto a map. Page numbers are used as a second dimension, so that an analysis of geospatially migrating topics in books is possible. *GapVis* does not avoid visual clutter on the map and a comparative view for books is also not provided.



(a) Distribution of named places **(b)** First and last book pages

Figure 3.4: Livy books: *Roman History* (red) and *The History of Rome, Vol. 5* (blue)

We use the same data and utilize our time view for the books' page dimension. Figure 3.4 shows an example for two books about the Roman empire, written by Livy approximately 2000 years ago: *Roman History* and *The History of Rome, Vol. 5*. The political map of 1BC reflects proper historian circumstances. Instantly we see quite different geospatial references. Figure 3.4a (bottom) clearly indicates that the second book thematizes the conflict between the ancient Rome and the Volscian territory starting at around page 100 (the placename *Volsci* occurs 103 times). In contrast, this

[14] http://nrabinowitz.github.com/gapvis

period is not discussed in *Roman History* (no *Volsci* occurrences), rather we detect a thematic migration (Figure 3.4b) from the ancient Italian (top) to the ancient Greek region (bottom).

III. Visualizing Medieval Places

Within the Visualizing Medieval Places project,[15] we used *GeoTemCo* to visualize thousands of place names extracted from nearly 550 medieval French texts. The toponyms (and their spelling variants) are being manually harvested from a canonical reference work, Flutre's *Table des noms propres* [Flu62]; they are subsequently disambiguated and geocoded. Since Flutre's work does not fully represent the variety of textual communities and genres of medieval French, we also extracted place names from name indices in selected critical editions.

Using the data has proved problematic since so many aspects of it are uncertain. Situating the composition of medieval texts in a specific time and place can be, at best, speculative. Date formats of traditional scholarship have been represented in idiosyncratic ways (e.g. "between 1095-1291," "first half of the 14th century," "before 1453"). Likewise, the toponyms found in these works are difficult for various reasons: they are unmappable, they can refer to multiple places, or they designate ancient Greco-Latin or medieval geographical zones no longer found on the contemporary map.

Drawing upon a long list of uncertainty types [GS06], a data item within our project might be said to embody two basic kinds of uncertainty. The first uncertainty is one of *lineage*, by which we mean the reliability of the text source. Certainty values for lineage can simultaneously affect the representation of data items in both dimensions, the geospatial and temporal. The second uncertainty is one of *accuracy* referring to the granularity of place or time, that is, to the distinct-sized intervals in which a value can lie. Again, granularity impacts both dimensions, the geospatial (with units such as landmarks, localities, regions, countries, continents) and the temporal (years, eras, as well as upper- or lower-bounded time declarations).

We adapted *GeoTemCo* in order to visualize occurring *accuracy* uncertainties according to the needs of the humanities scholar. First, we use distinct shapes to encode objects with distinct geospatial accuracies. Second, we replaced the time view with a ThemeRiver [HHN00] that visualizes uncertain datings of the underlying medieval

[15] https://visualizingmedievalplaces.wordpress.com/about/

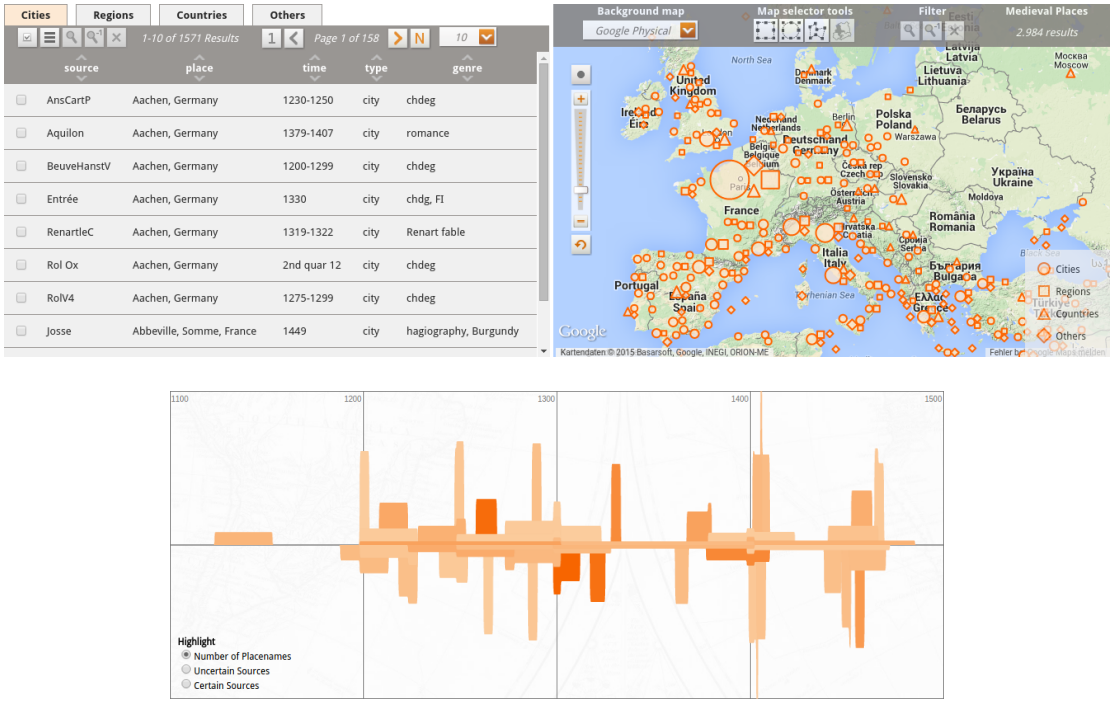


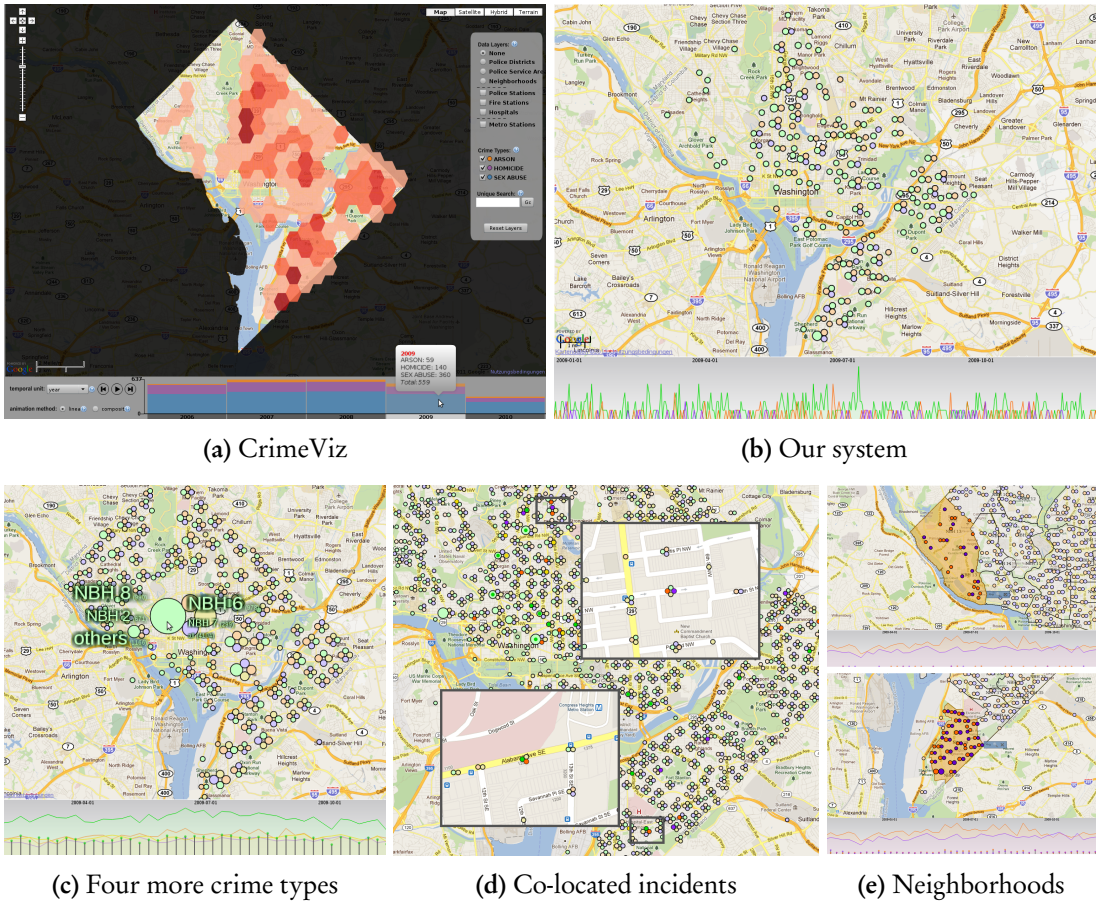
Figure 3.5: Adapted *GeoTemCo* design to visualize uncertainties.

French texts. Each text is represented with a stream. All streams cover the same area, and a stream's width reflects the uncertainty of the text's dating – the wider a stream, the more uncertain the dating. Precise datings occur in the form of peaks.

Figure 3.5 shows the extracted data visualized using the adapted *GeoTemCo* design. Although we could not solve the problem of visualizing all aspects of uncertainty within this project, we were able to design a valuable interface that supports the humanities scholar in analyzing space and time in the medieval French literature corpus for the first time digitally while considering it's ambiguous nature.

3.4.2 Further *GeoTemCo* Use Cases

To emphasize *GeoTemCo*'s adaptability, we illustrate three scenarios from different domains than the digital humanities. These were important examples when developing *GeoTemCo*'s generic design.



(a) CrimeViz **(b)** Our system

(c) Four more crime types **(d)** Co-located incidents **(e)** Neighborhoods

Figure 3.6: Crimes in Washington D.C. (Figure 3.6a reproduced with permission from [RRF+10])

I. Crime Incidents

We compared the *CrimeViz* [RRF+10] application (Figure 3.6a) to our visualization (Figure 3.6b) for crime incidents in 2009. Both show the distribution of homicide, arson, and sex abuse incidents in the Washington D.C. metropolitan area. *GeoTemCo* easily shows that there was neither homicide (blue) nor arson (red) incidents in the north western neighborhoods, which is not directly visible in the CrimeViz map because of the hexagonal binning that furthermore causes a loss of map context in dense regions. When considering further crime types, we find patterns for thefts and robberies near populated places like metro stations or shopping centers. Unlike CrimeViz, we aggregate preserving incident types, hence we find more relations,

e.g. by comparing the crime types stolen cars (red), burglaries (blue), thefts (green), and robberies (yellow) (Figure 3.6c). We detect relatively few stolen car (20%) and burglary incidents (23%) in comparison to thefts (45%) and robberies (35%) in Washington D.C.'s downtown.

Crime analysts can use our visualization to detect connected crime incidents of different type. For instance, a daily exploration reveals that stolen car incidents are often grouped with a theft, a burglary or a robbery incident, e.g., co-located incidents at 27th of June (Figure 3.6d). The number of correlations increases further by choosing a two day time range. The original *CrimeViz* did not allow such fine-grained selection of time ranges. The analyst can furthermore compare different districts using the administrative region selection offered by the map view. Another application could be a decision making support for apartment search, based on regions of low burglary probability, or where it would be safe to rent a garage for a car. Figure 3.6e indicates Neighborhood 13 (top) as substantially safer than Neighborhood 39 (bottom) with respect to stolen cars and burglaries.

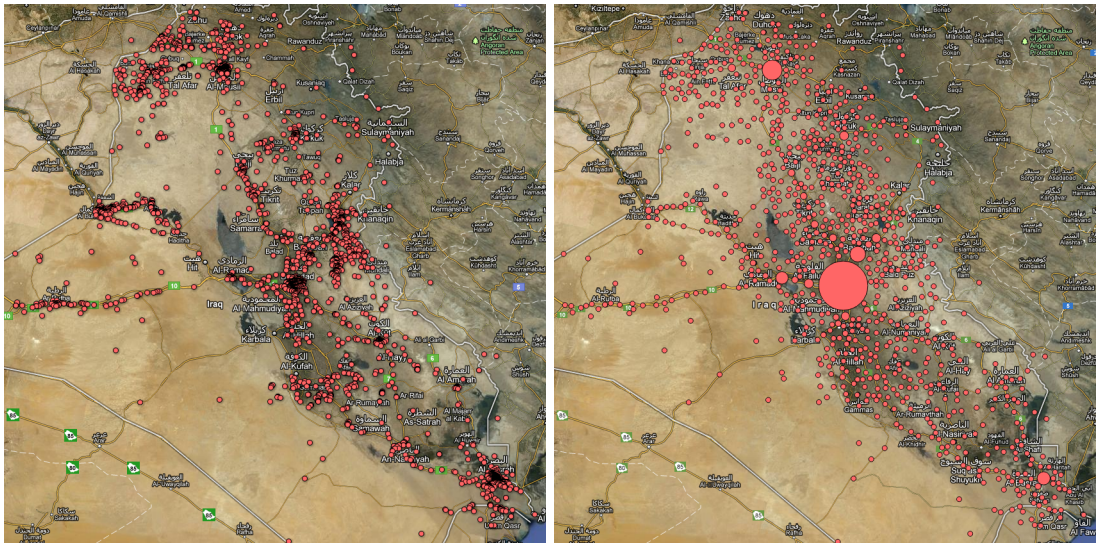
II. Guardian Data

The *Iraq war logs* dataset as published by *The Guardian*, contains around 60,000 entries; one for each incident with at least one casualty during the Iraq conflict from 2004 to 2009. Each entry states place and time, as well as the number of casualties by type (civilian, enemy, Iraq forces, coalition forces).

The Guardian visualization [Rog10] for the *Iraq war logs* is a map containing one circle for each incident. It produces a lot of clutter as the result of overlapping glyphs distorting perception of incident densities. In Figure 3.7a three conflict centers can only be guessed: Baghdad in the center, Al Mausi in the north, and Al Basrah in the south-east, but easily confirmed using *GeoTemCo*'s overlap removal algorithm (Figure 3.7b). Baghdad clearly stands out as the region with most incidents.

For comparison, we split all incidents based on the casualty type into four different datasets. Figure 3.8a shows an increased number of incidents with civilian casualties (red) in 2006 and 2007. A second histogram using logarithmic scale shows the number of incidents by casualty total. We discover the incident with most casualties, which is known as the *2005 Baghdad bridge stampede*. In contrast to newspapers reporting around 1,000 casualties, we see only 437 civilian and 7 Iraq forces casualties (green).

Furthermore, we prepared four datasets containing one item for each casualty to



(a) Guardian map **(b)** Our map view

Figure 3.7: Iraq War Logs: Map Clustering (Figure 3.7a reproduced with permission from [Rog10])

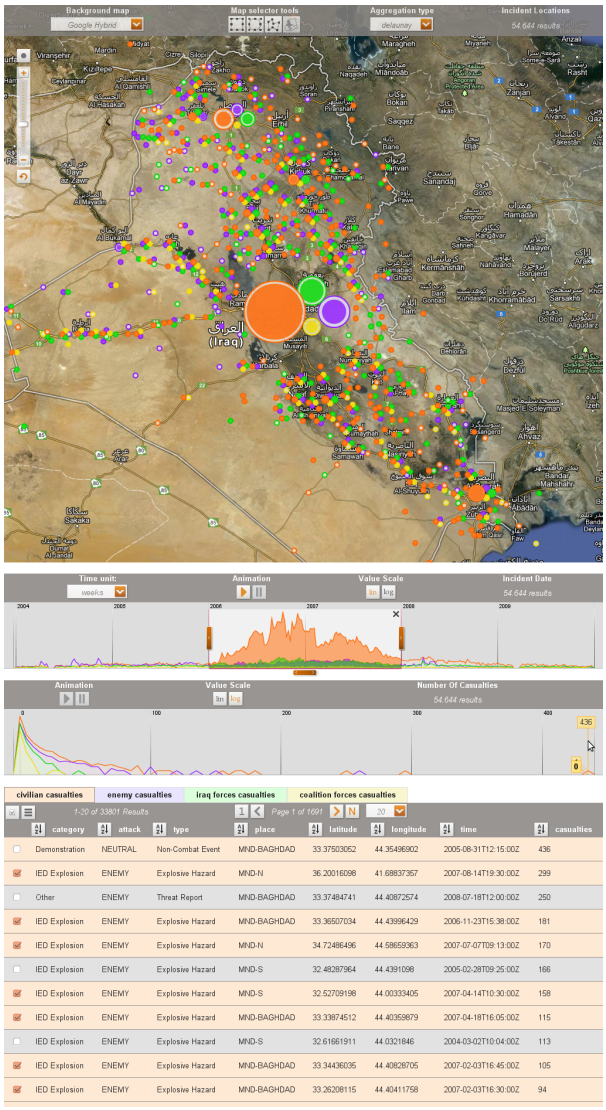
point out regions and time periods with lots of casualties. Figure 3.8b (top) shows an increased number of casualties in 2007 compared to 2004, except for enemy casualties, in particular in Baghdad. The *Operation Imposing Law* (February 14th - November 24th 2007) reduced the overall number of victims, most notably in Baghdad. On March 14th 2007, the Iraq military stated, that there were only 265 civilian casualties in the first month of *Operation Imposing Law*, which is a low compared to the month before the operation (1440). By filtering using the proper time ranges and clicking the tag cloud for Baghdad, we find 2,540 and 997 civilian casualties, respectively, for these time periods in Figure 3.8b (bottom).

III. Visualizing Climate Data with *GeoTemCo*

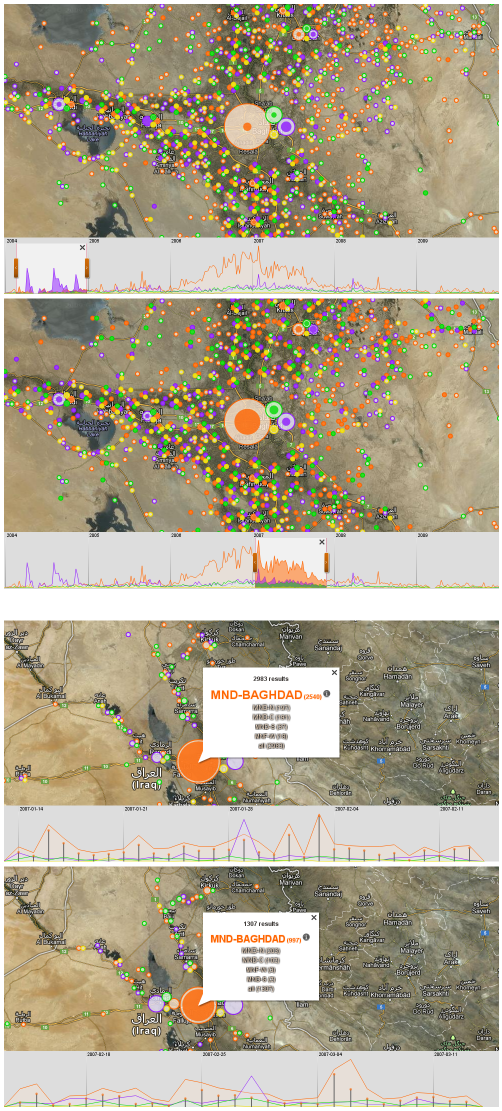
The Storm Events Database, provided by the National Climatic Data Center,[16] contains all significant weather phenomena in the United States from January 1996 to December 2013. For all entries, a date and the corresponding U.S. state is given and often also a precise location.

With the help of *GeoTemCo*, one is able to explore this data geospatially and tempo-

[16] http://www.ncdc.noaa.gov/stormevents/



(a) Incidents by casualty type



(b) Operation Imposing Law

Figure 3.8: Iraq War Logs: Analysis

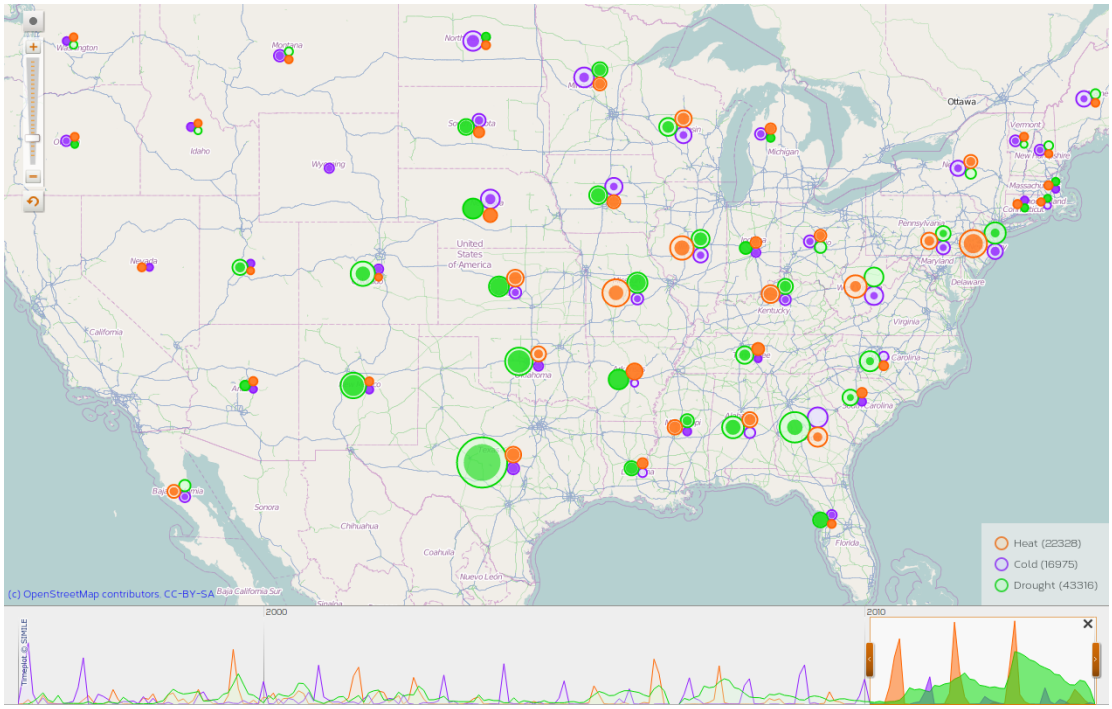


Figure 3.9: Periods of heat, cold and drought mapped

rally. Figure 3.9 shows all reported heat waves (red) and cold snaps (blue) in the given eighteen-year time frame in dependency of the state. Additionally, drought periods (green) are visualized. Before 2010, we can see a steady change between heat waves in summer and cold snaps in winter seasons of similar intensity. But from 2010 to 2012, this relationship breaks as we detect a vast number of heat waves in summer against a decreasing number of cold snaps in winter. Especially in the southern states (e.g., Alabama, Georgia), no cold snaps are reported anymore. As a consequence of the hot summer seasons, also the number of drought periods increased to a hitherto unknown extent (45% of all reported drought periods in the eighteen year time frame have been reported in the last four years). These facts might be an indication for global warming and the subsequent climate change in the United States.

Another weather phenomenon is visualized in Figure 3.10. For the occurrences of tornados (red) and funnel clouds (blue) exact locations are given (if a funnel cloud touches the ground and causes damage, it becomes a tornado). From March to July 2011, we discover a movement of the tornado season from south-east (Florida) to

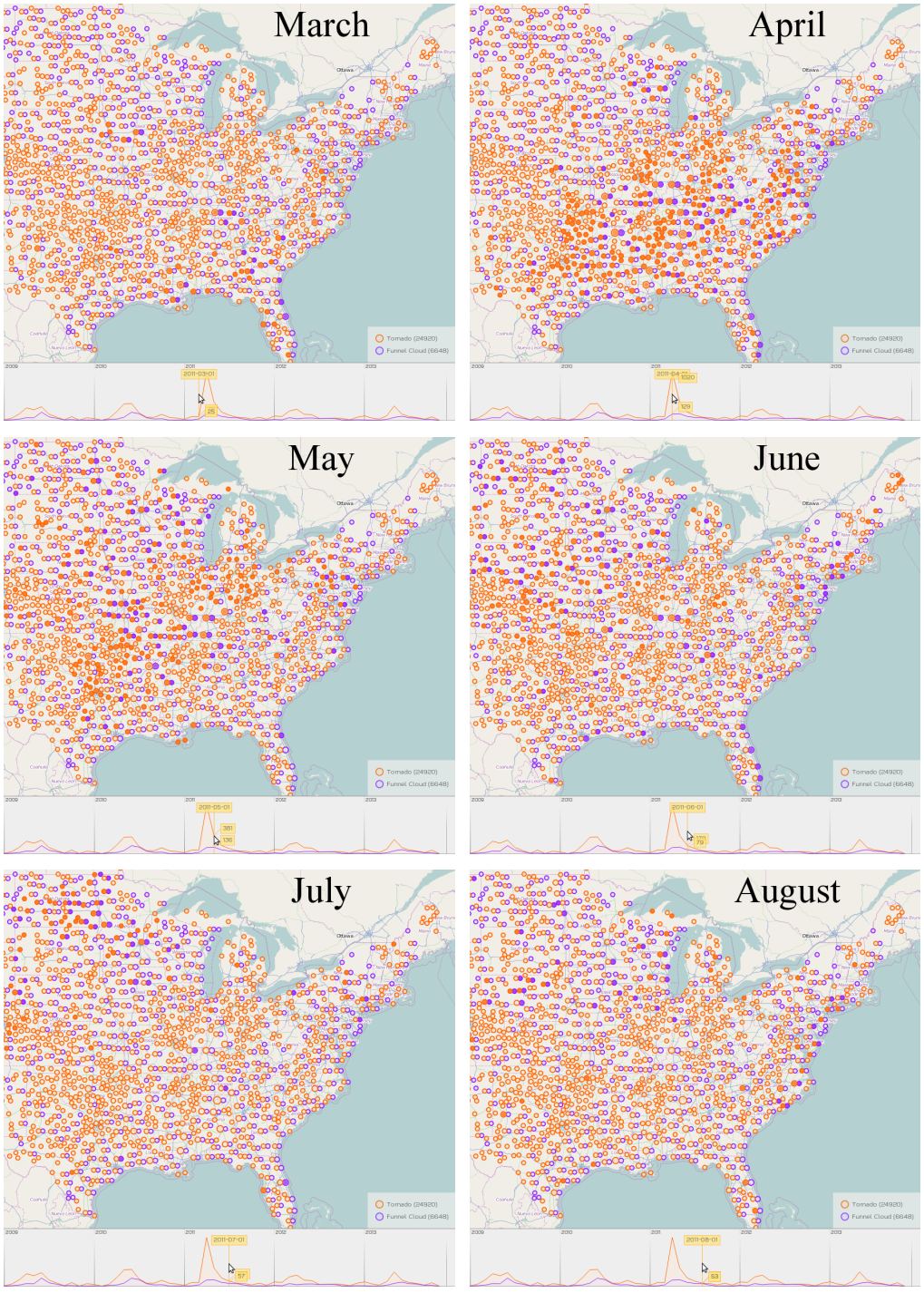


Figure 3.10: Tornados (red) and funnel clouds (blue) in 2011

north (North Dakota). In August, the season seems to be over. A similar behavior was also found for other years (2000, 2003, 2010). Furthermore, it seems that the farther away from the temporary center with most tornados, the lesser the probability that a funnel cloud turns into a tornado, since lots of individual funnel clouds appear without tornados in the immediate vicinity (e.g., individual funnel clouds in the south eastern states in June).

3.5 Summary

We presented a novel approach – implemented as *GeoTemCo* – to show, to compare and to explore multiple datasets in a geographical and temporal context. We were able to utilize, combine, and improve approaches from several prior works. In contrast to CrimeViz [RRF+10], which also offers comparative visualization, we display items without aggregating them into the same representatives for coarser zoom levels. We aggregate map glyphs to avoid visual clutter, which is an issue in the Guardian visualization for the Iraq war logs [Rog10]. A yet similar clustering approach [SvdWvW14] that removes geospatial clutter by aggregating multivariate glyphs was published after *GeoTemCo*. In contrast to our method, it uses pie charts to illustrate the proportions in aggregated glyphs. As the pie slices of different sized pie charts are hard to compare visually, we decided to use circle groups containing individual circles for data items in the same geographical area, but belonging to different datasets. Compared to the similar system VisGets [DCCW08], which only works for one dataset, we also made use of the linked views approach (map, timeline, detail view) to extend the user's exploration abilities. Furthermore, we enriched the filter capabilities in both geospatial (e.g., selecting all data items inside a country) and temporal dimension (e.g., selecting dynamic time ranges). Our method is limited to four datasets at a time mainly to ensure that the colors used for discrimination are properly distinguishable, the splitting of circles does not waste too much screen space, and the overlapping segmented area graphs do not occlude each other too much.

Our usage scenarios show that visually comparable datasets extend the exploration and analysis abilities of the user in an effective way. It helps to detect equalities and varieties between distinct data contents that unveil their relationships in space and time. We outlined *GeoTemCo*'s benefit for the digital humanities field, and furthermore illustrated it's adaptability to other research domains.

Clouds come floating into my life, no longer to carry rain or usher storm, but to add color to my sunset sky.

Rabindranath Tagore

4

Designing Tag Clouds to Analyze Faceted Textual Summaries

The usage of tag clouds to visualize textual data is a relatively novel technique, which was rarely applied in the past century. In 1976, Stanley Milgram was one of the first scholars who generated a tag cloud to illustrate a mental map of Paris, for which he conducted a psychological study with inhabitants of Paris, aiming to analyze their mental representation of the city [MJ76]. In 1992, a German edition of "Mille Plateaux," written by the French philosopher Gilles Deleuze, was published with a tag cloud printed on the cover to summarize the book's content [DG92]. This idea to present a visual summary of textual data can be seen as the primary purpose of tag clouds [SCH08]. But the popularity of tag clouds nowadays is attributable to a frequent usage in the social web community in the 2000s as overviews of website contents. Although there are known theoretical problems concerning the design of tag clouds [VW08], they are generally seen as a popular social component perceived as being fun [HR08]. With the simple idea to encode the frequency of terms to a given topic, tag clouds are intuitive, comprehensible visualizations, which are widely used metaphors (1) to display summaries of textual data, (2) to support analytical tasks such as the examination of text collections, or even (3) to be used as interfaces for navigation purposes on databases.

Figure 4.1: Result lists for text passages containing the Latin word *morbus* in PHI[2] (left) and Perseus Digital Library[3] (right).

In the recent years, various algorithms that compute effective tag cloud layouts in an informative and readable manner have been developed. One of the most popular techniques is Wordle [VWF09], which computes compact, intuitive tag clouds and can be generated on the fly using a web-based interface.[1] Although the produced results are very aesthetic, the different used colors do not transfer additional information and the final arrangement of tags depends only on the scale, and not on the content of tags or potential relationships among them. This chapter outlines the design process of tag clouds using color and position for encoding additional information in order to support humanities scholars in analyzing research questions in philology. Traditionally, scholars specialized in a specific topic read and analyze known texts containing related passages. But the digital age opens possibilities for the scholars to enhance their traditional workflows. Due to many digitization projects, humanities scholars nowadays have access to a vast amount of ancient texts through web portals like PHI Latin Texts[2] or the Perseus Digital Library.[3] The usual approach of humanities scholars working with digital corpora is a keyword based search. Often, they receive numerous results as plain lists of text passages (see Figure 4.1), which are hard to revise individually, and a generation of valuable hypotheses is a hard task.

[1] http://www.wordle.net/
[2] http://latin.packhum.org/
[3] Ed. Gregory R. Crane. Tufts University. http://www.perseus.tufts.edu

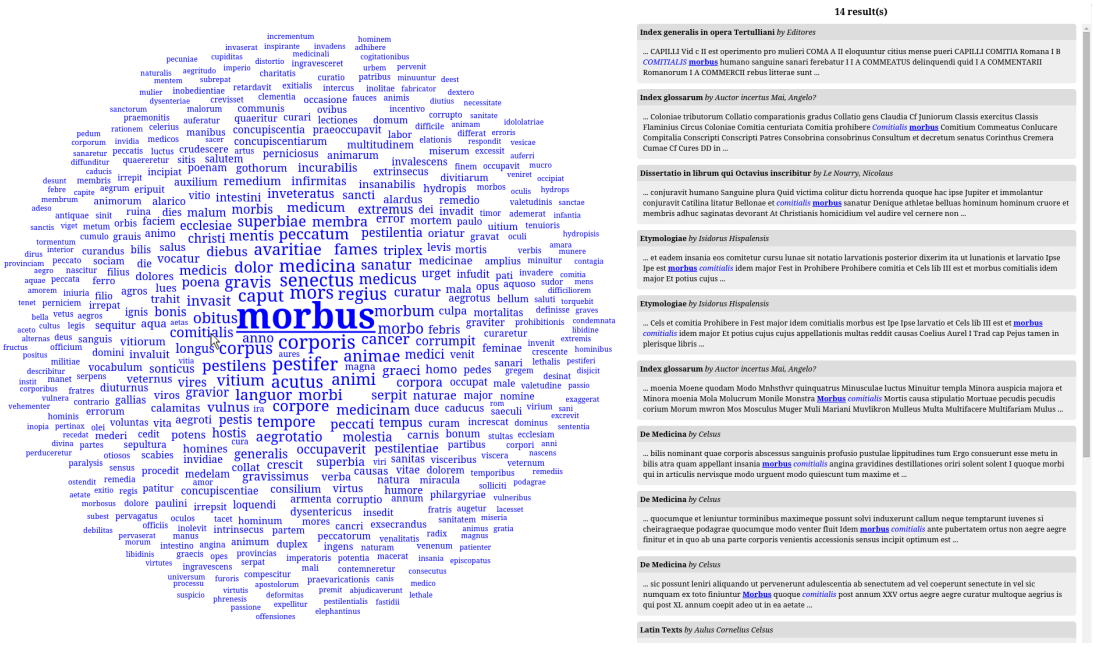


Figure 4.2: A tag cloud gives an overview of the co-occurrences of 1.558 occurrences of *morbus*. By clicking the term *comitialis*, the corresponding fourteen text passages are shown as previews.

Humanities scholars within the digital humanities project *eXChange*[4] also worked with a large corpus containing numerous Ancient Greek and Latin texts. In order to effectively query this corpus, the humanities scholars desired visualizations that give an overview of the results and can be used to directly jump to text passages containing specific co-occurrences of the searched term. As many humanities scholars are not used to work with complex tools, we decided to provide interfaces that are easy to understand. The usage of widely applied tag clouds was the means of choice. An example tag cloud showing the co-occurrences of the Latin term *morbus* (disease) and text passages containing both *morbus* and *comitialis* is shown in Figure 4.2.

In this chapter, we present two novel tag cloud designs tailored to support humanities scholars in exploring various research questions:

- **TagPies** – a tag cloud in a pie chart manner – facilitate the comparison of the results of multiple queries on the text corpus within one interface in order to discover similarities and differences of the contexts in which related terms were used.

[4] http://exchange-projekt.de/

- **TagSpheres** – a tag cloud that effectively visualizes hierarchies in textual summaries – support the analysis of the clause functions of an ancient term's co-occurrences.

4.1 Related Work

Although tag clouds rather became popular in the social media, research in visualization attended to the matter of developing various layout techniques in the last years. A basic tag cloud layout is a simple list of words placed on multiple lines [VWvH+07]. In such a list, tags are typically ordered by their importance to the observed issue, which is encoded by font size [Mur07]. An alphabetical order is also often used, but a study revealed that this order is not obvious for the observer [HR08]. Later, more sophisticated tag cloud layout approaches that rather emphasize aesthetics than meaningful orderings were developed. A representative technique is Wordle [VWF09, JLS15], which produces compact aesthetic layouts with tags in different colors and orientations, but both features do not transfer any additional information. A Wordle showing the most important terms in Edgar Allan Poe's *The Raven* is given in Figure 4.3.

Various approaches highlight relationships among tags by forming visual groups. In thematically clustered or semantic tag clouds, the detection of tags belonging to the same topic is supported by placing these tags closely [LZT09]. Traditional, semantic word lists place clustered tags successively [ST14]. More sophisticated layout methods often use force directed approaches with semantically close terms attracting each other [CWL+10, LSH14, WZG+14]. After force directed tag placement, tag cloud layouts can be compressed by removing occurring whitespaces [WPW+11].

Some methods generate individual tag clouds for each group of related tags, and combine the resultant multiple tag clouds to a single visual unity afterwards. An example is the Star Forest method [BKP14], which applies a force directed method to pack multiple tag clouds. Other approaches use predefined tag cloud containers, e.g., user-defined polygonal spaces in the plane [PTT+12], polygonal shapes of countries [NTST11], or Voronoi tesselations [SKG11]. Newsmap uses a treemap layout [SP98] to group newspaper headlines of the same category in blocks [Wes15]. Morphable Word Clouds morph the shapes of tag cloud containers in order to visualize temporal variance in text summaries [CLC+15]. For the comparison of the tags of various text documents, a ConcentriCloud divides an elliptical plane into sectors that list shared tags of several subsets of the underlying texts [LHB+15]. Due to

Figure 4.3: Wordle of Edgar Allan Poe's *The Raven*.

the rather independent computation of individual tag clouds – which often leads to large whitespaces in the final composition step – the above mentioned methods can be seen as sophisticated small multiples. A rather traditional small multiples approach is Words Storms [CS14] that supports the visual comparison of textual summaries of documents.

Tag clouds also have been used to visualize trends. Parallel Tag Clouds generate alphabetically ordered tag lists as columns for a number of time slices and highlight the temporal evolution of a tag placed in various columns on mouse interaction [CVW09]. In contrast, SparkClouds attach a graph showing the tag's evolution over time [LRKC10]. Other approaches overlay time graphs with tags characteristic for certain time ranges [SWL+10].

Only few approaches generate multifaceted tag cloud layouts in a single, continuous flow that includes the positioning of all tags belonging to various groups. RadCloud places the tags of various groups in a shared elliptical area [BLB+14]. In Compare Clouds, tags of two media frames (MSM, Blogs) are comparatively visualized in a single cloud [DES+15]. Both layout designs use color to highlight topical relations.

As all above approaches were not suitable to sufficiently support the humanities scholars of our project in their workflows, we designed TagPies and TagSpheres – both producing compact, aesthetic tag clouds in a single continuous flow.

4.2 TagPies

Within the *eXChange* project, humanities scholars were not only interested in the detection of related text passages containing a specific term, but also in the comparison of the contexts in which various terms in Latin and Ancient Greek were used. Typical research questions are discussed in Section 4.2.3. To support the generation of hypotheses, we present TagPies that visualize the results of various search requests – treated as various *data facets* – on the project database. In the following, we use the term *shared tag* for a tag that occurs in several data facets, and the term *unique tag* for a tag that occurs in only one data facet.

Several approaches for visualizing facets in tag clouds exist, but in contrast to TagPies they fail when it comes to preserving the compactness, intuitivity and readability as provided by approaches for single datasets like Wordle. In order to develop a beneficial visualization, we conducted two case studies. The purpose of the first case study (see Section 4.2.1) was to find out if state-of-the-art tag cloud layouts would be sufficient interfaces to support the humanities scholars in their workflows. Based upon the results of this study, we developed TagPies that aim to combine the opportunities and to avoid the drawbacks of the concurring state-of-the-art methods. The utility of our method was evaluated in a second case study (see Section 4.2.4). Eight collaborating humanities scholars (6 female, 2 male) participated both case studies. A formal user study using performance data was not viable due to the small number of humanities scholars, their diversified research interests and the exploratory nature of their tasks. In particular, we wanted to encourage the participants of the case studies to intensely work with the provided visualizations by allowing them to query the database with terms of their own interest (preference data). Thus, preparing examples and defining concrete tasks beforehand was not feasible.

4.2.1 Case Study 1: State-of-the-Art Tag Clouds

The goal of the first case study was to assess if existing state-of-the-art tag cloud layouts support answering the humanities scholars' research questions in a satisfactory way. As the key issue is the comparison of data facets with overlapping tag sets, we chose only those methods capable of displaying such relationships. In particular, we prepared variants of the following visualizations, all showing only upright tags and all using the same color map to identify tags belonging to specific data facets:

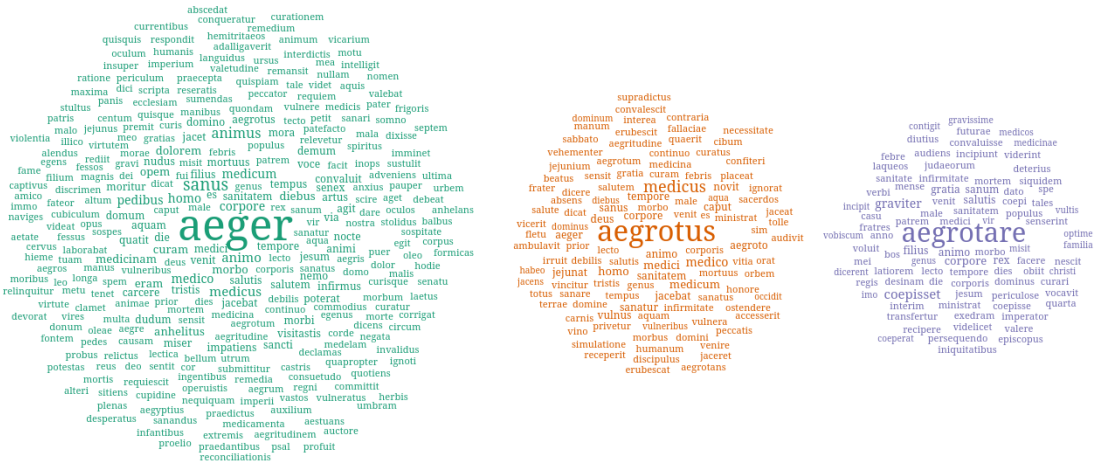


Figure 4.4: Small Multiples example.

- **Small Multiples**: For each data facet, we computed a Wordle [VWF09] tag cloud and used the idea of positioning equal tags at similar positions as suggested in WordStorms [CS14]. An example is given in Figure 4.4.

- **Parallel Tag Clouds** visualize data facets in a specific temporal order, which is not given in our use case. But as they provide a design for shared tags, we implemented a parallel tag cloud [CVW09] listing all tags alphabetically in various columns for various data facets. Shared tags are displayed on a gray background as shown in Figure 4.5.

- **Wordle**: Inspired by the Wordle [VWF09] algorithm, we designed a tag cloud that visualizes the tags of various data facets in one visual entity. All tags are ordered the way that shared tags $t_1, t_2, \ldots, t_n$ from n data facets are subsequently positioned. The spiral for the most frequent tag t_1 originates from the center of the cloud, for $t_2, \ldots, t_n$ we use the coordinates of t_1 as spiral origin. Consequently, shared tags are placed close to each other in small groups (see Figure 4.5), which enables the comparison of $t_1, t_2, \ldots, t_n$ among various data facets.

- **RadCloud**: We designed a RadCloud [BLB+14] variant for the digital humanities data merging shared tags and attaching a bar to each tag that shows its relative distribution among all data facets (see Figure 4.5).

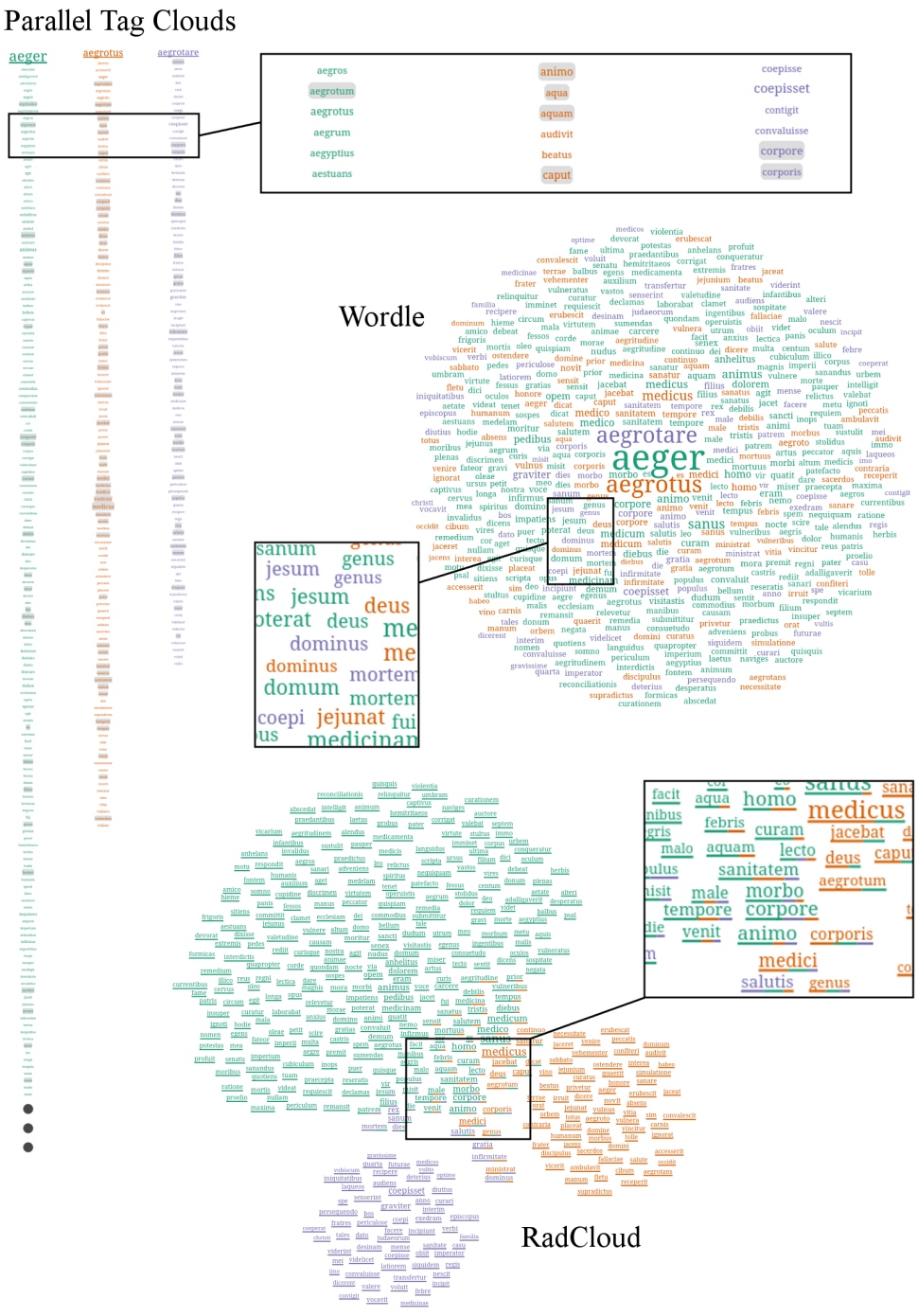


Figure 4.5: Variants of Parallel Tag Clouds, Wordle and RadCloud.

With these visualizations at hand, we conducted the first case study with the humanities scholars in individual face to face sessions. The various layouts were shown and briefly explained in random order, and the scholars were asked to try out the same example query on the database interesting for their research topic for each layout. We also encouraged the participants to "think aloud" [Lew82] to judge the skills of the participants on the one hand and to better understand the varying needs on the other hand. For each visualization, the humanities scholars were asked for a subjective rating regarding intelligibility of the tag cloud, utility for their work, and design and aesthetics of the visualization. For each decision on a Likert scale from 1 (very bad) to 7 (very good), we also required a justification by the participants. To gain further insights, we finally discussed opportunities and disadvantages of the various layouts regarding their applicability for humanities research. Based upon questionnaires and discussions, the evaluation can be summarized for each visualization:

- **Small Multiples** are most useful for exploring the context of single data facets in aesthetic tag clouds. However, the discovery of shared tags is hard without interaction means, so that a profound comparison of data facets is not possible.
- **Parallel Tag Clouds** were unexpectedly not favored, although their basic design is similar to word lists, with which humanities scholars are used to work. The major issues are the heights of the tag clouds, which lead to a lot of vertical scrolling during the exploration process, and the required interaction to gain additional information; the humanities scholars stated they want to see several information "at the first glance."
- **Wordle** was seen as being the most aesthetic layout. It helps to compare the importance of single tags for various data facets. However, the analysis of a single data facet's context or even the comparison of various contexts is a hard task.
- **RadCloud** is the best solution to discover words appearing in various data facets and to compare contexts. However, the analysis of a single data facet's context is not easy as relevant words are often spread across the whole layout as parts of other tag's bars. Although helpful, the bars reduce the aesthetics of the layout and confuse when trying to compare the frequency of terms, an information that is visualized in two concurring manners (bar, tag's font size).

The humanities scholars also wished a layout that approximately reflects the relative frequencies of data facets in the collective search result. The comparison of diameters of the individual tag clouds in the small multiples approach was most qualified for this task.

4.2.2 Designing TagPies

As a result of the first case study we generated a list of requirements for a tag cloud layout valuable for the collaborating humanities scholars. The layout should:

- support the analysis of a single data facet's context,
- support a comparison between the contexts of various data facets,
- visually separate shared from unique tags, and
- reflect the proportion of data facets in the collective result.

Furthermore, we took the often mentioned importance of aesthetics into account when developing our tag cloud approach. As postulated in [OG14], we designed TagPies based upon these requirements derived from the needs of the target group.

I. Design Decisions

For computing the tag cloud layout, we use several well-established design features. Evaluated as being the most powerful property [BGN08], we use font size to encode the number of occurrences of each tag. As suggested in [WSK+13], color is the best choice for distinguishing categories. Hence, we use qualitative color maps to assign distinctive colors to various data facets. For this purpose, we use those qualitative color maps provided by ColorBrewer [HB03] that contain solely saturated colors, as the tag clouds are displayed on white background in a web-based environment. Thereby, we consider not to assign red and green hues as well as colors with similar hues to adjacent TagPies sectors. Also stated in [WSK+13], users perceive rotated tags as "unstructured, unattractive, and hardly readable." Therefore, we do not rotate tags to keep the layout easily readable, thereby providing an interface that is beneficial for the collaborating humanities scholars. To avoid whitespaces, a problem addressed in [SKK+08], our method is based on the Wordle algorithm, which permits overlapping tag bounding boxes if the letters do not occlude. Thus, compact, uniformly looking aesthetic tag clouds even for multifaceted data are obtained.

II. Layout Algorithm

Given are n given categorical data facets $d^1, \ldots, d^n$ (n search result sets), each containing the co-occurrences for the queried search terms $T^1, \ldots, T^n$ (as main tags). The idea is to place the tags of a data facet in a specific circular sector. With the resultant tag cloud subdivision, the final layout is visually comparable to a pie chart, which helps the observer to immediately perceive the relative proportions of the various data facets. According to the actual proportions (the number of occurrences of the main terms in the database) and a maximum number of tags to be displayed (for the outlined examples we chose a maximum of 500 tags), we select the top co-occurring terms (tags) for each data facet. If the relative proportion of a data facet is too small, we leave a minimum of five tags to be displayed.

For each data facet d^i, we need to position the category's main tag T^i (the search term) and the tags $t^i_1, \ldots, t^i_{|d^i|}$ ($d^i = \{t^i_k | 1 \leq k \leq |d^i|\}$), which are co-occurrences of T^i. T^i encodes the number of the search term's occurrences in the database. It's visualization supersedes an additional legend and furthermore serves the purpose of accentuating the belonging of each individual tag to its facet. Each TagPies tag t^i_k has the following properties:

- a *multiplicity*, which is the number of data facets t_k belongs to,
- a *weight*, which is the number of co-occurrences of t_k with T^i, and
- an *aggregated weight*, which is the aggregated number of co-occurrences of t_k among all data facets.

Methodology. In preparation, we order the data facets according to their similarity in a queue *sq* with the goal to place as many similar tags as possible close to each other. The similarity $s(d^i, d^j)$ is defined by the number of shared tags in proportion to the number of unique tags between two data facets d^i and d^j as the Jaccard index

$$s(d^i, d^j) = \frac{|d^i \cap d^j|}{|d^i \cup d^j|}.$$

Initially, we put the most similar data facets in *sq*. Then, we iteratively determine the facet d^i with the highest similarity to either the first (then, we insert d^i at the start of *sq*) or the last element (then, we insert d^i at the end of *sq*) in *sq*. With the

resultant ordering at hand, we estimate the amount of space required to place the tags for each data facet d^i. This is achieved by mapping each tag in the corresponding font size dependent on a tag's frequency and adding up the bounding boxes for all tags of a data facet. Using a Cartesian coordinate system, we subdivide the plane at it's center $(0, 0)$ into circular sectors defined by the angles $\varphi^1, \ldots, \varphi^n$ based upon the above determined proportions for $d^1, \ldots, d^n$ as shown in Figure 4.6. The tags in the final uniform layouts are evenly distributed and the sector sizes approximately reflect the actual proportion, even though a heuristic approach is used.

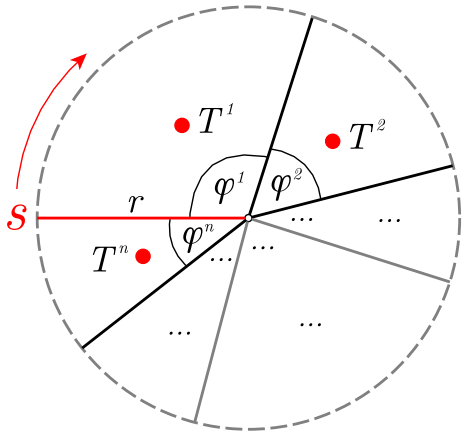


Figure 4.6: Defining circular tag cloud sectors.

Main Tag Placement. To position the main tags, we temporarily compute a Wordle tag cloud without sectors containing all tags using an Archimedean spiral to approximate the TagPies radius r, which is defined in dependency on the two most distant tags. Afterwards, we place the main tag T^i of each data facet d^i in the tag cloud in the corresponding sector at position $p(T^i) = (x^i, y^i)$ as illustrated in Figure 4.6. Starting with the orientation s, $p(T^i)$ is defined by

$$x^i = \gamma \cdot r \cdot \cos(\pi + \sum_{k=0}^{i-1} \varphi_k + \frac{\varphi_i}{2})$$

and

$$y^i = \gamma \cdot r \cdot \sin(\pi + \sum_{k=0}^{i-1} \varphi_k + \frac{\varphi_i}{2}).$$

With $\gamma = 0.5$, we position T^i at the center of the sector. Especially when several small sectors are adjacent, the corresponding main tags can occlude. To avoid these occlusions, we selectively modify γ in such cases.

Tag Ordering and Placement. To gain uniform tag cloud layouts, we order the tags of all data facets in a queue *tq* in two subsequent iterations. In the first iteration, we sort all tags by descending multiplicity, so that shared tags are placed in the center of the tag cloud. In case of even multiplicity, we sort first by descending aggregated weight, and then by descending tag weight. The second sort iteration slightly reorders the tags in *tq* according to the data facets' proportions, so that the tags of each data facet are uniformly distributed in *tq*. This procedure aims to enlarge all sectors uniformly to finally receive a smooth compact tag cloud in the form of a pie chart. With the final ordering, we iteratively position all tags following an Archimedean spiral originating from the tag cloud center at position $(0, 0)$. Thereby, a tag is placed if the determined position on the spiral lies in the corresponding sector and if the tag does not occlude other tags. At the borders, tags also overlap adjacent sectors. Therefore, the second sort iteration is especially important because of the limited space assigned to small facets. To forestall a complete overlapping of these small areas by tags of other facets, this approach specifically guarantees that tags belonging to small data facets are treated earlier during the layout algorithm.

III. Design Variants

In order to prepare the second case study (see Section 4.2.4), we developed design variants of our approach with the goal to discover the best TagPies design from the humanities scholars' point of view. All variants of a participant's case study example are shown in Figure 4.7:

- **Basic**: as described in this section without additional features
- **Bars**: Basic design enriched with bars for each shared tag showing its distribution among all data facets
- **Italics/Bold**: Basic design with shared tags in italics and unique tags in bold
- **Bold/Italics**: Basic design with shared tags in bold and unique tags in italics
- **Merged**: shared tags are merged as suggested in RadCloud [BLB+14], drawn in the color of the most contributing data facet and attached bars show tag

Figure 4.7: Six designs variants for TagPies. The humanities scholar working with this example preferred the **Bars** variant.

distributions among all data facets (to reduce the visual load, bars for unique tags are not drawn)

- **Merged Black**: Merged design variant with merged tags drawn in a separate color (black)

Although the bold main tags are already salient in their sectors due to font size, we underline main tags for variants without bars (Basic, Italics/Bold, Bold/Italics) to strengthen global salience.

IV. User Interface

Figure 4.8 shows a screenshot of the web-based user interface the humanities scholars of the *eXChange* project work with. TagPies, which are implemented as a JavaScript library that is based on the d3-cloud,[5] are embedded to visualize the results of multiple queries on the *eXChange* database, which contains a multitude of Latin and Ancient Greek texts. The humanities scholars can configure the visualization as desired. Next to the preferred design variant, the number of tags shown in TagPies and the maximum distance between a searched keyword term and considered co-occurrences can be defined (further information about all configuration possibilities can be found on the TagPies homepage[6]). After retrieving the results, stopwords are removed dependent on Latin and Ancient Greek stopword lists provided by the humanities scholars. The remaining co-occurrences are visualized with TagPies.

To facilitate navigation and exploration, we enhanced TagPies by basic means of interaction according to the humanities scholars' wishes. Of particular interest was the highlighting of spelling variants provided by the backend of the research platform. With mouse interaction, we enable the scholar to detect related tags more quickly. Hovering a tag highlights the remaining shared tags using a black font on transparent backgrounds having the data facet color as shown in Figures 4.8 and 4.9. Spelling variants retain their saturated font color, but gray, transparent backgrounds indicate relationships. Additionally, all related tags are listed in a tooltip (shown on mouse click), which illustrates the distribution using a bar chart. Here, bars receive saturated colors in case of similarity, the bars of spelling variants are displayed using unsaturated colors. Via mouse click, the humanities scholar is able to directly jump to text

[5] https://github.com/jasondavies/d3-cloud
[6] http://tagpies.vizcovery.org/

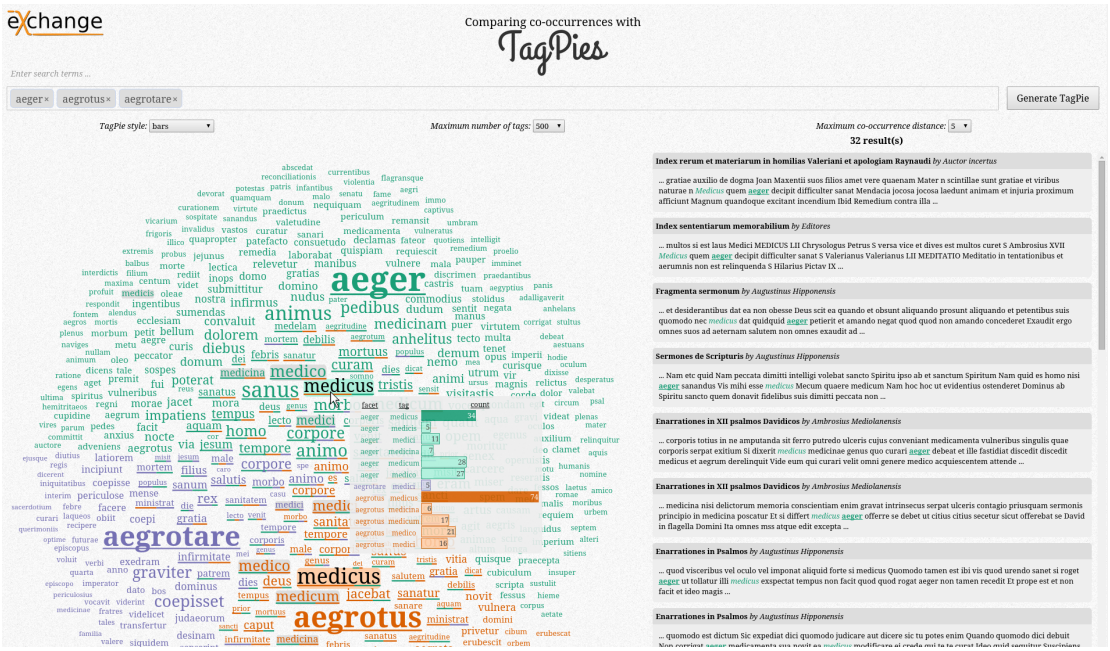


Figure 4.8: Screenshot of the user interface. The co-occurrence *medicus* of ***aeger*** is chosen.

passages containing a searched keyword and the selected co-occurrence. As shown in Figures 4.8, 4.10 and 4.11, corresponding text passages are displayed in the form of previews. This supports a quick inspection of potentially interesting results by the user, and a preview can be used to link to the given position in the corresponding source text.

4.2.3 Usage Scenarios

This section emphasizes the benefit of the proposed design for typical research questions in the digital humanities project *eXChange*. A main focus of this project is the analysis of the usage of medical terms in antiquity. We illustrate three such examples provided by the collaborating humanities scholars for whom TagPies turned out to be heuristically valuable for philological matters by directing the search for related text passages to interesting and potentially unexpected findings. An additional fourth example shall illustrate the applicability of TagPies to other fields.

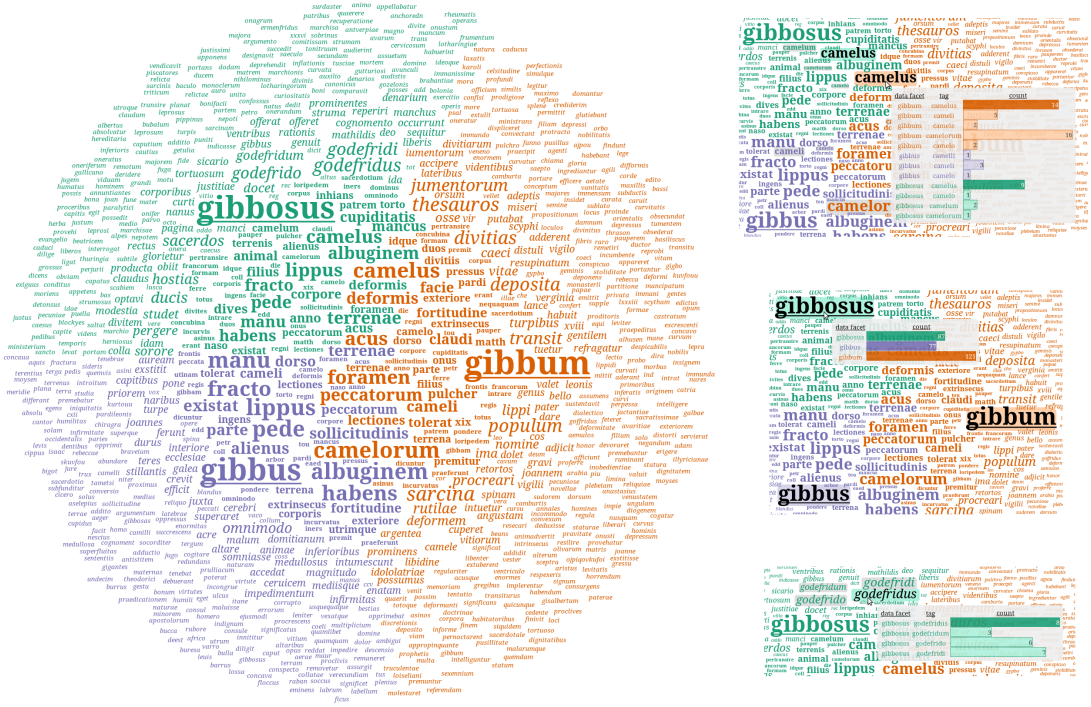


Figure 4.9: Comparing the co-occurrences of ***gibbum***, ***gibbus*** and ***gibbosus***.

I. Comparing *gibbum*, *gibbus* and *gibbosus*

The first example is given by the search for co-occurrences of the Latin words ***gibbum***, ***gibbus*** and ***gibbosus***, all meaning hump or humpy. With the help of the resultant TagPies in the Bold/Italics variant (Figure 4.9), the humanities scholar could generate hypotheses about the actual and improper usage of these terms by exploring their contexts.

gibbum is a frequently occurring Latin noun for "hump" and was most often used in its actual physical meaning. It describes the natural, not morbid hump of animals, shown by the co-occurrences of variants of camel (e.g., *camelorum*, *camelus*) and cattle (*jumentorum*). Often thematized was the transport (*transit*, *sarcina*, *lateribus*) of treasures (*thesauros*) with the help of these animals.

gibbus is an adjective meaning "humped" frequently used in medical contexts. On the one hand, it often appears when describing the natural or morbid contortion of a physical part of the body like the hand (*manu*), the back (*dorso*, *spina*), or the foot (*pede*, *pes*). On the other hand, the term is used in context of diseases, e.g., eye diseases

(*albuginem, lippus*), ulcer (*ulcus*), and broken bones (*fracto*).

gibbosus (a younger synonym for *gibbus*) is also, but less often used in the above mentioned contexts. Additionally, *gibbosus* describes moral contortion indicated by words like *cupiditatis* (lust), *curiositatis* (curiosity), and *glorietur* (boast). Due to coincident appearances of many Christian terms (e.g., *sacerdos, godefridus, hostias*) the humanities scholar hypothesized a local temporal usage of a monastic author, which could be verified by browsing the corresponding text passages (Frankish Empire, 9th century).

II. Comparing *morbus comitialis* and *morbus sacer*

The second example narrates an unexpected insight for the humanities scholar working with TagPies in the Bar variant displaying the co-occurrences of ***morbus comitialis*** and ***morbus sacer***, both terms were used to describe the disease epilepsy (Figure 4.10). The visualization supported the scholar in examining three major questions:

- **What is the semantic relationship between both terms?** In the center of the cloud the common terms *graeci*, *ιεραν* and *νοσον*, forming the phrase "the Greeks [call it] holy disease," can be seen. This relationship indicates that both terms were actually used as synonyms for epilepsy at that time.

- **Were the terms used to describe the disease in its medical or metaphorical meaning?** Whereas ***morbus sacer***, literally translated as "holy disease," was rather used as an euphemistic pseudonym for epilepsy, the co-occurrences of ***morbus comitialis*** instead hypothesize a medical disease, e.g., shown by *remedia* (medicine), *medici* (doctor), and *medentes* (curing).

- **How was the overall knowledge about the disease at that time?** The co-occurrences for both terms, e.g., *caput* (head), *insania* (insanity), *animi* (mind), *corporis* (body), *nervorum* (nerves), *mortis* (death), and *abit* (died), indicate that epilepsy was seen as a potentially lethal insanity with physical symptoms.

When examining the last question, the scholar discovered that Maurus Servius Honoratus, a popular grammarian in the 5th century, mistakenly conceived epilepsy as a feverish disease (see Figure 4.10, bottom), shown by co-occurring terms of ***morbus sacer***: *ignis* (fire), *carbunculi* (burning ulcer), and *papulae* (blister). As Pliny the Elder already ascertained in the first century that fever is not a symptom of epilepsy, the

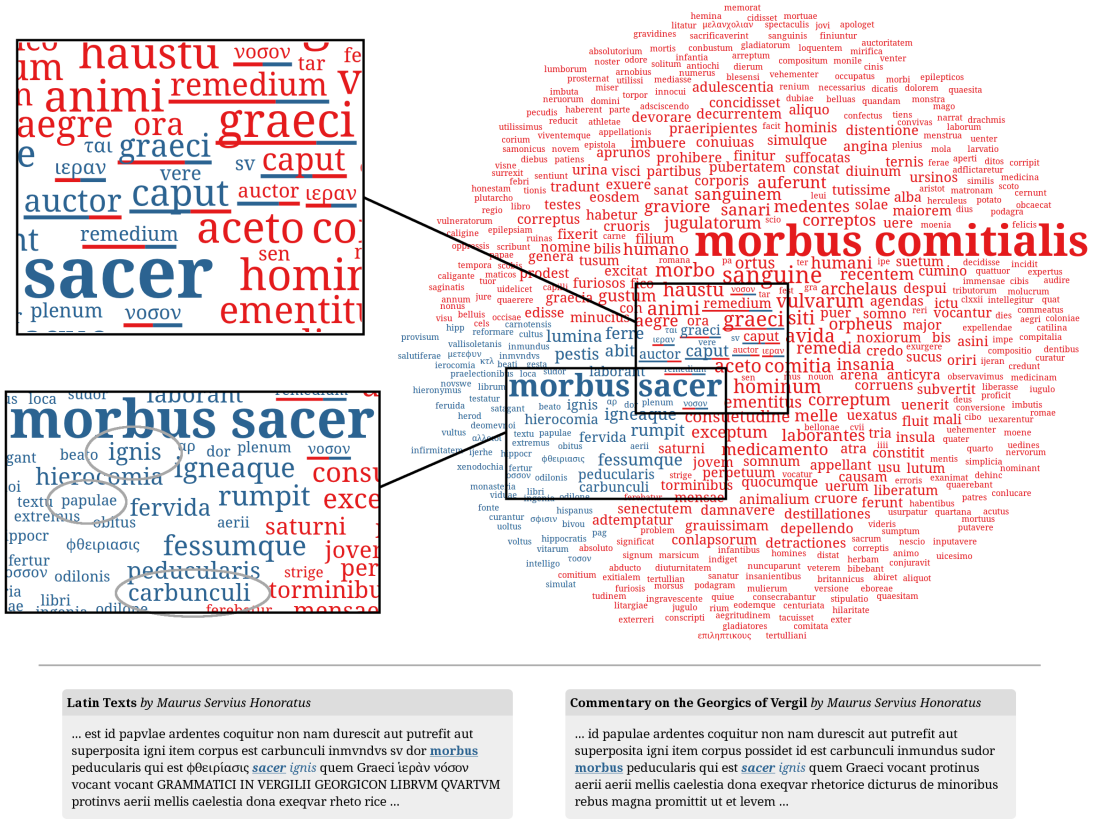


Figure 4.10: Comparing the co-occurrences of ***morbus comitialis*** and ***morbus sacer***.

humanities scholar denoted this discovery as not intuitive, so that it would have never been found using traditional methods.

III. Exploring τεχνη, υγιεια and νοσος

The third example investigates the meaning of *art* in antiquity, which is hard to describe nowadays. The idea at that time was that art can be taught as it includes knowledge. Therefore, art is related to many fields in Ancient Greek texts [All99]. Expectedly, these fields are visible in TagPies (see Figure 4.11) as co-occurrences of the Ancient Greek term for art, ***τεχνη***: *φυσικη* (natural science), *μαντικη* (art of prophecy), *γραμματικη* (grammar), *ανθρωπινη* (human art), *ρητορικη* (rhetoric), *ποιητικη* (poetics), *ιατρικη* (medicine), *μαγικη* (magic), *διαλεκτικη* (dialectic), etc.

The analysis of ***τεχνη*** in comparison to terms from one of these related fields was

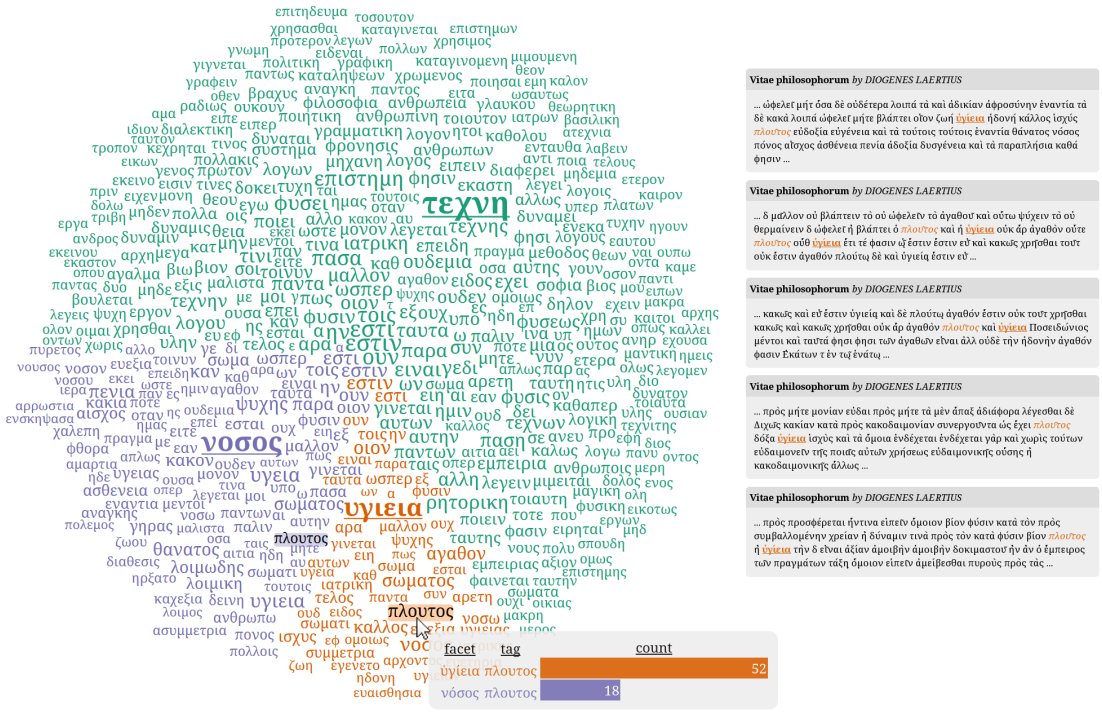


Figure 4.11: Comparing ***τεχνη***, ***υγιεια*** and ***νοσος***.

of particular interest. The art of physicians in the field *medicine* is visualized in two further sections of TagPies showing co-occurrences for the Ancient Greek terms for health (***υγιεια***) and disease (***νοσος***). In contrast to the diverse terms surrounding ***τεχνη***, the co-occurrences here are closely related to their main terms. Both terms co-occur with parts of the body, e.g., *σωμα* (body) and *ψυχη* (breath, life). Furthermore, ***υγιεια*** is related to positive terms like *καλλος* (beauty), *ισχυς* (strength) or *ηδονη* (enjoyment), whereas ***νοσος*** occurs together with rather negative terms like *λοιμικη* (plague), *ασθενεια* (weakness), *γηρας* (senility) or *θανατος* (death). Also, one of the known reasons of diseases, poverty (*πενια*), co-occurs 71 times.

Beyond that, the TagPies show unexpected relationships, e.g., the frequent usage of *πλουτος* (wealth) in connection with ***υγιεια*** (52 times). Looking at the references, five text passages in the biography of Zeno of Elea (*Vitae philosophorum*, written by Diogenes Laertius) are listed among others. Within this text, various things are denoted as *good* or *bad*. Thereby, Zeno of Elea categorizes neither health nor wealth as good, since both terms can be used also in a negative context. This example reveals

another aspect of ***υγιεια*** in a philosophical rather than a medical context. The humanities scholar working with this example expected a correlation between (medical) art and wealth as a consequence of the medical profession after the 5th century BC in the context of ***τεχνη***, but πλουτος does not appear as a significant co-occurrence.

IV. Comparing Movie Genres

To outline TagPies applicability to other domains and to test our proposed method for a larger number of data facets compared to the digital humanities examples, we generated a dataset containing characteristic tags of fourteen movie genres in the Internet Movie Database (IMDb). For this purpose, we extracted the words from all movie titles and determined a list of tags per genre by decreasing frequency. After removing stopwords for main languages, we chose the top tags for each genre dependent on the relative frequency of the genre compared to the others. The result for a total of 750 tags is shown in Figure 4.12. Especially for this use case, we can see the utility of the tag ordering before processing the layout. Although fourteen data facets are drawn, the tag cloud looks uniform and the various genres can be easily explored.

4.2.4 Case Study 2: TagPies

The purpose of the second case study was to evaluate the utility of TagPies for supporting humanities scholars in answering their research questions. The study was exactly designed and conducted as the first one. Additionally, we asked the participants for their two favorite visualizations, the ones they would prefer to work with regarding intelligibility, utility, design and aesthetics. Each participant could choose two approaches, ranked first and second, including those from the first case study. Again, the humanities scholars should justify their decision. The preferences of the humanities scholars are listed in Table 4.1.

1.	Small Multiples	Basic	Bars	Bars
2.	Italics/Bold	Bold/Italics	Bold/Italics	Merged Black
1.	Italics/Bold	Bold/Italics	Merged Black	Italics/Bold
2.	Bars	Merged Black	Bold/Italics	Bars

Table 4.1: Preferred design variants.

The overall perception of the TagPies design was very positive compared to the state-of-the-art methods. Only one scholar plans to use small multiples for her work

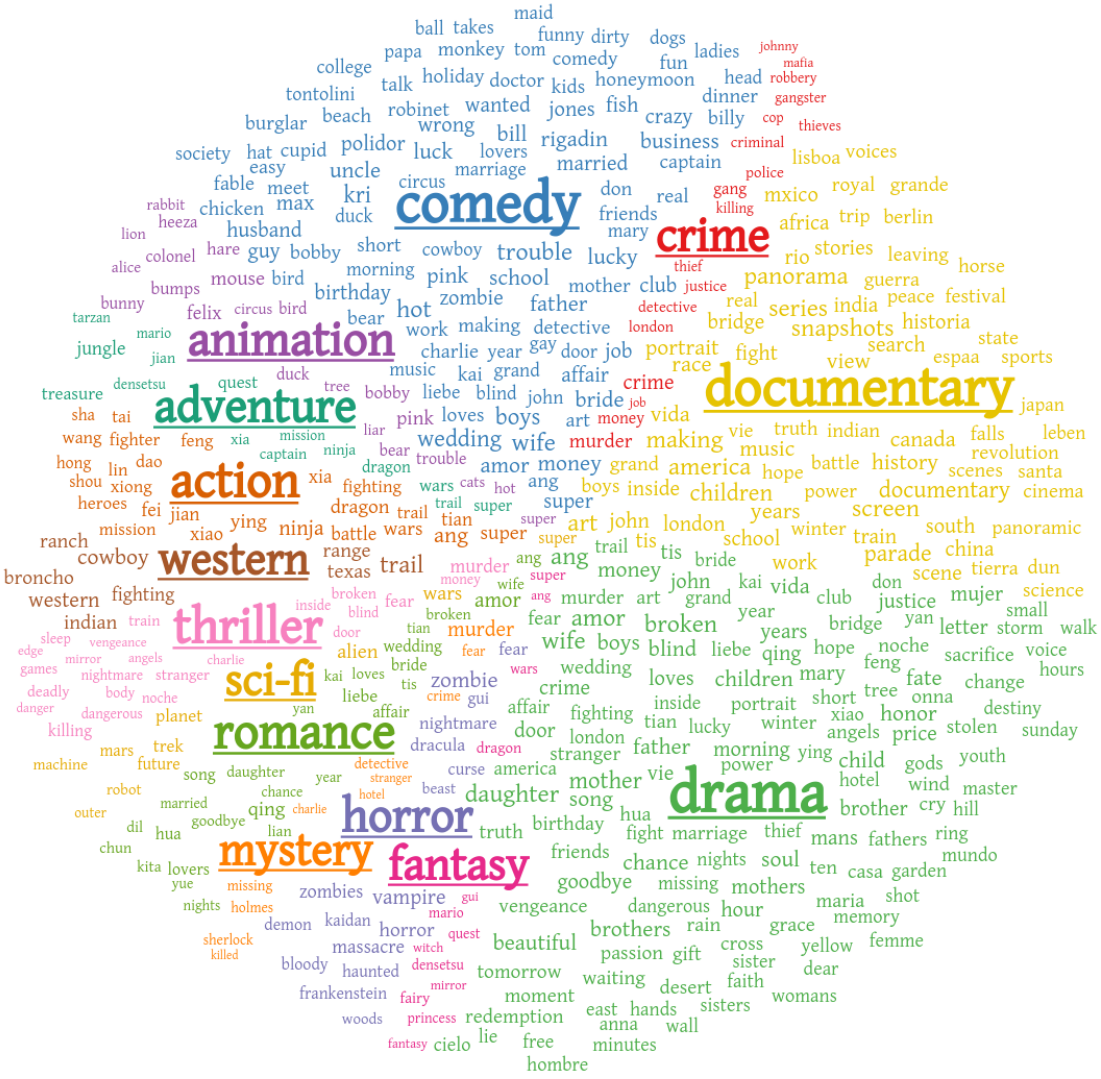


Figure 4.12: Characteristic tags of fourteen IMDb genres.

given “aesthetics and the best overview” as reasons for her decision. The other seven participants preferred TagPies without determining a clear favorite. But regarding the major design decisions (see Table 4.2), we could draw significant conclusions.

	1. choice		**2. choice**	
	bars	no bars	bars	no bars
multiple tags	2x	4x	2x	4x
merged tags	1x	N/A	2x	N/A

Table 4.2: Preferred TagPies features.

Most participants preferred TagPies visualizing multiple over merged tags giving reasons like “I can easily find all co-occurrences for one data facet!” and “The discovery of shared tags is simple as they are all placed in the center of the cloud!” Still, three participants also like to work with merged tags, but only with the Merged Black variant. As nobody preferred the Merged variant, coloring merged tags neutrally seems to work best in such a design. Also, one of the participants preferring Merged Black mentioned: “This design only works well for two categories [then, a merged tag always belongs to both data facets], for more categories it gets confusing!”

We could not spot a tendency regarding the highlighting of shared tags. Some participants preferred the variants with bars, probably due to the already learned metaphor in the first case study but also justified with “I can see similarities between categories with detailed information at first glance!” Although slightly distorting the perception of the tags’ font sizes, many participants preferred variants with bold and italics styles. Important for them was primarily the immediate visual separation between shared and unique tags. Participants favoring design variants without bars rather tended to use the provided interaction functionalities; e.g., the participant preferring the Basic variant stated: “The interactive tool provides all information I need on demand!”

Overall, the humanities scholars liked the comprehensible and aesthetic design. Especially the pie chart style was perceived as a suitable metaphor. In dependency on individual taste and research question, almost each TagPies design variant was valuable for any of the scholars. Already during the study, hypotheses could be generated by discovering interesting patterns. One possible reason for preferring different layout variants is that different tasks were performed and different data sets were used in the case study.

4.2.5 Limitations

Our main objective was to generate uniform looking tag clouds. Therefore, the spiral starts in all cases from the tag cloud origin when determining a tag's position. Especially if TagPies display many data facets and contain many shared tags, adjacent placement of shared tags cannot be guaranteed. Then, the discovery of shared tags placed far apart requires the usage of the provided interaction functionality. We experimented with moving the spiral origin to already placed related tags or to borders between sectors that share tags, but these approaches destroyed the intended unity in many cases.

Also, due to the nature of the Wordle based approach and the restriction of placing tags only in specific preassigned sectors, sometimes the tag clouds contain little holes. But the collaborating humanities scholars praised the good readability of the words – a positive side effect of this issue.

The proposed design variants are only a small number of possibilities from a potentially large design space (e.g., varying background color, orientation of tags). According to the suggestions made by the participants and our impressions gained during the first case study, we chose those variants that we considered to support the humanities scholars best.

Although the humanities scholars usually compare a limited number of data facets (< 6), we also tried examples with more than ten data facets to assess the scalability of our approach (an example is shown in Figure 4.12). Even though the results were satisfactory, TagPies produce best layouts for fewer data facets. For a high number of facets, pie sectors can be very small, so that tags are hard to position. Furthermore, qualitative color maps are then hard to define.

Sometimes, humanities scholars are interested in rare cases. TagPies aim to visualize the most significant co-occurrences of the given search terms. The more occurrences of a search term exist, the more co-occurrences need to be displayed. Then, rare but potentially interesting cases may be not shown due to the limited number of tags positioned in TagPies.

4.3 TagSpheres

The central idea of TagSpheres is the visualization of hierarchies in textual summaries. Although designed for humanities scholars to analyze the clause functions of an ancient term's co-occurrences, we designed TagSpheres in a way that various types of text hierarchies can be visualized in an intuitive, comprehensible manner. To emphasize the wide applicability of TagSpheres, we list several examples from digital humanities, sports and aviation (see Section 4.3.2). An overview of the characteristics of these examples is given in Table 4.3.

domain	I. digital humanities	II. sports	III. aviation
task	analyzing the clause functions of the co-occurrences of a search term T	comparing the performances of teams in championships	observing all direct flights from an airport or a city
H_1	search term T	best performing teams	departure airport/city
$H_2, \ldots, H_n$	co-occurrences in dependency on the word distance to T	teams grouped by decreasing performance	direct federal (H_2), continental (H_3) and worldwide flights (H_4)
n	4	6, 8	2..4
$w(t)$	number of (co-)occurrences of t	number of a team's appearances	inverse distance weighting between departure and arrival airports/cities
$p(t)$	equally labeled tag of a higher hierarchy level	same team if already placed on a higher hierarchy level	previously placed tag of the same country/continent
strong tag relations	equally labeled tags	same teams if placed on multiple hierarchy levels	departure/arrival airports/cities
weak tag relations	spelling variants	N/A	airports/cities of the same country/continent

Table 4.3: Characteristics of TagSpheres usage scenarios.

4.3.1 Designing TagSpheres

Given n hierarchy levels $H_1, \ldots, H_n$, the top hierarchy level H_1 contains tags representing the focus of interest of a usage scenario. All other tags are divided into $n-1$ groups in dependency on their hierarchical distance according to the observed topic, or to the tags on H_1. Each tag t in TagSpheres has a weight $w(t)$ reflecting its importance, and an optional predecessor tag $p(t)$ representing a relationship to another tag that was placed before t and usually belongs to a higher hierarchy level. In dependency on the observed topic, it might be necessary to place the same tag on several hierarchy levels to encode the change of a tag's importance among hierarchies. In such cases, predecessor tags help to visually link these tags.

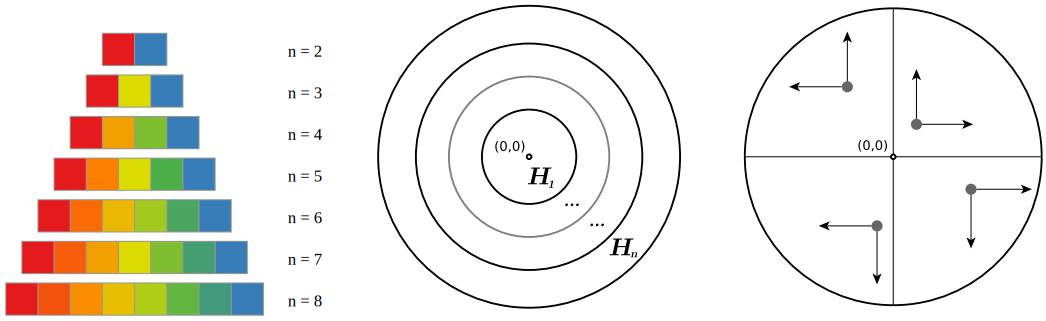
I. Design Decisions

Like used for TagPies, we also use well-established design features for tag clouds when designing TagSpheres. This includes font size [BGN08] to encode the weight $w(t)$ of a tag and the avoidance of rotations to keep the layout easily explorable [WSK+13]. Again, we use color to distinguish facets, in this case for tags belonging to different hierarchy levels. As TagSpheres encode the distance to a given topic, the usage of a categorical color map is inappropriate. Unfortunately, suitable sequential color maps as provided by the ColorBrewer [HB03] produce less distinctive colors even for a small number of hierarchy levels, so that adjacent tags belonging to different hierarchy levels are hard to classify. Following the suggestions given by Ware [War13], we defined a divergent cold-hot color map using red for the first hierarchy level and blue for tags belonging to the last hierarchy level n. To avoid uneven visual attraction of tags, we only chose saturated colors that are in contrast to the white background. Example color maps for up to eight hierarchy levels are shown in Figure 4.13a.

II. Layout Algorithm

In preparation, the tags are sorted by increasing hierarchy level, so that all tags within the same hierarchical distance to H_1 are placed successively. The tags of each hierarchy level are ordered by decreasing weight to ensure that important tags are circularly well distributed.

To avoid large whitespaces, a problem addressed by Seifert [SKK+08], our method follows the idea of the Wordle algorithm [VWF09] – permitting overlapping tag bounding boxes if the tags' letters do not occlude – to determine the positions of tags. So, we



(a) Resultant color maps for $n = 2, \ldots, 8$ hierarchy levels.

(b) Using spheres for the tags of different hierarchy levels.

(c) Vectors for occlusion check.

Figure 4.13: TagSpheres layout algorithm details.

obtain compact, uniformly looking tag clouds for the underlying hierarchical, textual data. To ensure well readable tag clouds, we use a minimal padding between letters of different tags.

As shown in Figure 4.13b, we aim to visually compose tags of the same hierarchy level in the form of spheres around the tag cloud origin at $(0, 0)$. Initially, we iteratively determine positions for the tags of H_1 in the central sphere using an Archimedean spiral originating from $(0, 0)$. An example is given in Figure 4.14a. For each tag t of the remaining hierarchy levels $H_2, \ldots, H_n$, we also use $(0, 0)$ as spiral origin, if $p(t)$ is not provided (see Figure 4.14b). If $p(t)$ is defined, we use the predecessor's position as spiral origin (see Figure 4.14c). As a consequence, hierarchically related tags are placed closely and visually compose in the form of rays originating from $(0, 0)$ as shown in Figure 4.17. In contrast to other spiral based tag cloud algorithms, we avoid to cover whitespaces with tags of hierarchy level H_i within spheres of already processed hierarchy levels $H_1, \ldots, H_{i-1}$. Dependent on the quadrant in the plane, in which a tag shall be placed, we search for already placed tags intersecting two vectors originating from the dedicated position as illustrated in Figure 4.13c. If no intersections between the vectors and tags of $H_1, \ldots, H_{i-1}$ are found, we place the tag. This approach coheres all tags of a hierarchy level as a visual unity outside the inner bounds of the previously processed hierarchy levels' spheres.

(a) Placing all tags of hierarchy level H_1.

(b) Placing a tag without predecessor.

(c) Placing a tag with predecessor.

Figure 4.14: Determining tag positions using an Archimedean spiral.

III. Interactive Design

Implemented as an open source JavaScript library, TagSpheres can be dynamically embedded into web-based applications. With mouse interaction, we enable the user to detect hierarchically related tags quickly. Thereby, we distinguish between strongly and weakly related tags, which are defined in dependency on the underlying usage scenario (see Table 4.3). Related tags are shown on mouseover (see Figure 4.15). For strongly related tags we use a black font on transparent backgrounds having the hierarchy level's assigned color. In contrast, weakly related tags retain their saturated font color, but gray, transparent backgrounds indicate relationships.

TagSpheres provide a configurable tooltip displayed when hovering or clicking a tag to be used, e.g., to list all related tags and their weights. The mouse click function can be used for displaying additional information, e.g., to link to external sources, or to show text passages containing the chosen tag.

4.3.2 Usage Scenarios

TagSpheres are applicable whenever statistics of unstructured text shall be visualized in the form of a tag cloud and a decent hierarchy among the tags exists. In the following, we illustrate usage scenarios of TagSpheres for text-based data from three different domains: digital humanities, sports and aviation.

I. Digital Humanities Scenario

Within the digital humanities project *eXChange*, humanities scholars use TagSpheres to analyze and classify a term's co-occurrences according to their clause functions. For this purpose, the scholars required four-level TagSpheres displaying the following tags:

H_1: search term T,

H_2: co-occurrences of T with word distance 1,

H_3: co-occurrences of T with word distance 2, and

H_4: co-occurrences of T with word distance 3 up to word distance m.

The font size of T on level H_1 encodes how frequent the search term occurs in the underlying text corpus; the font sizes of all other terms reflect their number of co-occurrences with T in dependency on the corresponding distance. On H_4, font sizes are normalized in relation to the distance range $m - 2$. A tag on hierarchy level H_i receives a predecessor tag if the corresponding term occurs on one of the previous layers $H_{i-1}, \ldots, H_1$.

A use case provided by one of the humanities scholars involved in the *eXChange* project shall illustrate the utility of TagSpheres to support the classification of a term's co-occurrences by their clause functions. Analyzing the co-occurrences of ***morbo*** (disease), terms in similar relationships to the given topic were discovered and classified (see Figure 4.15). In large distances, the humanities scholar found objects in the form of affected parts of the body, e.g., head (*caput*), soul (*animo*) and limbs (*membrorum*), affected persons, e.g., son (*filius*), woman (*mulier*) and king (*rex*), and related places, e.g., Rome (*romam*), church (*ecclesia*) and villa (*villa*). Closer to *morbo* (most often with distance 1 or 2), typical attributes and predicates can be found. Whereas attributes describe the type or intensity of the disease, e.g., pestilential (*pestifero*), heavy (*gravi*), deadly (*exitiali*) and acute (*acuto*), the occurring predicates illustrate the disease's progress, e.g., seize (*correptus*), disappear (*periit*) and worsening (*ingravescente*). Adjacent to *morbo*, specific terms for "moral" diseases, e.g., greediness (*avaritiae*), arrogance (*superbiae*) and lust (*concupiscentiae*), and actual diseases like jaundice ([morbo] *regio*), leprosy (*leprae*) and two common names for epilepsy (*[morbo] comitiali, [morbo] sacro*) occur.

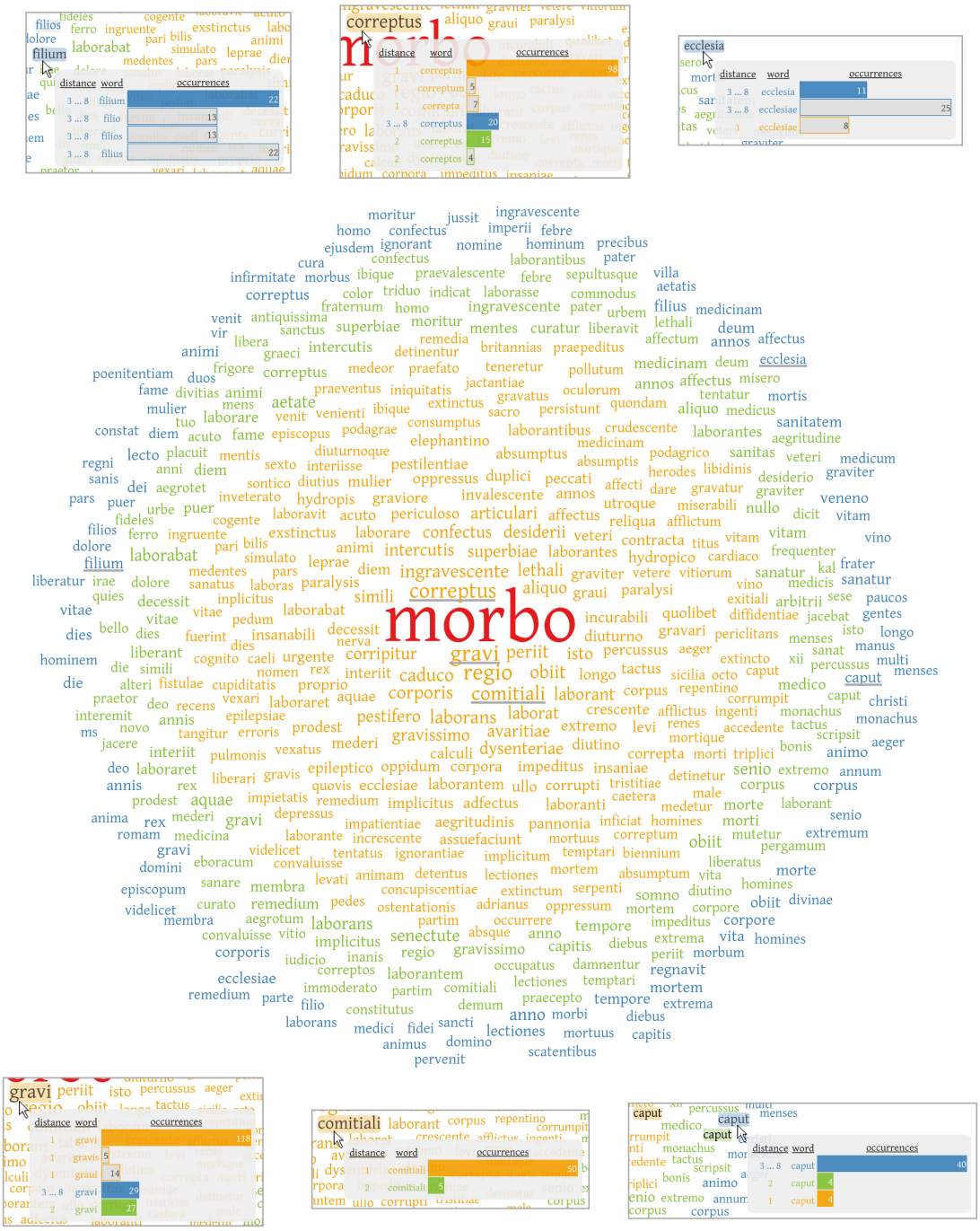


Figure 4.15: The analysis of co-occurrences of the Latin term *morbo* (disease) on word distance.

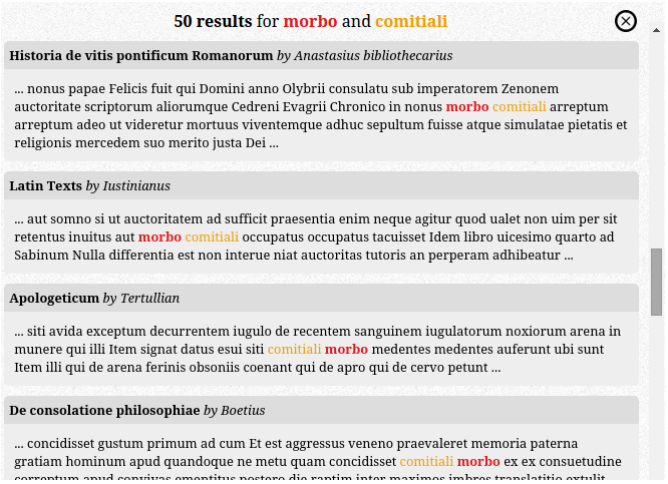


Figure 4.16: Close reading of text passages containing *morbo* and *comitiali* with word distance 1.

In this usage scenario, the interaction capabilities of TagSpheres are tailored according to the needs of the humanities scholars. Hovering a tag opens a tooltip showing the term's number of occurrences on all hierarchy levels as strongly related tags. Additionally, variant spellings or cases of the term are listed with their corresponding frequencies as weakly related tags to support the analysis process. An important requirement for the humanities scholars was the ability to close read the texts in order to discover potentially interesting passages. TagSpheres support close reading by clicking a tag, which displays the corresponding text passages containing the search term and the clicked term with the chosen distance. An example for text passages containing the adjacent terms *morbo* and *comitiali* is shown in Figure 4.16.

II. Championship Performances

This scenario illustrates how TagSpheres can be used to comparatively visualize performances in championships. Therefore, we processed a dataset containing the results of all national teams ever qualified for the FIFA World Cup. We receive the following six-level hierarchy:

H_1: FIFA World Champions,

H_2: second placed national teams,

H_3: national teams knocked out in the semifinal,

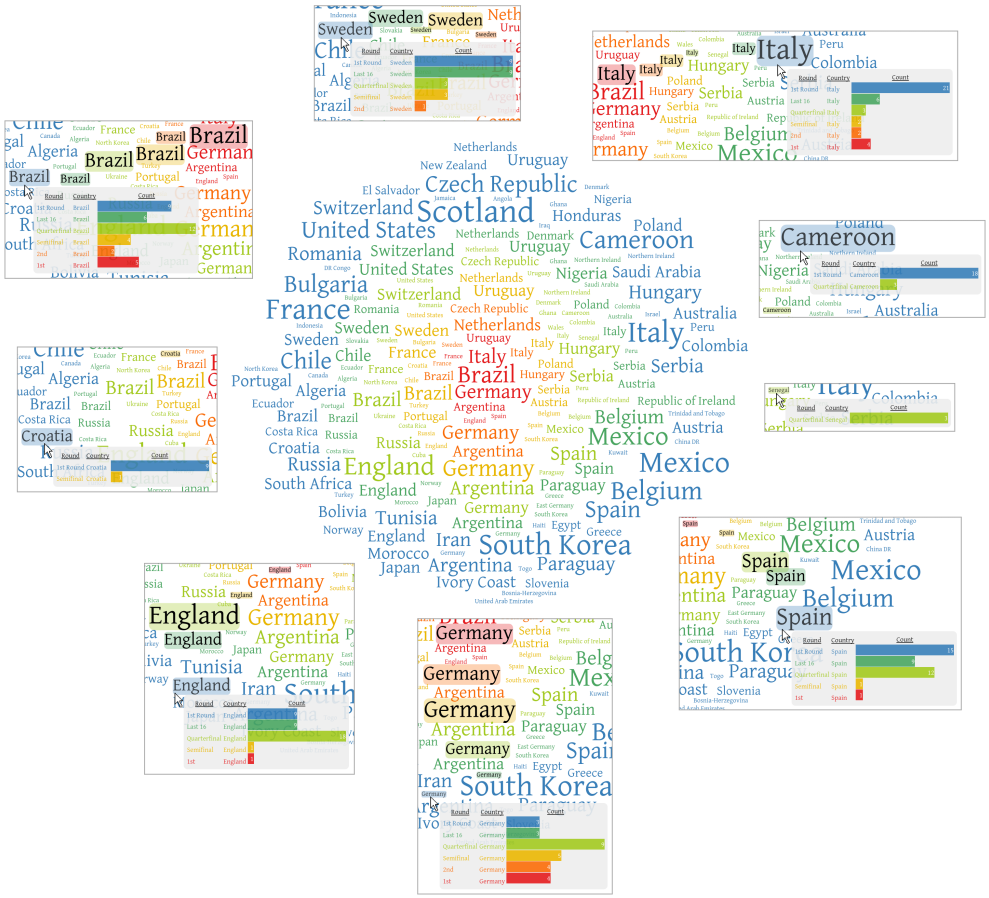


Figure 4.17: Performances of all nations qualified for the FIFA World Cup.

H_4: national teams knocked out in the quarterfinal,

H_5: national teams knocked out in the second round (second group stage or last 16), and

H_6: national teams knocked out in the (first) group stage.

The nations' names are used as tags and font size encodes how often a national team partook a championship round *without* reaching the next level. Therefore, most nations occur on various hierarchy levels. If a tag t for a nation to be placed on H_i was already placed at a higher hierarchy level $H_{i-1}, \ldots, H_1$, we use the corresponding tag as predecessor $p(t)$.

Figure 4.17 shows the resultant TagSpheres. Especially this scenario illustrates the benefit of using the positions of predecessor tags as spiral origins for successor tags. Mostly, multiple tags of a nation are closely positioned. Hovering a tag displays the all-time performance of a national team for all championship rounds in a tooltip. Expectedly, *Brazil* and *Germany* achieved very good results, especially in the last rounds. In contrast, *Italy* was often knocked out in the first round, but in case of reaching the semifinal (8x), *Italy* often became FIFA World Champion (4x). *England* and *Spain* show nearly equal performances. With the same number of appearances (38x), both nations reached the semifinal only twice. Few nations have a 100% success rate in the group stage. Qualified three times for the FIFA World Cup, *Senegal* always reached the quarterfinals. Most nations, e.g., *Sweden* and *Cameroon*, show the expected pattern "the higher the championship round, the lower the number of appearances."

Another example that illustrates the success of football clubs ever played in England's first league is given in Figure 4.18. Here, we use the average rank at the end of the seasons to cluster 68 clubs into eight hierarchy levels, and font size encodes the number of appearances.

Figure 4.18: Performances of England's first league football clubs from 1888/89 -- 2014/15.

III. Airport Connectivity

To analyze the federal, continental and worldwide connectivity of airports, we derived a dataset from the OpenFlights database,[7] which provides a list of direct flight connections between around 3,200 airports worldwide. With the selected departure airport d (or city) on H_1, all other airports (or cities) reachable with a non-stop flight cluster into three further hierarchy levels:

H_2: airports/cities in the same country as d,

H_3: airports/cities on the same continent as d, and

H_4: all other reachable worldwide airports/cities.

As tags we chose either airport names, the provided IATA codes,[8] or the corresponding city names. In this scenario, font size encodes the inverse geographical distance between the departure airport $d = \{lat_d, lon_d\}$ and an arrival airport $a = \{lat_a, lon_a\}$. To keep the deviation to the actual distance as small as possible, we apply the great circle distance G [Hea03], defined by

$$G = 6378 \cdot \arccos\Big(\sin(lat_d) \cdot \sin(lat_a) + \cos(lat_d) \cdot \cos(lat_a) \cdot \cos(lon_d - lon_a)\Big).$$

Predecessor tags are used to place airports or cities of the same country or continent closely. For a tag t to be placed on H_3, we choose the first placed tag with the same associated country as predecessor, if existent; for H_4, we choose the first placed tag with the same associated continent. Thus, a predecessor tag $p(t)$ in this scenario always belongs to the same hierarchy level as t.

Figure 4.19 shows TagSpheres for non-stop flights from various airports or cities. All examples show that airports/cities of the same countries/continents are placed closely in clusters. For Sydney, no tags are placed on H_3, and for Cagliari, no connections to airports outside Europe exist. When the user hovers a tag, the corresponding connection and the travel distance are shown in a tooltip. Clicking a tag redirects to Google Flights[9] listing possible flight connections.

[7] http://openflights.org/data.html
[8] http://www.iata.org/services/pages/codes.aspx
[9] https://www.google.com/flights/

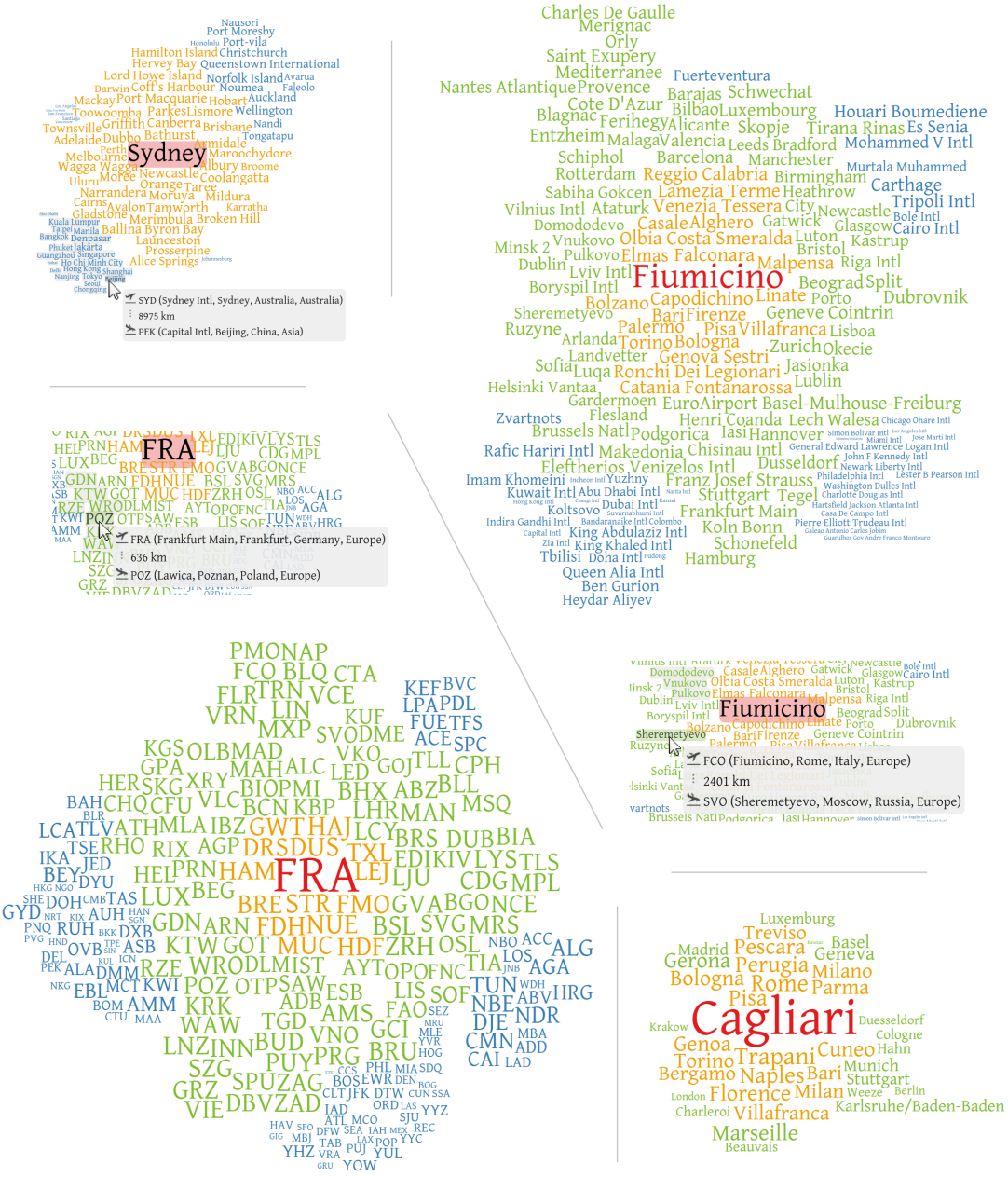


Figure 4.19: Direct flight connections from airports in Sydney, Rome, Frankfurt and Cagliari.

4.3.3 Discussion

The original motivation to design TagSpheres was to support humanities scholars in analyzing the clause functions of a search terms' co-occurrences. Some aspects of evaluating the design during the corresponding *eXChange* project are outlined below. Furthermore, we discuss general limitations of TagSpheres.

I. Evaluation

To ensure creating a valuable interface for our targeted user group, we closely collaborated with the humanities scholars in the design phase aiming to transform their notion of hierarchical distance as appropriate as possible. This included project workshops and regular meetings, where we demonstrated current prototypes, and the humanities scholars were able to suggest their ideas on the design, the interactivity and the embedding of TagSpheres into their research environment. Finally, we conducted a small evaluation with seven humanities scholars (five female, two male) – five of them were members of the *eXChange* project. Due to that small number of participants, diversified research interests and the exploratory nature of the humanities scholars' tasks, a formal user study with performance data was not viable. To encourage the participants to intensely work with TagSpheres, we allowed them to query the database with terms of their own interest (preference data). In a questionnaire, we asked the humanities scholars for subjective ratings on several aspects concerning TagSpheres. They needed to choose a value on a Likert scale from 1 (very bad) to 7 (very good), and we also asked them to justify their decisions. The results are shown in Figure 4.20.

During the case studies when developing TagPies (see Section 4.2), we found out that the aesthetics of tag clouds plays an important role for humanities scholars. The aesthetics of TagSpheres was generally perceived as good. Very important for us were the opinions of humanities scholars if our design would intuitively transmit their notion of hierarchical distance. Only two scholars were undecided, but four scholars gave the best rate stating that TagSpheres are *"easily understandable."* Especially, the chosen colors *"clearly visualize the word distance between co-occurrences and the search term."* As the tags are shown in different colors and varying font size, we further asked for the readability of tags, which was mostly justified positively. Although the humanities scholars stated that *"all important co-occurrences of the search term are visible at first glance,"* it was hard for them to detect (often closely positioned) simi-

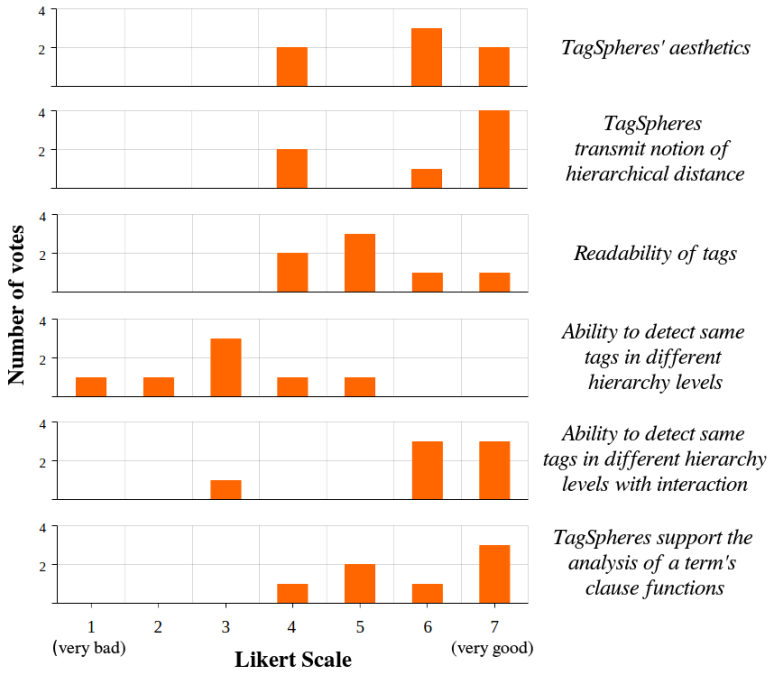


Figure 4.20: TagSpheres questionnaire results.

lar or related terms on different hierarchy levels. But all participants stated that the provided means of interaction facilitate this task and overall foster the understanding of the visualization and the explorative analysis of results. Finally, the utility of TagSpheres to support the humanities scholars in examining research questions regarding the clause functions of a search term's co-occurrences was also rated as good.

II. Limitations

The main objective of TagSpheres is to combine a hierarchical information of textual data with the aesthetics of tag clouds. In contrast to the usual layout algorithm to always initialize an Archimedean spiral at the tag cloud origin $(0, 0)$ when determining the position of a tag, the usage of predecessor tags as spiral origins slightly affects the uniform appearance of the result in some cases (e.g., see Figure 4.19). Occasionally, little holes occur, and – at the expense of visualizing the hierarchical structure of the underlying data – the tag cloud boundaries get distorted.

The proposed hot-cold color map used to visually convey hierarchical distance generates well distinguishable colors when the number of hierarchy levels is small. For

a larger number of hierarchies as displayed in Figure 4.18, closely positioned tags of different levels may become visually indistinct, especially when only few tags belong to a certain level.

The current TagSpheres design does not take the distribution of tags throughout different hierarchy levels into account. In use cases with a steadily increasing or decreasing number of tags per hierarchy level it gets possible that a considerable proportion of the color map's bandwidth is used for a comparatively small portion of tags. An assignment of colors taking the density distribution of the tags' weights into account could overcome this issue.

4.4 Summary

In close collaboration with humanities scholars from the *eXChange* project, we developed two novel tag cloud designs to support research tasks in digital humanities.

TagPies arrange the tags of multiple data facets in a pie chart manner to support humanities scholars in interpreting numerous results of multiple keyword search queries on large databases containing ancient texts. TagPies show the co-occurrences of the searched terms and facilitate the comparison of the contexts, in which these terms were used. Thereby, TagPies aim to direct the humanities scholar to potentially unknown but interesting text passages. We provide several TagPies design variants according to the preferences of the scholars. The definition of this design space as well as a comprehensive user study are future tasks. An example comparing typical tags of movie titles for different genres illustrates the applicability of TagPies to other domains.

TagSpheres arrange tags on several hierarchy levels to transmit the notion of hierarchical distance in tag clouds. We accentuate relationships between different hierarchy levels by placing hierarchically related tags closely. In *eXChange*, TagSpheres are used to explore the clause functions of the co-occurrences of a selected term. The design of TagSpheres was evaluated as aesthetic and intuitive, and the humanities scholars emphasized the utility of TagSpheres for their work. Further usage scenarios in sports and aviation outline the inherence of hierarchical textual information in various domains and the usefulness of TagSpheres as they provide an interesting view on this type of data.

Invaluable during the design processes of TagPies and TagSpheres were regular meetings to determine the needs of the humanities scholars and to provide proto-

types. The results reflect a beneficial collaboration between computer scientists and humanities scholars for both fields. On the one hand, we were able to construct novel layout algorithms for tag clouds capable of visualizing faceted text summaries. On the other hand, the resultant visualizations turned out to be powerful tools within the workflows of the humanities scholars who get a much more intuitive and dynamic access to search results in comparison to working with traditional result lists.

In quoting others, we cite ourselves.

Julio Cortázar

5 Visualization of Text Re-use

Text Re-use is defined as the oral or the written reproduction of textual content and is roughly divided into two categories [Büc13]. On the one hand, a text passage is re-used deliberately, like direct quotes and phrases like winged words and wisdom sayings. Translations of a text into other languages also count to this category and are called interlingual Text Re-use. A very popular form of deliberate Text Re-use is plagiarism, which has gained major attention in the recent years, mainly driven by plagiarism allegations in politics. On the other hand, Text Re-use may be unintended, like boilerplates, e-mail headers or the repetition of news agency texts when writing daily newspapers [CGPW02]. Further examples are idioms, battle cries and multi word units.

The analysis of Text Re-use among historic texts with the goal to explore known and discover unforeseen relationships in cultural heritage has become an important task within various digital humanities projects (see Section 5.1). Thereby, the humanities scholars are interested, which texts share patterns of consecutive similar units (systematic Text Re-use) and how frequent specific phrases occur and in which contexts they were used (repetitive Text Re-use).

This chapter shows how interactive visualizations help humanities scholars in understanding and interpreting Text Re-use occurrences. In particular, we present two visualizations that support the close and the distant reading of Text Re-use:

- **Text Re-use Grid**: A chart that juxtaposes all texts of a collection in relation to amount and type of Text Re-use (distant reading).
- **Text Re-use Browser**: A user interface consistent of an interactive Dot Plot View and a Text Re-use Reader that allows for the inspection and browsing through all Text Re-uses between two texts (close reading).

Various usage scenarios and experiences collected in digital humanities projects outline the benefit of the proposed visualizations for humanities scholars.

5.1 Related Work

The discovery of relationships between different texts and the alignment and visualization of the findings has been a challenging task in various works. Xanadu, founded in 1960, can be seen as one of the pioneer projects that attend to this matter [Nel99]. The current prototype shows the dedicated text in the center of the screen, related texts are positioned on both sides, and shared patterns are aligned and highlighted using various colors. John et al. [JHMK14] propose a focus and context approach for the visualization of texts sharing similar patterns. A vertical ribbon for each text shows the distribution of these patterns, and interactively, the user can drill down to regions of interest. Cheesman offers a visualization for the alignment of multilingual text passages in Shakespeare's Othello, where the user can interactively browse through the texts of two editions [CFRT14]. Likewise, related text entities are visually linked to each other. To illustrate computationally determined Text Re-uses in ancient Greek texts, Büchler suggests a graph to show the results for a certain author by number, citing authors, years of citing authors and passages of the book [BGEH10]. Additionally, the user can inspect individual text snippets with highlighted re-used passages. For plotting Text Re-use between Bible books [Lee07], Lee uses a static Dot Plot View, which was originally designed for bioinformatics to compare two genome sequences to each other [GM70]. A single dot marks a correlation between the genomes and multiple dots form patterns that indicate similar genomic segments. Lee utilizes this approach to highlight patterns of systematic Text Re-use. Various visualization methods highlight plagiarized passages of a given source text [RA00, Gut13]. A complete overview of the whole text is given and each page, chapter or plagiarized text passage receives its own block. Coloring is used to show the amount of re-used text or to indicate potential sources. Riehmann provides an interactive interface that supports the

analysis of alleged plagiarism cases by aligning plagiarized passages to their potential sources [RPSF15]. An alignment can be explored in detail with the help of so called difflines and a close reading view.

All the above visualization techniques focus on displaying relationships between a limited number of texts. Mostly, a certain source text is given and its correlations to other texts can be analyzed. A comparative overview between all texts of an arbitrary text collection is not provided. For this purpose, we propose the *Text Re-use Grid* as a distant reading solution to explore Text Re-use within a corpus (Section 5.3). In addition, we present the *Text Re-use Browser* that supports close reading of the Text Re-use between two selected texts (Section 5.4).

5.2 Theoretical Basis of Text Re-use

Let $A_1, \ldots, A_n$ denote a corpus of n texts. After splitting each text into units (e.g. sentences), the Text Re-uses between all text unit tuples are determined. Each detected Text Re-use $\{a_i, b_j\}$ consists of two corresponding Text Re-use units a_i (e.g. i-th sentence of text A) and b_j (e.g. j-th sentence of text B). The *Scoring value* $t(a_i, b_j)$ defines a weight for $\{a_i, b_j\}$ dependent on the text unit lengths of a_i and b_j and their *Re-use Overlap*, which is the proportion of matching to non-matching tokens. t is ranged in the interval $[0, 1]$; 0 means no similarity between the two units, 1 means that a_i and b_j are equal. The complete Text Re-use result list contains only relevant Text Re-uses above a certain threshold for t. A more detailed description of the underlying algorithms for Text Re-use detection and the computation of t can be found in Büchler's dissertation [Büc13].

Researchers working with Text Re-use have various research questions that require the definition of the following Text Re-use types:

Systematic Text Re-use. The consecutive occurrence of the same Text Re-use pattern is of particular interest for researchers when comparing different texts to each other. Such type of Text Re-use could be an indication for plagiarism. For instance, the pattern $\{a_i, b_j\}, \{a_{i+1}, b_{j+1}\}, \{a_{i+2}, b_{j+2}\}$ is a *Systematic Text Re-use* of three consecutive text units.

Repetitive Text Re-use. This type of Text Re-use appears when the researcher is interested in analyzing a text unit that is frequently used in a certain text. The goals in this use case are to explore the contexts, in which a text unit appears as well as to

what extent a specific text unit is spread in the text. *Repetitive Text Re-use* for a text unit a exists for a set of Text Re-use pairs in the form $\{a, b_1\}, \{a, b_2\}, \{a, b_3\}, \ldots$

Isolated Text Re-use. We classify a Text Re-use $\{a_i, b_j\}$ as isolated if it does not occur within a certain pattern, more precisely, if it is neither systematic nor repetitive. As systematic Text Re-use does not necessarily need to be consecutive in both texts due to potential insertions, deletions or changes in the ordering of the textual entities, we need to discriminate isolated from systematic Text Re-use. We define $\{a_i, b_j\}$ as isolated if there is no Text Re-use $\{a_u, b_v\}$ within a certain neighborhood ε, so that

$$\varepsilon = \sqrt{\frac{|i-u| + |j-v|}{2}} < 10.$$

Empirically, we determined 10 as the best value to separate systematic ($\varepsilon < 10$) from isolated Text Re-use ($\varepsilon \geq 10$).

5.3 Text Re-use Grid

The intention of this visualization is to give the researcher an overview of the Text Re-use distribution among all texts of a corpus. We transform the result of the Text Re-use detection algorithm into an intuitive, readable visual interface that immediately (1) reflects the amount of Text Re-uses between each pair of texts, and (2) provides evidence for the type of Text Re-use. For this purpose, we define three parameters:

1. Text Re-use Amount σ. σ is the number of Text Re-uses detected between two texts A and B.

2. Systematic Text Re-use Index λ. λ is an assessment for structures of systematic Text Re-use between two texts A and B with an ordered list of text units, so that $A = \{a_{first}, \ldots, a_i, \ldots, a_{last}\}$ and $B = \{b_{first}, \ldots, b_j, \ldots, b_{last}\}$. To detect these structures, we preliminary filter the Text Re-use results by removing all repetitive and isolated Text Re-use units. This filter process results in a decomposition of the remaining n Text Re-use units into m clusters $C = \{c_1, \ldots, c_h, \ldots, c_m\}$ containing more than one Text Re-use $\{a_i, b_j\}$ each. For each of these clusters c_h with $|c_h|$ Text

Re-uses in total, we compute a correlation coefficient $\rho(c_h)$ as

$$\rho(c_h) = \frac{\sum\limits_{\{a_i,b_j\}\in c_h} (i - \bar{i_h})(j - \bar{j_h})}{\sqrt{\sum\limits_{\{a_i,b_j\}\in c_h} (i - \bar{i_h})^2 \sum\limits_{\{a_i,b_j\}\in c_h} (j - \bar{j_h})^2}}$$

$$\text{with} \qquad \bar{i_h} = \sum_{\{a_i,b_j\}\in c_h} \frac{i}{|c_h|} \qquad \text{and} \qquad \bar{j_h} = \sum_{\{a_i,b_j\}\in c_h} \frac{j}{|c_h|}$$

to estimate the strength of the linear relationship between the Text Re-uses in c_h. Finally, the Systematic Text Re-use Index is defined as

$$\lambda = \sum_{h=0}^{m} \frac{|c_h|}{n} \rho(c_h).$$

λ ranges in the interval $[0, 1]$, whereas high values indicate that systematic Text Re-use patterns are contained.

3. Repetitive Text Re-use Index ω. ω is a measure for the amount of repetitive Text Re-use. Let N denote the number of Text Re-uses found between two texts A and B. To define ω, we remove each Text Re-use $\{a_i, b_j\}$, if both text units a_i and b_j occur only once within all Text Re-uses. Finally, we define ω in the interval $[0, 1]$ with the remaining n Text Re-uses as

$$\omega = \frac{n}{N}.$$

Grid Visualization. For the visual mapping, we construct a grid with each cell representing the Text Re-uses found between two texts of a corpus. For each cell, we compute σ, λ and ω for the corresponding two texts. The cells are displayed in the form of rectangles with bounds proportional to the lengths of the corresponding texts. Interactively, the user can change the display to equal-sized squares, so that even cells representing short texts are properly visible.

Because of the importance for the researchers to detect and analyze texts with extensive systematic or repetitive Text Re-use, we use a specific coloring for the grid cells, so that the type of Text Re-use (represented by λ and ω) and the amount of Text Re-use (σ) can be easily recognized. As the human ability to discriminate colors is limited, we chose a class based approach to compute a limited number of cell

colors. As proposed by Slocum et al. [SMKH09], we use an optimal classification method to group the cells into two sets of classes in dependency of σ, λ and ω. With the Jenks-Caspall-Algorithm [JC71] using reiterative cycling, we compute a configurable number of classes. We receive n classes $\alpha_1, \ldots, \alpha_n$ for the type of Text Re-use (systematic or repetitive), so that α_1 contains the cells with the smallest λ (or ω) and α_n contains the cells with the largest λ (or ω). Furthermore, we compute m classes $\beta_1, \ldots, \beta_m$ for the amount of Text Re-use with β_1 containing the cells with smallest σ and β_m containing the cells with the largest σ. We determine the color for a grid cell in the HSV color space dependent on these classes as follows:

$$H = 240 + \tfrac{i-1}{n-1} \cdot 120 \qquad S = \tfrac{j}{m} \cdot 100 \qquad V = 100$$

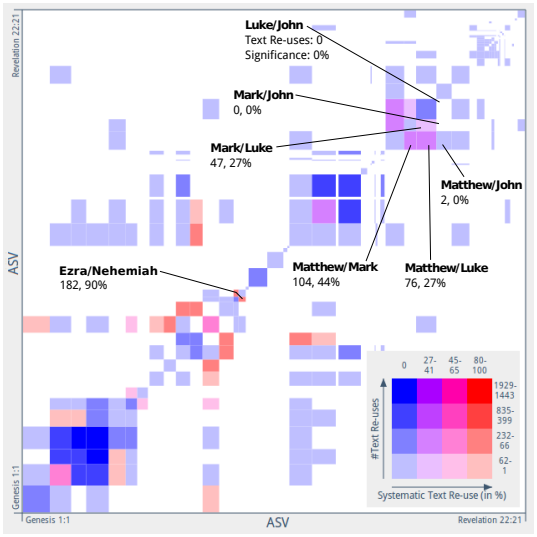
The type of Text Re-use defines the hue on the "Cold-Hot" color scale Diehl proposes [Die07] for the *EpoSee* tool from blue (cold) to red (hot). Thus, we receive cold hues for cell colors with few, and hot hues for cell colors with lots of systematic (or repetitive) Text Re-uses between the corresponding texts. The Text Re-use amount defines the saturation, so that cells with a high number of Text Re-uses receive highly saturated colors, and cells with few Text Re-uses receive lightly saturated colors. For the examples in this chapter we used $n = m = 4$.

In Figure 5.1, the resultant Text Re-use Grids for the Bible books of the American Standard Version (ASV) are juxtaposed highlighting systematic (Figure 5.1a) and repetitive Text Re-use (Figure 5.1b). With the help of a legend, the user is able to immediately categorize type and amount of Text Re-use between two Bible books. Interactively, the user can change from highlighting systematic to highlighting repetitive Text Re-use. By mouse clicking onto a cell, the user can switch from the distant reading grid view to the close reading Text Re-use Browser view that is explained in the next section.

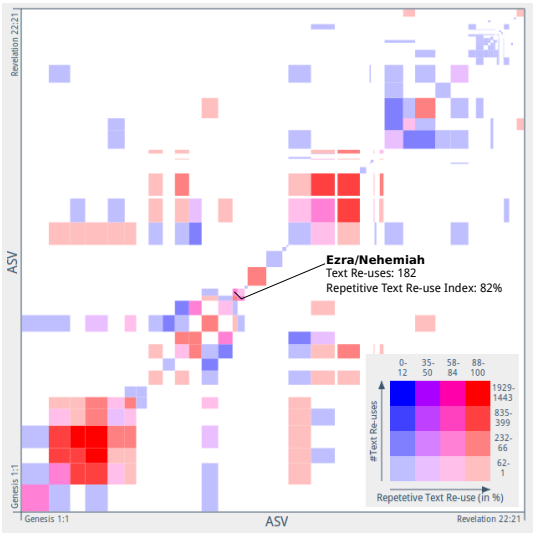
5.4 Text Re-use Browser

Whereas the Text Re-use Grid allows for distant reading of all Text Re-uses occurring within a text collection, the Text Re-use Browser provides a close look at the Text Re-uses found between two texts

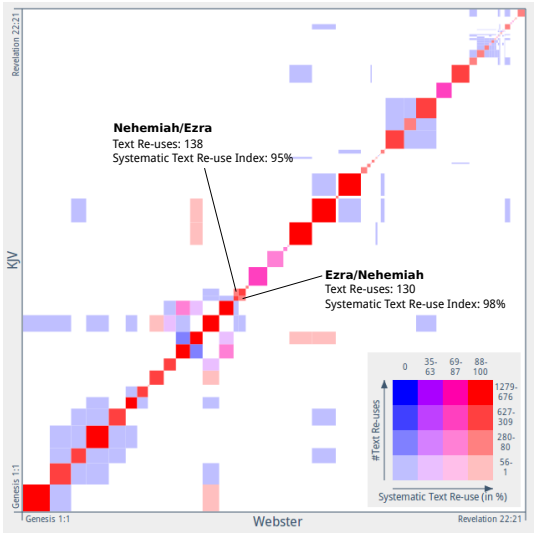
$$A = \{a_{first}, \ldots, a_i, \ldots, a_{last}\} \qquad \text{and} \qquad B = \{b_{first}, \ldots, b_j, \ldots, b_{last}\}.$$



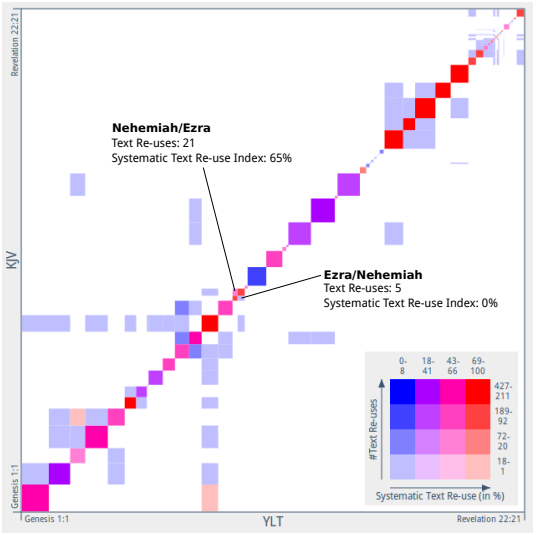
(a) ASV/ASV: Systematic Text Re-use



(b) ASV/ASV: Repetitive Text Re-use



(c) KJV/WBS: Systematic Text Re-use



(d) KJV/YLT: Systematic Text Re-use

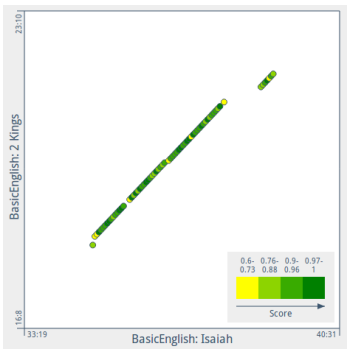
Figure 5.1: Text Re-use Grids for the juxtaposition of various Bible editions.

This still complies to the characteristics of distant reading, but the Text Re-use Browser can be used to drill down to a limited set of Text Re-uses, which supports close reading. In particular, the Text Re-use Browser provides two panels for this purpose:

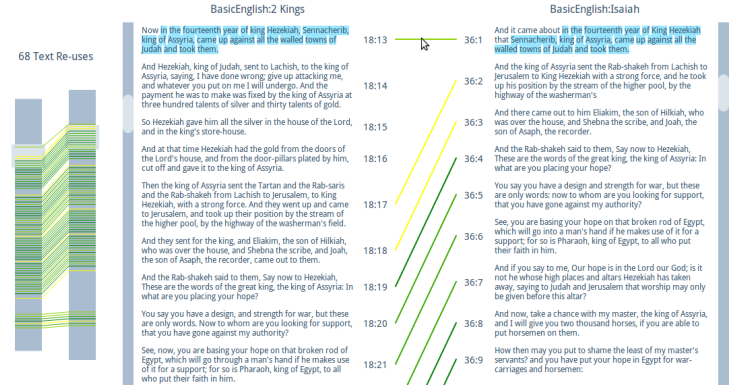
1. Dot Plot View. We adapted the idea of Lee's Dot Plot View [Lee07] to emphasize the types of Text Re-use between the given texts. In contrast to Lee's static variant, we provide an interactive chart, where the number $|A|$ of text units of A defines the range of the x-axis, and the number $|B|$ of text units of B defines the range of the y-axis. Each Text Re-use for a text unit pair is drawn as a single dot. As in bioinformatics, specific dot patterns indicate specific Text Re-use types. Diagonal patterns highlight sections that contain systematic Text Re-use (see Figure 5.3a), whereas vertical and horizontal patterns appear for phrase repetitions (see Figure 5.3c). In Figure 5.3e, we detect patterns for both types of Text Re-use. By selecting a dot via mouse click, a popup with the corresponding text units and a Text Re-use Alignment Visualization (see Chapter 6) is shown, an example is given in Figure 5.2. Via Drag-and-Drop, the user can zoom into a rectangular region of interest.

Figure 5.2: Text Re-use Alignment Visualization shows similarity between two Text Re-use units.

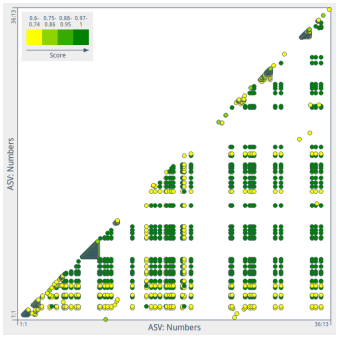
2. Text Re-use Reader. This panel allows for browsing A and B in two opposite windows. Whenever a re-used text unit appears in the viewport of one window, a connection to the opposite text unit is drawn. A click on a connection scrolls both texts, so that the text units of the corresponding Text Re-use are placed on the same horizontal level and a step-by-step exploration of consecutive Text Re-use is possible. A mouseover highlights matching tokens in both units. An additional overview for the texts gives an impression about all occurring Text Re-uses, and it can be used to



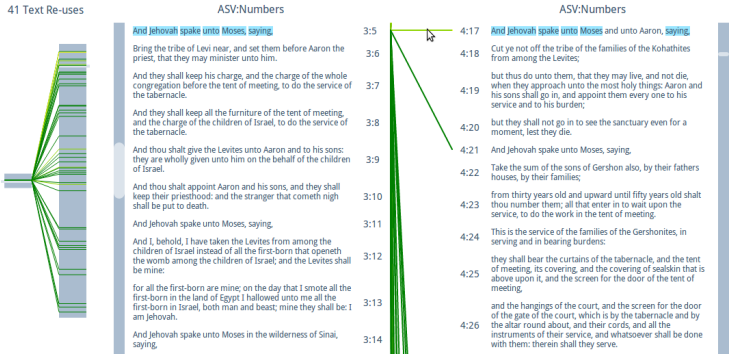
(a) Diagonal pattern in the Dot Plot View.



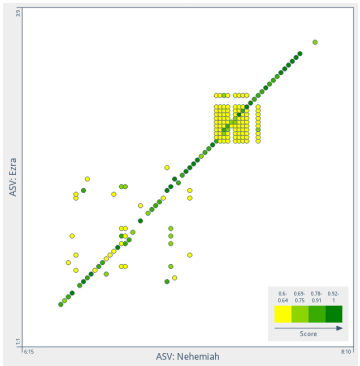
(b) Parallel lines in the Text Re-use Reader.



(c) Vertical and horizontal patterns in the Dot Plot View.



(d) Repetitive re-use of *Numbers 3:5* in the Text Re-use Reader.



(e) Various patterns in the Dot Plot View.

(f) Various line types in the Text Re-use Reader.

Figure 5.3: Panels of the Text Re-use Browser showing systematic and repetitive Text Re-use patterns.

directly jump to a dedicated position. In both views, an accumulation of parallel lines is an indication for systematic Text Re-use (see Figure 5.3b), and *hub*s (a single unit of one text that is connected to multiple units of the opposite text) occur for repetitive Text Re-use (see Figure 5.3d). Both features are present in the example shown in Figure 5.3f.

Both panels are linked to each other. A dot selection in the Dot Plot View triggers a scrolling of the texts to the corresponding positions, whereas selecting a connection in the Text Re-use Reader opens the popup for the corresponding dot. For coloring the Text Re-use glyphs (dots, connections), we use again a class based approach. We group the Text Re-uses in dependency on their scoring values (t) into p classes $\gamma_1, \ldots, \gamma_p$, so that γ_1 contains Text Re-uses with the smallest t, and γ_p these ones with the largest values for t. In order to avoid misinterpretations, we chose a different color scheme in comparison to the Text Re-use Grid. The glyph colors are defined in the HSV color space as:

$$H = 60 + \tfrac{k-1}{p-1} \cdot 60 \qquad S = 100 \qquad V = 100 - \tfrac{k-1}{p-1} \cdot 50$$

Thus, the hue of a glyph color for a Text Re-use with class γ_k ($1 \leq k \leq p$) ranges from yellow to green. To gain visually distinctive colors, the color value ranges between 100 and 50. For the examples in this chapter we used $p = 4$.

Some text juxtapositions contain lots of Text Re-uses that form various patterns. In order to help discovering and exploring specific patterns of repetitive or systematic Text Re-use, we provide visual filters to hide certain glyphs. Also, Text Re-uses with low scoring values can be hidden and a slider can be used to hide isolated Text Re-uses without adjacent Text Re-uses in a certain neighborhood.

5.5 Usage Scenarios

The presented Text Re-use visualizations were mainly designed for two digital humanities projects. During the development phase, the collaborating humanities scholars steadily evaluated the design of the Text Re-use visualizations. We wanted to ensure creating intuitive and flexible interfaces to be able to help answering a broad palette of research questions. The following usage scenarios for various English Bible editions and the Text Re-use among historic Arabic texts confirm the benefit of this iterative process.

5.5.1 Various English Translations of the Bible

The digital humanities project *eTRACES*[1] aimed to discover, analyze and evaluate intertextual similarities in the form of Text Re-use among historical texts of a given corpus. Since the Bible is known as one of the most often read and studied books, and therefore, potential findings are easily evaluable, it was chosen as a proof of concept for the project. The text corpus contained twenty-four different English translations of the Bible covering a time period from the 14th (Wycliffe Bible) to the 21th century (World English Bible). Since each Bible edition was translated with a specific intention, the humanities scholars of the *eTRACES* project had various research questions concerning the Bible corpus.

Of particular interest for the humanities scholars was the comparison of Bible books of the same edition regarding systematic Text Re-use. The Text Re-use Grid shows strong interdependencies for the three evangelists *Matthew*, *Mark* and *Luke*, whilst *John* has few or no Text Re-use at all with those three – confirming a well known fact by visualizing it. These interdependencies were detected for the ASV (Figure 5.1a) and for various other editions the visualization showed a similar pattern. The visualization reveals further insights by highlighting other cells of the grid. For example, there is an indication for vast systematic Text Re-use between the books *Ezra* and *Nehemiah*. Also, there is evidence for repetitive Text Re-use given (Figure 5.1b). Picking the corresponding cell in the Text Re-use Grid allows for close reading in the Text Re-use Browser (Figures 5.3e and 5.3f) and reveals (1) a rectangular cluster of repeatedly used phrases to be compared using the Text Re-use Alignment Visualization, and (2) a large systematic Text Re-use pattern between the beginning of *Ezra* and the middle section of *Nehemiah* (*Ezra 2:1/Nehemiah 7:6 - Ezra 2:70/Nehemiah 7:73*). When juxtaposing the King James Version (KJV) and its revision by Webster (WBS) the systematic Text Re-use pattern for *Ezra* and *Nehemiah* is still highlighted (Figure 5.1c). For the juxtaposition of the KJV and Young's Literal Translation (YLT), which uses Hebrew syntax, the overall number of Text Re-uses strongly decreases and a systematic Text Re-use pattern for *KJV:Ezra* and *YLT:Nehemiah* is not detected (Figure 5.1d). Interestingly, for the juxtaposition of *KJV:Nehemiah* and *YLT:Ezra* a small part of the systematic Text Re-use pattern is still preserved.

[1] http://etraces.e-humanities.net/

(a) *History* vs. *Fates of the Nations*

(b) Two al-Tabari texts juxtaposed

(c) Systematic Text Re-use pattern

(d) Close reading in Text Re-use Reader

Figure 5.4: Text Re-use Browser to discover systematic Text Re-use in Arabic texts

5.5.2 Text Re-use in Historic Arabic Texts

To explore the Text Re-use among historic Arabic texts *for the first time* digitally, historians from the Aga Khan University utilized Büchler's algorithm for detecting Text Re-use and the Text Re-use Browser for visualizing and exploring the results. Predominantly, the focus was on the analysis of systematic Text Re-use. On the one hand, known facts were confirmed to assess the reliability of the visualization, on the other hand, unexpected and unknown patterns were analyzed further.

Figure 5.4a shows evidence for systematic Text Re-use in the form of a diagonal pattern between two chronologies: *History*, called *Ta'rīkh al-rusul wa-l-mulūk* by al-Tabari (839–923), and *Fates of the Nations*, called *Tajārib al-umam* by Miskawayh (932–1030). Modern historians have often argued that Miskawayh relied heavily on al-Tabari's text. The analysis with the visualization suggests a more complex picture, namely, that Miskawayh more selectively copied al-Tabari's text. With the interaction capabilities of the Text Re-use Browser, the historians discovered that Miskawayh

copied al-Tabari's text directly for the Umayyad (661–750) and Abbasid (750–1517) periods but copied very little of it for the period up to 651, which includes Iran's pre-Islamic history and the history of the early Muslim community. It seems possible that Miskawayh wanted a fresh reading. To examine this judgment, the historians plan further Text Re-use analyses, including a comparison of Miskawayh's text against a larger pool of digitized Arabic texts.

Another research question tries to discover what conclusions can be drawn from common passages in a single author's works. In Figure 5.4b, the Text Re-use between al-Tabari's *History* and his *Commentary on the Qur'an*, called *Jāmi' al-bayān 'an ta'wīl āy al-Qur'ān*, is shown. After removing vast occurrences of repetitive Text Re-use, especially stock phrases, the remaining systematic Text Re-use patterns can be analyzed. An example is given in Figure 5.4c. The pattern begins with the statement *"According to what someone with knowledge claimed ..."* (Figure 5.4d). Read on its own, one might think this was al-Tabari's introduction to a topic or report. This might be the case, with al-Tabari repeating himself. It seems at least as likely, however, that this small bit of introduction derives from an original source, which al-Tabari (or perhaps a member of his editorial workshop) copied into both of his texts. Detecting chunks like this across his text and comparing them to textual units in other classical Arabic texts might give a sense of the size of units that passed through the tradition.

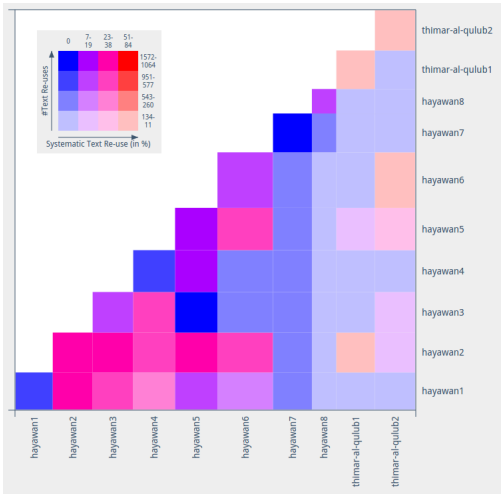
5.6 Summary

In close collaboration to humanities scholars, we designed close and distant reading visualizations that support the exploration and analysis of Text Re-use among historic texts. The *Text Re-use Grid* is a novel distant reading approach to discover type and amount of Text Re-use between each pair of texts of a given text corpus. At the researcher's convenience, one is able to highlight either grid cells with systematic or repetitive Text Re-use. The *Text Re-use Browser* facilitates a further exploration of such Text Re-use patterns between two texts and allows for close reading of individual text passages. This bridge between both perspectives turned out to be an important aspect for the collaborating humanities scholars.

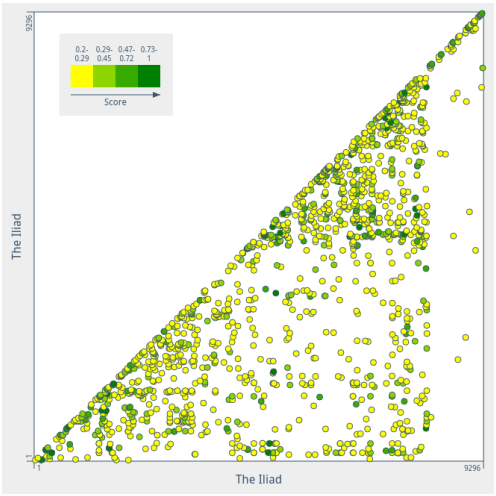
In regular, ongoing workshops on Text Re-use analysis organized by the *Göttingen Centre for Digital Humanities*,[2] both visualizations are used by humanities scholars to analyze automatically detected Text Re-use patterns visually. Although we did not

[2] http://www.gcdh.de/en/

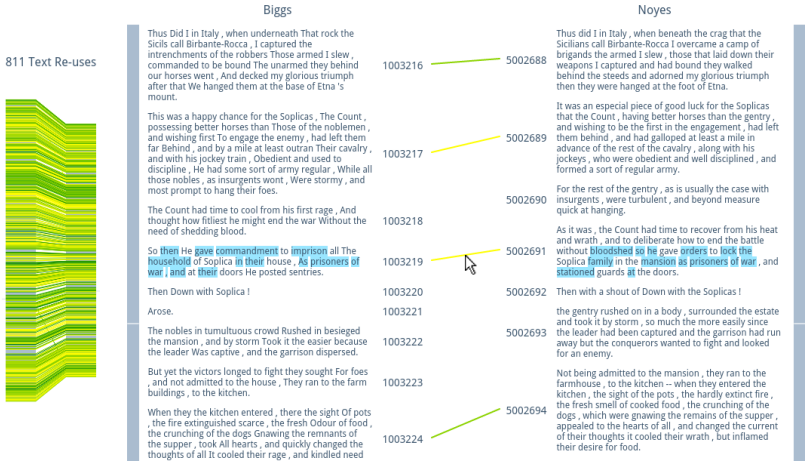
conducted a user study, the wide, language-independent applicability to a number of use cases underpins the benefit of the visualizations for the digital humanities community – as of now, we had use cases with texts in English, Arabic, German, Latvian and Latin. Representative examples are given in Figure 5.5.



(a) Text Re-use among Arabic text fragments



(b) Text Re-use patterns in Homer's *Iliad*



(c) Close reading two English editions of Adam Mickierwicz's *Pan Tadeusz*

Figure 5.5: Workshop examples

I should have no objection to go over the same life from its beginning to the end: requesting only the advantage authors have, of correcting in a second edition the faults of the first.

Benjamin Franklin

6

Visualization of Textual Variation

This research bears on Textual Criticism, a field of the humanities whose task it is to study the creation, distribution and dissemination of textual heritage. The standard publications in this field are critical editions of literary (or non-literary) works. One of the purposes of a text edition is to trace or reconstruct the archetype or original version of the text in order to better understand its evolution over time. To do so, the scholar examines and records the similarities and the differences between a number of exemplars, a practice known in the field as collation [Tan10]. Traditionally, variations are recorded in the edition's critical apparatus, a textually dense area at the bottom of the page mostly, if not only, intelligible to experts [Szp14]. In the process of collation, scholars select a number of editions they wish to compare, arrange these side by side, manually transcribe each exemplar (assuming no transcription already exists) and annotate variation between these. The more manuscripts one compares, the more complex and laborious the task becomes. Modern, digital methods are beginning to address this time-consuming, error-prone task in the form of semi-automatic transcription[1] and automatic collation. The development of TRAViz, which is presented in this chapter, falls under the latter effort but requires transcriptions in order to operate.

[1] Abbyy FineReader OCR (Optical Character Recognition) and ICR (Intelligent Character Recognition) technologies, for instance, are currently being developed to support computer recognition of handwritten script.

In 2009, Schmidt proposed the Variant Graph to represent multiple versions of a digital text [SC09]. A Variant Graph is a directed acyclic graph capable of modeling the similarities and differences among various editions of a text. Figure 6.1 shows an example of a Variant Graph in the design Schmidt uses for the three example variants: "the white dog chases a brown cat" (W1), "a brown cat chases the white dog" (W2) and "a brown cat chases the black dog" (W3).

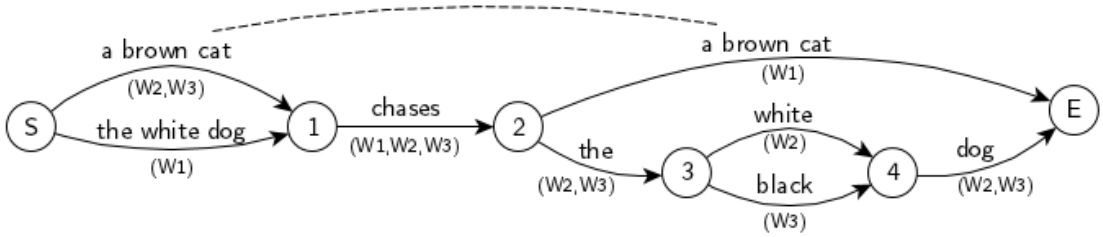


Figure 6.1: Example Variant Graph.

The reading direction of the above graph is sinistrodextral (Left-To-Right). All variants start at the S(tart) vertex and end at the E(nd) vertex. The vertices in the graph number and connect subsequent text-snippets; each variant appears as a label identifying an edge; the edition or variant identifiers are displayed in brackets. Additionally, transpositions of word groups ("a brown cat") are highlighted in the form of dashed connections. While informative, from a graph drawing standpoint this Variant Graph is not particularly easy to read: the textual information (text and edition identifiers) born by the edges puts a strain on comprehension. In 2011, Dekker introduced CollateX,[2] a web-based collation framework that generates Variant Graphs and facilitates work with electronic editions in the browser [HDvHM+15]. Unlike Schmidt, Dekker focused on improving the alignment structure between the various text editions. Along with output formats such as XML or JSON, CollateX uses the GraphViz[3] library, which computes non-interactive visualizations of the resultant Variant Graphs. Consequently, the readability of the graph in Figure 6.2 is made easier inasmuch as text and edition information are now split between vertices and edges.

[2]CollateX was designed as the successor to Peter Robinson's Collate software (see http://www.sd-editions.com/), which offered a primarily textual representation of variation. The development of CollateX was guided by the Interedition consortium (http://www.interedition.eu/)

[3]http://www.graphviz.org/

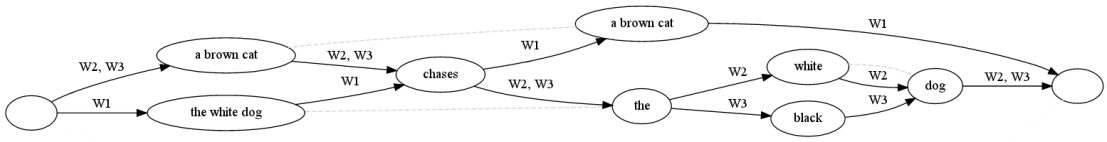


Figure 6.2: Example Variant Graph with CollateX.

The tool Stemmaweb [AM13] aims to support analyses of Variant Graphs. It extends the CollateX graph to enable user-driven annotation and modification of the graph structure (e.g., the merging and splitting of vertices). But for all its merits, Stemmaweb's straightforward adoption of the GraphViz visualization affects readability. TRAViz,[4] a web-based open source library, addresses this issue by implementing a layout algorithm for Variant Graphs and a set of design rules aimed at styling both vertices and edges, and thus supporting an intuitive reading of the collation. By extending the choice of interaction possibilities, TRAViz enhances support and allows users, for instance, to tweak the visualization to meet specific research questions. Figure 6.3 is a TRAViz reproduction of the above example.

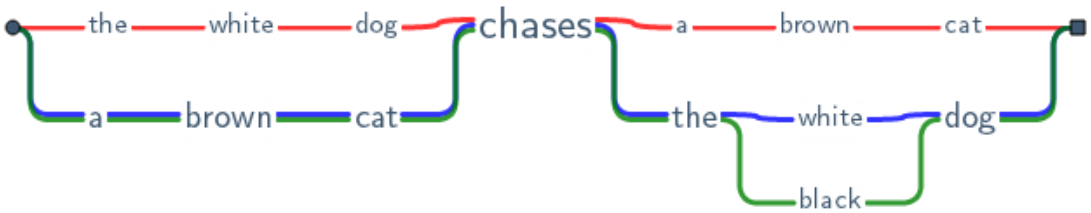


Figure 6.3: Example Variant Graph with TRAViz.

TRAViz' introduction of color – one per each edition (W1 = red, W2 = blue, W3 = green) –, the use of word-sizing, the linear alignment and the removal of unnecessary visuals (circular shapes), improve the readability of the Variant Graph. This chapter illustrates the TRAViz layout algorithm for Variant Graphs, aspects of graph design and interaction capabilities, and concrete applications to existing text-based projects. Finally, we outline an approach that supports the distant reading of Variant Graphs.

[4] http://www.traviz.vizcovery.org/

6.1 The Gothenburg model

The Gothenburg model[5] describes the steps required for the computer-supported collation of textual entities. In correspondence to this model, the TRAViz pipeline consists of the following five steps:

1. **Tokenization:** In order to align meaningful text patterns, we tokenize a text into words by splitting at whitespaces. Although a splitting on syllable level would be beneficial when aligning editions of poems, the required information is usually not provided.

2. **Normalization:** This step includes the transformation of all words into lower case as well as the removal of punctuation characters. With the Relative Edit Distance (RED) – introduced in Section 6.5 –, we furthermore provide a stemming mechanism to support aligning words sharing the same beginning.

3. **Alignment:** Let $\{w^1, \ldots, w^n\}$ denote a set of editions to be aligned. Various alignment algorithms exist (e.g. [NW70, BG07]), we use a brute force approach that borrows ideas from these algorithms and performs well for small text units. After tokenization and normalization, we insert each edition $w^i = w^i_1 \ldots w^i_{|w^i|}$ in the form of a directed path $v(w^i_1) \ldots v(w^i_{|w^i|})$ with vertices representing words in the initial Variant Graph. Then, we iteratively merge vertices of different paths with equal words and choose the alignment approach that reaches a maximum number of merge iterations while keeping the Variant Graph acyclic. Each vertex $v = \{w^i_s, w^j_t, w^k_u, \ldots\}$ of the graph is an alignment of the words $\{w^i_s, w^j_t, w^k_u, \ldots\}$. The *vertex degree* $|v|$ is the number of words assigned to v and $v(w^i_s)$ is the corresponding vertex for the s-th word of edition i.

4. **Analysis/Feedback:** Automatically, we provide a heuristic of transpositions at word level. As an alignment might not meet the expectations of the user, we provide an interactive mechanism to modify the alignment using split and merge operations on vertices; details can be found in Section 6.5.

5. **Visualization:** The resultant alignment is visualized as a directed acyclic graph in a layout tailored for Variant Graphs.

[5] http://wiki.tei-c.org/index.php/Textual_Variance

TRAViz primarily focuses on visualizing Variant Graphs. The following sections outline the graph layout algorithm and the graph design in detail.

6.2 Related Work

Variant Graphs are directed acyclic graphs that illustrate similarities and differences among text editions. A layered graph drawing as introduced by Sugiyama is the common drawing style used for directed acyclic graphs [STT81]. Typically, the vertices are placed on equally spaced horizontal (or vertical) layers and the edges are routed downwards (or rightwards) between the layers. Sugiyama's approach as well as many of it's variations [GKNpV93, UBSE98, Col01, ESK04] need to be adapted for the purpose of visualizing Variant Graphs because only single vertices of one path are usually placed on one layer. This complicates the required vertical alignment of synonyms consistent of various amounts of tokens (e.g., "swarmed" and "brought forth abundantly" in *Genesis 1:21*). Additionally, the widths of the vertices of a Variant Graph vary, so that a placing on vertical layers of equal width would further increase the distance between adjacent tokens. To remove the occurring clutter for layered graph drawings with lots of edges, some approaches bundle edges to improve the readability of the resultant layouts [EGM07, PNK11]. The *Word Tree* [WV08] is a visualization that cannot be directly applied to Variant Graphs since it only aligns shared beginnings of sentences in the form of a tree. The font size of a node label reflects the number of occurrences, and each variation splits a node of the tree into several leaves. The *Word Graph* [RGP+12] is a similar approach that visualizes the commonness of phrases. The layout is not applicable to Variant Graphs as it requires one or more predefined terms for construction. Also, the design does not support those humanities scholars who want to track individual text editions and who want to compare multiple text editions. A plain solution to align Text Re-use is given in [BGEH10]. The original text snippet is drawn as a main branch and variations of Text Re-use candidates are sub-branches with a certain color. This approach works fine for small examples with minor variations, but it fails for major differences, especially, when multiple Text Re-uses share the same sub-branches. A similar visualization for the uncertainty in lattice graphs also supports various sub-branches [CCP07]. But merging of multiple nodes of the same kind is not provided, although the metaphor for uncertainty could be used for this purpose. TRAViz utilizes some of the presented ideas in order to design a well readable layout for Variant Graphs.

6.3 Variant Graph Layout

The layout algorithm consists of three major steps: (1) placing the words on horizontal layers to minimize the crossings when (2) routing the edges between the words. The final step is (3) the removal of overlaps between vertices and edges.

6.3.1 Vertex Placement

We layout the vertices by placing the corresponding words onto horizontal layers $\{\ldots, l_{-2}, l_{-1}, l_0, l_1, l_2, \ldots\}$. The height of a layer depends on the maximum height of the words placed on it. We start by placing the words for the vertices $v(w_1^i), \ldots, v(w_{|w^i|}^i)$ of an arbitrary edition w^i in left-to-right order on layer l_0 (main branch). By default, we choose the edition w^i with the maximum value for

$$\sum_{s=1}^{|w^i|} |v(w_s^i)|,$$

which means that w^i has lots of words assigned to vertices with large vertex degrees (see step 3 in Section 6.1 for details). Afterwards, we iteratively determine the subpath of the graph with most vertices already assigned to layers. Each subpath $\{v_1, \ldots, v_n\}$ has assigned layers for v_1 and v_n and the layer for the vertices of $p = \{v_2, \ldots, v_{n-1}\}$ needs to be determined. Let l_i denote the layer of v_1 and l_j the layer of v_n. We aim to place p as close as possible to its adjacent vertices v_1 and v_n. Starting with layer l_k ($k = \lfloor (i+j)/2 \rfloor$), we iteratively search for a layer with enough free space for the words of the vertices of p in the order $l_k, l_{k+1}, l_{k-1}, l_{k+2}, l_{k-2}$, and so on. If the total width of the words of p is larger than the space between v_1 and v_n, we stretch the distance between v_1 and v_n. After a layer is found, we move all vertices of the graph horizontally, so that (1) the words do not overlap each other, (2) a minimum space of configurable width between all adjacent vertices is given, and (3) each vertex is placed in the barycenter of its neighbors. We perform this process for all subpaths to complete the vertex layout of the Variant Graph.

6.3.2 Edge Routing

In preparation, we insert a path layer p_i above each vertex layer l_i that is used to route horizontal links. Then, we initialize the type for each edge $e = \{t_l, t_r\}$ (l means *left*, r

means *right*) dependent on the corresponding layers l_l and l_r of the connected vertices t_l and t_r. We separate three different edge types (see Figure 6.4):

type 0: If $l_l = l_r$ and there is no other vertex placed on l_l between t_l and t_r, e is drawn as a straight horizontal line.

type 1: If $l_l = l_r$ and there are vertices placed on l_l between t_l and t_r, a path is routed above l_l, consistent of an upward vertical link v_l, a horizontal link h on path layer p_l and a downward vertical link v_r.

type 2: If $l_l \neq l_r$, a path consistent of an upward (or downward) vertical link v_l, a horizontal link h on path layer p_l (or p_{l+1}) and an upward (or downward) vertical link v_r. We always put h on the corresponding path layer with the higher absolute index.

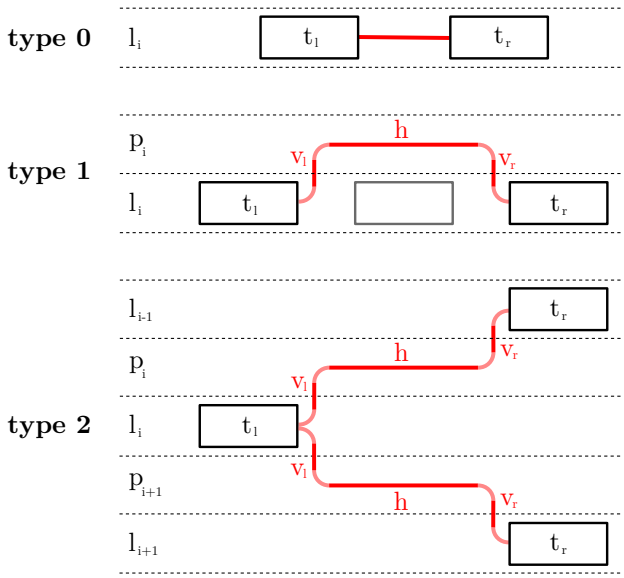


Figure 6.4: Different edge types

To smoothen the graph layout, we connect each vertical link to its adjacent vertex and to the horizontal link using bends with radius r_b. Since four bends are required to draw edges of type 1 and 2, all adjacent vertices receive a minimum gap of $4 \cdot r_b$. After initialization, horizontal and vertical links may occlude each other. Below, we outline bundling strategies to resolve these overlaps. The final step simplifies the graph layout by converting edges with type 2.

Bundling horizontal links For each path layer p, we receive a list of horizontal links $h_1, \ldots, h_n$ with $h_i = \overline{t_l t_r}$. We begin with constructing bundles $B = b_1, \ldots, b_n$ of horizontal links for all edges with the same left-hand vertex t_l and for all edges with the same right-hand vertex t_r. Thus, all horizontal links occur twice over all bundles. Afterwards, we sort B by decreasing number of horizontal links within the bundles. Iteratively, we insert the first bundle b_1 of B onto p. If b_1 overlaps with other already inserted bundles, we merge all these bundles into an overlap group. Then, we remove the duplicates of the horizontal links of b_1 from the remaining bundles of B and sort B again by decreasing number of horizontal links. After placing all bundles onto p, we place the bundled horizontal links of each overlap group parallel to each other, separated using a predefined gap g. Thereby, we order the bundles the way that the number of edge crossings is as minimal as possible. An example ordering is shown in Figure 6.5a. We perform this step iteratively by decreasing number of bundles in the overlap groups. Once a bundle that is part of multiple overlap groups is adjusted, it remains fixed for further ordering iterations.

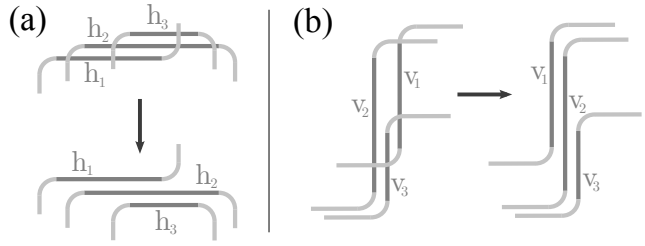


Figure 6.5: Ordering links of overlap groups

Bundling vertical links For each vertex that is linked to neighbors with edges of type 1 and 2, we create a total of four bundles for its upward/downward incoming vertical links and its upward/downward outgoing vertical links. Since bundles of distinct vertices can be too close or even overlap each other, we perform the following two steps to keep the graph layout uncluttered. Firstly, we insert the bundles stepwise by increasing x-value into the graph layout. If the minimum gap g to bundles that are already inserted is not given, we merge these bundles into overlap groups. Secondly, after all bundles are inserted, we order the vertical links of each overlap group – again, parallel to each other and separated using g –, so that the number of edge crossings is minimal as it is shown in Figure 6.5b. Finally, we test whether the required gap

between each vertical link v of the group and its subsequent glyph (right-hand vertical link v_r or vertex t_r) is large enough. If this is not the case, we slightly shift all subsequent edges and vertices of v to the right so that the requirement is fulfilled.

Converting edges with type 2 To improve the readability of the graph layout, we try to simplify edges of type 2 by removing one vertical link, and thereby, two of the four bends. Figure 6.6 illustrates an example of our approach. The upper connection $\{v_{l1}, h_1, v_{r1}\}$ between t_l and t_{r1} is replaced by $\{v_{new}, h_{new}\}$, because neither v_{new} nor h_{new} cause an overlap with a vertex of the layout. Since this is the case for the lower connection $\{v_{l2}, h_2, v_{r2}\}$ between t_l and t_{r2}, it cannot be replaced. There are two possibilities for each edge conversion, either the left hand vertical connection v_l gets removed and the right hand vertical connection v_r gets replaced by v_{new} or vice versa. If no overlaps are produced in both cases and $|l_l| > |l_r|$, we remove v_l and replace v_r. Otherwise, we remove v_r and replace v_l.

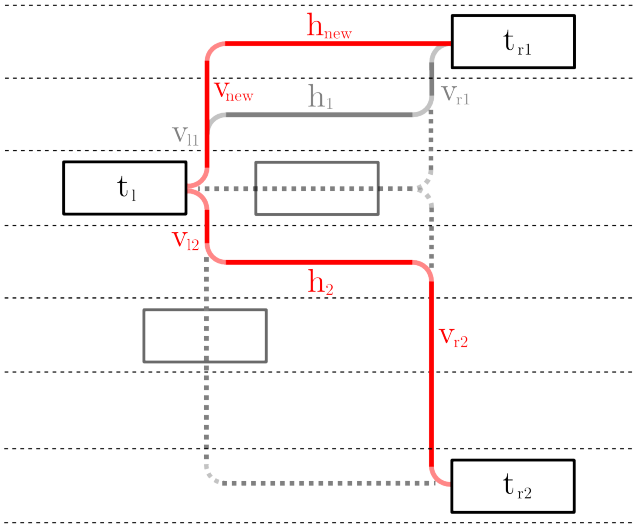


Figure 6.6: Converting edge type 2

6.3.3 Removing overlaps between vertices and edges

Although the graph is designed the way the observer follows the spreading of an edition in horizontal direction, potential overlaps between a vertical link v that crosses an intermediate vertex layer l_m and a vertex t placed on l_m may hamper the readability

of the graph. In such a case, we check if t can be moved horizontally without overlapping bends or other vertices by keeping the minimal required gaps to its neighbors.

An example can be seen in Figure 6.7a. A leftward movement of t is preferred, since the final position would be closer to its current position. Because this attempt fails, we move t to the right. If a horizontal movement of t is not possible (see Figure 6.7b), we move t and it's subsequent edges and vertices, so that the overlap gets removed.

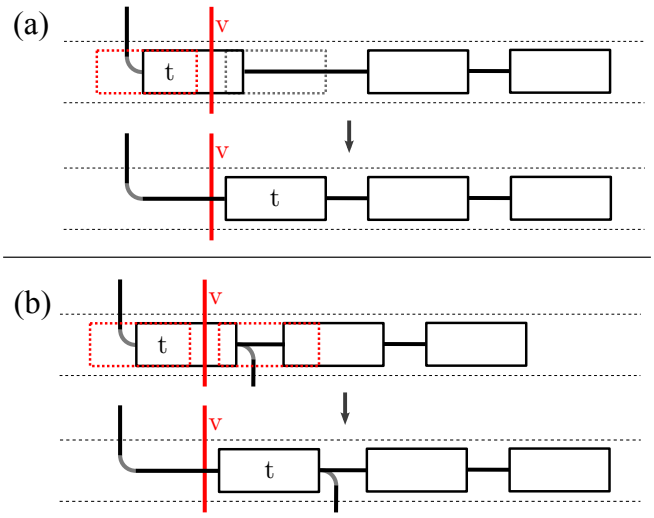


Figure 6.7: Overlap removal examples

6.4 Variant Graph Design

In mutual dependency to the aforementioned layout, we evolved a novel design tailored for Variant Graphs. CollateX was chosen as reference tool not only because it is one of the standard tools in the digital humanities, but also because it underpins many web-based extensions, including Stemmaweb. A CollateX Variant Graph representing five English translations of *Genesis 1:4* (see Figure 6.8) reveals a number of issues that hamper the readability of the collation. For this reason, we propose five Variant Graph design rules based on related works in information visualization and on guidelines for drawing graphs.

The *first rule* concerns the label size of a vertex. When looking at the CollateX graph in Figure 6.8, it is hard to determine the frequency of a word across all editions. Hence, it is difficult to study the use of synonyms and the occurrence of specific

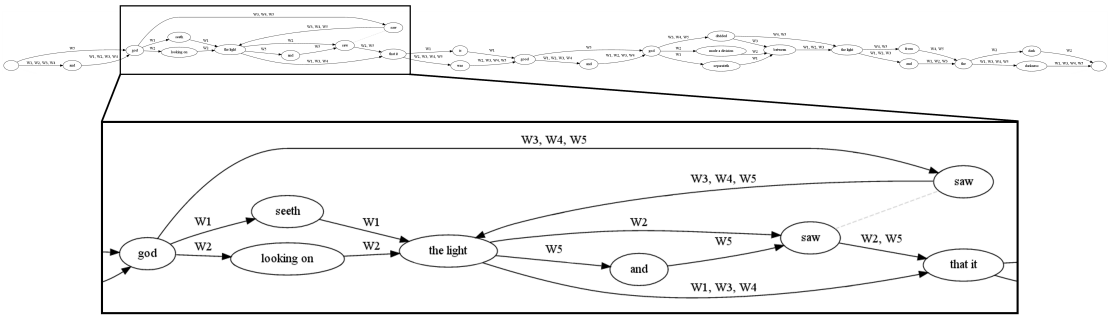


Figure 6.8: CollateX: Variant Graph (with zoom on opening segment) illustrating the relationships between five different versions of *Genesis 1:4*.

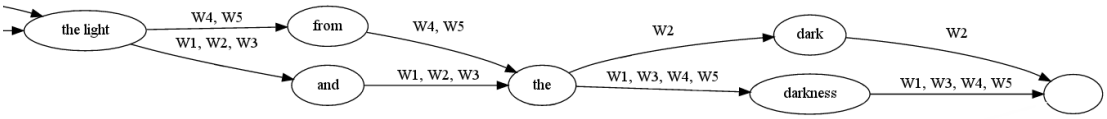


Figure 6.9: CollateX: large edge labels in the final part of *Genesis 1:4*.

text patterns. One workaround to this problem is to count the edition identifiers listed by the labels of the incoming edges. We can easily exploit this information by displaying vertex labels in varying font sizes, a common practice in tag clouds such as Wordle [VWF09], and a feature in visualizations similar to TRAViz [WV08, RGP+12]. In TRAViz, font size reflects the occurrence of individual text fragments, thus contributing to the clear identification of their frequency in the graph.

The *second rule* eliminates backward edges.[6] Instinct pushes us to draw the edges of a directed graph as arrows. But why do so if we recognize the reading direction of the text? Introducing bi-directional cues makes the graph counterintuitive. In graph theory, the accepted layout for a directed acyclic graph is a layered graph drawing (see Section 6.2) where all edges point in the same direction. When we turn the Variant Graph layout into a layered drawing as described in Section 6.3, we reduce the cognitive load of the visualization by replacing the arrows with undirected edges aligned to the writing direction.

The *third rule* avoids labeling edges. In CollateX, edges are labeled with edition identifiers. This leads to two problems. Firstly, edge labels interfere with the vertex labels (text fragments), forcing the reader to visually separate these information.

[6] In Figure 6.8, the direction of the edge from the word "saw" (right-top) to "the light" (bottom-left) is dextrosinistral (Right-To-Left). This is called a backward edge.

Secondly, if many editions pass an edge or if long edition identifiers are used, the corresponding edge labels significantly increase in size (Figure 6.9). As a consequence, adjacent vertices drift apart in order to gain enough horizontal space to accommodate the edge labels, thus forcing the width of the graph to the point where the reader rapidly loses the context of a text fragment. To avoid this outcome, TRAViz does not label edges when visualizing Variant Graphs. Instead, it draws an edge for each edition in a different color. As the human ability to distinguish colors is limited, this solution only works well with a small number of editions (< 10). However, following the suggestions made by Ware [War13] and Harrower et al. [HB03], a set of varying color hues is defined so as to increase this number to 24. Hence, a legend is required to map the colors to the corresponding edition.[7] Nevertheless, TRAViz is also capable of displaying edge labels upon interaction.

The *fourth rule* bundles major edges. When analyzing and comparing text editions, the user is often interested in those editions which deviate from the "standard reading." In Stemmaweb, edges that are passed by most editions are labeled with a "majority" tag and are accordingly bundled. As per the third rule, users are presented with multiple lines, which we bundle as major edges likewise to Stemmaweb. Resultant edge types – bundled and unbundled – are highlighted differently: unbundled edges are color-coded with inviting saturated hues, whereas bundled edges appear as thick gray strokes. By doing so, any deviations from the standard reading can be more easily detected. As per the present and third rules, TRAViz is able to reduce the number of edges – and, therefore, the cognitive load of the user's approach – to a minimum.

The final *fifth rule* inserts line breaks. One of the major problems emerging from the application of CollateX and TRAViz Variant Graphs to large portions of text is the enforced horizontal scrolling. In scrolling especially long texts, one might lose context and struggle to track the distribution of editions in the graph. Moreover, the small screen space occupied by the graph translates into an increase in whitespace. The outcome of a survey performed by the TAdER Project[8] on horizontal scrolling in the browser underpins our hypothesis that the user is accustomed to scrolling vertically. As a result, TRAViz inserts line breaks by splitting the Variant Graph when the width of the browser window is reached, thereupon mimicking the text layout of a book. As per the third rule, the user is presented with different colored edges or edge bundles at the end of each line, so that all paths are visually identifiable at the beginning of

[7] In the case of CollateX, a legend is also required to map the edition identifiers to edition titles.
[8] http://www.tader.info/scrolling.html

the next line. Additionally, line numbers support vertical orientation. The insertion of line breaks helps the user navigate large graphs, concurrently preserving context.

Figure 6.10 shows the TRAViz Variant Graph for *Genesis 1:4* in the aforementioned design. Unlike the CollateX Variant Graph, with TRAViz it is possible to evaluate the level of variation: the central string of larger consecutive tokens connected by thick edges bundles analogous translations, and the two texts containing the highest degree of variation are the Bible in Basic English (light green) and Young's Literal Translation (green).

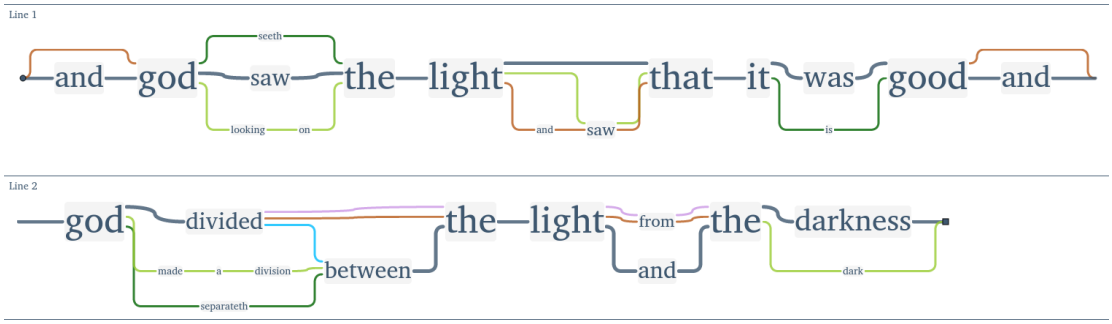


Figure 6.10: TRAViz: Variant Graph illustrating the relationships between five different versions of *Genesis 1:4*.

6.5 Means of Interaction

At the 2013 Digital Humanities Conference, Andrews stressed the importance for humanities scholars to be able to interact with a Variant Graph and to modify its structure [AVZ13]. TRAViz responds to this requirement by providing a wide range of interaction possibilities, whereby users can tweak the visualization in order to suit their particular needs.[9] Users may, for instance, prefer to visualize synonyms over spelling variation. In general, TRAViz only aligns exact word matches. This approach is useful, especially if the users want to focus on all types of variation (e.g., orthography). But if the users are not interested in orthographical variation, they are free to impose the alignment in order to better understand the correlation between the two

[9] Gibbs and Owens call for better and more user-friendly digital tools to support humanities scholarship [GO12]. TRAViz' focus on design and transparent documentation can be understood as a response to Gibbs' and Owens' recommendations.

words A and B. To this end, we define the Relative Edit Distance $RED(A, B)$ based upon the Levenshtein Distance $LevDist$ [Lev66] as

$$RED(A, B) = \frac{2 \cdot LevDist(A, B)}{|A| + |B|}.$$

$RED(A, B) = 0$ is an exact match, $RED(A, B) \leq 0.2$ allows for smaller variations (e.g. "beginnings" and "beginning"), and $RED(A, B) \leq 0.5$ allows for greater variations (e.g. "beginnings" and "bigynnyng"). Higher values should not be used as the probability that unrelated words cluster progressively increases. If various versions are aligned, we use the most frequent version as the label for the corresponding vertex in the graph. In the Bible usage scenario (see Section 6.6.1), the user can set the desired Relative Edit Distance by adjusting the provided slider.

The second means of interaction falls back to customary mouse behavior. For a thorough analysis of the visualized Variant Graph, TRAViz employs the customary mouse hover-and-click gestures. Hovering over a vertex in the graph highlights all editions passing through it and hides all information and connections pertaining to other versions. This helps to investigate the graph distribution of a subset of potentially similar translations and to explore the similarities and differences among them. Furthermore, this interaction singles out those editions forming majority edges. Two scenarios support the user in mapping colors to their corresponding editions (see Figure 6.14b), a particularly important functionality if working with a large number of editions. On the one hand, a mouseover on an edge shows the contributing editions in a tooltip; on the other hand, clicking on a vertex displays a pop-up window listing all editions and their corresponding tokens in the assigned colors.

Users may not always agree with machine-generated alignments. As stated by Andrews, humanities scholars want to be able to modify graph structure and are more likely to adopt a tool that helps them achieve their desired alignment. For this reason, TRAViz offers the option to split and merge vertices. To detach words, the user clicks on the corresponding button in the vertex pop-up window and creates a new branch in the graph. Merge operations are carried out through the customary Drag-and-Drop mouse functionality. As Variant Graphs are acyclic, two vertices can only be merged if this does not produce a cycle in the graph structure. When the user superimposes two vertices, the system calculates the feasibility of the merge: both words are highlighted in green if the merge is allowed or in red if not. If a merge is not possible in the first instance, a sequence of prior split operations can help to avoid

a potential cycle.

Next, a user may want to explore those editions containing a higher degree of variation. By default, TRAViz follows the Stemmaweb concept and draws a majority edge – not individual edges per edition – if the connection between the corresponding vertices is passed by at least half of the editions. But depending on the research question and the dataset to be examined, the definition of majority may vary. For this reason, TRAViz allows the user to reduce the majority value if the overall variation among the texts is high. For editions sharing little variation, this threshold can be increased. In both cases, editions which do not make the majority group earn more visual presence.

Finally, TRAViz offers the option to visualize potential transpositions. As Schmidt points out, the algorithmic detection of transposed text passages is extremely complex [Sch09] and thus hard to manage in the web-browser. Nevertheless, TRAViz attempts to provide leverage points for potential transpositions by visually connecting related vertices. Two vertices are related if they share at least one word. Especially for large graphs, the number of potential transpositions is high. To avoid unnecessary clutter, the potential transpositions are only displayed when the user hovers over a vertex for which transpositions have already been established. To visually separate transpositions from the graph's connections, we repurpose the concept of CollateX and visualize transpositions in the form of dotted black lines (see Figure 6.16). This means of interaction is particularly useful when employing the relative edit distance merging various versions into one vertex. All of these vertices are labeled with only the most frequent version, so that transpositions are not always visible at first glance.

6.6 Usage Scenarios

The following case studies serve to demonstrate the flexible application and advantages of using TRAViz with a view to emphasizing its potential in the field of textual criticism in particular and of digital humanities in general.

6.6.1 Various English Translations of the Bible

The first application of TRAViz involves the visualization of twenty-four English translations of the Bible. The Bible editions used for this example are listed in Table 6.1.

1380	Wycliffe Bible	1885	English Revised Version
1535	Coverdale Bible	1890	Darby Bible
1537	Matthew Bible	1901	American Standard Version
1539	Great Bible	1949	Bible In Basic English
1560	Geneva Bible	1995	God's Word Translation
1568	Bishop's Bible	1999	American King James Version
1610	Douay-Rheims Bible	2000	Updated King James Version
1611	King James Version	2000	King James 2000 Version
1749	Douay-Rheims-Challoner Bible	2000	World English Bible
1833	Webster's Revision	2004	A Voice in the Wilderness
1863	Young's Literal Translation	2009	Catholic Public Domain Version
1876	Smith's Literal Translation	2009	Lighthouse Bible

Table 6.1: Used Bible editions in chronological order (datings taken from [Tal12]).

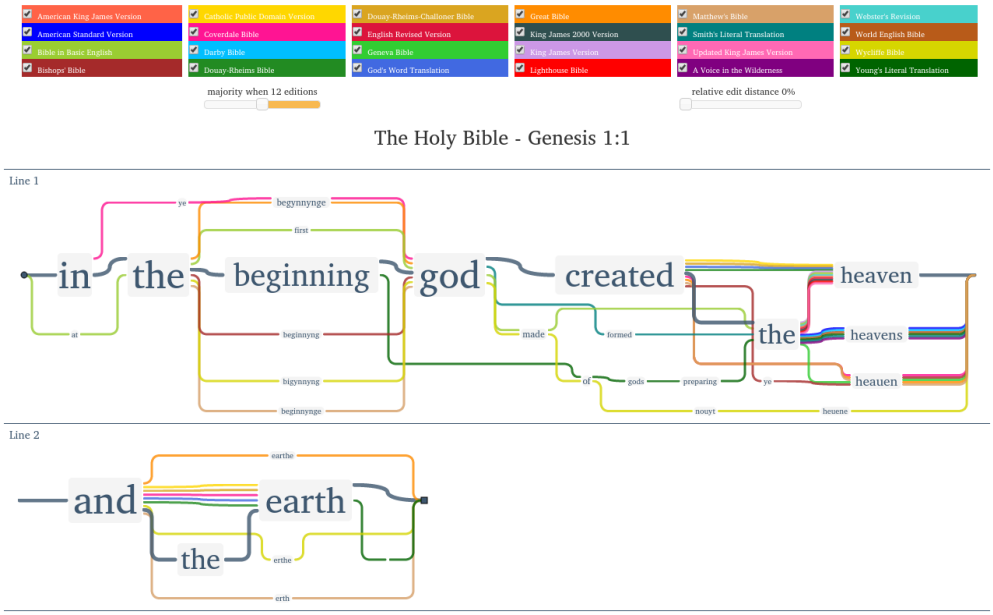


Figure 6.11: *Genesis 1:1* in twenty-four English translations with a majority of twelve and $RED = 0$.

Figure 6.11 shows a web interface embedding TRAViz. The top panel lists the different Bible versions with their assigned colors. Sets of translations can be analyzed at once by checking or unchecking their corresponding boxes in the panel. Two sliders can be used to further modify the graph structure, either by defining the majority

value or by re-aligning the underlying verses (dependent on a relative edit distance) so as to highlight the most obvious differences. The Variant Graph for the first Bible verse is visualized using a majority of twelve and a relative edit distance of 0% ($RED = 0$). Figure 6.12 shows the same verse with a joint application of a majority of at least six editions and a relative edit distance of 40% ($RED = 0.4$), aligning various versions of "beginning," "heaven" and "earth" into one vertex each. The following are three examples illustrating how this visualization supports distinct research questions.

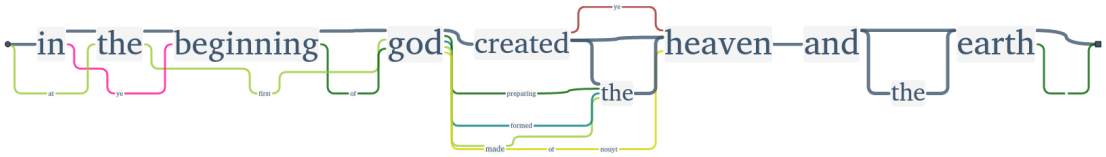


Figure 6.12: *Genesis 1:1* in twenty-four English translations with a majority of four and $RED = 0.4$.

The *first example* concerns itself with the sixth commandment (*Exodus 20:13* and *Deuteronomy 5:17*), commonly recited as "You shall not kill." But there exist numerous different translations of this verse, which can be crucial for its interpretation. The Variant Graph in Figure 6.13 displays the sixth commandment with a majority of four: the thick gray lines chain synonymous terms which have been used by the translations at least four times. The variation of the verb is of particular interest. Twelve translations choose the word "kill" and four editions its older form "kyll"; five authors write "murder," one "sle" and another "put anyone to death without cause." The latter variation belongs to the Bible in Basic English (light green), a non-literal translation, and it is easily distinguishable by its length. Another version, significantly shorter, is contained in God's Word Translation (blue): "Never Murder."

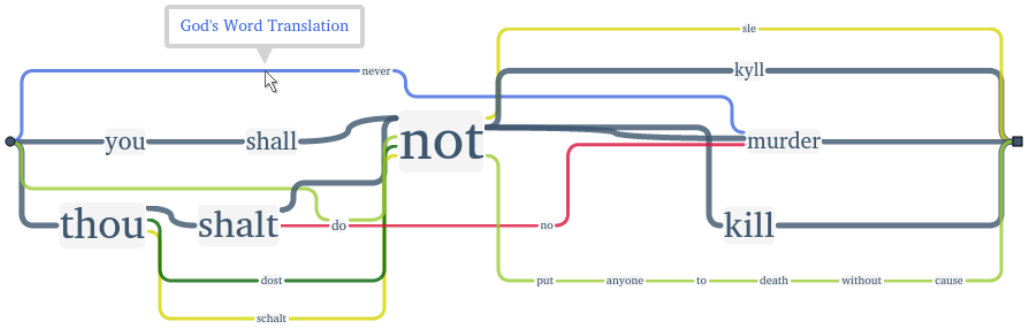
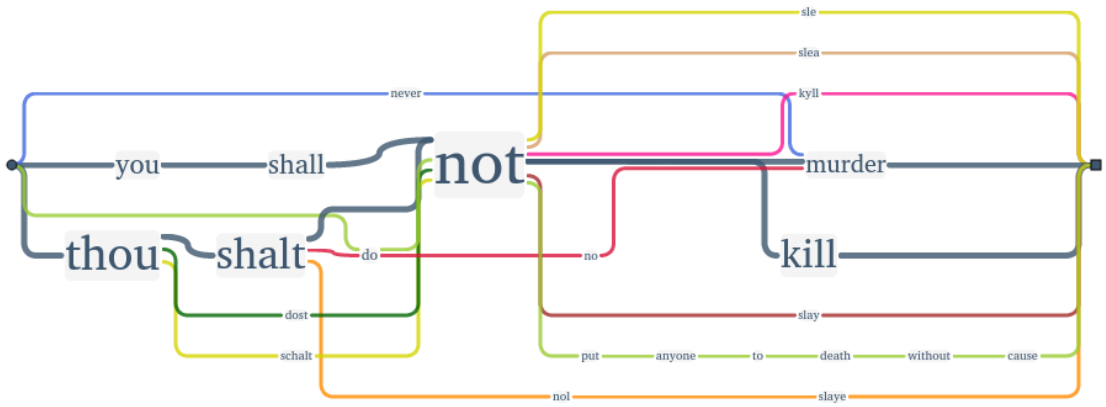
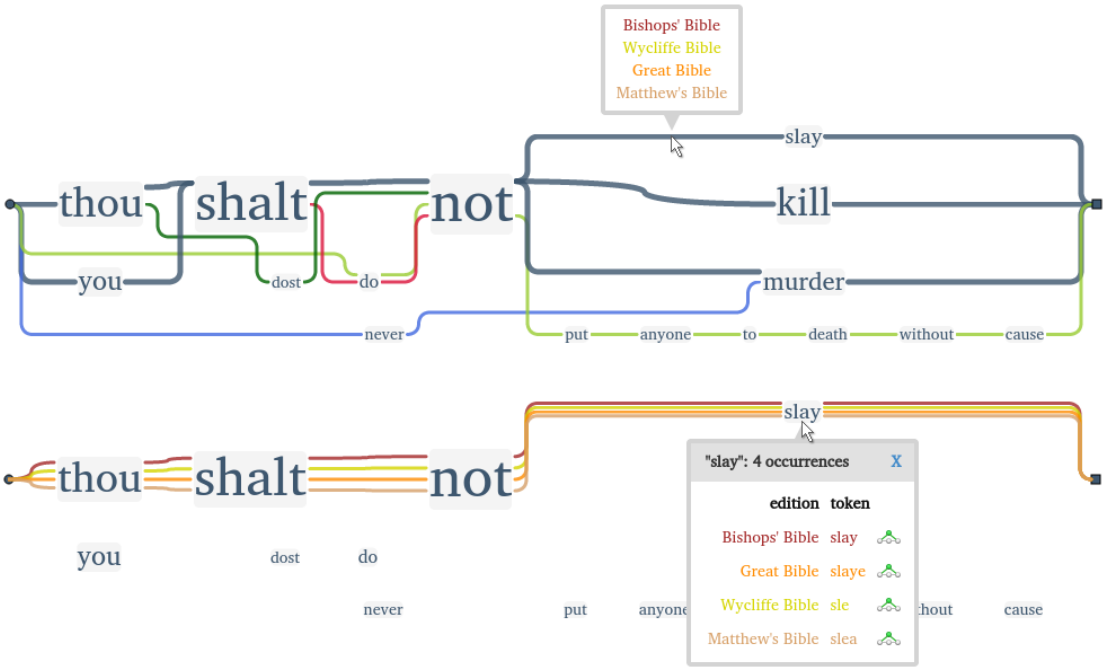


Figure 6.13: *Exodus 20:13* in twenty-four English translations with a majority of four.



(a) Using $RED = 0$.



(b) Using $RED = 0.4$. Hovering an edge reveals corresponding edition labels (top), clicking a vertex additionally lists orthographic variations (bottom).

Figure 6.14: *Deuteronomy 5:17* in twenty-four English translations using a majority of four.

Users not interested in orthographical differences may cluster those variants by manually dragging and dropping vertices, or by automatically enabling a relative edit distance of 40% ($RED = 0.4$). Not only does the latter approach save time but the user learns to quickly define the optimal percentage for the relative edit distance inasmuch as the results are simultaneously displayed. The higher the percentage, the more words tend to cluster. For the chosen examples, a relative edit distance of no more than 50% seems to yield the best results without clustering unwanted words. A higher percentage is also possible if the user unravels this cluster by manually splitting unrelated words into separate vertices. In the example above, the orthographic variations "kill/kyll" and "shalt/shall/schalt" are clustered, thus drawing more attention to the semantic dissimilarities. The sixth commandment as described in *Deuteronomy 5:17* (Figure 6.14a) showcases more variance compared to the *Exodus 20:13* version. By selecting a relative edit distance of 50% ($RED = 0.5$), we notice a high frequency of the verb "slay" and its variants "slaye/sle/slea" (Figure 6.14b). Clicking the corresponding vertex reveals no semantic dissimilarities between the contributing editions.

A *second example* (Figure 6.15) displays an output scenario for *Luke 2:1*, generated by selecting a majority of four, a relative edit distance of 40% ($RED = 0.4$), and by manually merging words and splitting vertices.
The result brings out synonyms and matching expressions in the different Bible editions. For instance, the opening expression has many variations, including "it came to pass(e)/about" and "it happened/chanced/fortuned/was (don)." The God's Word Translation (blue) does not translate this expression at all. Moreover, the translations vary between "(all) the (whole/habitable) world/globe" (the term "globe" only appears in Smith's Literal Translation) and "the roman empire," an interpretative variant of the God's Word Translation. Different expressions are also used to describe the emperor's decree, namely those stacked at the end of the verse stating that everybody should be "taxed," "enrolled," "discyrned" or "registered." Whilst 'misplaced,' variations containing the words "census" and "numbering" are nevertheless clearly discernible owing to their divergent branch structure. When brought together, all of these variant readings help convey the meaning of the decree as understood by various translators. Differences in word order can also be better understood by enabling the transposition option (see Figure 6.16): the position of the word "decree," for instance, changes between translations, from "a decree went out" to "went out a decree." Hovering the cursor over the term "augustus" displays two transpositions, "caesar augustus" and "augustus caesar." This detection leads to yet another transpo-

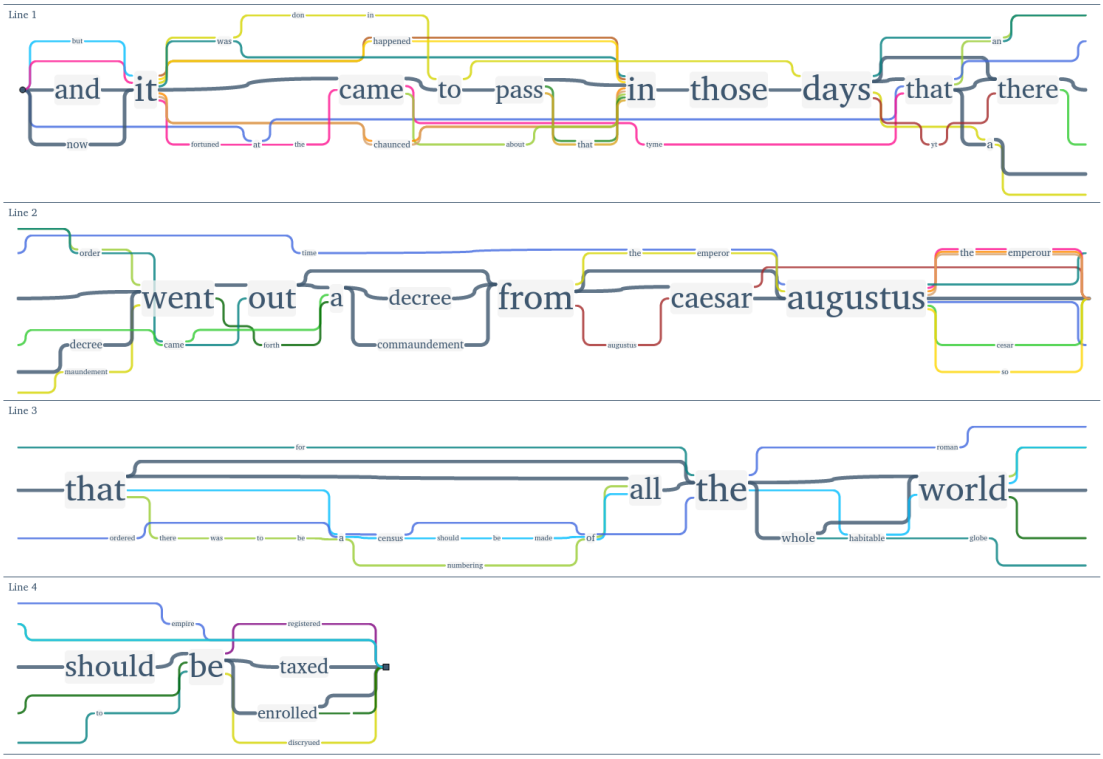


Figure 6.15: *Luke 2:1* in twenty-four English translations with $RED = 0.4$.

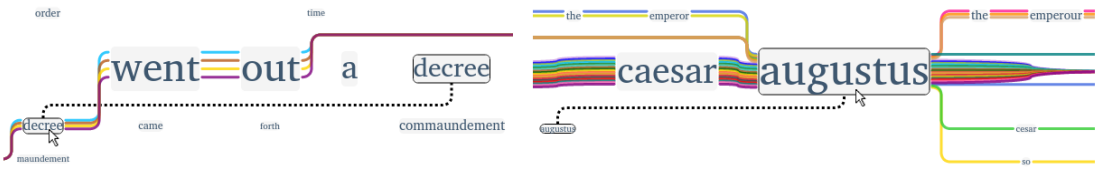


Figure 6.16: Transpositions in *Luke 2:1*.

sition pertaining to the Roman emperor stemming from the word "augustus." Due to different orthography, the terms "augustus cesar," "(the) emperor augustus" and "augustus the emperour" are not shown as variations, but can easily be identified, because they are marked as transpositions.

The *third example* scrutinizes the most influential English Bible translation: the King James Version [Ryk11]. In order to understand the development of biblical variants before the King James Version became so important, eight translations dating between 1500 and 1800 are chosen as a representative sample: the Coverdale Bible,

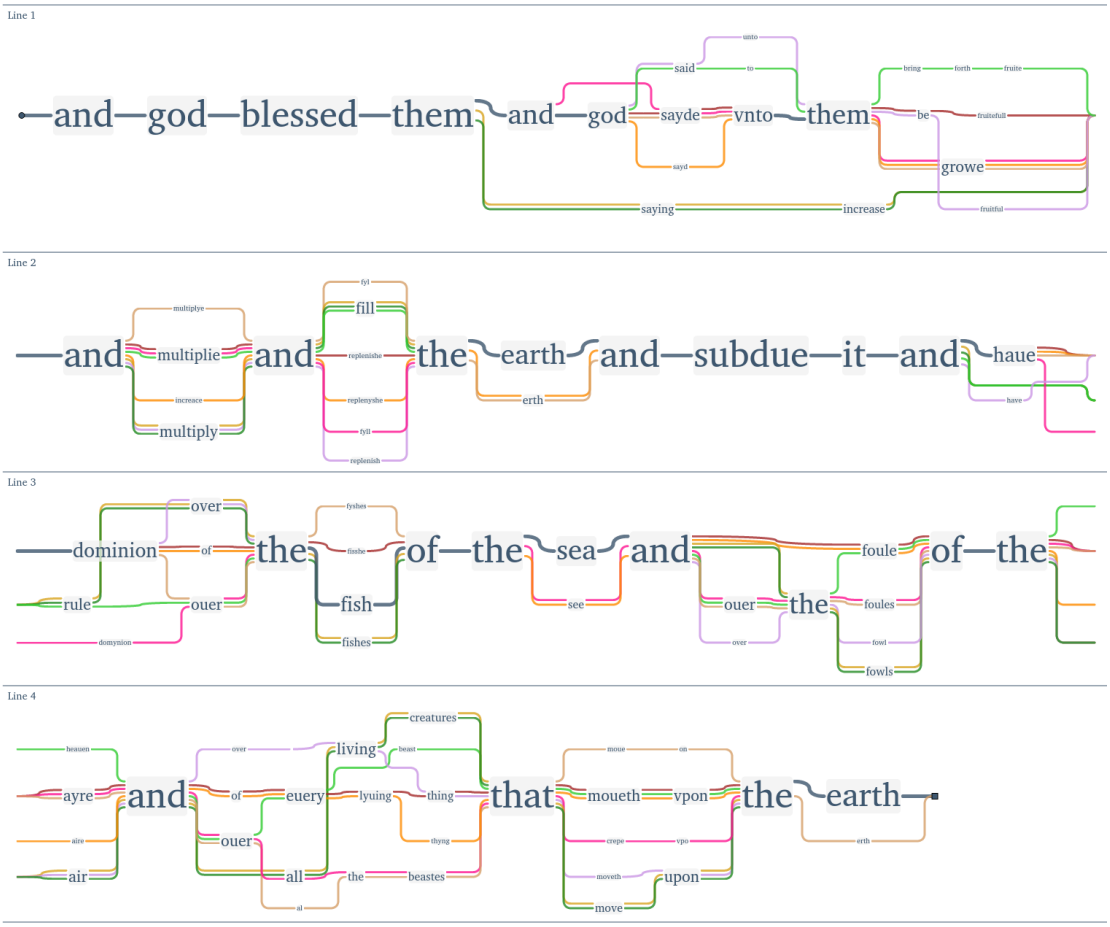


Figure 6.17: Eight English translations for *Genesis 1:28*.

the Matthew Bible, the Great Bible, the Geneva Bible, the Bishop's Bible, the Douay-Rheims Bible, the King James Version and the Douay-Rheims-Challoner Bible. Figure 6.17 shows the resulting Variant Graph applied to *Genesis 1:28*. From an orthographical standpoint, one notices the emergence of clusters. Unsurprisingly the Douay-Rheims- and the Douay-Rheims-Challoner-Bible form a cluster. It is evident, that those two editions tend to deviate from the majority. Although the Douay-Rheims-Challoner Bible is said to have been influenced strongly by the King James Version [New59], a specific closeness to this edition is not visible, at least not in this verse. The Bishop's Bible has an independent writing style and as a revision of the Great Bible, it tends to keep its content-based variations (e.g. "of the fish" instead

of “over the fish”) but uses a more contemporary orthography. But hovering for instance over the word “thing,” which only these two Bible translations share, shows how similar the Bishop’s Bible and the King James Version are in translating this verse. Figure 6.18 juxtaposes the Variant Graphs of old (1500-1800) and modern (after 1800) Bible translations both including the King James Version, whose words appear in the vertical center of the graphs. We applied a majority of five and a relative edit distance of 50% ($RED = 0.5$). Comparing the results of this particular verse with other verses corroborates the impression that the majority of both modern and older translations employed similar if not identical terminology to that used by the King James Version. The graph shows a strong line of transmission among older translations (Figure 6.18, left) and even the deviations from it seem to be done in a similar way. Here the more modern editions (Figure 6.18, right) deviate significantly from the King James Version, especially the Bible in Basic English (light green), the God’s Word Translation (blue), Smith’s Literal Translation (turquoise) and the World English Bible (brown).

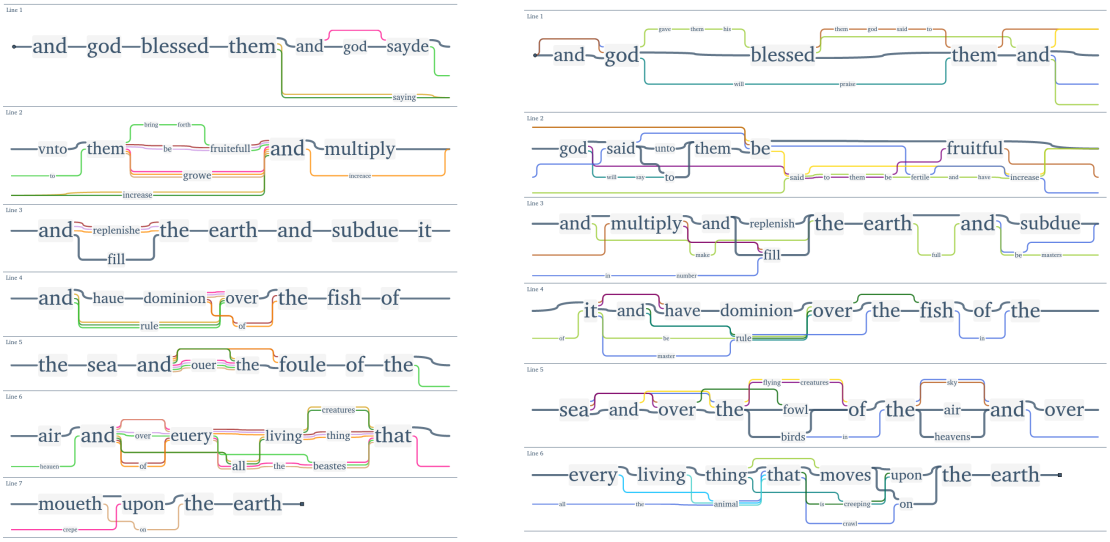


Figure 6.18: Old (left) and modern (right) English translations of *Genesis 1:28* compared to the King James Version visualized with a majority of five and $RED = 0.5$.

6.6.2 Comparing German Shakespeare Translations

The year 2014 marked the 450th anniversary of William Shakespeare's birth and the 300th anniversary of the German Shakespeare Society.[10] The joint celebration was accompanied by a multidisciplinary symposium – held at the Mainz Academy of Sciences and Literature – on the German reception of Shakespeare's works. To complement contributions from the literary and cultural sciences, the Leipzig Max Planck Institute for Mathematics in the Sciences and the Natural Language Processing Group of Leipzig University were invited to present newly developed language analysis methods and their application to Shakespearean works [EHJ15].

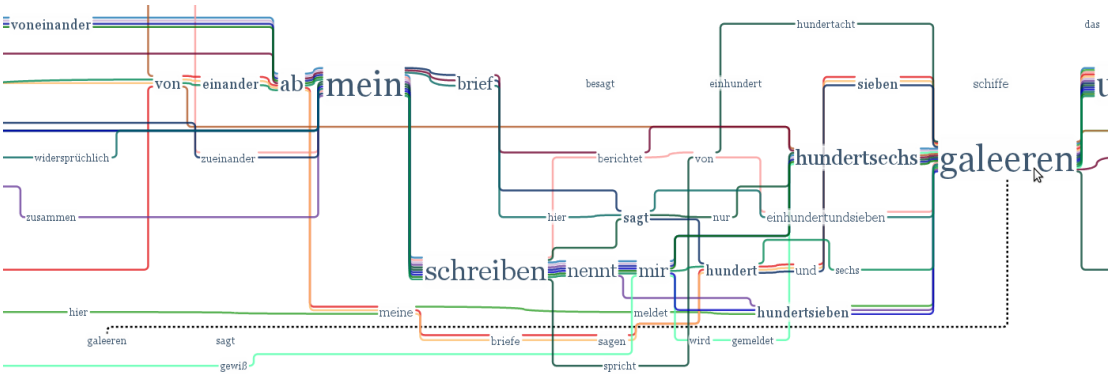


Figure 6.19: Part of Shakespeare's *Othello*, Act 1, Scene 3.

TRAViz was presented as a means of visualizing linguistic and structural aspects expressing the dramatic progression in Shakespeare's plays, and of highlighting differences and similarities between various German translations of his works. Figure 6.19 shows how varied the language can be across twenty-two German editions of the same *Othello* excerpt. Along the vertical axis, one notices word substitutions and preferred alternatives – constituting the so-called paradigmatic level – such as "brief" and "schreiben" for "letter," or "galeeren" (galleys) and "schiffe" (ships). Interestingly, the original "hundred and seven galleys" was changed to '106' nine times. This circumstance is most likely attributable to a smaller syllable count and to the softer pronunciation of "ga-lee-ren." With regards to word position and order, the so-called syntagmatic level bears fixed and grammar-related expressions, such as the plural "meine briefe" (my letters) versus the singular "mein brief" (my letter), or the passive con-

[10] Deutsche Shakespeare-Gesellschaft, http://shakespeare-gesellschaft.de/

struction “mir wird gemeldet” (is reported to me) versus the active “nennt mir” (tells me), which can be visually tracked along the horizontal axis. The overall syntax of the visualized *Othello* scene is relatively stable with the exception of the transposition of “galeeren.” The corresponding edition by Schaller (1959) chooses “a hundred and seven galleys say my letters” over the accepted “my letter says a hundred and seven galleys.”

6.6.3 Exploring the Multiple Meanings of a Term in Ancient Greek Texts

In the digital humanities *eXChange* project,[11] TRAViz is being used to align and visualize Ancient Greek text snippets containing specific terms. The humanities scholars on the team explore the various meanings of the given term defined by a set of descriptors. Their intent is to learn how these meanings were transmitted and how they changed over time.

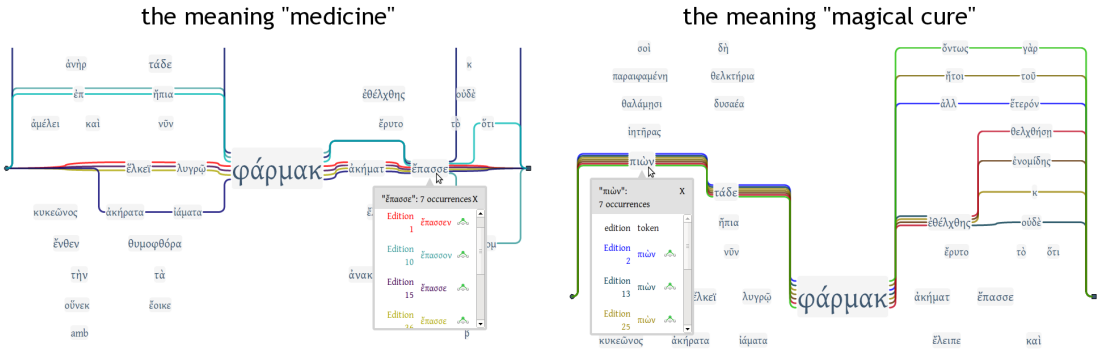


Figure 6.20: The two meanings of *φαρμακον*.

Eva Wöckener-Gade, a classical philologist, is interested in the multiple meanings of the term *φαρμακον*. A visualization of text snippets containing the truncated form *φαρμακ* yields clusters representing the different meanings, all easily traceable by hovering over vertices and highlighting branches that include related words (Figure 6.20). The word *ελκει* (wound) and the variations of *ακεσματα* (cure) interpret *φαρμακον* as “medicine,” whereas the word *εθελχθης* (aorist form of the verb ’to enchant’) supports the meaning “magical cure.” Both meanings are to be found in Ancient Greek poetry. By selecting a relative edit distance of 30% ($RED = 0.3$), TRAViz bundles up various versions of the verb “to apply” (*επασσεν*, *επασσον*, *επασσε*), a term often

[11] http://exchange-projekt.de/

used in the context of medication. Similarly, *πιων* suggests that the magical cure was drunk.

Contrarily to the traditional methods of analyzing the contexts of a given term's occurrence, this visualization facilitates a rapid comprehension of the term's different meanings by clustering the related tokens that define them.

6.6.4 Syllable stress

Dr. Michael Cade-Stewart, a British Academy Postdoctoral fellow at Kings College London whose main interest lies within the digital exploration of Poetic Rhythm between 1800 and 1970, analyzes and visualizes different ways of orally performing poems. An example of six different possibilities for a pentameter line (Shakespeare's *Sonnet 18*, line 1) is shown in Figure 6.21. The stressed syllables are in upper-case, the unstressed in lower-case. The likelihood that a syllable will be stressed in performance is indicated by the size of the word, and by the number of colored strands that run through it. The words "a summer's day" are performed equally in all versions, and can therefore be regarded as relatively objective; the other syllables are more subjective. If one wanted to look at a feature in the stressed syllables, one might restrict one's focus to the more objective variants or, in this case, weigh them more strongly.

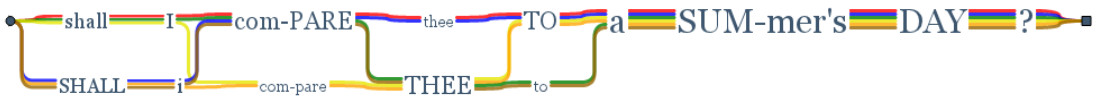


Figure 6.21: Six ways of stressing Shakespeare's *Sonnet 18*, line 1.

6.7 Distant Reading Visualization for Variant Graphs

Designing a Variant Graph with TRAViz as outlined in the previous sections, the humanities scholar is able to analyze both similarities and differences among various text editions on sentence level. Although TRAViz can also be used for larger text entities such as sections or chapters, the resultant visualizations are hardly readable and an analysis of the Variant Graphs becomes a laborious task. Therefore, TRAViz remains a close reading visualization tool only to be used when the user has a specific research question for a desired text part, such as the analysis of different Bible translations of the sixth commandment (see Section 6.6.1). Research questions like "For which Bible books the various translations are most similar?" or "For which chapters

of book X the rather similar translations A, B and C differ most and why?" cannot be answered with TRAViz. As Moretti suggests, such kind of questions require distant reading approaches to be answered [Mor05]. This section proposes a visualization that offers a distant view on Variant Graphs in order to support the humanities scholars in detecting and analyzing occurring patterns on higher text hierarchy levels as the sentence level.

We extend the Bible use case from Section 6.6.1 by designing a distant reading visualization for Variant Graphs. The Bible is a very good use-case for this purpose. Firstly, it is a very influential and well-known text, which supports easy evaluation of results. Secondly, it's structure includes a four-level hierarchy that makes views of varying distance on the text possible: the Bible (level 1) consists of books, each book (level 2) is divided into chapters, and each chapter (level 3) is composed of verses. Each verse (level 4) can be visualized using TRAViz, but all other hierarchy levels require a distant reading solution. Thirdly, our Bible corpus includes editions in Middle and Modern English, a great variety of translation dependencies (e.g., editions based on the King James Version) as well as a diversity of translation motivations (e.g., simplified language in the Bible in Basic English). The versatility of the corpus allows for a multitude of research questions to be asked by the collaborating humanities scholars.

6.7.1 Visualization Design

Figure 6.22 shows a screenshot of the distant reading visualization. Arranged in columns, the top panel lists all available Bible editions either sorted by year of publication or in alphabetical order. On demand, the user is able to compose a desired set of editions by clicking the corresponding checkboxes. Below the Bible editions, the humanities scholar is able to tweak the visualization dependent on the given research question. In particular, the user can:

- define a threshold value maj for the majority of editions,
- either highlight text passages that are similar or dissimilar to a certain extent from maj, and
- adjust the percentage of words p that need to be similar or dissimilar regarding maj to highlight a text passage.

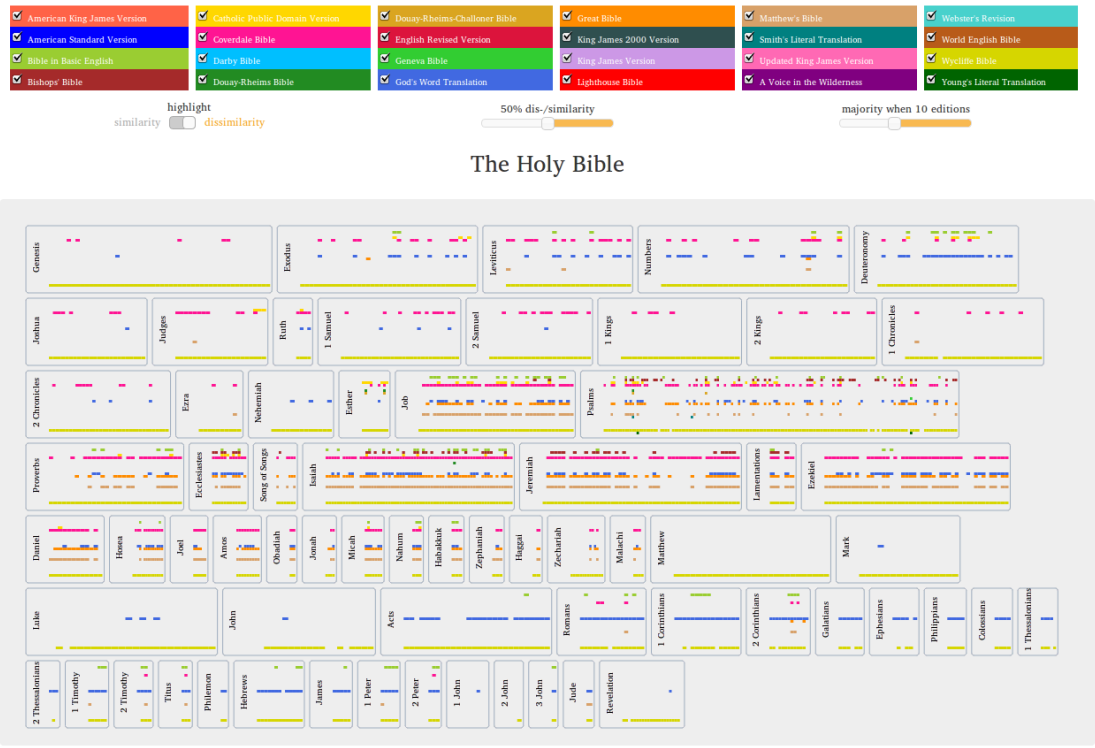


Figure 6.22: Distant reading of twenty-four Bible editions.

The bottom panel visualizes a "(dis-)similarity fingerprint" for all twenty-four editions of the Bible on level 1; the user can interactively navigate between the different hierarchy levels. Each Bible book receives a rectangular block with its width reflecting the number of chapters. According to the configuration in this example, a tiny rectangle for a chapter of an edition is drawn in its assigned color, if at least 50% of its contained words differ from the majority of at least ten editions. The resultant pattern reflects, e.g., three salient editions for the New Testament, which reveals individual translation styles: the Wycliffe Bible (1380, dark yellow) is the only translation in Middle English, whereas the God's Word Translation (1995, blue) and the Bible in Basic English (1949, light green) aim to be understood very easily nowadays, and thus, choose to deviate from other editions that tend to be more antiquated and sophisticated in language and style.

The distant reading visualization is based on the Variant Graphs computed for each Bible verse. With the user-defined majority setting, we receive two types of vertices

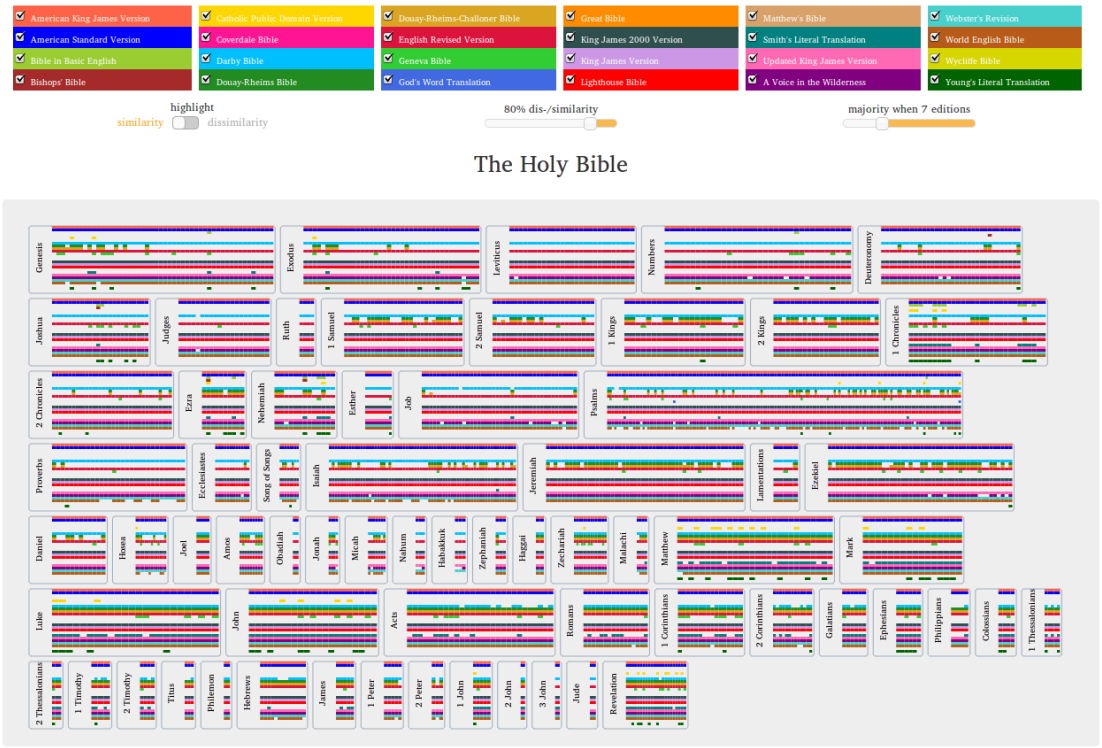


Figure 6.23: The Bible (level 1) visualizing similarity.

according to the graph's alignment structure (see Section 6.1, step 3): (1) majority vertices v_{maj} with $|v_{maj}| \geq maj$, and (2) vertices v with $|v| < maj$. Suppose, the humanities scholar analyzes the dissimilarity among the Bible editions. On chapter level 3, each block is a verse and the tiny colored rectangles then represent words that are not part of the Variant Graph's majority vertices. On book level 2, each block is a chapter and a rectangle, which represents a whole verse, is now colored if the percentage of majority vertices of the verse is smaller than p. The same procedure is done on level 1. Each block is a Bible book and small colored rectangles are whole chapters with a percentage of majority vertices smaller than p.

6.7.2 Usage Scenario

Concentrating on similarity (80% similarity, majority of seven) highlights Bible editions that are very similar in almost all chapters of all books (Figure 6.23). Most of these editions are based on the King James Version, which is known as the most

influential translation [Ryk11]. But surprisingly, albeit claiming to be "as exact a translation as possible" in Modern English from the original languages Hebrew and Greek, the Darby Bible (1890, light blue) is also highlighted as very similar.

To determine the role of the Darby Bible, we remove all Bible editions not (apparently) connected to the King James Version. Now, clicking the book *Matthew* and highlighting dissimilarity (50% dissimilarity, majority of seven), we detect a cluster of derivations for the verses 16-18 of chapter 7 (Figure 6.24). As implied on the book-level, the chapter-level confirms the Darby-deviations by highlighting the corresponding passages in the individual verses based upon a majority of four (Figure 6.25).

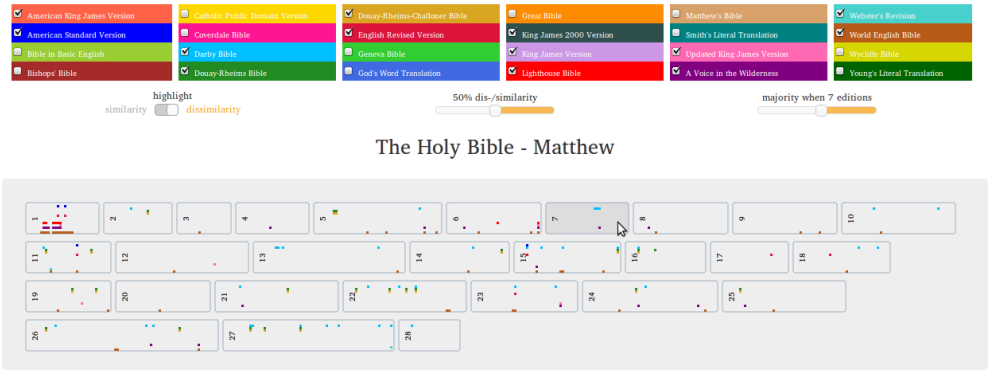


Figure 6.24: Dissimilarity among selected editions in *Matthew* (level 2).

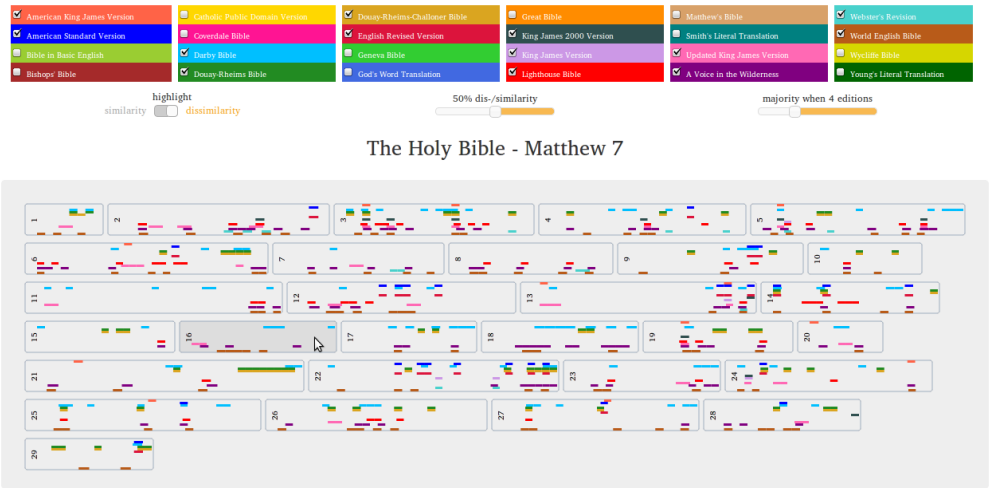


Figure 6.25: Dissimilarity among selected editions in *Matthew 7* (level 3).

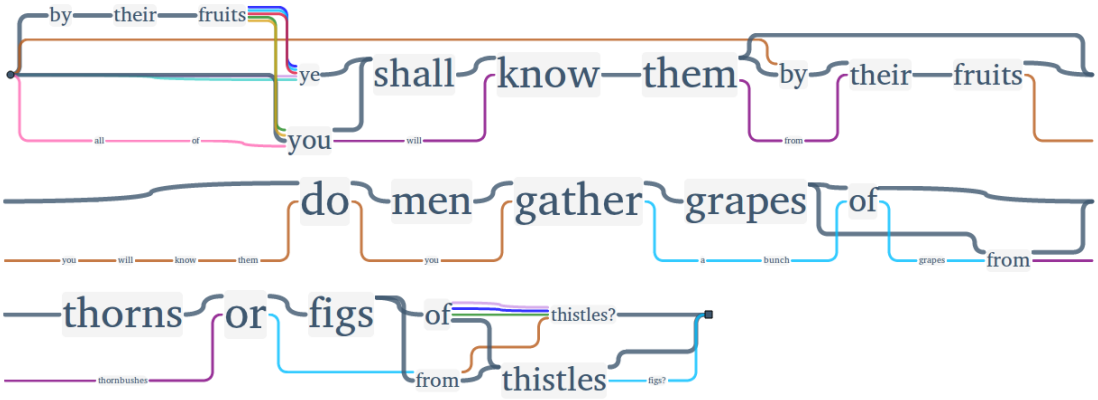
Indeed in *Matthew 7:16* (Figure 6.26a), Darby and some other editions differ in word order (but not so much in the translations); among others he choses the elder word "ye" instead of "you" and as the only translation he writes "a bunch of grapes" instead of "grapes" and instead of "figs of thistles" he writes "thistles figs?" In *Matthew 7:17* (Figure 6.26b) and *Matthew 7:18* (Figure 6.26c) the main structure of the verse remains in Darby, but some words are translated differently: instead of "bring forth" Darby uses the word "produce" (followed by World English Bible [2000, brown] and A Voice in the Wilderness [2004, purple]), the word "nor" instead of "neither" (followed by A Voice in the Wilderness), and when fruit are described, Darby uses the not morally associated adjective "bad" instead of "evil," and "worthless" instead of "corrupt".

All in all, with this setting even when choosing passages marked as dissimilar, the close view confirms what the distant view implied, which is, how close the Darby Bible sticks to the other editions concerning word order, language and most of the words themselves. Most deviations are single words that are substituted by synonyms or something similar.

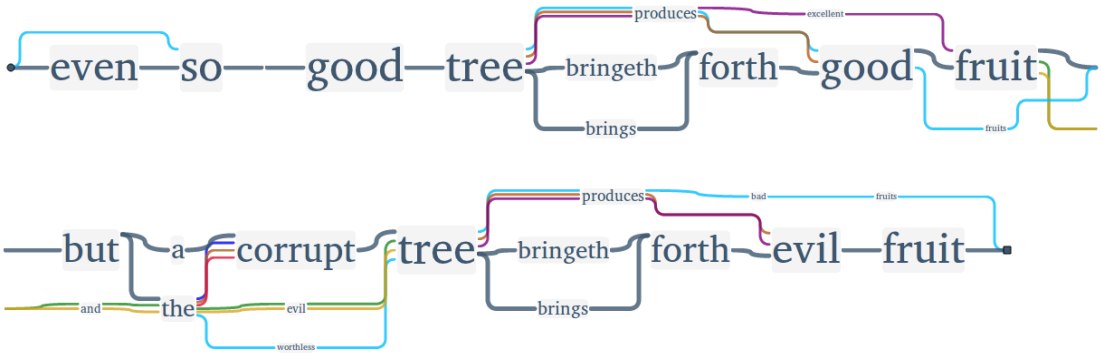
Nevertheless, the Darby variations tend to be the very obvious ones, next to those of A Voice in the Wilderness, but rarely seem to change the meaning of the text in a significant way. Thus, it seems that the Darby Bible, which tried to translate the ancient languages as exactly as possible, may have had a significant impact on later translations based (not only) on the King James Version.

6.8 Summary

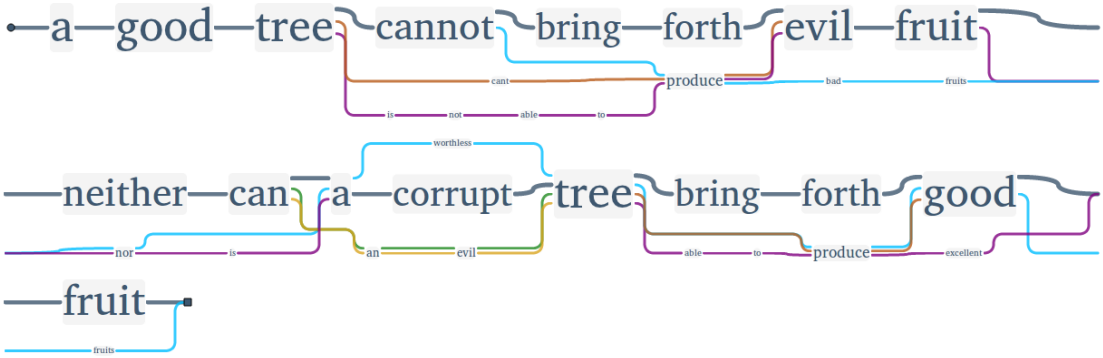
This section introduced close and distant reading visualizations for Variant Graphs. The web-based library TRAViz implements a layout algorithm and a novel design, both tailored to support the close reading of Variant Graphs. Furthermore, we demonstrated TRAViz' various means of interaction, e.g. the relative edit distance, to facilitate an on-demand modification of the underlying alignment. Finally, the use cases discussed have served to prove the intuitiveness of the design and the applicability of the tool to distinct studies in the humanities. It has been shown that TRAViz is suited for working with various text corpora. An approach to work interlingually should prove very interesting as well, aligning words of similar meaning in different languages, for example translations of the same text (e.g. the Bible) into different languages. This could enrich dictionary databases, show different translation approaches as well as reception or even be useful for language acquisition.



(a) *Matthew 7:16*



(b) *Matthew 7:17*



(c) *Matthew 7:18*

Figure 6.26: Close reading of three verses of *Matthew 7* (level 4).

By and large, the Variant Graph model works well on smaller text entities. The Bible use case adequately fits this model inasmuch as the highly structured text facilitates verse by verse collation. In addition, transpositions can be analyzed with tools such as TRAViz or CollateX. But when the text passages are long, the probability of transpositions drifting apart and of unaligned text snippets appearing rapidly increases. While the structure of the Variant Graph still works, an exhaustive visualization and analysis is almost impossible. One solution would be to modify the model so that it accepts cycles, thus ensuring that similar patterns retain alignment independently of their order in the corresponding source texts. As this approach is only applicable to moderately larger text entities, we instead suggest visualization methods based upon Text Re-use detection such as introduced in Chapter 5.

In addition to the close reading of Variant Graphs with TRAViz, we presented a novel technique for the distant reading of Variant Graphs for a potential high number of editions. Reading all those editions would require a huge amount of time, and comparing them would take even longer. Being able to see similarity and dissimilarity on various text hierarchy levels enables the user not only to save time by directing to passages that could be of interest, but it can even raise new worthy research questions. This way of visualizing bears even the potential to cause serendipity by showing relations that have not been seen yet. With the provided top-down combination strategy, the scholar is able to detect specific patterns in the distant reading view, and to investigate generated hypotheses with TRAViz in the close reading view. Although designed for the Bible use case, this technique is applicable to other hierarchically structured texts. Visualizing the (dis-)similarity between various editions of Homer's epics could be one of the interesting examples. Being an important work for philologists, we could also extend the visualization to display the occurring transposed verses.

The music is not in the notes, but in the silence between.

Wolfgang Amadeus Mozart

7

Interactive Visual Profiling of Musicians

The digitization age changed strategies and methods to gain knowledge in the humanities essentially. In terms of philology, humanities approaches were traditionally oriented primarily on language and text. As retrieval strategies in printed encyclopedias were mostly based on the alphabetical order of contained names, the now available access to large relational databases provides the opportunity for humanities scholars to filter groups of entities not only based on names, but also on various other data facets. But the pure access to a database containing a multitude of information also reveals limitations when it comes to investigating concrete research questions beyond only showing lists of entries that match a given database query. Especially for humanities scholars who are not used to exploit the cardinality of query languages, it becomes hard to navigate large databases and filter results accordingly [GO12]. Also, with experiences of the so-called visualistic turn [SH93] in the early 1990's, humanities scholars more and more wished to explore complex issues not only based on texts, but also based on images.

An example of this process is given by the Bavarian Musicians Encyclopedia Online[1] (German: "Bayerisches Musiker-Lexikon Online (BMLO)"), a powerful web-based interface for musicologists, as it provides access to information about around 30,000 musicians, extracted from various digitized printed media. Being an invalu-

[1] ed. Josef Focht. http://www.bmlo.lmu.de/

able tool in the daily work of musicologists, this innovative access to music history heritage raised new research questions in musicology. Besides the visualization of a musician's profile, the major desire was the ability to find similar musicians based upon the attributes of a musician of interest, in short: *profiling*.

In order to develop a valuable visual analytics system capable of determining similar musicians interactively, we closely collaborated with those musicologists who originally raised the profiling research question. Initially, we discussed possible profiling tasks and various aspects of how the musicologists imagined profiling workflows. We transformed the provided data for profiling purposes, mapped required musicians' attributes to visual features, and defined similarity measures to be used as the basis for profiling similar musicians. In summary, the contributions of this chapter are:

- **The similarity of person attributes**: For various attributes of musicians, we designed similarity measures in accordance to musicological conceptions.
- **Profiling system**: Based upon eight similarity measures and a semi-automatic weighting of attribute dimensions, we designed the visual analytics system *MusikerProfiling*[2] that is used by musicologists to iteratively search for musicians similar to a musician of interest.
- **Column Explorer**: For the comparison of temporal and textual metadata of musicians, we provide a column based representation that borrows ideas from Jigsaw's list view [SGL08] and Parallel Tag Clouds [CVW09] in order to explore correlations among various attributes.
- **Comparative visualization of musician profiles**: Consistent of the Column-Explorer, a map and a social network graph, the system allows for the analysis of similarities and differences between the biographical characteristics of various musicians.
- **Temporal uncertainty**: We consider existent temporal uncertainties when fixing a musician's activity time, and we provide a design that communicates these uncertainties to the musicologists.

We emphasize the utility of our visual analytics profiling system for musicologists by providing various usage scenarios. In a storytelling style, each scenario exemplifies

[2] http://profiling-musicians.vizcovery.org

how the profiling system can be used to discover unexpected insights. Additionally, we report experiences gained during our digital humanities project. This includes the iterative evaluation of our profiling system with musicologists, limitations due to the nature of humanities data and future prospects.

7.1 Related Work

The database of our visual profiling approach is based upon the textual contents of numerous digitized documents about music history. Many visualization techniques with a motivation from the humanities also provide abstract views on digital text contents, an overview is given in Chapter 2. This work on a profiling system for musicians is related to two further research areas. As the usual method of profiling is based upon similarities to a given subject, we first take a look at recommendation systems, where recommendations are determined in dependency on an object of interest. Second, we list visual analytics and information visualization papers that (1) are motivated by similar research questions to ours, or (2) present similar visualization techniques to the ones our profiling system is composed of.

7.1.1 Recommendation Systems

Tapestry, the first recommender system developed in 1992, was based on collaborative filtering [GNOT92]. The system aggregates recommendations provided by users, and directs them to appropriate recipients. In the last two decades, many different recommendation algorithms have been developed [KR12]. Basically, the initial step of our profiling approach imitates recommendation based on item similarity [LMY+12]. In our case, the item is a musician, and the system determines similar musicians as recommendations. The further profiling steps of our system are rather comparable to content-based recommendation systems [PB07]. An item is recommended based upon description and the "profile of the user's interests," which are user-defined similarity weights and mandatory musician attributes in our case. The benefit of using weights to reflect the importance of certain attributes for users has already been shown for discovering related movies [DGM08]. Other applications include the development of recommender systems for YouTube videos [DLL+10], music files [CC01], or related people in social networks [CGD+09]. The recommendation of social media data proposed by Guy et al. [GZR+10] is based on relationship information among people. Our system uses this idea and determines a relationship similarity between musicians.

7.1.2 Related Visual Analytics & InfoVis Techniques

The visualization of relational, geospatial, temporal and nominal textual data is a common task for researchers in visual analytics and information visualization. In accordance to the features of our profiling system, we outline the most related works under various aspects.

Visualizing recommendations Many works provide visualizations for recommendations of different type. Choo et al. present a recommendation system for a vast collection of academic papers [CLK+14]. Recommended documents can be explored in a scatterplot. Overview is a system that also supports the systematic analysis of large document collections to be used by investigative journalists [BISM14]. Documents are hierarchically clustered based on content similarity, thus, can be recommended during investigation. A recommendation of notes from past analysis tasks when operating a visual analytics system is explained by Shrinivasan et al. [SGL09]. For the visualization of recommended movies, Vlachos uses a radial layout [VS12]. Gansner offers geographic maps that use the metaphor of neighborhoods and clusters to group related recommendations [GHKV09]. The latter approach communicates relations among recommendations. That a recommendation system is not seen as a black box from the user's point of view [VPBD13], we provide visual and textual cues for the collaborating musicologists to explain the existence of a recommendation.

Social network visualizations In our profiling system, relationship graphs visualize the social networks of musicians under inspection. Especially interesting are unknown musicians that connect unacquainted musicians. Applied to the relationships of characters in literary works, Euler diagrams can be used to visualize groups of social networks in clusters while also showing relations among clusters [RD10]. Weaver embeds an attribute relationship graph in his visual analytics environment to visualize the relationships between movie actors [Wea10]. Quite often, visualizations attend to the matter of visualizing large social networks and their navigation by various interaction means [PS06]. The design of such visualizations is especially important for the exploration of online social networks [HB05, GOB+10].

Multiple views Our system consists of multiple views that visualize the musician's metadata information, e.g., we provide visual representations of temporal, geospatial, and nominal textual data. The multiple views concept is often used to communicate

various data aspects. An overview of various visual analytics approaches to dynamically explore spatio-temporal data with the help of maps and timelines is given by Andrienko et al. [AA05]. The purpose of such visual analytics environments ranges from the analysis of crime incidents [JME+12] to the extraction and characterization of significant places from mobility data [AAH+11], and the visualization of semantic web data [CDC+07]. Also popular is the additional visualization of contextual keywords – next to map and timeline – in the form of tag clouds, for instance, to visualize topical metadata [DCCW08] or to support the discovery of meaningful events in news and social media data [DWS+12]. Many approaches are based on textual data sources. The cross-filtered views for multidimensional data sets as proposed by Weaver include various interfaces to be used as filters to determine potentially interesting events in newspaper article collections [Wea08]. Heimerl et al.'s visual analytics system is also composed of many views with the purpose of interactively training classifiers to be used for document retrieval on large text collections [HKBE12].

Layered Textual Metadata One of the major components of our system is the Column Explorer (see Section 7.4.2) with a column per data facet showing textual attributes of musicians. Jigsaw's list view [SGL08] provides the basic idea for such a visualization. The user can select an arbitrary entry and correlations to other attributes are shown. Related data entries in adjacent columns are linked. Similarly, this is done in Parallel Tag Clouds [CVW09]. Each column lists tags of a certain time slice, and equal tags are linked on selection. PivotPaths is a yet similar approach to ours [DRRD12]. After selecting a research paper, links are drawn to related authors and paper keywords. The related attributes in our Column Explorer are connected with colored streams. Often, streams express a temporal evolution of events [BW08, CLWW14]. Tags at certain positions in streams can be used to illustrate contextual information [SWL+10].

7.2 Digital Humanities Background

This research bears on musicology, a field of the humanities that observes musicians and their achievements. This includes not only composers, although a *composition* is seen as the fruit of a musical process. Moreover, many other musical professions are in the focus of interest of musicological research, e.g., instrument makers, conductors, singers, instrumentalists, music publishers, etc.

Motivation The Bavarian Musicians Encyclopedia Online (BMLO) project was initiated in 2004 with the goal to create a database that combines a multitude of biographical information about musicians of various professions. In cooperation with the Bavarian State Library and the Society for Bavarian Music History, musicologists of the Ludwig Maximilian University of Munich searched, collected and digitized related documents on music history. They combined biographical information about musicians extracted from various sources such as encyclopedias, periodicals, and series concerning musicology as well as research papers from musicology, history and science of art. A web-based platform provides access to the database, which contains musicians who are part of the Bavarian music history; musicians with an active lifetime period living in Bavaria as well as musicians with a considerable influence on the Bavarian music history are included. Despite the prior focus on Bavaria, the BMLO is a valuable tool for many musicologists as it provides information about 28,137 musicians from all musical eras, spanning a time range from 4AD to the present. Working with the BMLO, the main interests are not examinations of the musicians' achievements – musicologists rather explore the features of musical professions or analyze the biographies of musicians. This includes generic research questions concerning the geographical or temporal evolution of musical professions as well as precise research questions that focus on an individual musician. One such research question – the profiling of musicians with similar careers to a musician of interest – is interesting for musicologists for a long time. Traditionally approaching an answer to this type of question, musicologists solely refer to musicological editions and monographs. But musicology primarily focuses on fifty musicians – mostly composers – and their main works. Due to this inhomogeneous state of research (what we call the *popularity* of musicians), a traditional similarity analysis usually ends within this limited set of musicians. Although the BMLO provides an immense diversity of information about a large number of musicians, the profiling of musicians is not supported. Maybe the database could be used to address some research questions, but for musicologists complex database queries are hard to formulate [GO12] and the musicians' attributes in the query result are hard to analyze and to compare. Therefore, musicologists desired a system that allows to approach a profiling task interactively with the aid of visual interfaces that pictorially illustrate the provided information.

Digital Humanities Project In collaboration with musicologists using the BMLO, we developed a visual analytics system that supports the profiling of similar musicians

based on a selected musician of interest. For the implementation of this project, we adopted several suggestions made by Munzner [Mun09] to ensure designing a beneficial, powerful tool that supports answering the posed research questions. We furthermore took collaboration experiences from other visualization researchers who worked together with humanities scholars into account to avoid typical pitfalls of such interdisciplinary projects. Additionally, we worked through related works in the digital humanities, which provide valuable suggestions and guidelines for designing interfaces for humanities scholars, e.g., as outlined in [GO12]. To avoid making assumptions for the design of a profiling system that is hard to comprehend and does not solve the concerned musicological research questions, we initially discussed the needs of the musicologists, their workflows and challenges in the targeted domain in several meetings. Furthermore, we presented and discussed related visualization techniques to convey an impression of the capabilities and challenges within our research field. The musicologists explained how they use the BMLO in their workflows and communicated their fascination about this unique type of encyclopedia invaluable for their daily work. This get together turned out to be important to understand each others mindsets. A major outcome was a set of research questions on the profiling of musicians and the analysis of musician profiles.

Project Data The provided database and aspects of data transformation were also discussed with the musicologists. This included data anomalies, the conversion of temporal metadata to a uniform scheme while considering occurring uncertainties as well as defining popularity values by counting a musician's references. In discussions about the provided musician attributes we could separate attributes worth to integrate into the profiling process – a musician's sex, lifetime data, places of activity, musical and further professions, relationships, divisions and denomination – from irrelevant ones. For instance, the potentially interesting attribute *nationality* is only provided for 416 musicians (1.5%) as most musicians lived in a time when the assignment of a nationality to a person did not exist. Therefore, we decided to exclude nationalities from the profiling process. The musicologists argued that most research interests concern musicians without nationality attributes. We also asked for the relevance of each attribute dimension for a profiling task and the comparative analysis of musicians in order to push the development of the profiling system the way that predominant attributes receive more attention. Additionally, we gained information how musicologists imagined to operate with the musicians' attributes. For example,

they wanted to see how attributes of different facets correlate, and they wanted to detect links between unrelated musicians.

Project Challenges To solve the profiling task, we faced two main challenges. On the basis of relevant musician attributes, we first needed to define various similarity measures that determine the similarity of musicians (see Section 7.3). Second, we needed to design visual interfaces that communicate these similarities intuitively. In preparation, we looked at related visualizations and collected possible representations to map relevant attributes of musicians to visual attributes. In meetings with the musicologists, we argued on opportunities and drawbacks when applying various visualization techniques. The resultant visualization design is explained in Section 7.4. Finally, various usage scenarios illustrate the utility of the profiling system for the collaborating musicologists (see Section 7.5), now capable of detecting similar musicians without the bias of popularity. As further demands included the visual exploration of individual musician profiles and the comparative analysis of multiple profiles, we also provide an example besides profiling.

7.3 The Similarity of Musicians

Based on various biographical information, the similarity $S(m_i, m_j)$ between the musician of interest m_i and a similarity candidate m_j is determined as

$$S(m_i, m_j) = w_p \cdot P(m_j) + \sum_{k=1}^{8} w_k \cdot S_k.$$

w_k is a weight for the relevance of the corresponding similarity S_k. To mimic the traditional profiling approach by referring to musicological editions and monographs, we insert the popularity $P(m_j)$ of m_j as a further component into the similarity equation. The collaborating musicologists define $P(m_j)$ in dependency on the number of publications (articles, editions and media) from and about m_j. Thus, the popularity reflects the current state of research on a musician. According to this heuristic, Wolfgang Amadeus Mozart is the most popular musician with around 150,000 publications. Taking the musicologists' suggestions into account, we group all musicians with the same number of publications into groups $g_1, \ldots, g_n$ sorted by ascending publication count. The popularity $P(m_j)$ is then defined in dependency on m_j's

popularity group g_k as

$$P(m_j) = \frac{k}{n}.$$

w_p can be used to adjust the influence of popularity during the profiling process. Using $w_p = 0$ disregards popularities and $w_p = 1$ mimics the traditional profiling approach. All weight values are defined interactively during the profiling process. In the following, we outline the calculation for each of the eight contributing similarities. Some of the similarity measures are defined by the Jaccard index like

$$S_i(m_i, m_j) = J\Big(f(m_i), f(m_j)\Big) = \frac{|f(m_i) \cap f(m_j)|}{|f(m_i) \cup f(m_j)|}.$$

7.3.1 Sex Similarity S_1^{sex}

For some research questions of the collaborating musicologists, the sex of a musician plays an important role when determining similar musicians. Such an information in the form of *male* or *female* is provided for nearly all musicians (27,403 $\hat{=}$ 97.4%). If the sexes of m_i and m_j are equal, we define $S_1^{sex}(m_i, m_j) = 1$. For unequal sexes or if the sex of one musician is unknown, we use $S_1^{sex}(m_i, m_j) = 0$.

7.3.2 Activity Time Similarity S_2^{tem}

The activity time of a musician is defined in dependency on the temporal metadata provided for nearly all musicians of the database (27,681 $\hat{=}$ 98.4%). Three various datings may be given for a musician: a dating of birth B (provided for 27,357 musicians $\hat{=}$ 97.2%), a first mentioned dating F (25,592 $\hat{=}$ 91%), and/or a dating of death D (18,610 $\hat{=}$ 66.1%)

The musicologists exploited the underlying textual sources of the database the way that the first mentioned dating is always an evidence for an active phase of a musician, thus, always ranges between birth and death. The granularity of the given datings ranges from date to year. Due to uncertain information in the textual sources, the given datings are often imprecise. Three types of uncertainty occur: *before* datings (e.g., before 1745), *around* datings (e.g., around March, 1745), and *after* datings (e.g., after September 22, 1745).

In order to process uncertain datings for the purpose of defining and visualizing the activity time for each musician, the collaborating musicologists provided a taxonomy – based on state-of-the-art knowledge in musicology – to map uncertainties to years

uncertainty	dating year	difference
before/after	≤ 1700	$-/+30$ years
	$1701-1800$	$-/+25$ years
	$1801-1900$	$-/+10$ years
	>1900	$-/+5$ years
around	≤ 1500	± 20 years
	$1501-1600$	± 15 years
	$1601-1700$	± 8 years
	$1701-1800$	± 5 years
	$1801-1900$	± 3 years
	>1900	± 2 years

Table 7.1: Mapping of uncertain datings.

as approximate datings. Table 7.1 lists how various uncertain datings are resolved in dependency on centuries. For all *before* and *after* datings we add or subtract the given difference value. For datings with an *around* uncertainty, we subtract the difference value for births, add the difference for deaths, and for first mentioned datings we define F_{min} by subtracting and F_{max} by adding the difference value. In few cases, irregularities occur after resolving uncertain datings. In case of $F_{min} < B$ (or $F < B$) we set $F_{min} = B$ (or $F = B$), and if $F_{max} > D$ (or $F > D$) we set $F_{max} = D$ (or $F = D$).

The activity time $t(m) = \{t_{min}(m), t_{max}(m)\}$ of a musician m is determined based upon the given dates as follows:

- **if** F or F_{min} and D are defined and unequal, we set $t_{min}(m) = F$ or $t_{min}(m) = F_{min}$ and $t_{max}(m) = D$
- **else if** F_{min} and F_{max} are defined, we set $t_{min}(m) = F_{min}$ and $t_{max}(m) = F_{max}$
- **else if** F is provided, we define F_{max} by applying the *after* uncertainty to F and use $t_{min}(m) = F$ and $t_{max}(m) = F_{max}$
- **else if** B and D are provided, we use $t_{min}(m) = B + 20\,years$ and $t_{max}(m) = D$

In the rare cases if only B or only D are provided, the definition of an activity time range is too hypothetical according to the musicologists. In such cases, the corresponding similarity is always $S_2^{tem}(m_i, m_j) = 0$. In case of two valid activity time

ranges, we define $S_2^{tem}(m_i, m_j)$ using the Jaccard index as

$$S_2^{tem}(m_i, m_j) = J(t(m_i), t(m_j)).$$

7.3.3 Activity Region Similarity S_3^{reg}

The database contains places of activity where musicians lived or worked for a certain period of time. At least one such place is provided for 26,101 musicians (92.8%). For the most often occurring 1,661 places, geographical coordinates as longitude/latitude pairs and hierarchical place IDs for the contemporary political belonging of a place are given.

The activity region of a musician consists of all places of activity. The similarity $S_3^{reg}(m_i, m_j)$ between the activity regions of m_i and m_j is determined taking the political belongings as well as the geographical positions of the musicians' associated places into account. For this purpose, we define the two measures Political Distance D_{pol} and Geographical Distance D_{geo}.

Political Distance D_{pol} The (contemporary) political distance $D_{pol}(p_1, p_2)$ between two places p_1 and p_2 is defined in dependency on hierarchical place identifiers provided for each place. The level of detail of such an identifier varies from one (only continent) to seven. Examples are listed in Table 7.2. For most places, at least five hierarchy levels are given. Therefore, we define $D_{pol}(p_1, p_2)$ dependent on k first equal hierarchy levels as

$$D_{pol}(p_1, p_2) = \frac{k}{5}.$$

Geographical Distance D_{geo} To determine the geographical distance $D_{geo}(p_1, p_2)$ between two places $p_1 = \{x_1, y_1\}$ and $p_2 = \{x_2, y_2\}$ in kilometers, we use the great circle distance G, taken from [Hea03]:

$$G = 6378 \cdot \arccos\Big(\sin(y_1) \cdot \sin(y_2) + \cos(y_1) \cdot \cos(y_2) \cdot \cos(x_1 - x_2)\Big).$$

$D_{geo}(p_1, p_2)$ is then defined as

$$D_{geo}(p_1, p_2) = \frac{d_{max} - G}{d_{max}}.$$

place, id, hierarchy levels	1.	2.	3.	4.
1. Bonn XA–DE–05–3–14 Europe–Germany–North Rhine-Westphalia–Cologne (county)–Bonn	1.0	0.4	0.4	0.4
2. Munich XA–DE–09–1–62 Europe–Germany–Bavaria–Upper Bavaria–Munich	0.4	1.0	0.6	0.8
3. Nuremberg XA–DE–09–5–64 Europe–Germany–Bavaria–Middle Franconia–Nuremberg	0.4	0.6	1.0	0.6
4. Erding XA–DE–09–1–77–117 Europe–Germany–Bavaria–Upper Bavaria–Erding (county)–Erding	0.4	0.8	0.6	1.0

Table 7.2: Political identifiers of four German cities and their political distances.

Specified by the musicologists, d_{max} is the maximum distance allowed for two places to be geographically related in former times. For the examples in this chapter, we used $d_{max} = 500$km, a value empirically determined by the musicologists. In case of $G > d_{max}$, we define $D_{geo} = 0$.

Given two sets P_i and P_j of places of activity for m_i and m_j, we use the iterative closest point algorithm [BM92] to calculate the activity region similarity $S_3^{reg}(m_i, m_j)$. For each place p_i^k in P_i, we determine the distance $d(p_i^k)$ to the "closest place" in P_j, which we define as

$$d(p_i^k) = \max_{p_j^l \in P_j} \Big(D_{pol}(p_i^k, p_j^l) \cdot D_{geo}(p_i^k, p_j^l) \Big).$$

Likewise, we determine the distance $d(p_j^l)$ to the "closest place" in P_i for each place $p_j \in P_j$. Finally, $S_3^{reg}(m_i, m_j)$ is defined as

$$S_3^{reg}(m_i, m_j) = \frac{\sum_{k=0}^{|P_i|} d(p_i^k) + \sum_{l=0}^{|P_j|} d(p_j^l)}{|P_i| + |P_j|}.$$

7.3.4 Musical Profession Similarity S_4^{mus}

For 26,695 musicians (94.9%), the database contains information about their musical professions such as composer, conductor or pianist. Musical professions are of special importance for the musicologists as they substantially define the emphasis of a musi-

cian's activity. They are given as lists $mus(m)$ for each musician m, and the similarity $S_4^{mus}(m_i, m_j)$ between the musical professions of m_i and m_j is defined by the Jaccard index as

$$S_4^{mus}(m_i, m_j) = J(mus(m_i), mus(m_j)).$$

In case of $|mus(m_i)| = |mus(m_j)| = 0$, we define $S_4^{mus}(m_i, m_j) = 0$.

7.3.5 Further Profession Similarity S_5^{pro}

The database also provides information about professions unrelated to music (e.g., philosopher, teacher, soldier) for 7,920 musicians (28.1%). As above, the Jaccard index is used to determine the similarity $S_5^{pro}(m_i, m_j)$ for the further professions $pro(m_i)$ and $pro(m_j)$ of m_i and m_j as

$$S_5^{pro}(m_i, m_j) = J(pro(m_i), pro(m_j)).$$

In case of $|pro(m_i)| = |pro(m_j)| = 0$, we define $S_5^{pro}(m_i, m_j) = 0$.

7.3.6 Relationship Similarity S_6^{rel}

One of the key features of the database are the inherent social relationships. For the collaborating musicologists, these information generate an invaluable social network that reflects interpersonal relationships of the most important musicians in the musical landscape, although relationships are only provided for 9,739 musicians at the moment (34.6%). Nevertheless, the resultant social network contains large communities as connected components. The largest community is composed of 5,065 musicians. Each relationship has a specific type and a role is assigned to both connected musicians. The musicologists also defined the strength s_{rel} for each relationship type (see Table 7.3). Taking all relationships provided by the database into account, the relationship similarity $S_6^{rel}(m_i, m_j)$ of m_i and m_j is derived from the shortest path $p(m_i, m_j) = \{m_i, \ldots, m_j\}$ connecting both musicians in the social network graph and its length $|p(m_i, m_j)|$ using Dijkstra's algorithm [Dij59]:

$$S_6^{rel}(m_i, m_j) = \prod_{k=0}^{|p(m_i,m_j)|-1} \frac{1}{k+1} \cdot s_{rel}(p[k], p[k+1]).$$

category	relationship	s_{rel}
family of origin	parents, children, siblings grandparents, grandchildren	1
partnership	partners	1
education	fellow students, teachers, students	0.8
relatives	cousins, nephews, nieces, uncles, aunts, great uncles, great aunts, grandnephews, grandnieces	0.6
godparenthood	godparents, godchildren	0.6
affinity	parents in law, children in law brothers/sisters in law	0.4
personal relationships	network, patrons, protégés	0.4
working environment	colleagues, predecessors, successors	0.2
dedication	dedication donors & recipients	0.2

Table 7.3: Relationships and their strength s_{rel}.

Thus, the similarity between acquainted musicians is $S_6^{rel}(m_i, m_j) = s_{rel}(m_i, m_j)$ and the similarity for musicians unconnected in the graph is $S_6^{rel}(m_i, m_j) = 0$.

7.3.7 Division Similarity S_7^{div}

Further important characteristics are the divisions where musicians worked (e.g., court, theater). These information are given for 17,062 musicians (60.6%). We determine the similarity $S_7^{div}(m_i, m_j)$ between the known divisions $div(m_i)$ and $div(m_j)$ of the musicians m_i and m_j using the Jaccard index as

$$S_7^{div}(m_i, m_j) = J(div(m_i), div(m_j)).$$

In case of unknown divisions $|div(m_i)| = |div(m_j)| = 0$, we define $S_7^{div}(m_i, m_j) = 0$.

7.3.8 Denomination Similarity S_8^{den}

Especially in former times, the denomination(s) of a musician influenced her activity in a particular manner. Although this information is not provided for 21,302 musicians of the database (75.7%), research questions may include references to a

musician's denomination(s). One denomination is given for 6,733 musicians, two denominations for 117 musicians, and for two musicians even three denominations are provided. Therefore, we use the Jaccard index also to determine the denomination similarity $S_8^{den}(m_i, m_j)$ between m_i and m_j in dependency on the musicians' denominations $den(m_i)$ and $den(m_j)$ as

$$S_8^{den}(m_i, m_j) = J(den(m_i), den(m_j)).$$

In case of unknown denominations $|den(m_i)| = |den(m_j)| = 0$, we define the denomination similarity as $S_8^{den}(m_i, m_j) = 0$.

7.4 The Profiling of Musicians

The idea of musician profiling is to detect a user-defined number N of similar musicians $s_1, \ldots, s_N$, who shared similar attributes with a given musician m of interest. The profiles of all observed musicians are visualized in three different views: Column Explorer, Relationship Graph and Map.

7.4.1 Profiling Workflow

Initially, the musicologist enters the musician m of interest for whom the profile is visualized in the visual interfaces outlined in this section. Observing the various attributes of m, the scholar is able to define mandatory profiling attributes. A similar musician s then requires to share this attribute. Possible mandatory attributes are:

- **Musical & further professions, divisions, denomination(s)**: s shares all mandatory attributes of m in these categories.
- **Activity time**: The intersection of the activity time ranges of m and s is not empty.
- **Place(s) of Activity**: All mandatory places of activity of m were also places of activity of s.

The selection of mandatory attributes supports specific research questions like "Find the most similar musicians to Wolfgang Amadeus Mozart with the musical profession *concertmaster* who worked at a *court* and who had *Salzburg* as place of activity!" Mozart worked as a concertmaster at the court of Salzburg between 1772 and 1777.

Although the database does not contain information if a musician worked at a specific place in a certain profession, the system is capable of providing hints to investigate such questions.

After selecting mandatory attributes, the musicologist performs the first profiling iteration based on all similarity measures defined in the previous section. The weight w_i of a similarity measure S_i is automatically determined in dependency on the diversity of available attributes in relation to the attributes of m in the corresponding dimension. With the number n of musicians $m_1, \ldots, m_n$ with the given attribute ($m \notin m_1, \ldots, m_n$), we define w_i as

$$w_i = 1 - \frac{\sum_{k=1}^{n} S_i(m, m_k)}{n}.$$

An example is given by the weight w_1 for sex similarity S_1^{sex}. The database contains 23,865 male musicians (84.8%), 3,538 female musicians (12.6%) and 751 musicians (2.7%) without a sex information. When profiling similar musicians for Wolfgang Amadeus Mozart, the initial weight for sex similarity is $w_1 = 0.13$ as the database contains mostly male musicians. An initial profiling on Wolfgang Amadeus Mozart's wife Constanze Mozart would use $w_1 = 0.87$ due to the comparatively small number of female musicians.

The result of the first profiling iteration are N similar musicians $s_1, \ldots, s_N$. As outlined above, the profiles of $s_1, \ldots, s_N$ are visualized alongside the profile of m. That individual attributes are easy to track, a certain color is assigned to each musician. As N is usually small – less than ten similar musicians –, we use the ColorBrewer [HB03] to generate a qualitative color map that provides solely saturated colors to be used on the bright website background. In further iterations of the profiling process, the musicologist can gradually modify mandatory attributes and similarity weights as desired in order to receive similar musicians with certain attributes relevant to the posed research question. In Section 7.5, we illustrate several usage scenarios with interesting findings to emphasize the benefit of this interactive visual analytics approach for the collaborating musicologists.

7.4.2 Column Explorer

Inspired by Jigsaw's list view [SGL08] and Parallel Tag Clouds [CVW09], we designed an interface that allows for the exploration of various metadata information provided.

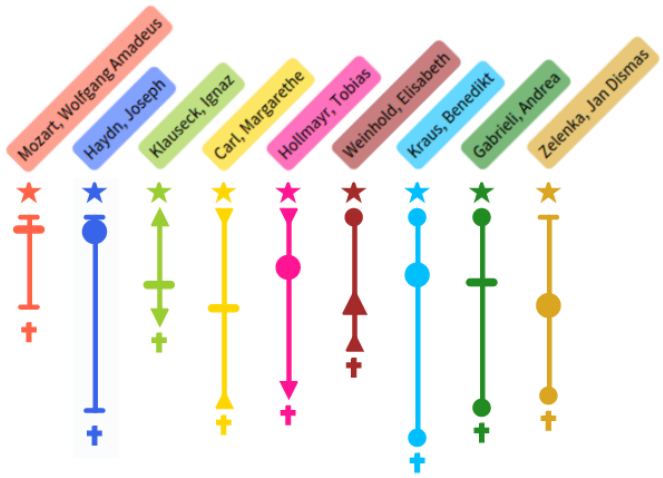


Figure 7.1: Lifetime data examples: various shapes encode uncertain birth, death and first mentioned datings.

The Column Explorer consists of various columns that serve various purposes.

LEGEND All observed musicians are shown in the form of a legend in the leftmost column. m is positioned at the top, and $s_1, \ldots, s_N$ are listed below, ordered by descending similarity to m. The background of a musician's name is drawn in the musician's assigned color. Hovering a musician lists the following attributes in a popup: sex, popularity rank, nationalities and BMLO identifier.

LIFETIME DATA In a vertical timeline, the temporal metadata of all observed musicians is visualized in vertical slots. If provided, we put marks for the date of birth (additionally highlighted with a star symbol ★), the first mentioned date (slightly larger mark), and the date of death (additionally highlighted with a cross symbol †). The shapes used as marks reflect the precision of the provided dating. A small horizontal line — is used for precise datings, and circles ● highlight *around* datings. Triangles ▲ mark *before* datings as they point to the start of the vertical timeline, thus, upside down triangles ▼ illustrate *after* datings. The lifetime of a musician is shown with a vertical line in the corresponding color that connects dates of birth and death. Various examples are shown in Figure 7.1.

NOMINAL TEXTUAL METADATA Four columns list the occurring musical and further professions, divisions and denomination(s) of m and $s_1, \ldots, s_N$. In each column, the

attributes of m are listed first in alphabetical order. By descending similarity, further attributes of the determined similar musicians are listed. Being the most powerful metaphor of tag clouds [BGN08], we use variable font size of labels to encode the number of attribute occurrences. If a musician does not have an attribute in a certain column, we put a "no information" label to communicate this information – an often mentioned demand of the musicologists of our project. Clicking an attribute label toggles its mandatory selection for the profiling process. Only attribute labels (except "no information" labels) belonging to m can be selected.

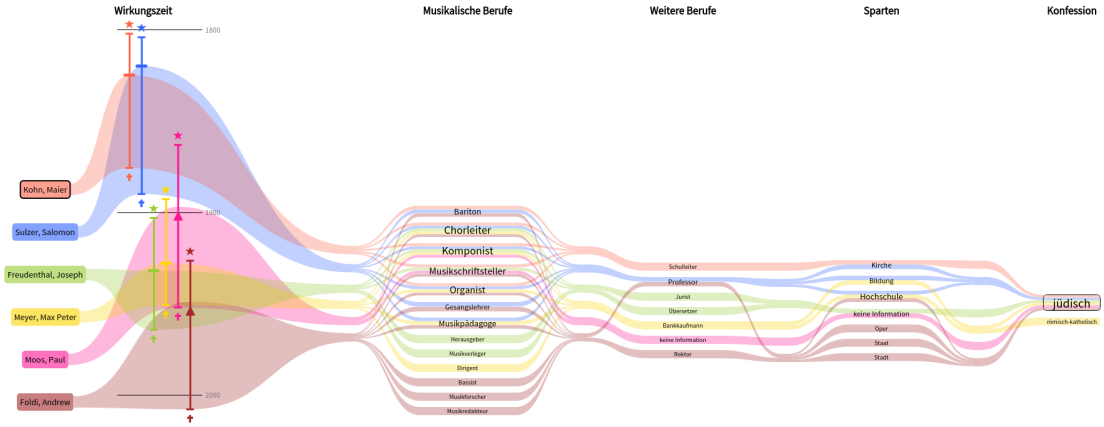


Figure 7.2: The Column Explorer compares the activity time (Wirkungszeit) of musicians and shows correlations among various text-based information about musical professions (Musikalische Berufe), further professions (Weitere Berufe), divisions (Sparten) and denomination (Konfession). The profiling scenario for the Jewish cantor Maier Kohn using mandatory Jewish denomination detects other Jewish cantors as similar musicians with multifaceted interpretations of the cantor profession.

In Jigsaw's list view, attributes from different columns are connected if they belong to the same data entity. Also, in Parallel Tag Clouds various tags are connected after selection. We use these ideas and display the coherence of the attributes of each musician as colored streams passing all columns of the Column Explorer. Starting from the legend, a stream marks the activity time of the corresponding musician in the timeline. Thereby, the occlusion of streams illustrates similar activity times. After passing the timeline, a stream runs through all related attributes of a musician. In case of multiple attributes, a stream splits and passes the corresponding attributes. This metaphor aims to visualize occurring correlations among various attributes of musicians, and to further facilitate the visual comparison of different profiles. A ColumnExplorer example is given in Figure 7.2.

7.4.3 Relationship Graph

The relationship graph of our profiling system is invaluable for the collaborating musicologists as it provides the view on a musician's social network for the first time visually. Furthermore, musicians that connect two observed musicians of interest become visible. For many musicians, a list of relationships to other musicians in the database is provided. Possible relation types between two musicians and the strength of each relationship are shown in Table 7.3. Taking all relationships of m and $s_1, \ldots, s_N$ forms a social network graph with vertices representing musicians and edges connecting related musicians. In order to facilitate an easy exploration of the graph, we only take the direct relationships of each musician into account. Furthermore, we add the relationships between related musicians $\notin \{m, s_1, \ldots, s_N\}$ to receive a closed social network. We use a force-directed algorithm to generate the network graph. We thereby map the strengths of a relationship between two musicians m_i and m_j to intended ideal edge lengths when computing the layout as

$$\frac{1}{s_{rel}(m_i, m_j)}.$$

An example social network is shown in Figure 7.3. The observed musician vertices for m and $s_1, \ldots, s_N$ are drawn in the corresponding color, and their full names are shown next to it. To keep the social network explorable, all additional musicians are drawn as gray vertices, and only the first four letters of their names are shown. The latter design decision reduces the occurrences of occluding labels to a minimum while alongside providing an "adequate information" for the musicologist, who is usually aware of the social relations of the observed musician(s). More detailed information can be shown using mouseover interaction. Hovering a gray vertex pops up the full name of the corresponding musician, whereas hovering an edge provides the roles of the two connected musicians in their relationship. Unobserved, but potentially interesting musicians shown in the social network graph can be added to the profile visualization via mouse click. A mouse click onto the vertices representing m and $s_1, \ldots, s_N$ visualizes the shortest paths to all other musicians under investigation. This feature supports the musicologists in examining the channels through which information was most probably transferred in former times.

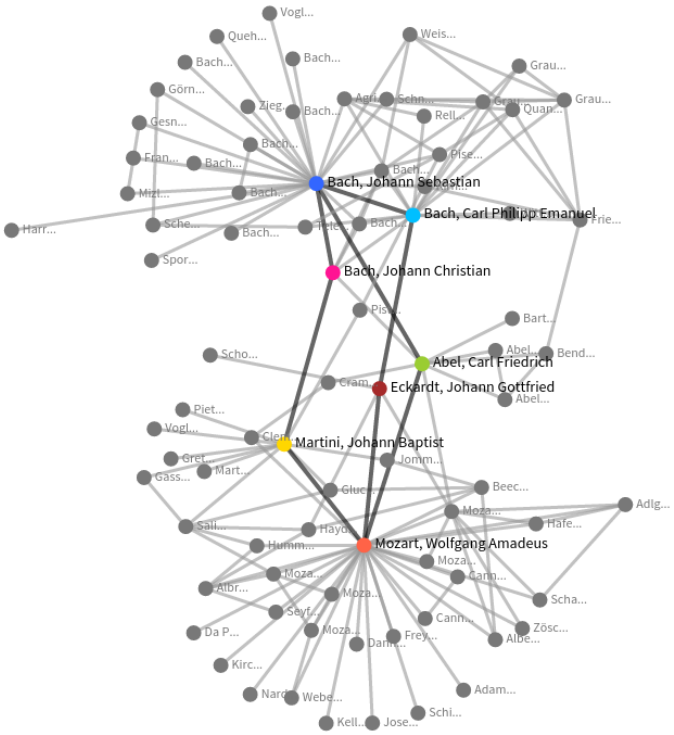


Figure 7.3: The social network shows potential pathways how Bach's score was transferred to Mozart (related links are highlighted for illustration).

7.4.4 Map

The map of the profiling system visualizes all places of activity provided for m and $s_1, \ldots, s_N$. The focus of interest is to facilitate the visual interpretation of a certain activity region and to support the comparison of different activity regions. A location that was only the activity place of one of the musicians under inspection is displayed as a single circle with a radius r_c drawn in the corresponding color. Quite often, musicians shared the same places of activity. For example, Munich was a place of activity for 10,558 musicians of the database (37.5%). To forestall the misinterpretation of activity regions through occluding individual circles, we draw a pie chart for each shared place. We scale the radius r_p of a pie chart dependent on the number k of associated musicians to avoid visual distortion. To receive pie slices with the same

area as an individual circle, we define r_p as

$$r_p = \sqrt{k} \cdot r_c.$$

All shapes are drawn slightly transparent to avoid losing the geographical context in dense regions. An example is given in Figure 7.4. Hovering a shape displays a popup that shows the place name and a list of related musicians.

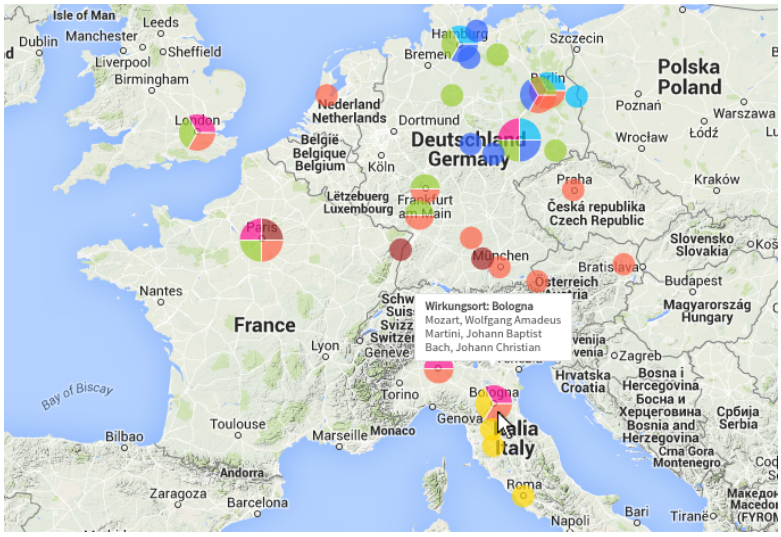


Figure 7.4: The map shows places of activity from Mozart, Bach, and relationships who connect both musicians in the social network to analyze potential places where related musicians met each other or worked together. The city Bologna marks a shared place of activity on one possible transfer path of Bach's score to Mozart.

7.5 Usage Scenarios

The traditional approach of searching for similarities between musicians is biased due to the inhomogeneous state of research (popularity). Rather observing the similarities among the other musicians' attributes, the usage of the profiling system revealed substantial anomalies in contrast to this traditional approach. The first out of four scenarios provided by musicologists using our system exemplifies this issue.

Profiling Georg Friedrich Handel Handel is one of the most popular court composers. A musicologist used the profiling system to iteratively discover court com-

posers of the same era (mandatory activity time, division *court* and musical profession *composer*) with similar careers. The initial profiling result shows similar musicians from different generations with first mentioned datings ranging from 1691 to 1731. Now increasing the weight for activity time ($w_2^{tem} = 1$) and ignoring activity regions ($w_3^{geo} = 0$) better models the musicologists imagination of a "same era" by narrowing this time range (1691-1708). Then, the musicologist tests various combinations of weights for popularity and denomination disregarding relations ($w_4^{rel} = 0$) with interesting insights (in all combinations the era range remains small). The set of similar musicians for various popularity settings and $w_8^{den} = 1$ changes only slightly and always contains popular Evangelical-Lutheran musicians like Johann Sebastian Bach and Georg Philipp Telemann. By further applying varying denominational significance and using $w_p = 0$, the musicologist discovers unexpected, very similar profiles to Handel in terms of musical professions and divisions for rather unknown musicians with activity places in southern European regions; especially, two Italian musicians are in the result set. With Venice, Rome and Naples, Handel had an active period in three Italian cities. As $w_3^{geo} = 0$ was used, this correlation hypothesizes mutual influences between Handel and the musicians found as well as an Italian influence on Handel's work. Now mimicking the traditional profiling approach based upon print media by applying $w_p = 1$, most of the rather unknown musicians are replaced by popular ones with a lesser similarity regarding musical professions and divisions (see Figure 7.5). Especially the similarity between Handel and Antonio Vivaldi – both sharing only few characteristics – seemed accidental to the musicologist. Thus, this use case exemplifies the biased influence of popularity. But the tool opens new research perspectives by focusing rather unknown but more similar musicians as opposed to focusing popular musicians.

Profiling Meinrad Spieß Starting a profiling for musicians similar to monastery composer Spieß, a musicologist would predominantly refer to musicological editions and monographs. According to them, the results of this traditional approach would be again biased due to the inhomogeneous state of research (popularity). Our profiling system was used to search for musicians similar to Spieß disregarding popularity ($w_p = 0$). The initial profiling step shows a list of other southern German Catholic church musicians of the early modern era. To further specify the profile scheme, the musicologist increases the weight for activity region, activity time, division and denomination similarity, whereas the weight for relationship similarity is lowered. As

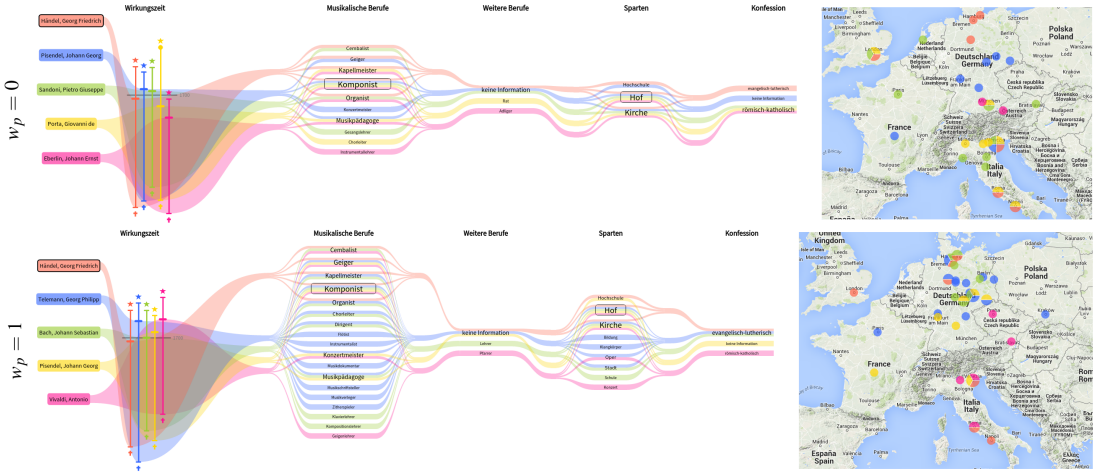


Figure 7.5: A profiling for musicians similar to court composer Georg Friedrich Handel. The musical profession *composer* (Komponist) and the division *court* (Hof) are marked as mandatory. The results are shown disregarding popularity ($w_p = 0$) and taking popularity into account ($w_p = 1$).

the result differs only slightly, the weights for denomination and activity region are set to 0. Then, middle German Protestant contemporaries occur – a comprehensible fact as Spieß (1) was an active member of Lorenz Mizler's musical circle in Leipzig, and (2) corresponded frequently with academy colleagues outside his denominational bounds and activity region. A particular observation is an obvious similarity to musicians belonging to a generation that essentially characterized the *bandmaster* profession. Similarities are discovered to the known musicians Johann Sebastian Bach and Georg Friedrich Handel, and other representatives of this first bandmaster generation – a fact hardly recognized in previous music research. Approaching insights this way is a novel technique in musicology. An also known, but never visualized phenomenon in musicology are homogeneous subcultures of the early modern era. By only regarding relational similarities for Spieß, the musicologist detects a closed Benedictine network composed of Spieß' students and the relatives of his own teacher Bernabei.

Profiling Maier Kohn The cantor is one of the most multifarious musical professions in cultural history. The characteristics of this profession equally depend on chronology, cultural area and denomination. Consequently, the responsibilities of a cantor are widespread and the term *cantor* is impractical to be used as musical vocabulary. Subsequently, cantors cannot be easily retrieved using the Bavarian Musicians

Encyclopedia Online. The musicologist requires an analysis in dependency on the various duties of the cantor profession. Using the profiling system, this multifacetedness can be visualized and analyzed in individual cases. The Jewish cantor Maier Kohn had many musical responsibilities in clerical music and in school, as a singer, organist, composer and choirmaster. Similar profiles regarding musical professions that compose the cantor profession can be found in various contexts for musicians of different generations, with different activity regions and denominations. Limiting the profiling to Jewish musicians with similar profiles provides a list of Jewish cantors (e.g., Salomon Sulzer and Joseph Freudenthal) with many varying musical professions (see Figure 7.2). This brief meta-analytic test reveals a contradiction to the anti-Semitic influenced state of research of the early 20th century. At that time, musicologists claimed a monotonous interpretation of the cantor profession for Jewish musicians. But the multifaceted musical professions of all cantors in the result suggest a diversity similar to known Christian cantors. According to the anti-Semitic research, Maier Kohn was therefore not a "typical Jew." His strong similarity to Christian cantors when disregarding Jewish denomination in the profiling underpins this fact. This example outlines the utility of the profiling system to transform a musicological issue – the multifacetedness of the musical profession cantor –, which is not existent in the database, into a representation visualized as the Column Explorer.

The missing link between Johann Sebastian Bach & Wolfgang Amadeus Mozart
This example illustrates the usage of the system without its profiling capacity. Mozart was born few years after Bach's death, but Mozart played Bach's music. Mozart primarily worked at southern German Catholic residences where the trade with music supplies was unincisive. An interesting research question for the musicologist arose: "What was the connection through which Bach's score was transferred to Mozart?" First, the musicologist visualizes the profile of both musicians. Second, the relationship graph is explored and candidates on probable pathways between Bach and Mozart are added to the profile visualization (see Figure 7.3). Taking all visualized attributes into account, the musicologist is now able to measure possible pathways, especially by observing shared places of activity (see Figure 7.4). Although the musicologist requires additional literature to examine this question more precisely, the system provides valuable evidence to narrow the number of possibilities.

7.6 Discussion

The proposed profiling system was designed to support answering a novel type of research question in musicology. Some aspects of the collaborative work are outlined below.

Evaluation When developing the profiling system, we closely collaborated with four musicologists – a professor, a PhD student and two M.Sc. students –, who iteratively evaluated current prototypes. One of the key features of the profiling system was the design of similarity measures for relevant musician attributes included in the profiling process. Some of the similarity measures were refined step-by-step to incorporate musicological knowledge in order to gain results that meet the expectations of the musicologists. For example, when designing the activity region similarity, we always provided a list of place tuples with their calculated similarities to the musicologists to ensure an appropriate representation of musicological imagination of space. When determining relationship similarity – first defined for two musicians only by their distance in the social network graph – we mapped relationship strengths, provided by the musicologists, to edge lengths. As a result, familial relationships form clusters, which was an important requirement of the musicologists. Activity time similarity – first defined as lifetime similarity by birth and death of musicians – was also iteratively modified. Here, the inclusion of *first mentioned* dates and the mapping of uncertainties allowed us to define this similarity measure more precisely. The visualization of the profiling system was also iteratively improved and evaluated by the musicologists to meet their needs. This included both aspects of visual representation and interaction design. We could communicate our own concerns as well. For instance, we thought that overlapping streams in the vertical timeline are too confusing. But the musicologists prevented us from changing this representation arguing that it perfectly reflects their imagination of activity time similarity. As there is no ground truth regarding the profiling of musicians, the accuracy of our approach is not easy to measure. Sometimes, surprising and unexpected results occur. Being involved in all development stages, the musicologists assess individual similarity measures as well as entire profiling results as reasonable, which underpins the benefit of our method for musicology.

Limitations Being a challenge for developing the visual analytics system on the one hand, the existence of uncertain temporal metadata slightly affects the reliability of a profiling result. As the BMLO gets updated gradually, the removal of uncertainties requires future effort for musicologists using the database. A further limitation concerns the missing consideration of historical circumstances when calculating activity region similarities. First, the meaning of a geographical distance varies for different ages. Whereas a travel between European cities required several weeks in the Renaissance era, such a trip takes only few hours nowadays. Second, the usage of contemporary political conditions cannot be applied appropriately to historical contexts, although our collaborative solution turned out to be heuristically valuable for musicologists. But the elaboration of historical place identifiers could further improve the profiling result. As the provided textual metadata is not linked, e.g., the existence of "London" as activity place, "bandmaster" as musical profession and "church" as division does not imply that a musician indeed was a bandmaster at a church in London. The interpretation of such information still requires a musicologist's knowledge or the usage of further sources. In terms of scalability, our proposed system is designed to compare the profiles of a rather small number of musicians – distinguishable through various colors –, usually less than ten. Therefore, general research interests like analyzing and comparing all *court composers* is not supported.

Future Work The BMLO is an ongoing digital humanities project under crowdsourcing aspects. Next to potential future data transformation and data representation challenges, the collaborating musicologists suggested several improvement prospects to determine similar musicians more precisely. First, the inclusion of hierarchical information into the profiling process was an often discussed issue. At the moment, a hierarchy is only given for musical professions. According to the musicologists, hierarchical representations for other text-based metadata dimensions (further professions, divisions), hierarchical relationship strengths or the calculation of activity region similarities taking historical circumstances into account could further strengthen the result of a profiling process. Consequently, we would need to adapt similarity measures and the visualization, especially the Column Explorer. Another future work is the profiling for coupled musicians – a novel type of research question stimulated through our profiling system. For instance, the profiling result for musicians similar to Wolfgang Amadeus Mozart and Johann Sebastian Bach in a single request could answer the question if a found musician would be more similar to Mozart or to Bach. As a

straightforward adaption of the similarity measures does not anticipate adequate results, we require further interdisciplinary sessions to discuss required implementation steps.

7.7 Summary

As of the late 19th century, musicology focuses primarily on circa fifty musicians and their main works in a traditional philological manner. The achievements of other musicians only obtain less attention. The proposed profiling system aims to change this imbalance by rather throwing the spotlight on less popular musicians. Based upon a musician of interest – potentially one of the popular ones – musicologists are now capable of discovering less popular musicians with similar careers.

During the development, we closely collaborated with musicologists, who state that the resultant profiling system is a valuable analysis instrument that serves a novel type of research interest and provokes new research questions. Thereby, we designed the profiling system the way that it can easily be adapted to other historical groups of people.

Our presented approach facilitates comparative methods and research questions concerning musicians – for the first time with the aid of visual means. As the visualization indicates historical circumstances and cultural contexts, it gets possible to review time-dependent ideological opinions about individual musicians. Usage scenarios showcasing Handel's, Spieß' and Kohn's careers demonstrate this capability of the profiling system.

The outcome of any serious research can only be to make two questions grow where only one grew before.

Thorstein Veblen

8 Discussion

Computer scientists and humanities scholars seemingly do not have many things in common. Although they share some methodologies, they are geared towards different goals. But the digital age created a platform that brings people from two research areas together: the digital humanities.

During our literature research in Chapter 2, we had the opportunity to take a look at various fascinating digital humanities projects proposing visualization techniques for the close and distant reading of texts. In this process and our own efforts in designing beneficial visualizations for the comparative analysis of digital humanities data derived from textual sources, we derived insights into a research field that requires intuitive interfaces in order to support answering a broad palette of research questions. But, visualization techniques for textual data as part of the cultural heritage are rarely published in the visualization community. Taking the publication year of Moretti's first book on distant reading [Mor05] as a starting point, Figure 8.1 shows the temporal distribution of the papers addressing related research questions published in the visualization and the digital humanities communities, including our own ones. The trend within the digital humanities reflects the increasing value of visualizations for the close and the distant reading of text in the recent years. Until now, the visualization community did not notice or consider these needs. The reason for this may lie in the obstacles encountered in publishing application papers with a digital humanities

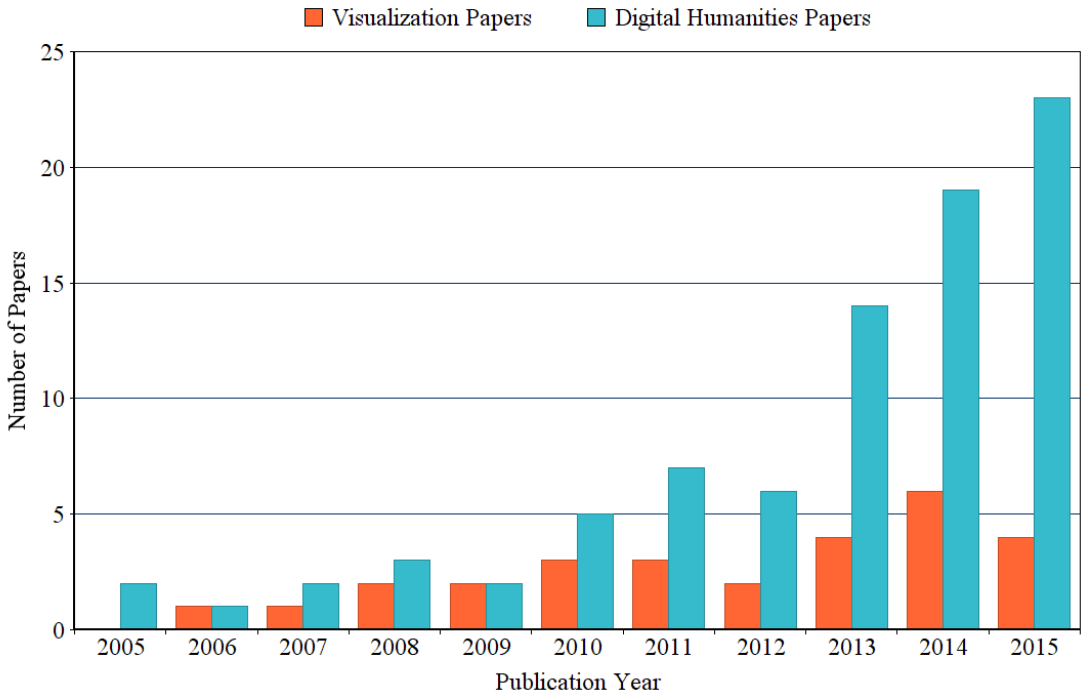


Figure 8.1: Papers published in visualization and digital humanities communities by year.

background because the often demanded quantitative evaluations are hard to perform due to the usually limited number of collaborating humanities scholars – a problem we also faced in some of our projects.

To strike a balance between these shortcomings, this chapter outlines our own and the collaboration experiences reported by other visualization researchers working in the field of digital humanities as a means of singling out the important ingredients for a successful project. Furthermore, we list future challenges concerning the design of comparative visualizations in particular and unsolved problems in general to support humanities scholars with close and distant reading methods.

8.1 Collaborating with Humanities Scholars

To ensure that a developed visualization is beneficial for collaborating humanities scholars and to overcome the problem of having only few participants for evaluation purposes, we suggest to apply a user-centered design study [Mun09] when designing visualizations for digital humanities research tasks. This leads to a very close col-

laboration between researchers of the different fields, a better understanding of each others mindsets, and an iterative evaluation of the visualization that avoids gearing the development into false directions. We carried out most of our projects using that methodology, but some other works also publish valuable insights into collaboration experiences. Abdul-Rahman [ARLC+13] also outlines the benefit of a user-centered design study when designing the Poem Viewer,[1] a tool that supports the analysis of sound in poems. She reports insights "of a fascinating collaboration between computer scientists and literary scholars." Other publications also share important experiences. Taking together our own experiences gained during the development of close and distant reading visualizations and the experiences of other visualization researchers working on solutions for the digital humanities, this section provides suggestions that might help visualization researchers new to the field of digital humanities to develop successful visualizations. Various aspects of development are outlined below.

Project start. One of the most important, initial tasks of a digital humanities project are discussions about the research questions and perspectives for which a visualization, be it for close or distant reading, can be beneficial. These discussions include the analysis of the data features as well as the setup of regular project meetings to work on and extend a collaborative idea. A typical issue of digital humanities projects is reported by McCurdy et al. [MLCM16]. The "initial conversations [between visualization and humanities scholars] were broad and open-ended," also, because the humanities scholars "did not have specific goals" in mind. Furthermore, the humanities scholars were sceptical that visualization can support their research, and there was also an "anxiety that the computer would inhibit the qualitative experience of the poetic encounter." After humanities scholars presented examples of interesting features and computer scientists "established methods for computationally detecting and analyzing the devices that most interested them," a common project basis and tasks could be generated. We had a similar problem during the *eXChange* project. As the collaborating humanities scholars could not explicitly tell us their desires, it took some time for us to comprehend their needs and to suggest valuable visualizations (see Chapter 4). In such circumstances, special workshops helped us computer scientists and humanities scholars get acquainted with each others' tasks, mindsets and workflows. Also, Abdul-Rahman [ARLC+13] reports the importance of visualization researchers participating in poetry readings and in-depth discussions with

[1] http://ovii.oerc.ox.ac.uk/PoemVis/

literary scholars to discover "a variety of interesting problems that might be subject to visualization solutions." A small corpus generated for literary scholars was helpful for Abdul-Rahman to examine research questions without the aid of existent visualizations. In contrast to the problem of forming a synergy between the scholars of both fields in order to develop valuable visualizations, we faced a yet different situation when developing the visual analytics profiling system for musicians (see Chapter 7). Already in the initial meeting, the musicologist precisely posed the profiling research questions, and we could quickly suggest applicable visualizations.

Iterative development of prototypes. The involvement of humanities scholars in various stages of the development is necessary to ensure creating an intuitive visualization that will be used. For example, regular face-to-face sessions between computer scientists and humanities scholars can help to identify problems and potential enhancements of the prototype design. Such a session should be composed of a demonstration and trials of the visualization prototype as well as intense discussions in order to gather the levels of detail and complexity that a visualization should ideally reach. In the development of *TRAViz* (see Chapter 6), my digital humanities colleague Annette Geßner stated that such a process finally helps to gain an intuitive result "even for the inexperienced, maybe sceptical user." Regular meetings when designing the musicians profiling system (see Chapter 7) were important for us to communicate our own concerns and to iteratively redesign the underlying mathematical basis (similarity measures) – thereby ensuring that aspects of data transformation retain comprehensible for the musicologists. Another example is given by scholars involved in the development of Neatline [NMG+13], which is based upon Omeka,[2] a content management system for online digital collections. The iterative development of Neatline led to advancements of Omeka itself, thus benefiting a far wider audience than originally anticipated.

Evaluating visualizations with humanities scholars. The evaluation sessions during our projects provided important insights into design, intuitiveness, the utility of visualizations and into potential enhancements. Sometimes, the humanities scholars working with the visualizations suggested further enhancements, some of which strengthen the importance of close reading solutions. This was especially an important aspect when designing tag clouds as distant views for summaries of historical texts (see Chapter 4). The provided immediate access to corresponding text passages

[2] http://www.omeka.org/

when working with the tag clouds built trust in the visualization method and fostered hypotheses generation. The importance of close reading is also pointed out in other works. For example, when providing close and distant views, "users stressed that it is preferable to see the actual words" rather than abstract overviews [JRS+09]. When working with the VarifocalReader [KJW+14], a user liked to see "the digitized image of a book's page and mentioned that this would increase his trust in the approach." This metaphor of a digitized text we also used when designing an interface (see Figure 8.2) to be used by the humanities scholars for the comparison of various English translations of the Bible by displaying *TRAViz* Variant Graphs for each Bible verse (see Chapter 6). According to the humanities scholar who worked with the system, this would "remind the user that it is a book to be read, not just some string of letters." We also recognized the importance of aesthetic appeal to engage in information exploration, which is also reported by Hinrichs et al. [HSC08]. Furthermore, visualizations for humanities scholars should be designed to meet their work practices. For example, we took a set of representative workflows collected by collaborating humanities scholars into account when designing the user interface embedding the tag clouds presented in Chapter 4. During evaluation, humanities scholars also have the opportunity to mention issues or limitations with the presented tools necessary to mark future prospects. For example, future extensions of the visualizations that compose the musicians profiling system (see Chapter 7) require efforts from us as well as from the collaborating musicologists. Scientists involved in [HKTK14] stated that collaborative work helped to reactivate and to regenerate traditional literary methodologies rather than abandon them. The turn from initial scepticism when starting the digital humanities project to enthusiam when using the resultant visualization is reported by McCurdy et al. [MLCM16]. Finally, the utility of visualization in the humanities is corroborated by the fact that, already during the evaluation phase, many humanities scholars make surprising discoveries, generate new hypotheses or suggest further usage scenarios.

8.2 Future Challenges

The visualizations for the comparative analysis of digital humanities data presented in this dissertation and the related works listed in Chapter 2 are tailored to support the investigation of various digital humanities research questions. Yet, there are still major challenges in the digital humanities where the visualization community can

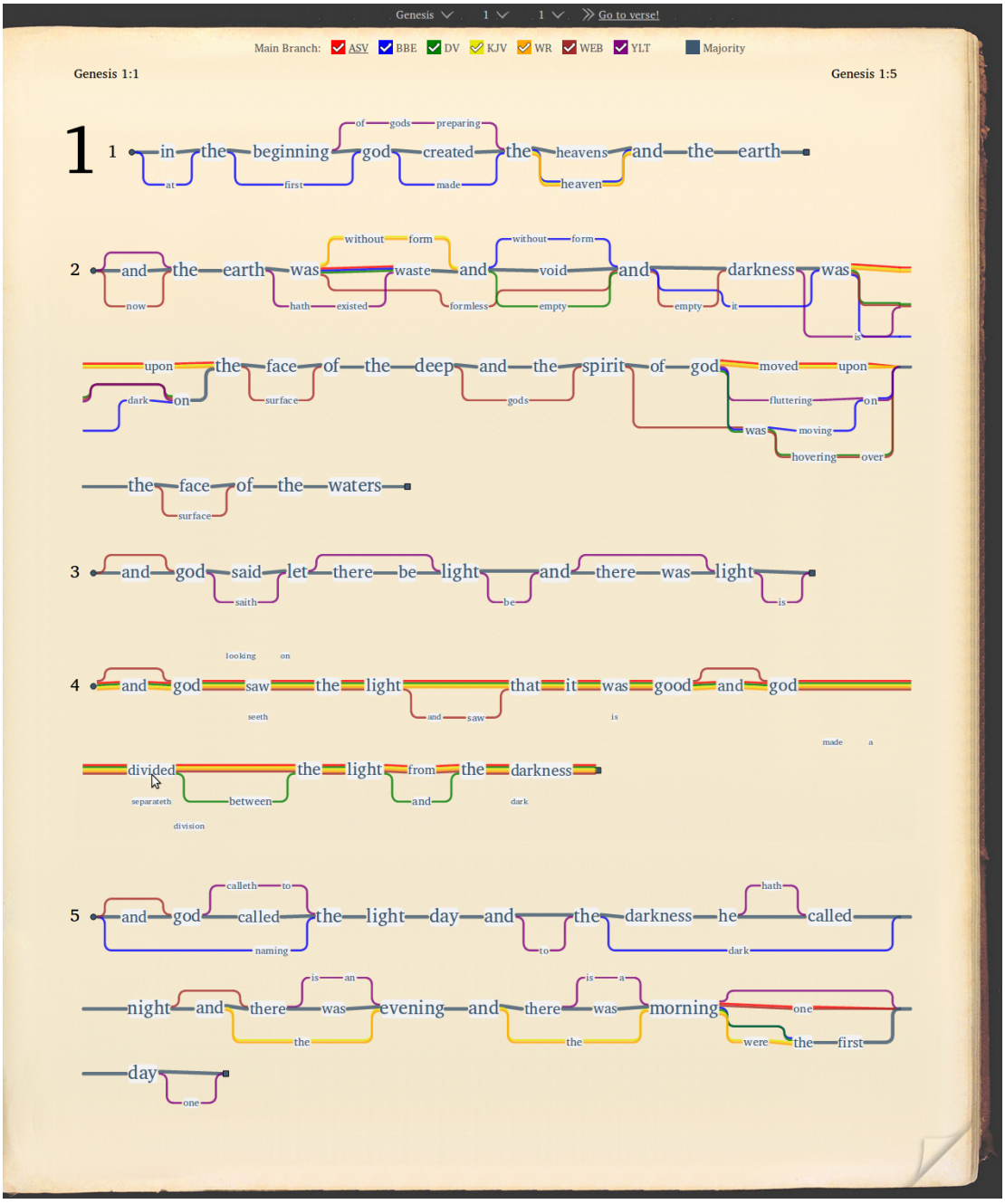


Figure 8.2: Utilizing the metaphor of a digitized text for close reading.

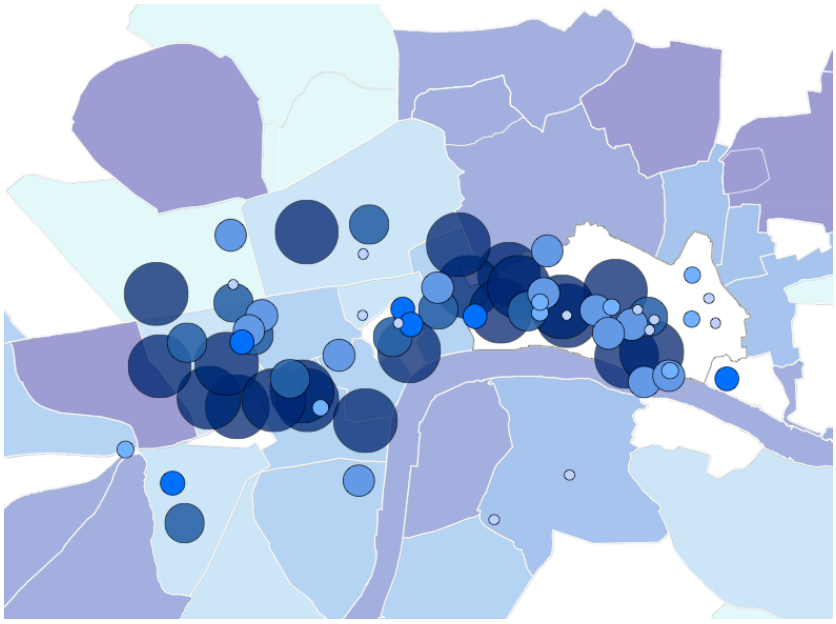


Figure 8.3: Colored shapes encode emotions about London places (Figure reproduced with permission from Heuser et al. [HAHT+15]).

contribute valuable research. Although all of the following open challenges came up during our research, some of them are closely related to designing comparative visualizations, and some of them are rather general future challenges for the visualization community regarding the design of visualizations for digital humanities scholars.

Geospatial uncertainty. Many visualizations deal with placenames extracted from literary texts to illustrate the geographical knowledge of a particular era (see paragraph *Maps* in Section 2.5). Within our project *Visualizing Medieval Places*, various mapping issues arose (see Section 3.4.1). The source texts contained placenames of varying granularity (e.g., country, region, city) or type (e.g., points for cities, polygons for areas, polylines for rivers) or even fictional placenames, which are hard to represent. We provided a workable solution for the geographical granularity by using different glyphs for different place types. A further solution is presented by Heuser [HAHT+15], who uses circles to map discrete London places occurring in fictional literature and polygons for wider spaces such as neighborhoods or districts of London (see Figure 8.3). In addition to the granularity issue, some placenames themselves carried uncertainty of varying degree in our dataset, e.g., the exact locations of "Sparta" and "Atlantis" have yet to be discovered. Another form of uncertainty is defined by contextual informa-

tion, e.g., expressions like “in Sevilla” and “close to Sevilla” cover various geospatial ranges. The development of a design space providing solutions to visualize these various types of geospatial uncertainty is one of the current primary challenges when dealing with geospatial references in digital humanities data. Such a design space could be built upon the ideas of MacEachren for visualizing geospatial uncertainty [MRO+12].

Temporal uncertainty. The visualization of temporal uncertainty is an equally important future task. Such uncertainties occur, for instance, when dating cultural heritage objects, such as historical manuscripts [BESL14]. The temporal metadata in *Visualizing Medieval Places* (see Section 3.4.1) was provided in multifarious manners, e.g., 1450, before 1450, after 1450, around 1450, 15th century, first half of the the 15th century, etc. We transformed these formats into machine-parsable time ranges, and experimentally used a ThemeRiver to visualize occurring uncertainties. In general, the visualization of such uncertainties is a crucial issue as it comprises considerable risks of misinterpretation. This happened to us when presenting our approach at the digital humanities conference, but the ThemeRiver approach was still valuable for the collaborating humanities scholar. Applying methods capable of appropriately visualizing temporal uncertainty as proposed by Slingsby [SDW11] can be a first step, but their utility for humanities applications needs to be investigated.

Novel techniques for close reading. In our projects, we marked that the humanities scholar’s close reading task generally benefits from visualization, e.g., by highlighting searched keywords (see Chapter 4) or displaying textual features and structural relationships (see Chapters 5 and 6). Other works also support the essential close reading task for humanities scholars [ARLC+13, KG13, WMN+14], but in most cases only simple visualization techniques, such as color coding textual entities, are provided. Few works attend to the matter of enhancing close reading in a beneficial way. For example, the work on word scale visualizations is a promising technique [GWFI14] from which many humanities scholars may profit. But despite the proposed annotations of individual words with statistics or of country names with polygons, the concept needs to be expanded to annotating other kinds of named entities, which are important for humanities research. For example, providing supplementary information about (1) acting persons and their relationships, (2) artifacts mentioned in texts, or (3) occurring references could be interesting features for close reading. Future work in visualization should include the development of design methods to meet such use cases, and studies that measure the benefit of glyph based approaches for close reading in comparison to using color or font size to express certain

text features.

Visualizing transpositions in parallel texts. When observing similarities and differences among various editions of a text, one focus is to detect transpositions of textual entities. Such transpositions may occur on various text hierarchy levels, e.g., changed word order, modified argumentation structures, or even when exchanging whole paragraphs or sections. In Chapter 6 we presented with *TRAViz* a suitable method for visualizing changed word order, and in Chapter 5 we introduced the Text Re-use Reader that highlights transpositions of re-used text passages. Both solutions were developed independently, so that – as of now – there are no visualization techniques capable of coherently visualizing transpositions on all hierarchy levels by combining means of close and distant reading. The development of a uniform design would certainly improve the humanities scholars' workflows as the comparative analysis of text editions is a very common task in digital humanities.

Reconstructing workflows with visualization. During our collaborations with humanities scholars, we learned that an emulation of the scholar's workflow helps to build trust in the visualization – the Bible interface in Figure 8.2 being only one example. Other researchers report similar experiences. In [KJW+14], users desired the display of digital copies of books, and, when working with genealogy visualizations [BDF+10], historians "insisted on redundant representation of gender ... that is consistent with their current practices." These situations illustrate the challenge of inventing visualization techniques for digital humanities applications that humanities scholars can easily adapt. An important task for a computer scientist is not only to incorporate a scholar's workflow when designing the visualization, but to also communicate all aspects of data transformation, so that a scholar is able to generate trustworthy hypotheses. The importance of this issue is documented in [GO12].

Adapting existing visualization techniques. For some of the research questions posed in the digital humanities, the adaption of existent techniques proposed in visualization research papers is beneficial. For example, we adapted the ThemeRiver idea to visualize temporal uncertainties in the *Visualizing Medieval Places* project (see Section 3.4.1). Another positive example is the *Trading Consequences* project [HAC+15]. Involved visualization scholars designed a system inspired by VisGets [DCCW08] and made use of Parallel Tag Clouds [CVW09] to explore texts about commodity trading. Both visualization techniques were not developed for digital humanities data, but they were beneficially adapted to support humanities scholars. Occasionally, new techniques for close and distant reading are designed while appropriate, sophisticated

visualizations unrelated to digital humanities data already exist. For future research tasks, the inclusion of these visualizations into the workflows of humanities scholars could lead to faster hypotheses generation due to the limited time for development. Representative examples can be found in our survey on close and distant reading techniques [JFCS15].

Usability studies. Although the utility of most visualizations considered in Chapter 2 was illustrated by usage scenarios, we found little evidence about conducted usability studies to, for example, justify taken design decisions. The number of humanities scholars participating in such studies is potentially very small due to the multifarious research interests scholars may have on a large body of texts belonging to different eras and genres. For specific research tasks, applying a user-centered design study [Mun09] as we did in some of our projects can lead to satisfying results for both the visualization and the digital humanities communities. But generating a user study format that caters for the interests of many different scholars is required to gain valuable insights into guidelines for designing visualizations for the digital humanities. When it comes to tool building, in fact, the digital humanities community poses interesting and complex challenges by virtue of its interdisciplinary nature. It embraces a wider range of disciplines, so the techniques it offers should address the larger scope. It also welcomes contrasting mindsets, methods and cultures. While sharing similar logical and analytical methods, computer scientists tend towards problem solving, humanities scholars towards knowledge acquisition and dissemination [Hen14]. No one community should operate in subservience to the other but together, complementing each others' approaches. For these reasons and in this context, specialist terminology, assumptions and technical barriers should all be avoided. It is in this sense that tool usability should be understood not only as improved functionality or aesthetics but as a transparent guide to utility [GO12].

A story has no beginning or end: arbitrarily one chooses that moment of experience from which to look back or from which to look ahead.

Graham Greene

9 Summary

Within this dissertation, we outlined how visualizations can be used to facilitate the access and the exploration of cultural heritage data sources. In particular, we presented solutions for the comparative analysis of multiple datasets to support the investigation of various research questions.

At first, we introduced the system *GeoTemCo* that supports the comparative analysis of multiple geospatial-temporal datasets representing multiple topics, using a map to visualize the geospatial distributions and a timeline to show trends. This allows not only for a comparison of geospatial and temporal characteristics, the provided means of interaction also support the discovery of cooccurring items among various datasets, the comparison of geospatial migrations over time, and, in the best case, to detect causalities between datasets. Although not particularly designed for humanities applications, *GeoTemCo*'s frequent usage in the digital humanities underpins its benefit to support investigating research questions in that field.

At second, we presented two tag cloud designs that visualize faceted textual summaries. *TagPies* – a tag cloud arranged in a pie chart manner – allow for the comparison of tags belonging to different data facets in different pie slices. *TagSpheres* are designed to visually encode the notion of distance to an underlying topic. Tags are grouped in spheres around the topic of interest in the center of the cloud. Both designs were developed for humanities scholars who analyze historic texts and want to discover

text passages related to their observed topic. Although tag clouds summarize textual content and thereby dissolve textual structure, which is required for interpretation, our visualizations are valuable interfaces to get an overview of the contexts in which ancient terms were used, and to get a quick access to potentially related text passages.

At third, we developed visualizations to interactively explore re-used text passages in a text corpus. The *TextReuseGrid* is a distant reading view that juxtaposes all texts in the form of a grid. The cells of the grid are colored according to the amount and the type of Text Re-use among the two texts of a tuple in order to direct the humanities scholar to interesting Text Re-use patterns. In the *TextReuseBrowser*, which juxtaposes the two texts of a tuple in close reading mode, these Text Re-use patterns can be analyzed in detail. This combination between close and distant reading of Text Re-use turned out to be valuable for the collaborating humanities scholars.

At fourth, we proposed *TRAViz* that provides a layout and a design for Variant Graphs. It is tailored to appropriately reflect a Variant Graph's features in order to facilitate the comparative analysis of text editions. A multitude of use cases outlines the benefit of this enhanced close reading method for humanities scholars to explore syntactic and linguistic similarities and differences. As our approach is chiefly valuable for small text entities such as sentences, we additionally introduced a distant reading method for Variant Graphs to allow posing a broader range of research questions concerning textual variation.

At fifth, we designed the visual analytics system *MusikerProfiling* used by musicologists to facilitate profiling a musician – the discovery of similar musicians –, a task which formerly has been done referring to print media. In contrast to this traditional profiling approach that usually generates a list containing only popular musicians, our analysis is based on the data of the Bavarian Musicians Encyclopedia Online, which includes around 28,000 musicians. This fosters the discovery of rather unknown musicians with a high similarity according to biographical characteristics. Several usage scenarios illustrated this – for the musicologists essential – capability of our system.

We extended the numerous experiences we gained during these interdisciplinary projects with collaboration aspects from other visualization researchers working in the digital humanities field to provide guidelines for conducting a successful project and developing visualizations that are valuable for the involved humanities scholars. During our research we marked general future prospects for the visualization community and open challenges concerning the comparative analysis of digital humanities data.

One of the most important lessons learned during our interdisciplinary cooperations was that a visualization is usually not used at the end of a research process in a digital humanities workflow. Although a visualization is capable of triggering hypotheses generation, a humanities scholar will unlikely proof a hypothesis based predominantly on an image. A statement from a humanities scholar working with a poem visualization, taken from [ARLC+13], points out this fact by mentioning that "they would not likely look for insight from the tool itself ... they would look for enhanced poetic engagement, facilitated by visualization." The reasons for this issue often relate to *uncertainty* and *incompleteness*. Basing a visualization on vague or on partial information can lead to misinterpretations. A visualization should communicate these aspects and a humanities scholar should be able to access the underlying data in order to build trust in the method at hand. During our projects, we tried to follow this guideline at best.

Bibliography

[AA05] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[AA06] G. Andrienko and N. Andrienko. Visual Data Exploration: Tools, Principles, and Problems. In *Classics from IJGIS: twenty years of the International journal of geographical information science and systems*, pages 475–479. CRC Press, 2006.

[AAH+11] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 161–170. IEEE, 2011.

[AGL+07] L. Auvil, E. Grois, X. Llorà, G. Pape, V. Goren, B. Sanders, and B. Acs. A Flexible System for Text Analysis with Semantic Networks. In *Proceedings of the Digital Humanities 2007*, 2007.

[AGZH15] B. Alex, C. Grover, K. Zhou, and U. Hinrichs. Palimpsest: Improving Assisted Curation of Loco-specific Literature. In *Proceedings of the Digital Humanities 2015*, 2015.

[AKV+14] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 173–182, Oct 2014.

[All99] J. Allen. s.v. Kunst. *Der Neue Pauly (DNP)*, Bd. 6:Sp. 915–919, 1999.

[AM13] T. L. Andrews and C. Macé. Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*, 2013.

[ARLC+13] A. Abdul-Rahman, J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, C. R. Johnson, A. Trefethen, and M. Chen. Rule-based Visual Mappings–with a Case Study on Poetry Visualization. In *Computer Graphics Forum*, volume 32, pages 381–390. Wiley Online Library, 2013.

[Arm14] F. Armaselu. The Layered Text. From Textual Zoom, Text Network Analysis and Text Summarisation to a Layered Interpretation of Meaning. In *Proceedings of the Digital Humanities 2014*, 2014.

[ARR+12] O. Arazy, S. Ruecker, O. Rodriguez, A. Giacometti, L. Zhang, and S. Chun. Mapping the Information Science Domain. In *Proceedings of the Digital Humanities 2012*, 2012.

[AVZ13] T. L. Andrews and J. J. Van Zundert. An Interactive Interface for Text Variant Graph Models. In *Proceedings of the Digital Humanities 2013*, 2013.

[BB15a] E. Beshero-Bondar. Visualizing the Digital Mitford Project's Prosopography Data. In *Proceedings of the Digital Humanities 2015*, 2015.

[BB15b] E. Beshero-Bondar. World-View from Poetic Structure: An "Anti-Social" Network Analysis of Robert Southey's and Eramus Darwin's Epic Poems. In *Proceedings of the Digital Humanities 2015*, 2015.

[BDF+10] A. Bezerianos, P. Dragicevic, J. Fekete, J. Bae, and B. Watson. GeneaQuilts: A System for Exploring Large Genealogies. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1073–1081, Nov 2010.

[Bea08] D. Beavan. Glimpses though the clouds: collocates in a new light. In *Proceedings of the Digital Humanities 2008*, 2008.

[Bea11] D. Beavan. ComPair: Compare and Visualise the Usage of Language. In *Proceedings of the Digital Humanities 2011*, 2011.

[Bea12] D. Beavan. DiaView: Visualise Cultural Change in Diachronic Corpora. In *Proceedings of the Digital Humanities 2012*, 2012.

[Bea14] M. Beals. TEI for Close Reading: Can It Work for History?, 2014. http://tinyurl.com/nvdndsb (Retrieved 2016-02-01).

[Ben14] D. C. Benner. "The Sounds of the Psalter: Computational Analysis of Soundplay". *Literary and Linguistic Computing*, 29(3):361–378, 2014.

[Ber83] J. Bertin. *Semiology of graphics*. University of Wisconsin Press, 1983.

[BESL14] F. Binder, B. Entrup, I. Schiller, and H. Lobin. Uncertain about Uncertainty: Different ways of processing fuzziness in digital humanities data. In *Proceedings of the Digital Humanities 2014*, 2014.

[BG07] J. Bourdaillet and J.-G. Ganascia. Practical block sequence alignment with moves. In R. Loos, S. Z. Fazekas, and C. Martín-Vide, editors, *LATA*, volume Report 35/07, pages 199–210. Research Group on Mathematical Linguistics, Universitat Rovira i Virgili, Tarragona, 2007.

[BGEH10] M. Büchler, A. Geßner, T. Eckart, and G. Heyer. Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2), 2010.

[BGHE10] M. Büchler, A. Geßner, G. Heyer, and T. Eckart. Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project. In *Proceedings of the Digital Humanities 2010*, 2010.

[BGHJ+14] T. Bögel, V. Gold, A. Hautli-Janisz, C. Rohrdantz, S. Sulger, M. Butt, K. Holzinger, and D. A. Keim. Towards visualizing linguistic patterns of deliberation: a case study of the S21 arbitration. In *Proceedings of the Digital Humanities 2014*, 2014.

[BGN08] S. Bateman, C. Gutwin, and M. Nacenta. Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, pages 193–202. ACM, 2008.

[BHG08] M. Büchler, G. Heyer, and S. Gründer. eAQUA - Bringing modern Text Mining approaches to two thousand years old ancient texts. In *Proceedings of the 4th International Conference on e-Science (IEEE08)*, 2008.

[BHW11] M. Bingenheimer, J.-J. Hung, and S. Wiles. Social network visualization from TEI data. *Literary and Linguistic Computing*, 26(3):271–278, 2011.

[BIBK11] E. Barker, L. Isaksen, K. Byrne, and E. Kansa. GAP: a neogeo approach to classical resources. In *European Conference on Complex Systems 2010*, 2011.

[BISM14] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2271–2280, Dec 2014.

[BJ14] J. M. Binder and C. Jennings. Visibility and meaning in topic models and 18th-century subject indexes. *Literary and Linguistic Computing*, 29(3):405–411, 2014.

[BKP14] L. Barth, S. Kobourov, and S. Pupyrev. Experimental Comparison of Semantic Word Clouds. In *Experimental Algorithms*, volume 8504 of *Lecture Notes in Computer Science*, pages 247–258. Springer International Publishing, 2014.

[BLB+14] M. Burch, S. Lohmann, F. Beck, N. Rodriguez, L. Di Silvestro, and D. Weiskopf. RadCloud: Visualizing Multiple Texts with Merged Word Clouds. In *Information Visualisation (IV), 2014 18th International Conference on*, pages 108–113, July 2014.

[BM92] P. J. Besl and N. D. McKay. Method for Registration of 3-D Shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[Boo13] A. Booth. Documentary Social Networks: Collective Biographies of Women. In *Proceedings of the Digital Humanities 2013*, 2013.

[Boy13] N. Boyles. Closing in on Close Reading. *Educational Leadership*, 70(4):36–41, 2013.

[BPBI10] E. Barker, C. Pelling, S. Bouzarovski, and L. Isaksen. Mapping the World of an Ancient Greek Historian: The HESTIA Project. In *Proceedings of the Digital Humanities 2010*, 2010.

[Bra12] A. J. Bradley. Violence and the Digital Humanities Text as Pharmakon. In *Proceedings of the Digital Humanities 2012*, 2012.

[Büc13] M. Büchler. *Informationstechnische Aspekte des Historical Text Re-use*. 2013.

[BW08] L. Byron and M. Wattenberg. Stacked Graphs – Geometry & Aesthetics. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1245–1252, Nov 2008.

[BWKK00] M. Q. W. Baldonado, A. Woodruff, A. Kuchinsky, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Advanced Visual Interfaces*, pages 110–119, 2000.

[CC01] H.-C. Chen and A. L. Chen. A music recommendation system based on music data grouping and user interests. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 231–238. ACM, 2001.

[CCP07] C. Collins, S. Carpendale, and G. Penn. Visualization of Uncertainty in Lattices to Support Decision-Making. In *Proceedings of the 9th Joint Eurographics / IEEE VGTC conference on Visualization*, EUROVIS'07, pages 51–58, Aire-la-Ville, Switzerland, Switzerland, 2007. Eurographics Association.

[CDC+07] M. Cammarano, X. Dong, B. Chan, J. Klingner, J. Talbot, A. Halevy, and P. Hanrahan. Visualization of Heterogeneous Data. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1200–1207, Nov 2007.

[CDP+07] T. Clement, A. Don, C. Plaisant, L. Auvil, G. Pape, and V. Goren. 'Something that is interesting is interesting them': Using Text Mining and Visualizations to Aid Interpreting Repetition in Gertrude Stein's The Making of Americans. In *Proceedings of the Digital Humanities 2007*, 2007.

[CEJ+14] H. Craig, M. Eder, F. Jannidis, M. Kestemont, J. Rybicki, and C. Schöch. Validating Computational Stylistics in Literary Interpretation. In *Proceedings of the Digital Humanities 2014*, 2014.

[CFRT14] T. Cheesman, K. Flanagan, J. Rybicki, and S. Thiel. Six Maps of Translations of Shakespeare. In B. Wiggin, C. Macleod, D. DiMassa, and N. Theis, editors, *Un/Translatables: New Maps for Germanic Literatures*. Northwestern University Press, 2014.

[CGD+09] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. "Make New Friends, but Keep the Old" – Recommending People on Social Networking

Sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 201–210. ACM, 2009.

[CGM+12] M. Chaturvedi, G. Gannod, L. Mandell, H. Armstrong, and E. Hodgson. Myopia: A Visualization Tool in Support of Close Reading. In *Proceedings of the Digital Humanities 2012*, 2012.

[CGPW02] P. Clough, R. Gaizauskas, S. S. L. Piao, and Y. Wilks. METER: MEasuring TExt Reuse. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 152–159, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[CL13] K. Coles and J. G. Lein. Solitary Mind, Collaborative Mind: Close Reading and Interdisciplinary Research. In *Proceedings of the Digital Humanities 2013*, 2013.

[CLC+15] M. Chi, S. Lin, S. Chen, C. Lin, and T. Lee. Morphable Word Clouds for Time-Varying Text Data Visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 21(12):1415–1426, Dec 2015.

[CLK+14] J. Choo, C. Lee, H. Kim, H. Lee, Z. Liu, R. Kannan, C. Stolper, J. Stasko, B. Drake, and H. Park. VisIRR: Visual Analytics for Information Retrieval and Recommendation with Large-Scale Document Data. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 243–244, Oct 2014.

[CLN86] D. B. Carr, R. J. Littlefield, and W. L. Nichloson. Scatterplot matrix techniques for large n. In *Proceedings of the Seventeenth Symposium on the interface of computer sciences and statistics on Computer science and statistics*, pages 297–306, New York, NY, USA, 1986. Elsevier North-Holland, Inc.

[CLT+11] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards Better Understanding of Evolving Topics in Text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421, Dec 2011.

[CLWW14] W. Cui, S. Liu, Z. Wu, and H. Wei. How Hierarchical Topics Evolve in Large Text Corpora. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2281–2290, Dec 2014.

[CM84a] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[CM84b] W. S. Cleveland and R. McGill. The Many Faces of a Scatterplot. *Journal of the American Statistical Association*, 79(388):807–822+, 1984.

[Cob05] A. Coburn. Text Modeling and Visualization with Network Graphs. In *Proceedings of the Digital Humanities 2005*, 2005.

[Col01] R. Cole. Automated Layout of Concept Lattices Using Layered Diagrams and Additive Diagrams. In *Proceedings of the 24th Australasian Conference on Computer Science*, ACSC '01, pages 47–53, Washington, DC, USA, 2001. IEEE Computer Society.

[CRS+14] A. Christie, S. Ross, J. Sayers, K. Tanigawa, and I.-M. R. Team. Z-Axis Scholarship: Modeling How Modernists Write the City. In *Proceedings of the Digital Humanities 2014*, 2014.

[CS14] Q. Castellà and C. Sutton. Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 665–676. ACM, 2014.

[CSV08] A. Ciula, P. Spence, and J. M. Vieira. Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project. *Literary and Linguistic Computing*, 23(3):311–325, 2008.

[CTA+13] T. Clement, D. Tcheng, L. Auvil, B. Capitanu, and J. Barbosa. Distant Listening to Gertrude Stein's 'Melanctha': Using Similarity Analysis in a Discovery Paradigm to Analyze Prosody and Author Influence. *Literary and Linguistic Computing*, 28(4):582–602, 2013.

[CVW09] C. Collins, F. Viégas, and M. Wattenberg. Parallel Tag Clouds to explore and analyze faceted text corpora. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 91–98, Oct 2009.

[CWG11] M. Correll, M. Witmore, and M. Gleicher. Exploring collections of tagged text for literary scholarship. *Computer Graphics Forum*, 30(3):731–740, 2011.

[CWL+10] W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu. Context preserving dynamic word cloud visualization. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, pages 121–128, March 2010.

[DCCW08] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1205–1212, Nov 2008.

[Den99] B. D. Dent. *Carography: Thematic Map Design*. McGraw-Hill, 5th edition, 1999.

[DES+15] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, and K. Hofland. Compare Clouds: Visualizing Text Corpora to Compare Media Frames. In *Proc. of IUI Workshop on Visual Text Analytics*, 2015.

[DFM+08] O. Dyens, D. Forest, P. Mondou, V. Cools, and D. Johnston. Information visualization and text mining: application to a corpus on posthumanism. In *Proceedings of the Digital Humanities 2008*, 2008.

[DG92] G. Deleuze and F. Guattari. *Tausend Plateaus. Kapitalismus und Schizophrenie II*. Merve Verlag, 1992.

[DGM08] S. Debnath, N. Ganguly, and P. Mitra. Feature Weighting in Content Based Recommendation System Using Social Network Analysis. In *Proceedings of the 17th international conference on World Wide Web*, pages 1041–1042. ACM, 2008.

[Die07] S. Diehl. *Software Visualization: Visualizing the Structure, Behaviour, and Evolution of Software*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[Dij59] E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[DLL+10] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[DNCM14] D. J. Dye, J. B. Napolin, E. Cornell, and W. Martin. Digital Yoknapatawpha: Interpreting a Palimpsest of Place. In *Proceedings of the Digital Humanities 2014*, 2014.

[DRRD12] M. Dörk, N. Riche, G. Ramos, and S. Dumais. PivotPaths: Strolling through Faceted Information Spaces. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2709–2718, Dec 2012.

[DWS+12] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. Zhou. LeadLine: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102, Oct 2012.

[Ede14] M. Eder. Stylometry, network analysis, and Latin literature. In *Proceedings of the Digital Humanities 2014*, 2014.

[EGM07] D. Eppstein, M. T. Goodrich, and J. Y. Meng. Confluent Layered Drawings. *Algorithmica*, 47(4):439–452, 2007.

[EHJ15] T. Efer, G. Heyer, and J. Jost. Text Mining am Beispiel der Dramen Shakespeares. In C. Jansohn, W. Habicht, D. Mehl, and P. Redl, editors, *Shakespeare unter den Deutschen*, pages 217–230, Stuttgart, 2015. Akademie der Wissenschaften und der Literatur, Mainz, Franz Steiner Verlag.

[EJ14] C. Evans and B. Jasnow. Mapping Homer's Catalogue of Ships. *Literary and Linguistic Computing*, 29(3):317–325, 2014.

[ESK04] M. Eiglsperger, M. Siebenhaller, and M. Kaufmann. An Efficient Implementation of Sugiyama's Algorithm for Layered Graph Drawing. In *Proceedings of the 12th International Conference on Graph Drawing*, GD'04, pages 155–166, Berlin, Heidelberg, 2004. Springer-Verlag.

[ESK14] J. Eisenstein, I. Sun, and L. F. Klein. Exploratory Thematic Analysis for Historical Newspaper Archives. In *Proceedings of the Digital Humanities 2014*, 2014.

[EX10] M. Esteva and W. Xu. Finding Stories in the Archive through Paragraph Alignment. In *Proceedings of the Digital Humanities 2010*, 2010.

[FKT14] P. Fankhauser, H. Kermes, and E. Teich. Combining Macro- and Micro-analysis for Exploring the Construal of Scientific Disciplinarity. In *Proceedings of the Digital Humanities 2014*, 2014.

[Flu62] L.-F. Flutre. *Table des noms propres avec toutes leurs variantes: Figurant dans les romans du Moyen Age écrits en français ou en provençal et actuellement publiés ou analysés.* Poitiers: Centre d'études supérieures de civilisation médiévale., 1962.

[FS11] C. Forstall and W. J. Scheirer. Visualizing Sound as Functional N-Grams in Homeric Greek Poetry. In *Proceedings of the Digital Humanities 2011*, 2011.

[GCL+13] Z. Geng, T. Cheesman, R. S. Laramee, K. Flanagan, and S. Thiel. ShakerVis: Visual analysis of segment variation of German translations of Shakespeare's Othello. *Information Visualization*, 2013.

[GDMF+14] I. Gregory, C. Donaldson, P. Murrieta-Flores, C. Rupp, A. Baron, A. Hardie, and P. Rayson. Digital approaches to understanding the geographies in literary and historical texts. In *Proceedings of the Digital Humanities 2014*, 2014.

[GH11a] J. Goodwin and J. Holbo. *Reading graphs, maps, trees: responses to Franco Moretti*. Parlor Press, Anderson, SC, 2011. Book, Whole.

[GH11b] I. N. Gregory and A. Hardie. Visual GISting: bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistic Computing*, 26(3):297–314, 2011.

[GHKV09] E. Gansner, Y. Hu, S. Kobourov, and C. Volinsky. Putting Recommendations on the Map: Visualizing Clusters and Relations. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 345–348, New York, NY, USA, 2009. ACM.

[GKNpV93] E. R. Gansner, E. Koutsofios, S. C. North, and K. phong Vo. A Technique for Drawing Directed Graphs. *IEEE Transactions on Software Engineering*, 19(3):214–230, 1993.

[GM70] A. J. Gibbs and G. A. McIntyre. The Diagram, a Method for Comparing Sequences. Its Use with Amino Acid and Nucleotide Sequences. *Eur J Biochem*, 16(1):1–11, 1970.

[GNOT92] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[GO12] F. Gibbs and T. Owens. Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly*, 6(2), 2012.

[GOB+10] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllererk. Smallworlds: Visualizing social recommendations. In *Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'10, pages 833–842, Aire-la-Ville, Switzerland, Switzerland, 2010. Eurographics Association.

[GS06] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156, 2006.

[GTAHS15] S. J. Gray, M. Terras, R. Ammann, and A. Hudson-Smith. Textal: Unstructured Text Analysis Workflows Through Interactive Smartphone Visualisations. In *Proceedings of the Digital Humanities 2015*, 2015.

[Gut13] GuttenPlag. GuttenPlag Wiki Visualizations, 2013. http://de.guttenplag.wikia.com/wiki/Visualisierungen (Retrieved 2016-02-01).

[GWFI14] P. Goffin, W. Willett, J.-D. Fekete, and P. Isenberg. Exploring the Placement and Design of Word-Scale Visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2291–2300, Dec 2014.

[GZ12] S. Gedzelman and J.-C. Zancarini. HyperMachiavel: a translation comparison tool. In *Proceedings of the Digital Humanities 2012*, 2012.

[GZR+10] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel. Social Media Recommendation Based on People and Tags. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 194–201, New York, NY, USA, 2010. ACM.

[HAC+15] U. Hinrichs, B. Alex, J. Clifford, A. Watson, A. Quigley, E. Klein, and C. M. Coates. Trading Consequences: A Case Study of Combining Text Min-

ing and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*, 2015.

[HAHB15] R. Heuser, M. Algee-Hewitt, and J. Bender. Knowledge Networks, Juxtaposed: Disciplinarity in the Encyclopédie and Wikipedia. In *Proceedings of the Digital Humanities 2015*, 2015.

[HAHT+15] R. Heuser, M. Algee-Hewitt, V. Tran, A. Lockhart, and E. Steiner. Mapping the Emotions of London in Fiction, 1700-1900: A Crowdsourcing Experiment. In *Proceedings of the Digital Humanities 2015*, 2015.

[Har99] R. L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, 1999.

[Haw00] J. Hawthorn. *A Glossary of Contemporary Literary Theory*. Oxford University Press, 2000.

[HB03] M. Harrower and C. A. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[HB05] J. Heer and D. Boyd. Vizster: visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 32–39, Oct 2005.

[HBBS11] G. Heyer, M. Büchler, V. Boehlke, and C. Schubert. Aspects of an Infrastructure for eHumanities. *ACM Journal on Computing and Cultural Heritage*, 2011.

[HCC14] J. Hsiang, L. Chen, and C.-H. Chung. A glimpse of the change of worldview between 7th and 10th century China through two leishu. In *Proceedings of the Digital Humanities 2014*, 2014.

[HDvHM+15] R. Haentjens Dekker, D. van Hulle, G. Middell, V. Neyt, and J. van Zundert. Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, 30(3):452–470, 2015.

[Hea03] K. Head. Gravity for Beginners. *University of British Columbia*, 2053, 2003.

[Hen14] C. Henseler. Minecraft Anyone? Encouraging A New Generation of Computer Scientists and Humanists, 2014. http://tinyurl.com/lk58xlv (Retrieved 2016-02-01).

[HFM16] U. Hinrichs, S. Forlini, and B. Moynihan. Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):429–438, Jan 2016.

[HHN00] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: Visualizing Theme Changes over Time. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 115–123. IEEE, 2000.

[HK10] F. Hardisty and A. Klippel. Analysing Spatio-Temporal Autocorrelation with LISTA-Viz. *Int. J. Geogr. Inf. Sci.*, 24:1515–1526, October 2010.

[HKBE12] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2839–2848, Dec 2012.

[HKTK14] S. Howell, M. Kelleher, A. Teehan, and J. Keating. A Digital Humanities Approach to Narrative Voice in The Secret Scripture: Proposing a New Research Method. *Digital Humanities Quarterly*, 8(2), 2014.

[Hoc04] S. Hockey. The History of Humanities Computing. *A Companion to Digital Humanities*, pages 3–19, 2004.

[HPR14] E. Hoyt, K. Ponto, and C. Roy. Visualizing and Analyzing the Hollywood Screenplay with ScripThreads. *Digital Humanities Quarterly*, 8(4), 2014.

[HR08] M. Hearst and D. Rosner. Tag Clouds: Data Analysis Tool or Social Signaller? In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 160–160, Jan 2008.

[HSC08] U. Hinrichs, H. Schmidt, and S. Carpendale. EMDialog: Bringing Information Visualization into the Museum. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1181–1188, Nov 2008.

[Jas01] J. Jasinski. *Rhetoric and Society: Sourcebook on Rhetoric: Key Concepts in Contemporary Rhetorical Studies*, volume 4. Sage Publications, 2001.

[JBR+16] S. Jänicke, J. Blumenstein, M. Rücker, D. Zeckzer, and G. Scheuermann. Visualizing the Results of Search Queries on Ancient Text Corpora with Tag Pies. *To appear in Digital Humanities Quarterly*, 2016.

[JBS14] S. Jänicke, M. Büchler, and G. Scheuermann. Improving the Layout for Text Variant Graphs. In *VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 41–48, 2014.

[JC71] G. F. Jenks and F. C. Caspall. Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 61(2), 1971.

[JEBS15] S. Jänicke, T. Efer, M. Büchler, and G. Scheuermann. Designing Close and Distant Reading Visualizations for Text Re-use. In S. Battiato, S. Coquillart, J. Pettré, R. S. Laramee, A. Kerren, and J. Braz, editors, *Computer Vision, Imaging and Computer Graphics - Theory and Applications*, volume 550 of *Communications in Computer and Information Science*, pages 153–171. Springer International Publishing, 2015.

[JFCS15] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli, and I. Viola, editors, *Eurographics Conference on Visualization (EuroVis) - STARs*. The Eurographics Association, 2015.

[JFCS16] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. Visual Text Analysis in Digital Humanities. *Computer Graphics Forum*, 2016.

[JFS16] S. Jänicke, J. Focht, and G. Scheuermann. Interactive Visual Profiling of Musicians. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):200–209, Jan 2016.

[JG15] S. Jänicke and A. Geßner. A Distant Reading Visualization for Variant Graphs. In *Proceedings of the Digital Humanities 2015*, 2015.

[JGBS14a] S. Jänicke, A. Geßner, M. Büchler, and G. Scheuermann. 5 Design Rules for Visualizing Text Variant Graphs. In *Proceedings of the Digital Humanities 2014*, 2014.

[JGBS14b] S. Jänicke, A. Geßner, M. Büchler, and G. Scheuermann. Visualizations for Text Re-use. *GRAPP/IVAPP*, pages 59–70, 2014.

[JGF+15] S. Jänicke, A. Geßner, G. Franzini, M. Terras, S. Mahony, and G. Scheuermann. TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities*, 30(suppl 1):i83–i99, 2015.

[JHMK14] M. John, F. Heimerl, A. Müller, and S. Koch. A Visual Focus+Context Approach for Text Comparison Tasks. In *VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 29–32, 2014.

[JHS13] S. Jänicke, C. Heine, and G. Scheuermann. GeoTemCo: Comparative Visualization of Geospatial-Temporal Data with Clutter Removal Based on Dynamic Delaunay Triangulations. In *Computer Vision, Imaging and Computer Graphics. Theory and Application*, pages 160–175. Springer, 2013.

[JHSS12] S. Jänicke, C. Heine, R. Stockmann, and G. Scheuermann. Comparative Visualization of Geospatial-temporal Data. In *GRAPP/IVAPP*, pages 613–625, 2012.

[JKH+15] M. John, S. Koch, F. Heimerl, A. Müller, T. Ertl, and J. Kuhn. Interactive Visual Analysis Of German Poetics. In *Proceedings of the Digital Humanities 2015*, 2015.

[JLS15] J. Jo, B. Lee, and J. Seo. WordlePlus: Expanding Wordle's Use through Natural Interaction and Animation. *Computer Graphics and Applications, IEEE*, 2015.

[JME+12] Y. Jang, A. Malik, D. S. Ebert, R. Maciejewski, W. Huang, and N. Elmqvist. A correlative analysis process in a visual analytics environment. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 33–42, Washington, DC, USA, 2012. IEEE Computer Society.

[Joc12] M. Jockers. Computing and Visualizing the 19th-Century Literary Genome. In *Proceedings of the Digital Humanities 2012*, 2012.

[Joc13] M. Jockers. *Macroanalysis: Digital Methods & Literary History*. University of Illinois Press, 2013.

[JOL+15] P. Jähnichen, P. Oesterling, T. Liebmann, G. Heyer, C. Kuras, and G. Scheuermann. Exploratory Search Through Interactive Visualization of Topic Models. In *Proceedings of the Digital Humanities 2015*, 2015.

[JRS+09] C.-H. Jong, P. Rajkumar, B. Siddiquie, T. Clement, C. Plaisant, and B. Shneiderman. Interactive Exploration of Versions across Multiple Documents. In *Proceedings of the Digital Humanities 2009*, 2009.

[JS14] S. Jänicke and G. Scheuermann. Utilizing GeoTemCo for Visualizing Environmental Data. In O. Kolditz, K. Rink, and G. Scheuermann, editors, *Workshop on Visualisation in Environmental Sciences (EnvirVis)*. The Eurographics Association, 2014.

[JS16] S. Jänicke and G. Scheuermann. Tagspheres: Visualizing hierarchical relations in tag clouds. In *GRAPP/IVAPP*, 2016.

[JW13] S. Jänicke and D. J. Wrisley. Visualizing Uncertainty: How to Use the Fuzzy Data of 550 Medieval Texts? In *Proceedings of the Digital Humanities 2013*, 2013.

[Kau15] M. Kaufman. 'Everything on Paper Will Be Used Against Me': Quantifying Kissinger. In *Proceedings of the Digital Humanities 2015*, 2015.

[KBK11] M. Krstajic, E. Bertini, and D. Keim. CloudLines: Compact Display of Event Episodes in Multiple Time-Series. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2432–2439, Dec 2011.

[KG13] A. Kehoe and M. Gee. eMargin: A Collaborative Textual Annotation Tool. *Ariadne*, 71, July 2013.

[KJW+14] S. Koch, M. John, M. Worner, A. Muller, and T. Ertl. VarifocalReader – In-Depth Visual Analysis of Large Text Documents. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1723–1732, Dec 2014.

[KKL+11] H. Kim, B.-m. Kang, D.-G. Lee, E. Chung, and I. Kim. Trends 21 Corpus: A Large Annotated Korean Newspaper Corpus for Linguistic and Cultural Studies. In *Proceedings of the Digital Humanities 2011*, 2011.

[KLB14] A. Kochtchi, T. v. Landesberger, and C. Biemann. Networks of Names: Visual Exploration and Semi-Automatic Tagging of Social Networks from Newspaper Articles. *Computer Graphics Forum*, 33(3):211–220, 2014.

[Kle12] L. F. Klein. Social Network Analysis and Visualization in 'The Papers of Thomas Jefferson'. In *Proceedings of the Digital Humanities 2012*, 2012.

[KMS91] T. Kao, D. M. Mount, and A. Saalfeld. Dynamic Maintenance of Delaunay Triangulations. Technical report, University of Maryland at College Park, College Park, MD, USA, 1991.

[KO07] D. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 115–122, Oct 2007.

[KOTM13] F. Kimura, T. Osaki, T. Tezuka, and A. Maeda. Visualization of relationships among historical persons from Japanese historical documents. *Literary and Linguistic Computing*, 28(2):271–278, 2013.

[KR12] J. A. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.

[KZ14] T. Krause and A. Zeldes. ANNIS3: A new architecture for generic corpus query and visualization. *Literary and Linguistic Computing*, 2014.

[Lee07] J. Lee. A Computational Model of Text Reuse in Ancient Literary Texts. In Association for Computational Linguistics, editor, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479. 2007.

[Lev66] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Technical Report 8, 1966.

[Lew82] C. Lewis. *Using the "Thinking Aloud" Method in Cognitive Interface Design*. Research report. IBM T.J. Watson Research Center, 1982.

[LHB+15] S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl. ConcentriCloud: Word Cloud Visualization for Multiple Text Documents. In *19th International Conference on Information Visualisation*, 2015.

[Llo82] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[LMY+12] L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou. Recommender Systems. *Physics Reports*, 519(1):1–49, 2012.

[LRKC10] B. Lee, N. Riche, A. Karlson, and S. Carpendale. SparkClouds: Visualizing Trends in Tag Clouds. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1182–1189, Nov 2010.

[LSH14] X. Liu, H.-W. Shen, and Y. Hu. Supporting multifaceted viewing of word clouds with focus+context display. *Information Visualization*, 2014.

[LWW+13] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. StoryFlow: Tracking the Evolution of Stories. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2436–2445, Dec 2013.

[LZT09] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In *Human-Computer Interaction - INTERACT 2009*, volume 5726 of *Lecture Notes in Computer Science*, pages 392–404. Springer Berlin Heidelberg, 2009.

[Mar06] G. Marchionini. Exploratory Search: From Finding to Understanding. *Commun. ACM*, 49(4):41–46, April 2006.

[Mar12] S. Marche. Literature is not Data: Against Digital Humanities, 2012. https://lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities (Retrieved 2016-02-01).

[MBL+06] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial Analysis of News Sources. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):765–772, Sept 2006.

[McC15] M. M. McCabe. *Platonic Conversations*. Oxford University Press, USA, 2015.

[MFM13] L. Meneses, R. Furuta, and L. Mandell. Ambiances: A Framework to Write and Visualize Poetry. In *Proceedings of the Digital Humanities 2013*, 2013.

[MH13] A. Muralidharan and M. A. Hearst. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing*, 28(2):283–295, 2013.

[MJ76] S. Milgram and D. Jodelet. Psychological Maps of Paris. *Environmental Psychology*, pages 104–124, 1976.

[MLCM16] N. McCurdy, J. Lein, K. Coles, and M. Meyer. Poemage: Visualizing the Sonic Topology of a Poem. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):439–448, Jan 2016.

[MLSU13] B. Miller, F. Li, A. Shrestha, and K. Umapathy. Digging into Human Rights Violations: phrase mining and trigram visualization. In *Proceedings of the Digital Humanities 2013*, 2013.

[Mor05] F. Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, July 2005.

[Mor13] F. Moretti. *Distant reading*. Verso, 2013.

[MRMK15] A. Medek, J. Ritter, P. Molitor, and S. Kösser. Interactive Similarity Analysis of Early New High German Text Variants. In *Proceedings of the Digital Humanities 2015*, 2015.

[MRO+12] A. MacEachren, R. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual Semiotics & Uncertainty Visualization: An Empirical Study. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2496–2505, Dec 2012.

[MSR+15] J. Montague, J. Simpson, G. Rockwell, S. Ruecker, and S. Brown. Exploring Large Datasets with Topic Model Visualizations. In *Proceedings of the Digital Humanities 2015*, 2015.

[Mun09] T. Munzner. A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):921–928, 2009.

[Mur07] S. Murugesan. Understanding Web 2.0. *IT Professional*, 9(4):34–41, July 2007.

[Mur11] A. Muralidharan. A Visual Interface for Exploring Language Use in Slave Narratives. In *Proceedings of the Digital Humanities 2011*, 2011.

[Nel99] T. H. Nelson. Xanalogical Structure, Needed Now More than Ever: Parallel Documents, Deep Links to Content, Deep Versioning, and Deep Re-Use. *ACM Computing Surveys (CSUR)*, 31(4es):33, 1999.

[New59] J. H. Newman. The Text of the Rheims and Douay Version of Holy Scripture. *The Rambler*, 1, 1859.

[NMG+13] B. Nowviskie, D. McClure, W. Graham, A. Soroka, J. Boggs, and E. Rochester. Geo-Temporal Interpretation of Archival Collections with Neatline. *Literary and Linguistic Computing*, 28(4):692–699, 2013.

[Nov04] M. Novotny. Visually effective information visualization of large data. In *In 8th Central European Seminar on Computer Graphics (CESCG 2004*, pages 41–48. CRC Press, 2004.

[NTST11] D. Q. Nguyen, C. Tominski, H. Schumann, and T. A. Ta. Visualizing Tags with Spatiotemporal References. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 32–39, July 2011.

[NW70] S. B. Needleman and C. D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[OG14] D. Oelke and I. Gurevych. A Study on Human-Generated Tag Structures to Inform Tag Cloud Layout. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces (AVI 2014)*, pages 297–304. ACM, 2014.

[OGH15] S. Odat, T. Groza, and J. Hunter. Extracting structured data from publications in the Art Conservation Domain. *Digital Scholarship in the Humanities*, 30(2):225–245, 2015.

[OKK13] D. Oelke, D. Kokkinakis, and D. A. Keim. Fingerprint Matrices: Uncovering the dynamics of social networks in prose literature. In *Computer Graphics Forum*, volume 32, pages 371–380. Wiley Online Library, 2013.

[ÓML14] T. Ó Murchú and S. Lawless. The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data. In *Proceedings of the Digital Humanities 2014*, 2014.

[Ope09] OpenBible.info. Phrase Net Bible Visualizations, 2009. http://www.openbible.info/blog/2009/03/phrase-net-bible-visualizations/ (Retrieved 2016-02-01).

[OST+10] P. Oesterling, G. Scheuermann, S. Teresniak, G. Heyer, S. Koch, T. Ertl, and G. Weber. Two-stage framework for a topology-based projection and visualization of classified document collections. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 91–98, Oct 2010.

[Pal02] W. B. Paley. TextArc: Showing word frequency and distribution in text. In *Poster presented at IEEE Symposium on Information Visualization*, volume 2002, 2002.

[PB07] M. J. Pazzani and D. Billsus. Content-Based Recommendation Systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[PBD14] E. Peña, M. Brown, and T. Dobson. On Metaphor in Text Visualization Prototypes. In *Proceedings of the Digital Humanities 2014*, 2014.

[Pet14] N. Peterson. Visualization As a Bridge to Close Reading: The Audience in The Castle of Perseverance. In *Proceedings of the Digital Humanities 2014*, 2014.

[Pie10] W. Piez. Towards Hermeneutic Markup: An architectural outline. In *Proceedings of the Digital Humanities 2010*, 2010.

[Pie13] W. Piez. Markup Beyond XML. In *Proceedings of the Digital Humanities 2013*, 2013.

[PMMR15] M. Pöckelmann, A. Medek, P. Molitor, and J. Ritter. _CATview_ - Supporting The Investigation Of Text Genesis Of Large Manuscripts By An Overall Interactive Visualization Tool. In *Proceedings of the Digital Humanities 2015*, 2015.

[PNK11] S. Pupyrev, L. Nachmanson, and M. Kaufmann. Improving Layered Graph Layouts with Edge Bundling. In U. Brandes and S. Cornelsen, editors, *Graph Drawing*, volume 6502 of *Lecture Notes in Computer Science*, pages 329–340. Springer Berlin Heidelberg, 2011.

[Poi15] T. Poibeau. Generating Navigable Semantic Maps from Social Sciences Corpora. In *Proceedings of the Digital Humanities 2015*, 2015.

[Pos07] S. Posavec. Literary Organism, 2007. http://www.stefanieposavec.co.uk/ (Retrieved 2016-02-01).

[PS06] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):693–700, September 2006.

[PTT+12] F. V. Paulovich, F. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato. Semantic Wordification of Document Collections. In *Computer Graphics Forum*, volume 31, pages 1145–1153. Wiley Online Library, 2012.

[RA00] R. L. Ribler and M. Abrams. Using Visualization to Detect Plagiarism in Computer Science Classes. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, INFOVIS '00, pages 173–178, Washington, DC, USA, 2000. IEEE Computer Society.

[RARC+15] G. Roe, A. Abdul-Rahman, M. Chen, C. Gladstone, R. Morrissey, and M. Olsen. Visualizing Text Alignments: Image Processing Techniques for Locating 18th-Century Commonplaces. In *Proceedings of the Digital Humanities 2015*, 2015.

[RD10] N. Riche and T. Dwyer. Untangling Euler Diagrams. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1090–1099, Nov 2010.

[RFH14] N. Reiter, A. Frank, and O. Hellwig. An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, 29(4):583–605, 2014.

[RGP+12] P. Riehmann, H. Gruendl, M. Potthast, M. Trenkmann, B. Stein, and B. Froehlich. WORDGRAPH: Keyword-in-Context Visualization for NETSPEAK's Wildcard Search. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1411–1423, Sept 2012.

[Rog10] S. Rogers. The Guardian - Wikileaks Iraq war logs: every death mapped, 2010. http://www.guardian.co.uk/world/datablog/interactive/2010/oct/23/wikileaks-iraq-deaths-map (Retrieved 2016-02-01).

[RPSF15] P. Riehmann, M. Potthast, B. Stein, and B. Froehlich. Visual Assessment of Alleged Plagiarism Cases. *Computer Graphics Forum*, 34(3):61–70, 2015.

[RRF+10] R. E. Roth, K. S. Ross, B. G. Finch, W. Luo, and A. M. MacEachren. A User-Centered Approach for Designing and Developing Spatiotemporal Crime Analysis Tools. In R. Purves and R. Weibel, editors, *Proceedings of GIScience*, 2010.

[RRRG05] S. Ruecker, S. Ramsay, M. Radzikowska, and A. Galey. Interface Design. In *Proceedings of the Digital Humanities 2005*, 2005.

[RSDCD+13] J. Roberts-Smith, S. DeSouza-Coelho, T. M. Dobson, S. Gabriele, O. Rodriguez-Arenas, S. Ruecker, S. Sinclair, A. Akong, M. Bouchard, M. Hong, D. Jakacki, D. Lam, A. Kovacs, L. Northam, and D. So. Visualizing Theatrical Text: From Watching the Script to the Simulated Environment for Theatre (SET). *Digital Humanities Quarterly*, 7(3), 2013.

[Ryk11] L. Ryken. *The Legacy of the King James Bible: Celebrating 400 Years of the Most Influential English Translation*. Crossway, 2011.

[SC09] D. Schmidt and R. Colomb. A Data Structure for Representing Multi-version Texts Online. *Int. J. Hum.-Comput. Stud.*, 67(6):497–514, June 2009.

[SCH08] J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, 2008.

[Sch09] D. Schmidt. Merging multi-version texts: a general solution to the overlap problem. 2009.

[SDW11] A. Slingsby, J. Dykes, and J. Wood. Exploring Uncertainty in Geodemographics with Interactive Graphics. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2545–2554, Dec 2011.

[SGL08] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[SGL09] Y. Shrinivasan, D. Gotz, and J. Lu. Connecting the Dots in Visual Analysis. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 123–130, Oct 2009.

[SH93] K. Sachs-Hombach. Das Bild als kommunikatives Medium. *Elemente einer allgemeinen Bildwissenschaft*, 1993.

[Shn96] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Visual Languages, Proceedings*, pages 336–343, 1996.

[SKG11] C. Seifert, W. Kienreich, and M. Granitzer. Visualizing Text Classification Models with Voronoi Word Clouds. In *Proceedings of the International Conference Information Visualisation (IV), London*, 2011.

[SKK+08] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer. On the Beauty and Usability of Tag Clouds. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 17–25, July 2008.

[SMKH09] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard. *Thematic Cartography and Geovisualization*. Prentice Hall Series in Geographic Information Science. Prentice Hall, 3rd, international edition, 2009.

[SOI10] S. Saito, S. Ohno, and M. Inaba. A Platform for Cultural Information Visualization Using Schematic Expressions of Cube. In *Proceedings of the Digital Humanities 2010*, 2010.

[SP98] B. Shneiderman and C. Plaisant. Treemaps for space-constrained visualization of hierarchies, 1998.

[ST14] J. Schrammel and M. Tscheligi. Patterns in the Clouds - The Effects of Clustered Presentation on Tag Cloud Interaction. In *Building Bridges: HCI, Visualization, and Non-formal Modeling*, Lecture Notes in Computer Science, pages 124–132. Springer Berlin Heidelberg, 2014.

[STT81] K. Sugiyama, S. Tagawa, and M. Toda. Methods for Visual Understanding of Hierarchical System Structures. *Systems, Man and Cybernetics, IEEE Transactions on*, 11(2):109–125, Feb 1981.

[SvdWvW14] R. Scheepens, H. van de Wetering, and J. van Wijk. Non-overlapping Aggregated Multivariate Glyphs for Moving Objects. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 17–24, March 2014.

[SWL+10] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 99–106, Oct 2010.

[Szp14] R. Szpiech. Cracking the code: Reflections on manuscripts in the age of digital books. *Digital Philology: A Journal of Medieval Cultures*, 3(1):75–100, 2014.

[Tal12] B. B. Taliaferro. *Encyclopedia of English Language Bible Versions*. McFarland, Incorporated, Publishers, 2012.

[Tan10] G. T. Tanselle. *A Rationale of Textual Criticism*. University of Pennsylvania Press, 2010.

[TFK15] P. Trilcke, F. Fischer, and D. Kampkaspar. Digital Network Analysis of Dramatic Texts. In *Proceedings of the Digital Humanities 2015*, 2015.

[TKK11] S. Tsipidis, A. Koussoulakou, and K. Kotsakis. Geovisualization and Archaeology: supporting Excavation Site Research. In A. Ruas, editor, *Advances in Cartography and GIScience. Volume 2*, volume 6 of *Lecture Notes in Geoinformation and Cartography*, pages 85–107. Springer Berlin Heidelberg, 2011.

[Tót13] G. M. Tóth. The computer-assisted analysis of a medieval commonplace book and diary (MS Zibaldone Quaresimale by Giovanni Rucellai). *Literary and Linguistic Computing*, 28(3):432–443, 2013.

[Tra09] C. Travis. Patrick Kavanagh's Poetic Wordscapes: GIS, Literature and Ireland, 1922-1949. In *Proceedings of the Digital Humanities 2009*, 2009.

[Tuk77] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[UBSE98] J. Utech, J. Branke, H. Schmeck, and P. Eades. An Evolutionary Algorithm for Drawing Directed Graphs. In *Proceedings of the International Conference on Imaging Science, Systems and Technology*, pages 154–160. CSREA Press, 1998.

[VCPK09] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. What's being said near "Martha"? Exploring name entities in literary text collections. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 107–114, Oct 2009.

[vHWV09] F. van Ham, M. Wattenberg, and F. Viégas. Mapping Text with Phrase Nets. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1169–1176, Nov 2009.

[VPBD13] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362, New York, NY, USA, 2013. ACM.

[VS12] M. Vlachos and D. Svonava. Recommendation and visualization of similar movies using minimum spanning dendrograms. *Information Visualization*, page 1473871612439644, 2012.

[VW08] F. Viégas and M. Wattenberg. TIMELINES: Tag Clouds and the Case for Vernacular Visualization. *interactions*, 15(4):49–52, July 2008.

[VWF09] F. Viégas, M. Wattenberg, and J. Feinberg. Participatory Visualization with Wordle. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1137–1144, Nov 2009.

[VWvH+07] F. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: a Site for Visualization at Internet Scale. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1121–1128, Nov 2007.

[War13] C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2013.

[Wea08] C. Weaver. Multidimensional visual analysis using cross-filtered views. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 163–170, Oct 2008.

[Wea10] C. Weaver. Multidimensional data dissection using attribute relationship graphs. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 75–82, Oct 2010.

[Wes15] M. Weskamp. Newsmap, 2015. http://newsmap.jp/ (Retrieved 2016-02-01).

[WH11] J. A. Walsh and W. Hooper. Computational Discovery and Visualization of the Underlying Semantic Structure of Complicated Historical and Literary Corpora. In *Proceedings of the Digital Humanities 2011*, 2011.

[Wil05] L. Wilkinson. *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[Wil15] T. Wills. Relational data modelling of textual corpora: The Skaldic Project and its extensions. *Digital Scholarship in the Humanities*, 30(2):294–313, 2015.

[WJ13a] S. Weingart and J. Jorgensen. Computational analysis of the body in European fairy tales. *Literary and Linguistic Computing*, 28(3):404–416, 2013.

[WJ13b] D. Wheeles and K. Jensen. Juxta Commons. In *Proceedings of the Digital Humanities 2013*, 2013.

[WMN+14] B. Walsh, C. Maiers, G. Nally, J. Boggs, and P. P. Team. Crowdsourcing individual interpretations: Between microtasking and macrotasking. *Literary and Linguistic Computing*, 29(3):379–386, 2014.

[Wol13] M. Wolff. Surveying a Corpus with Alignment Visualization and Topic Modeling. In *Proceedings of the Digital Humanities 2013*, 2013.

[WPW+11] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-Preserving Word Clouds by Seam Carving. In *Computer Graphics Forum*, volume 30, pages 741–750. Wiley Online Library, 2011.

[WSK+13] M. Waldner, J. Schrammel, M. Klein, K. Kristjánsdóttir, D. Unger, and M. Tscheligi. FacetClouds: Exploring Tag Clouds for Multi-dimensional Data. In *Proceedings of Graphics Interface 2013*, GI '13, pages 17–24. Canadian Information Processing Society, 2013.

[WV08] M. Wattenberg and F. B. Viégas. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228, November 2008.

[WZG+14] J. Wang, J. Zhao, S. Guo, C. North, and N. Ramakrishnan. ReCloud: Semantics-based Word Cloud Visualization of User Reviews. In *Proceedings of the 2014 Graphics Interface Conference*, GI '14, pages 151–158. Canadian Information Processing Society, 2014.

[YMSJ05] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & Magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

[ZNMS15] T. Zahora, D. Nikulin, C. J. Mews, and D. Squire. Deconstructing Bricolage: Interactive Online Analysis of Compiled Texts with Factotum. *Digital Humanities Quarterly*, 9(1), 2015.