Visualizing a Thinker's Life

Patrick Riehmann, Dora Kiesel, Martin Kohlhaas, and Bernd Froehlich

Abstract—This paper presents a visualization framework that aids readers in understanding and analyzing the contents of medium-sized text collections that are typical for the opus of a single or few authors.We contribute several document-based visualization techniques to facilitate the exploration of the work of the German author Bazon Brock by depicting various aspects of its texts, such as the *TextGenetics* that shows the structure of the collection along with its chronology. The *ConceptCircuit* augments the *TextGenetics* with entities – persons and locations that were crucial to his work. All visualizations are sensitive to a wildcard-based phrase search that allows complex requests towards the author's work. Further development, as well as expert reviews and discussions with the author Bazon Brock, focused on the assessment and comparison of visualizations based on automatic topic extraction against ones that are based on expert knowledge.

Index Terms—Glyph-based Techniques, Text and Document Data, Coordinated and Multiple Views

1 Introduction

In the Big Data era, most text-based visualizations and analysis tools are intended to cope with larger and larger text corpora. As these usually contain tens of thousands of documents, the influence of a single document on the resulting visualization is often negligible and out of focus. Thus, most approaches prefer summarizing information visually (e.g., Collins [1], Wei [2], Cui [3]). However, for scholars working with the opus of a single author or another medium-sized text collection, a document-based visualization is desirable or even necessary. One can argue that the information need of such scholars differ compared to researchers focusing on larger text collections. Scholars have to be able to work with particular texts, yet also benefit from visual information gained from the entire collection or single documents. They have to compare texts regarding different aspects as well as search through and filter the documents accordingly. Verifying hypotheses about the internal structure of the text collection – like when certain topics occurred or vanished – or finding singular relationships between certain documents are common tasks while exploring such a medium-sized corpus.

The opus of Bazon Brock, an artist and profound thinker on art history and aesthetics, is a perfect use case for developing interactive visualization techniques to fulfill the information need of such scholars. His entire opus is publicly accessible on his website [4] consisting of more than 2.7 million words, 1 200 pictures, 33 videos, and 70 audio files written, recorded, and transcribed in over 55 years. He developed the method of "Action Teaching" and launched the documenta-schools for visitors. Even though Brock turned 80 in 2016, he is more active than ever.

For the visual analysis of Brock's work, we derived a particular design vocabulary basing on the ideas of Russian Constructivism [5]; following form and appearance of the Suprematism paintings of Kazimir Malevich(1878-1935), known for introducing geometric abstracts into art, especially seen in his famous picture the Black Square (see Baier and Dümpelmann [6] or Boersma et al. [7]). The principal visual concepts and colors were approved by Bazon Brock to be used for the visualization, which he refers to as *computer-aided drawing*, since in his opinion "visualization is a process that only happens in the latter stages of the visual cortex of the human mind, *obviously*".

Based on interviews with experts familiar with the matter, we identified aspects of Brock's work that needed to become accessible and explorable for apprehending the structure and the characteristics of the entire text collection as well as the relationships between single texts. In other words, as one expert put it, "What can Brock tell us?" and "What questions do we have towards Bazon?" As a result, we developed and evaluated appropriate visualization and interaction techniques enabling users to explore different aspects of Brock's work.

- The *TextGenetics* presents the oeuvre's time and structure.
- The *ConceptCircuit* expresses entities such as persons and locations important to his opus.
- The *TopicAssembly* focuses on Brock's topics.
- The *TopicGenetics*, *TopicCalendar* and *TopicTrends* show expert-curated topic occurrences in Brock's periods of creativity.
- A wildcard-based phrase search for answering questions towards Bazon Brock's work by visually adjusting the visualizations, which enables the readers to interactively explore the search results as well as the documents relevant to the query.
- Comparisons of modeled topics against expert-curated ones indicate that neither of over 90 tested LDA (Latent Dirichlet Allocation) configurations nor the HDP (Hierarchical Dirichlet Process) results are aligned to expert-curated topics.

During the development we conducted reviews with Bazon

- *P. Riehmann, D. Kiesel, and B. Froehlich are with theVirtual Reality and Visualization Research Group at Bauhaus-Universität Weimar. E-mail: <first>.<last>@uni-weimar.de*
- *Martin Kohlhaas is with Kohlhaas&Kohlhaas Agency in Weimar. E-mail: m.kohlhaas@kohlhaas-kohlhaas.de*

Manuscript received April 19, 2005; revised August 26, 2015.

Brock himself and with three experts being very familiar with his work. The reviews provided very positive feedback and inspirations for further research and development. In particular, due to dissatisfaction with the automatic topic extraction, we created a catalog of expert-curated topics and had our experts tag each document of Brock's corpus. A comparison of the automatically extracted topics against the expert-curated ones indicated that automatic topic extraction does not work well for collections involving sophisticated text and language such as Bazon Brock's work.

2 Related Work

Well designed (static) data graphics are often a good starting point for getting familiar with a particular artist. Accurat [8] provided such a visualization focusing on the life and work of the painter Kandinsky. Pellegrini [9] and Ciuccarelli [10] present the work of Immanuel Kant using word lists and stream graphs to show word frequencies from each year of his opus. The system presented in this paper will extend these approaches by providing multiple, interactive views to explore different aspects of Brock's work. Additionally, some visual metaphors were meant to become real objects to be used in exhibitions about Brock. As works on a single author or other mid-sized text corpora are rare, we drew from publications on larger corpora as well.

2.1 Topic-based Visualization

Visualizations based on topics support abstraction beyond simple word lists and tag clouds. *Themail* [11] illustrates topics in email archives by plotting keywords and emails in columns along a discrete time line. Similarly, Collins et al. [1] visualized differences in topics between subsets of documents. Gansner et al. [12] introduce *TwitterScope*, that utilizes a map metaphor to depict clusters of tweets and their relationships. *InfoSky* [13] uses the telescope metaphor; the documents are shown as stars while the hierarchical structure is represented as constellations. The *FacetAtlas* [14] visualizes the relationships between documents grouped by a chosen primary facet. The *TopicFlower* [15] combines topic modeling with known categories to create a visualization that reveals the connections between them. Wei et al. [2] and Leskovec et al. [16] rely on stream graphs to depict the topics' evolution. *TextFlow* [17] extends this idea using a river metaphor, where topics are able to split and merge. Xu et al. [18] and Havre et al. [19] focus on the visualization of topic competition. The *Serendip* [20] allows for the exploration of topics on the corpus level, using a highly reorderable document-topic-matrix. We propose a new way of arranging documents in space according to their topical representation: barycentric coordinates. Additionally, we depict distinguishable document representations for comparing topic occurrences across documents and over time.

2.1.1 Topic Modeling

A common tool to model topics within text is Latent Dirichlet Allocation (LDA) [21] and its many extensions [22]: e.g. dynamic topic models [23] for evolving topics, or hierarchical topic models [24]. However, Boy [25] shows that the output of basic LDA is more meaningful to humans than for example that of correlation-based LDA [26]. Hierarchical Dirichlet Processes (HDP) [27] are closely related to LDA and infer the number of topics as well as the topics themselves. To improve the results of topic modeling, Choo et al. [28] enables the user to semi-supervise the process. Our work reveals shortcomings of LDA and HDP applied to a non-trivial and highly elaborate corpus by allowing a detailed comparison of modeled topics against topics that were curated by experts.

2.1.2 Layout of Topic Data

Multi-dimensional scaling (MDS) [29] is often used to generate a two-dimensional layout for n-dimensional data such as the results of an LDA. As alternative we use barycentric coordinates, which express a point as a weighted average of the vertices of a simplex. Similarly, the LDA represents a document as a weighted mix of the extracted (latent) topics and thus barycentric coordinates appear to be a natural paradigm for visualizing the results of an LDA. Cheng and Mueller [30] provide an overview of barycentric coordinates in data visualization, which have been used by several bioinformation systems [31], [32]. We augmented the MDS layout with a heuristic placement of the topics around the canvas and display the documents as glyphs instead of simple points, which provide an impression of the weights of the topics in each document. To our knowledge, we are the first to suggest the use of a barycentric coordinate system for visualizing the results of an LDA, which seems a rather natural mapping and also allows the use of glyphs as a document representation.

2.2 Named-entity-based Visualization

Named entity based visualizations try to connect raw text to known entities, such as the names of persons or locations. Woodward et al. [33] extract locations, people and events of historical Wikipedia articles and display them along a time line and on a world map. Similarly, *GeoTracker* [34] and *FEMARepViz* [35] place RSS feeds and national situation reports in time and space. The *LeadLine* [36] additionally uses vertical violin plots to depict the topics of the events. Harrison [37] presents an biblical association graph that displays people, places, and their relationships. With the *ConceptCircuit*, we contribute novel glyph-based visual metaphors for connecting entities to their occurrences in the text collection.

2.3 Tag clouds

Tag clouds and word lists are popular for showing rough ideas about documents or text collections. A study about which words to select for a tag cloud was proposed by Venetis et al. [38]. Cui et al. [3] and Lee et al. [39] extended tag clouds to depict a time series of documents. The *Word Cloud Explorer* [40] provides additional information like co-occurring terms for each tag. Endert's [41] *Typograph* arranges texts as well as tags in space, transitioning between them with semantic zooming. *Compare Clouds* [42] can compare different corpora or facets of one corpus by showing the overlap as well as the differences in words connected to a selected anchor term. *Gist Icons* [43] are glyphs that

present the term histogram of a large number of documents. Our system incorporates word clouds as representatives of automatically extracted topics in the *TopicAssembly*. Additionally, we extended the *Gist Icons* to work as document representations in a topic space.

2.4 Pixel-based Visualizations

Although our visualizations were not designed as pixel-oriented techniques [44] per se, we deal with numerous small glyphs. Moreover, we incorporate small (near)-pixel-based miniatures that serve as overviews by providing a distinctive pixel pattern per topic. Literature fingerprinting [45] suggests a pixel-based technique that visualizes local document features as a sequence of color-mapped blocks each representing the feature value for a chunk of text. So far, we focus on global document features but our work could benefit from an analysis of local document features to support the detection of changes in writing style over time.

2.5 N-gram-based Visualizations

N-gram based visualizations support the analysis of phrases and words (their variations and frequencies). The *Word Tree* [46] uses a tree metaphor to break down all possible phrases beginning with a given root term. Harrison [47] used the same system to create the *Web Trigrams*, which show the differences in the usage of pronouns. Luz et al. [48] extended this to form a keyword-in-context (KWIC) technique with a bi-directional hierarchical view that arranges preceding and following words according to their frequency. A more sophisticated KWIC visualization is proposed by Riehmann et al. [49]. The *WordGraph* is able to display not only the context of a single selected keyword but also of key phrases that contain wild-cards. We did not aim at creating an advanced KWIC visualization, but our combination of wildcard-based search with interactive visualizations significantly enriches the expressiveness of our interface.

3 Approaching Bazon Brock Visually

The corpus of Bazon Brock was brought to our attention by the design agency Kohlhaas&Kohlhaas who was responsible for Brock's webpage. Being aware of the uniqueness of the corpus, they aimed at a joint project in order to advance support for Brock's readers - novices, as well as experienced readers. Brock's corpus consists of 2 160 documents, including articles and book chapters as well as transcripts of discussions, films or TV reports, that have been published in 946 works – some publications consist of more than one document – and cover 50 years spanning 1958 to 2008. A document, thus, can either be a standalone article or a chapter of a larger work. The cosine similarity applied to the document's word vectors could barely reveal any similarities among them. Only a few clusters of 8 to 10 documents exist that contain pair-wise similarities above a threshold of 0.8. The rest are not similar at all.

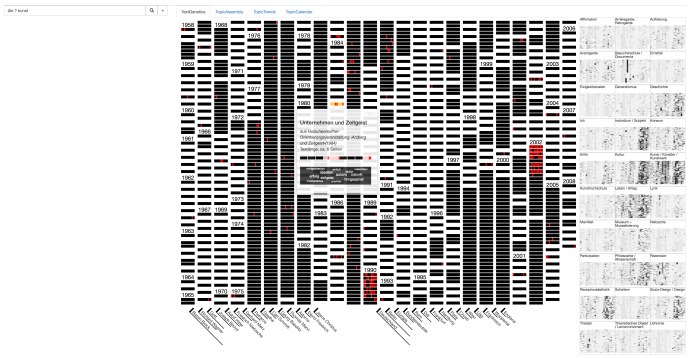
In order to know more about how scholars are going to approach such a writer, we did initial interviews with two experts at different stages of their PhDs about Brock's work (we hadn't been in touch with the third expert yet). Only the *TextGenetics* existed back then as a first draft without any further interaction - it only presented links to other webpages showing the respective texts. Our experts were very fond of the general idea of examining his work over time and assessed it as particularly useful for novices since *TextGenetics* shows the aspect of time and structure by aligning document-representing glyphs along a particular timeline wrapped into columns. It was supposed to present every document as a separate visual gylph which was crucial in our eyes for a mid-sized corpus. This visualization was all set from the beginning and was intended to be the visualization that all other visualizations originate their shapes and appearance from.

Although they were very fond of the *TextGenetics*, they stated it lacked information about work periods and associated topics. This was most important to them – identifying, surveying and comparing potential topics Brock was dealing with in his writings and, especially, how they have changed over time. They hoped to be able to verify hypotheses more fluently regarding occurrences of different topics across the opus or to find singular relationships between certain documents. So, what changes or patterns exist in his choice of topics over the decades? Due to the importance of the topic issue to our experts, we designed the *TopicAssembly* as another main visualization representing the modeled and later manually curated topics by positioning so-called *TopicStars* according to their affinity towards a certain topic in order to reveal relations and patterns among Brock's topics (see Section 3.3).

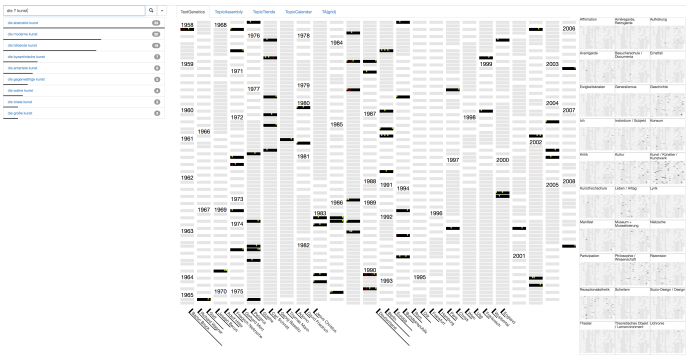
The topics are addressed even further as so the called *TopicGenetics* – small topic-wise multiples of *TextGenetics* – intended as a temporal topic overview visualization for filtering and exploring *TextGenetics* topic-wise (see Section 7). Aiming to support the experts wishes to relate and compare topic patterns and topic progression over time, *TopicTrends* and *TopicCalendar* were created; the former one a very tight display for recognizing and comparing document belongings directly in close proximity. The latter is of a more loose appearance and presents a hierarchically explorable display taking advantage of the inherent topic hierarchy provided by organizing Brock's opus bottom up from documents and works up to years and decades (see Section 7).

The experts' second (and third) point was having proper search capabilities available for words and text, but they were extremely interested in searching for persons such as colleagues (other philosophers), contemporaries, or historic persons Brock was engaged with. Who were the people he mentioned most and in which topical contexts and decades did this happen? Do co-occurrences of two or more people exist that appear multiple times? The *ConceptCircuit* augments the *TextGenetics* with specialized glyphs derived from circuit lanes to route the user to documents containing entities – people and places – important to Brock (see Section 3.2). The search for persons seemed to be so essential to our experts that it was directly attached to the glyphs of the main visualization instead of simply providing a specialized search facility.

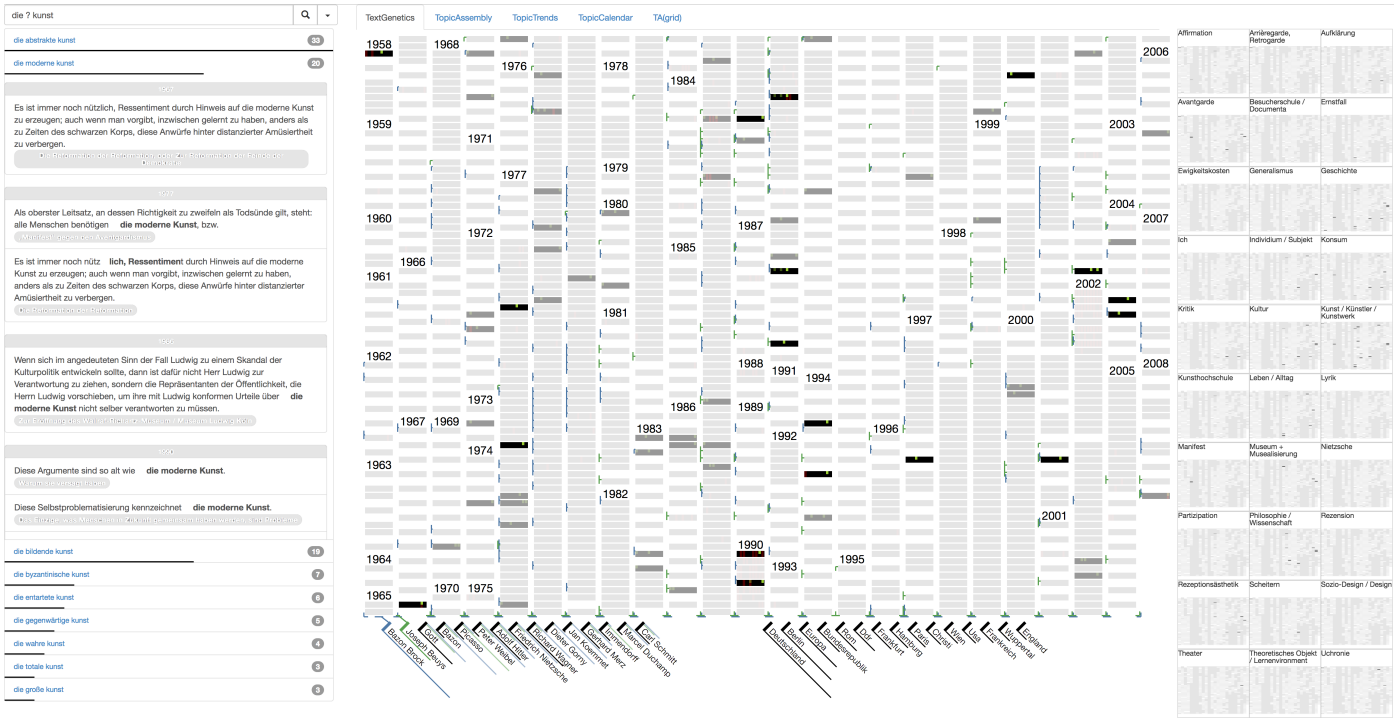
Furthermore, all visualizations were intended to be (and are) sensitive to a wildcard-based phrase search engine in order to accommodate the experts' wishes. The search enables the user to enrich a search term with wildcards like



(a) The TextGenetics and ConceptCircuit



(b) The TextGenetics and ConceptCircuit



(c) The entities Bazon Brock and Joseph Beuys selected in the ConceptCircuit

Fig. 1. Subfigure (a) shows the initial display that consists of the search bar (top), the *TextGenetics* (center), and the most important entities of the *ConceptCircuits* (below). The miniatures so called *TopicGenetics* on the right give insights about Brock's topics (see Section 3). Additionally, detail is provided on demand about a certain document. The search results of the query *die ? kunst* (*the ? art*) yield a phrase list (left) along with the respective answering phrases and the filtered *TextGenetics*can be seen in Subfigure (b). In Subfigure (c) a certain phrase has been selected, which shows its embeddings within the text in the list and narrows down the search results in the *TextGenetics*. Additionally, the entities Bazon Brock and Joseph Beuys selected in the *ConceptCircuit* for comparing their influence towards the filtered documents.

die ? kunst * (*the ? art* *) in which the question mark stands for exactly one arbitrary word and the asterisk substitutes zero to many words. Phrases that match search queries are displayed on the left in our interface (Figure 1(b)) and are ordered according to their number of occurrences top down, showing their absolute and relative frequency as a black bar respectively as a label. Selecting one or more items in the list (as shown in Figure 1(c)) reduces the number of highlighted documents further (see Section 4).

In order to fulfill another element of feedback from our experts – having information about the texts as well as the text itself more easily accessible than linking to another website – the list of phrases can be augmented by phrase embeddings showing the context in which the phrases occur (see Figure 1(c)). The embeddings change according to phrases found and selected as well as to the user's interaction with the visualizations in order to gain specific knowledge by the different aspects of topics, structure, time, and entities or phrases.

Writing style analysis and the evolution of his typical writing style were other issues the experts would like to examine. However, that could not be met in our developments so far.

The entire web-based application employs simple, yet powerful, metaphors to visualize the written words of Bazon Brock. The design vocabulary is influenced by the paintings of Kazimir Malevich [6], [7], especially by his Suprematistic cycle such as Suprematist Composition (Oil on canvas) [51]. The Russian Avant-Garde at the dawn of the 20th century with their simple forms, clear lines, particular

angles and strong colors worked to our advantage since it led to succinct visual metaphors. We followed the principles of Malevich's paintings rather than those of other abstract artists like Kandinsky for example, "an experimental artist, [who] approached abstraction tentatively and visually, by gradually and progressively concealing forms drawn from nature, whereas Malevich, a conceptual innovator, plunged precipitously into abstraction, by creating symbolic elements." [50] While Kandinsky's work appeared too playful and lavish to us, Malevich's is very stringent and clear.

3.1 TextGenetics

The *TextGenetics* visualizes the text collection based on the aspects of time and structure by arranging the documents along a vertical timeline wrapped into multiple columns. We consider the documents as being the most important entity, hence they are designed to be the most "prägnant" (Gestalt Law) elements. We opted for black bars on white ground to represent the documents of the text collection, since they also represent the most eye-catching elements used in Malevich's paintings. In exhibitions of Brock's work, this design has also been used for large prints to facilitate distant reading. It is even planned to use the design for 3D objects in future exhibitions.

In order to facilitate the match between a position on the timeline and an actual point in time, each year with at least one publication is depicted as a label, followed by a sequence of black bars representing each document written during this year. Two different vertical spacing steps indicate whether a document is published on its own or belongs to a larger publication. Some of the documents are augmented with red bars showing the "worthiness of reading" expressing so-called *essences* – passages that were manually tagged by readers stating that they cover essential thoughts of Brock.

Fig. 2. The different states of document glyphs of the *TextGenetics*. (a) Fade Level 0: There are no answering phrases in the document. (b) Fade Level 1: There are answering phrases, but outside the user's phrase selection. (c) Fade Level 2: There are answering phrases within the document. (d) Highlighted: The document was selected.

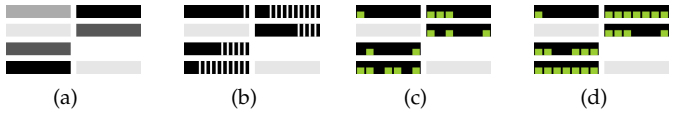


Fig. 3. Different types of search result visualizations. (a) Fading, (b) using ticks, (c) naive binning, and (d) overflow binning

TextGenetics is sensitive to the wildcard-based search, which aids the user in narrowing down by providing a set of found phrases on the left to select from. The matching documents in the *TextGenetics* (see Figure 1(b)) remain black while the others are grayed out. Three gray levels show whether a document contains no phrases (Figure 2(a)), it contains phrases outside the user's selected subset (Figure 2(b)), or it contains phrases within the user's selection (Figure 2(c)). We chose to introduce an intermediate gray level to remind the user that the emphasized documents in black represent only a subset of the original wildcard-based search results. Particular selected documents turn yellow (Figure 2(d)). Green markers express the location of a matching phrase (similar to the red *essences*) in a document bar as a *Naive Binning* inspired by TileBars [52]. We tested different alternatives (Figure 3) starting from using different gray levels or thin ticks for showing absolute number of contained phrases. However, both lacked information about the location of the found phrase in the document. The *Naive Binning* is able to provide the phrase position by horizontally dividing the glyph in n bins (depending on screen resolution); representing n chunks of the text. If at least one phrase is found in the respective text chunk its bin is filled in green (red as an essence). Thus, users are able to derive the importance of a document by assessing the distribution of contained answering phrases. Yet, it tends to underrate especially long documents, as each bin covers large portions of text and multiple answering phrases per chunk are highly probable. On the other hand, documents containing fewer but evenly distributed phrases receive a high visual rating. This fact may confuse readers that are not familiar with the concept of binning. We tested an *Overflow Binning* method to circumvent this problem by forcing phrases to overflow into a bin, if it is already highlighted, to the nearest untriggered bin. This ensures that two results are actually perceived as two highlighted bins no matter how close they are within the text. Generally this works quite well for most reasonable queries but for queries returning many hits in a single document, it is possible that we run out of bins. However, we did not add a particular overflow tag or color to avoid further complexity. Even though we consider *Overflow Binning* as an interesting alternative for avoiding bin overplotting, we chose *Naive Binning* as the default in order to preserve location accuracy instead of focusing on scalability.

Detailed information on the textual level is provided on demand (Figure 4 left), such as document title, the title and year of the publication as well as an approximate number of pages. Besides those facts, we provide a more detailed document glyph for the *essences* and answering phrases, respectively. The detailed glyphs are binned and highlighted in the same manner as in the *TextGenetics*. Additionally, more exact positions of both are given on the detail glyphs in a darker shade of the respective color. Multiple answering phrases contained in the document are presented by derived shades of green below the detailed glyphs and above a miniature word cloud giving a more detailed impression of the document's content.

3.2 ConceptCircuit

The *ConceptCircuit* augments the *TextGenetics* by visualizing important entities such as places and persons most frequently occurring in Brock's work. The glyphs that connect the entity labels to their occurrences are derived from electronic circuit diagrams. The wires are usually oriented either vertically or horizontally and running side by side when heading in the same direction (see Figure 5). In order to further facilitate visual mapping, glyphs and labels are colored the same (see Figure 1(c)).

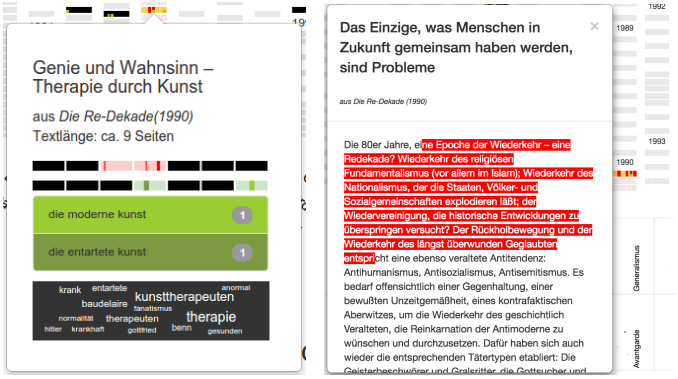


Fig. 4. Details on demand provide the positions of *essences* and answering phrases at better granularity or give access to the text level.

Eventually, we decided against showing a complete wire visualization due to certain unintended effects. The wires cover a large amount of screen space which results in significant visual clutter and such visual prominence that the *TextGenetics* literally becomes background, which contradicts our intention. Besides this, complete wires could indicate false positives. Due to the Gestalt Law of Proximity, lanes near a particular document glyph could be interpreted as markings. Lanes along the columns of the *TextGenetics* seem to mark all document glyphs right to them and not only those that they are actually connected to. We solved these issues by reducing the connections, stripping the lanes, and keeping glyphs only at routing positions: at bends resulting in ⌞-shaped glyphs, splits resulting in ⊤-shaped glyphs and the connections to the document glyphs and entity. Even this reduced forms resemble the typical rectangular line crossing often appearing in Malevich's work (see also Section 3). The ⊥-shaped glyph tells the user that there are other columns to the right and left in which the entity occurs, as well. ⌞- and ⌟-shaped glyphs, on the other hand, work like brackets and tell the user that either there are no occurrences of the entity on the further right (⌟) or on the further left (⌞). Therefore the user surely knows which columns contain entities and which do not. Similar to the glyphs on the horizontal line, the glyphs along the columns show which and how many of the selected entities are contained in the document they point at. As before, for each entity contained in the text, a glyph is drawn that can be either ⊢- or ⌜-shaped. A ⊢-shaped glyph shows the user that there are other documents containing the entity further up, whereas an ⌜-shaped one marks the upmost document glyph in the column. Using the glyphs, the user is able to tell, at a glance, which documents contain selected entities.

The resulting glyphs are smaller in comparison to the bars of the *TextGenetics*, since they were meant to augment the *TextGenetics*. Increasing the circuit glyphs further would promote the circuits to the foreground and demote the *TextGenetics* to the background. This has not been our intention. We are also limited to the space between the elements of the *TextGenetics*, which is designed to fit on a single screen and its elements have to maintain a particular aspect ratio. We believe that appropriate colors enable users to recognize and distinguish up to 3 or 4 different entities (see Figure 5). Color is crucial here, yet our reduced overall color scheme limits our choices. However, during our expert reviews we never experienced a situation where more were needed.

Albeit the full-wire version is not suitable for the initial and static display of the *TextGenetics*, it is useful for leading the user's attention when interacting with the system. We temporarily restore the full circuit lanes of a particular entity while hovering above it (Figure 6 right).

The entities are depicted below the *TextGenetics* simply showing the most frequent locations and persons Brock has written about. Selecting one or more entries changes the list of presented entities (see in Figure 5). List entries are replaced with entities that are related to the chosen ones and are most likely to coincide with the reader's interest. The strength of a relationship between two entities is measured in proportion to the number of documents both entities appear in. Persons and locations can be added by searching and selecting them using a particular search mode.

In order to extract the entities, we use the Stanford Named Entity Recognizer [53] trained on a labeled corpus of a German newspaper (220k tokens) and generalized with either HGC (Huge German Corpus, 175M tokens) or deWac (German Web Crawl, 1.71B tokens). To increase precision, we applied both in conjunction with one another such that only tags are accepted that have been agreed upon by both classifiers. Subsequently, a refinement stage was needed as the resulting tags are assigned token-wise and, thus, entities consisting of two words – such as Bazon Brock – are not a recognized as one. Consecutive tokens tagged with the same label were merged and the correctness of the resulting entities was manually verified. Rare combinations were removed. This process reduced 30k person tags and 4.5k location tags to 643 person names and 359 locations in Brock's opus. This reduced set of entities functions as a database out of which the two lists of the *ConceptCircuit* are formed.

3.3 TopicAssembly

The *TopicAssembly* visualizations as shown in Figure 6 were designed to layout the documents of the text collection according to the topics they cover. Each *TopicAssembly* provides an overview of the contents of a single document as well as the whole text collection. We employ Latent Dirichlet Allocation (LDA) [21], [22] to extract n latent topics from the document collection using the LDA implementation of the Stanford Topic Modeling Toolbox [54]. As a result, each document is represented by an n-dimensional topic vector which represents the influence of each latent topic on the document. The *TopicClouds* introduce readers to the contents of the respective topic by displaying the most important words in texts covering that topic. The words are sized according to their probability provided by the LDA. For the *TopicAssembly*, the *TopicClouds* are arranged around a rectangular or polygonal canvas.

The documents are depicted on the canvas as so called *TopicStars*. Each glyph represents a single document and visually encodes its corresponding n-dimensional topic vector. The central shape of each *TopicStar* resembles the shape of the canvas and serves as an anchor for the spikes of the glyph. Each of the spikes is directed towards its respective topic surrounding the canvas and encodes the influence of

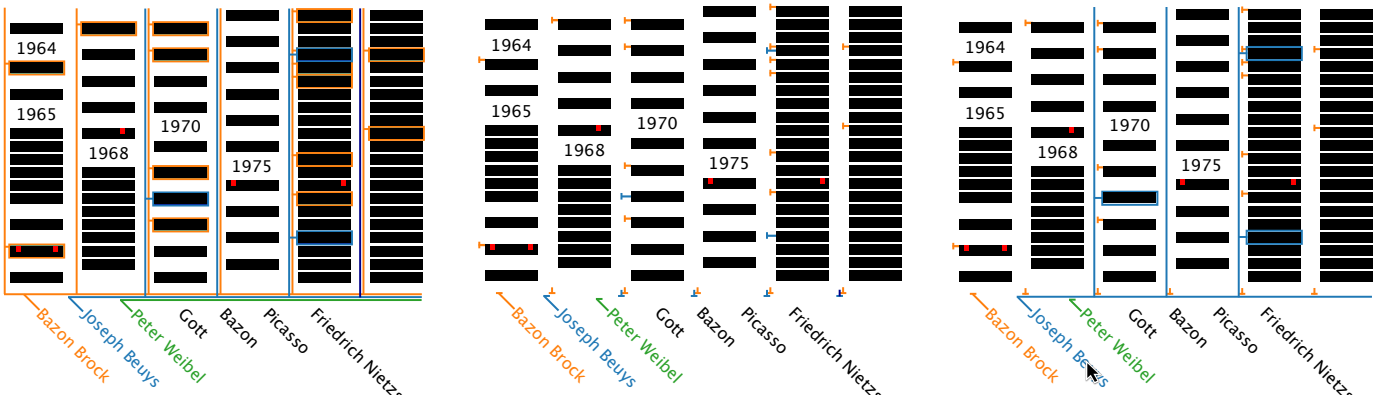


Fig. 5. The full wire (left), the reduced (center) and the interactive (right) version of the *ConceptCircuit* with the entities "Bazon Brock", "Joseph Beuys" and "Peter Weibel" selected. In the interactive version hovering above "Joseph Beuys" temporarily reveals the full circuit lanes for this particular entity.

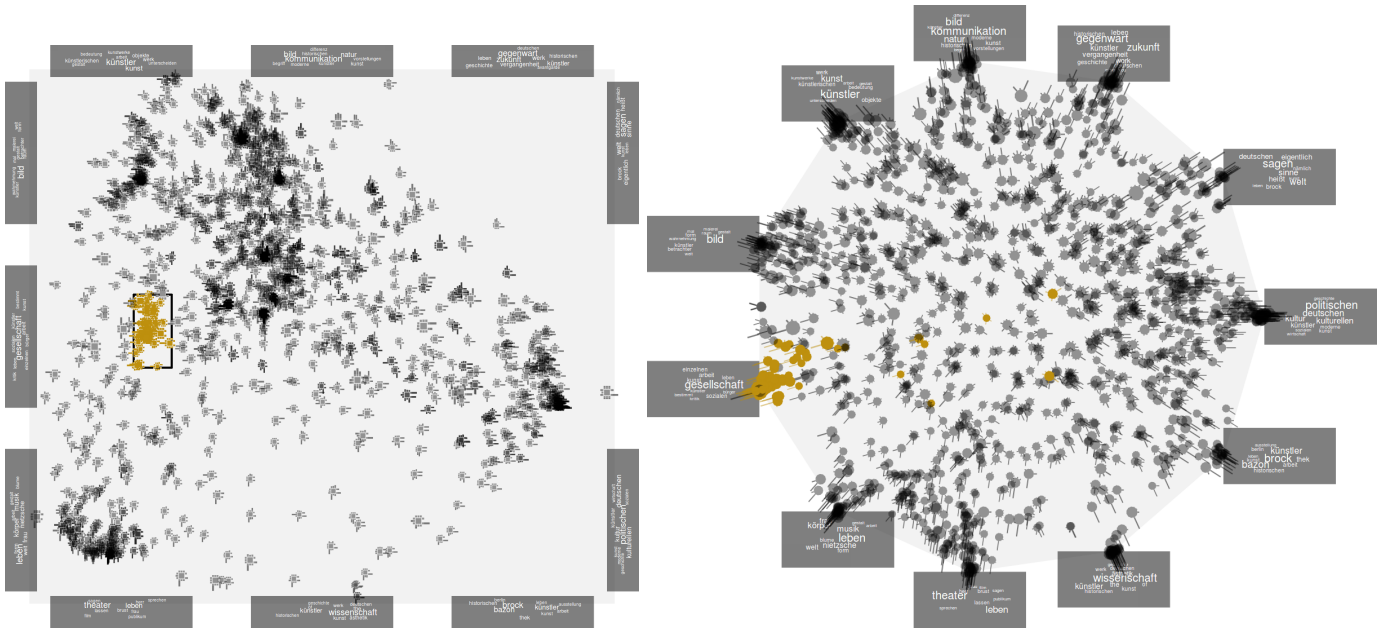


Fig. 6. The two versions of the *TopicAssembly* (using MDS on the left, using a barycentric coordinate system on the right). Both are showing the same highlighted document glyphs. The glyphs were selected by a rectangular selection tool applied to the document cluster in the *TopicAssembly* on the left, which corresponds to the topic mostly concerned with "Gesellschaft". The barycentric layout shows some outliers, which reveal that the MDS lumped together documents with rather different topic vectors.

that topic on the document by its length (as for 2 --- 3 4 — 5 and for 7 topics for instance). The size of a glyph center provides a cue of the length of the document.

For creating the layout of the *TopicStars* in the *TopicAssembly* we used two different methods. Our first implementation uses multi-dimensional scaling (MDS) [29] to compute a 2D-embedding of the *TopicStars* such that the distance between each pair of 2D-coordinates tries to match the dissimilarity between both documents regarding their topic distribution (Figure 6, left). The *TopicClouds* are distributed as evenly as possible around the inner canvas. We use a simple distance-based heuristic to try placing the individual *TopicClouds* close to a region of the canvas where corresponding *TopicStars* are placed by the MDS. The MDS concept works quite well in less crowded regions of the canvas, however, prominent clusters of similar document glyphs introduce overplotting, which makes the explicit topical information via the glyphs less recognizable.

As an alternative to the MDS-based layout we used a barycentric coordinate system since the topic vectors produced by the LDA contain the weights of each latent topic (Figure 6, right). Thus, the topic vectors already have the properties of barycentric coordinates and it seemed natural to visualize the documents in a barycentric coordinate system. Here, the topics are placed around a regular n-sided polygon.

Clusters of documents related to mainly two topics appear as a lines or edges of *TopicStars* connecting the two topics in the barycentric coordinate system, while glyphs representing a more even distribution among the topics tend to be positioned towards the center. As the visibility of correlations between topics is strongly dependent on the positions of the *TopicStars*, which are defined by the placement and the order of the topics around the polygon,

interactive reordering of topics is supported via Drag and Drop.

As both the *TopicStars'* positions within the *TopicAssembly* and their glyph shapes are derived based on the same information, *TopicStars* that have similar shapes tend to be in the proximity of each other. This results in areas of similar structure (Gestalt Law of Similarity) that help users to identify documents with similar content. Additionally, these areas automatically draw the user's attention towards the most relevant topic clouds, since the most prominent spikes of the *TopicStars* point toward their corresponding *TopicClouds*.

Initially, the *TopicAssembly* shows all *TopicStars*. Entering a search query removes those that do not contain any of the answering phrases. The remaining ones can be either unaltered, meaning that they belong to the currently selected subset of the resulting phrases from a wildcard query. *TopicStars* outside this selection (yet containing answering phrases) are grayed out. A *TopicStar* that is directly selected changes to yellow.

4 Search Backend

We refined the *Netspeak* [49] search technology to enable the user expressing advanced search queries e.g. for finding different variations of phrases Brock wrote about in several documents. Those uncertainties or ambiguities can be expressed by wildcards such as a question mark "?" (placeholder for exactly one word), an asterisk "*" (arbitrary number of words), and brackets "[]" (compare alternatives of the words within). An entirely new n-gram corpus had to be constructed based on the Bazon Brock writings containing all n-gram from 1 to 10 words that can be found in the texts along with their number of occurrences and, for our backtracking mechanism, in which documents they occurred. We extracted all n-grams from every sentence within the text collection by splitting it into all possible phrases of n consecutive words. We chose to perform the method on single sentences in order to prevent n-grams from extending across sentences and mixing subsequent paragraphs.

The resulting corpus consists of two separate parts. Each part contains one file for each type of n-gram – 1-grams, 2-grams, 3-grams and so on – listing all n-grams of that type and one additional piece of information. One of the parts additionally contains the number of occurrences for each n-gram and is used by the *Netspeak* technology to build the index for the search mechanism. The second part links each n-gram to its character positions within the texts it occurs in. This information is necessary in order to find the documents related to the answering phrases from a user query. In total, all data takes about 1.2 GB of disk space. Almost 500 MB of those are used to create the inverted index of Netspeak and are, thus, attended by the Netspeak technology; the other 700 MB are used by LevelDB to re-link phrases to positions in text.

5 First Expert Review Cycle

We performed reviews with three external Brock experts thoroughly familiar with his topics, his work and his life. One expert published a documentary and other material about Brock. The other experts are currently writing their PhD theses about certain aspects of Brock's work. They were each introduced to our system, the main visualizations *TextGenetics*, *ConceptCircuit* and MDS-based *TopicAssembly* (the barycentric version was developed later on) were explained along with test queries that showed the capabilities of the search interface.

The experts found the overall design and layout of the framework to be pleasing to work with and clearly structured. All our experts were able to identify his main works in the *TextGenetics* at a glance. Two experts immediately started a discussion about parts of these central publications. They showed us certain samples by clicking on the documents in the *TextGenetics*, they considered being crucial for this particular publication. This led to discussions about his writing style. Using the *TextGenetics* they showed us samples how his style changed from the early nineteen-sixties as compared to the eighties. The experts used the search to pose some queries they were interested in. The interactive visualization of search results was instantly understood and appreciated by all participants. The second fade level showing documents that contain answering phrases outside the user's selection and the green markings showing the position of answering phrases, however, needed a second explanation, since there was some confusion about differences of the red and green markings. The *ConceptCircuit* was also understood at once. One of the experts stated that she liked the visual impression of the frequency of the connected name and of its temporal placement provided by the visualization. Additionally, the used color scheme was mentioned to be aesthetically pleasing but not able to separate the symbol sets of different selected names.

The MDS-based*TopicAssembly* – though praised for the aesthetics of the clusters formed by the *Topic Stars* – was perceived to be the most complex visualization and, therefore, the most difficult to understand. The automatically extracted topics shown in the *Topic Clouds* were not convincing. Furthermore, one expert suggested offering the original printed view since the typography is often an integral part of Brock's text itself. The two of others agreed that visualizing page breaks would be helpful for citing from these texts since the page number is usually needed to complete citations, however automatically generated citations of any selected text would be even better. Incorporating a tutorial for each visualization and the search function was proposed in order to facilitate the first steps.

6 Topics Matter

We presented the same version of the software to Bazon Brock himself as we did to our other experts. He was very impressed by the advances of interaction and search capabilities incorporated into the *TextGenetics* and also appreciated the visual impression of glyph stars. However, when it came to the algorithmically extracted topic suggestions by the LDA, Bazon Brock was somewhat disappointed and even annoyed at the vagueness of the results, stating that his core topics are very clear (to him and his devoted readers) and that the extracted ones are not connected in the least to his real topics.

TABLE 1
Purity and soft purity of HDP and LDA results against our ground truth of expert-curated topics and assigned documents. HDA resulted in 35 clusters. LDA was computed several times for up to 40 clusters (cluster 35-40 are shown in the table). Purity and soft purity of HDP seem to outperform all LDA clusterings, however are false positive (Figure 7).

Method	LDA-35	LDA-36	LDA-37	LDA-38	LDA-39	LDA-40	HDP(35)
Purity	0.1189	0.1231	0.1251	0.1210	0.1169	0.1189	0.8800
S. Purity	0.0470	0.0506	0.0504	0.0495	0.0484	0.0483	0.7204

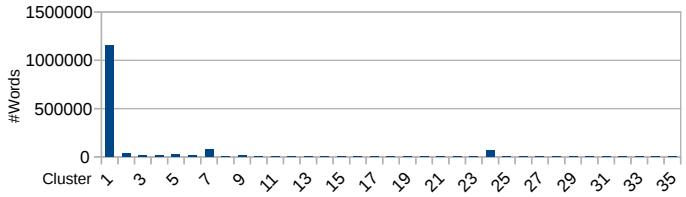


Fig. 7. The HDP lumps almost all the documents into a single cluster.

Given the obviously incompatible results of LDA topic extraction vs. Bazon Brock's and the experts' view on the core topics of his work, we decided to experiment with other probabilistic topic models and also to invest in manual topic extraction and assignment. We approached two of our experts to aid us by providing a catalog of topics (resulting in 38 topics, which both agree on) and by assigning each individual document to the suggested topics with a particular weight. Additionally, a Hierarchical Dirichlet Process (HDP) was performed. Other alternatives such as pLSI (Probabilistic Latent Semantic Indexing) and CTM (correlated topic model) were not considered further, since they were evaluated for instance by Chang [25] with the result that they generate even less comprehensible topic sets than those resulting from LDA. For obtaining a quality measure for the used probabilistic topic modeling algorithms, we analyzed the similarity between the topics the two experts annotated within Bazon Brock's work and the clusterings produced by the LDA and HDP algorithms using the Purity measure.

For creating an appropriate variety of LDA-clusterings we opted for generating and comparing with LDA-clusterings resulting in 5 to 100 clusters. We hoped for at least one configuration of clusterings that somehow shows a slight resemblance to the experts' assessment. However, both the purity and soft purity values remained low for all parameters (Table 1), which indicates that the modeled LDA results do not cover the expert-curated topics at all. HDP identified 35 topics, which is close to our 38 expert topics. Furthermore both values of HDP against our expert-curated and assigned topics showed encouraging results (see Table 1) outperforming the LDA results by far. A closer look at the coherence matrix, however, revealed that they were false positives since the HDP outcome consists mainly of a single cluster containing almost all documents (Figure7). Bazon Brock is a writer who barely repeats himself and is eloquent to such a high degree that approaches mainly based on the analysis of word frequencies are seemingly not capable of finding salient patterns of similar vocabulary.

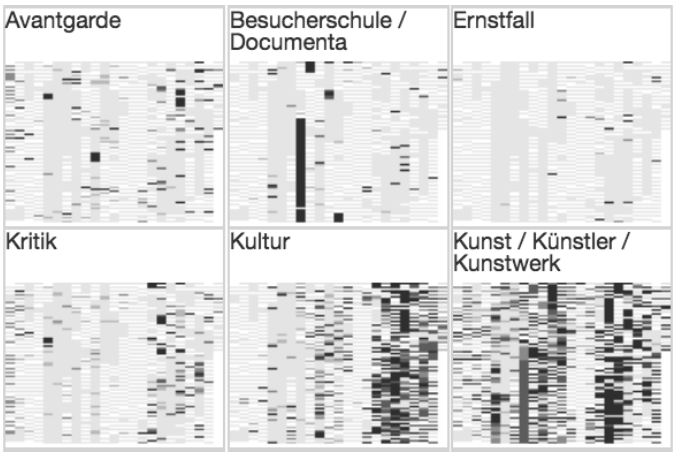


Fig. 8. Six topics expressed as *TopicGenetics* provide an idea about their importance in Brock's work.

7 Visualizing Expert-Curated Topics

Considering the unsatisfactory results of the algorithmic approaches, we eventually turned the human expertise into visualizations which focus on different aspects such as documents, topics and time. Remarkably, yet somewhat expected and hoped for and in contrast to the automatically extracted ones, the topic assignments of our experts showed clear visual patterns regarding the works over time, as one can see in the so-called *TopicGenetics* (Figure 8, one visualization of three, particularly designed for gaining insight into the expert-curated topics for answering questions such as: What was the hot phase of a topic? Does it occur in individual documents before or after? Does the context of a topic change over time? Which topics co-occur and in which periods? The *TopicGenetics* consists of one miniature of the *TextGenetics* for each expert-curated topic. Although only relying on a small number of pixels, every miniature provides a clear impression of the importance of a particular topic during certain work periods in Brock's opus and is able to filter the main *TextGenetics* on selection.

While the *TopicGenetics* and the *TextGenetics* use a 2D layout for the temporal ordering, the *TopicTrends* (Figure 9) use a 1D layout of the time axis and facilitate the comparison of evolutions of different topics over many decades. Each tick marks a document published at a given time that belongs to the respective topic. A reader is therefore able to tell, at a glance, which topics show a similar trend or progression whereas others are more anti-cyclic to one another.

In order to reduce the number of visual items to consider at first and, even more crucial, for providing an overview of the topics important in different decades, we developed the hierarchical *TopicCalendar* (Figure 10). The visualization goes beyond the topic×word (or document in our case) matrix of Termite [55] by introducing hierarchical exploration of the temporal aggregation. Initially, it shows the aggregated topic assignments of all documents grouped by their respective publication decades. Upon expansion, however, the aggregated topic assignments of publication years, publications and the individual documents become visible. On the document level the glyphs express the magnitude of a document's assignment towards a topic (.i.e. how influential

Fig. 9. The *TopicTrends* provide an overview of the temporal progression of each topic. The selected topics show their mutual documents in two levels. If both topics are assigned to a certain document with high ratings the document appears in red and with lower ratings in yellow. Documents belonging only to one topic remain in their gray value.

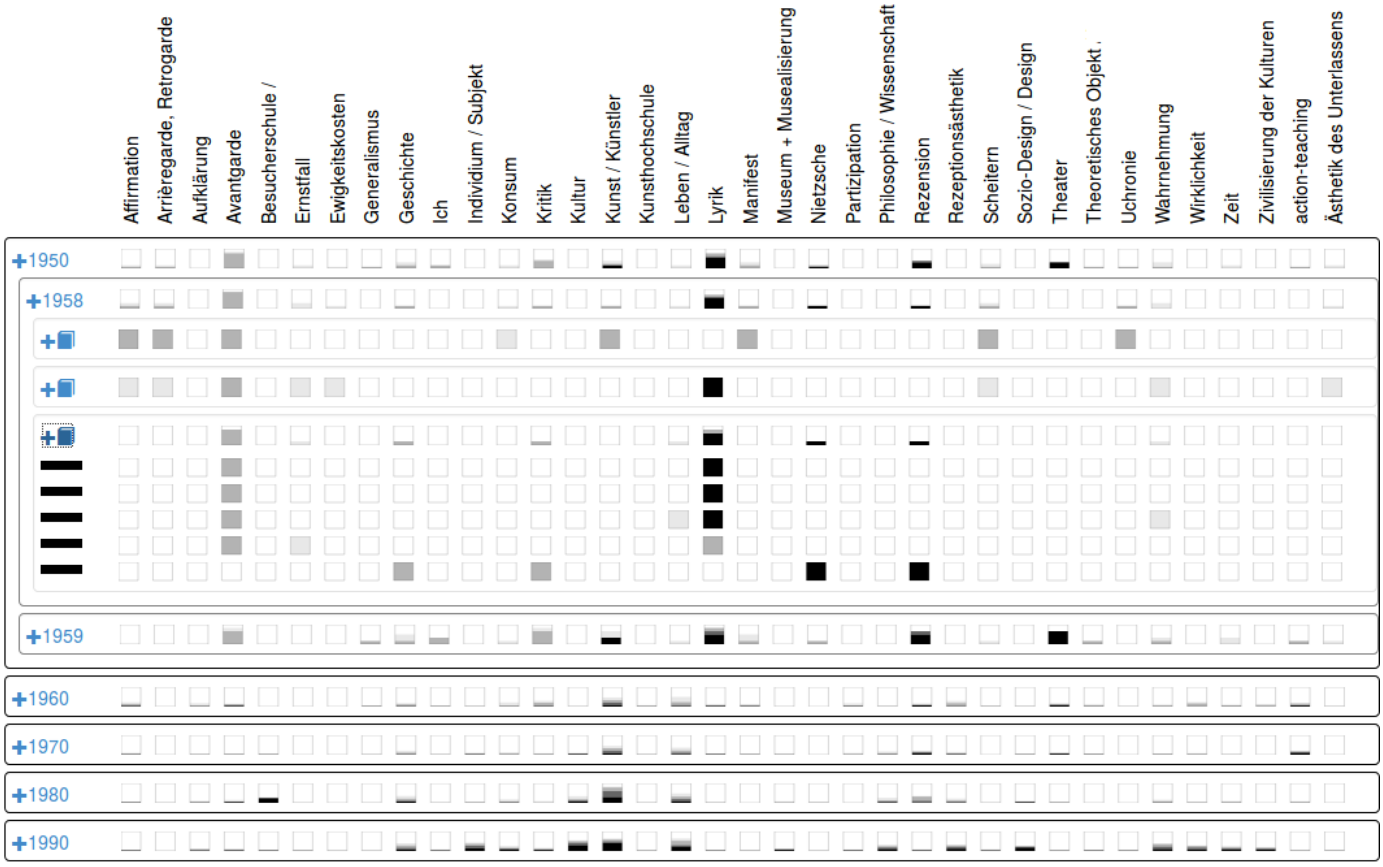


Fig. 10. The opened *TopicCalendar* shows the strength of each topic for different (hierarchically explorable) temporal levels. At first, decades are shown. They can be drilled down to year, publication, and document level. Each glyph visually aggregates the distribution of the topic assignments in the temporal levels below.

a topic for the given document was) in five discrete levels of gray ranging from white depicting that it does not belong to this topic at all to black meaning this document is crucial for this topic (see Figure 10).

For the higher levels the glyphs are supposed to visually aggregate the distribution of topic assignments of the elements from the levels below. This concerns the interesting issue of how a distribution of assignments on the leaf level of a hierarchy can be visually propagated to upper levels up to the root. In our case, common aggregation functions like mean, median or mode appeared to be unsatisfying, since they mostly resulted in some form of medium gray which is not helpful in judging if it is worthwhile to investigate further. We opted for encoding the distribution of topic assignments as a glyph, since it reveals if there is a fraction of documents with high ratings in lower levels.

We tested several glyph designs including regular pie charts ( ) and pie charts that center and split the gray shades symmetrically ( ). Eventually, we opted for a stacked bar-chart approach that matched our overall design ( ) by stacking the gray values bottom-up. Besides the better recognition of length differences over angles numerous pie charts at once appeared irritating. It resulted in the impression that the angles would encode a second kind of information that we (as well as our second expert) could not make sense of. Contrary to Streit et al. [56] we argue that our glyph outlines (especially for empty glyphs) guide the user within the topic columns and the respective temporal level (or documents at the lowest level) and enable the users to judge the ratio of important to unimportant documents.

8 Second Expert Review

In order to assess the effectiveness of our tool after the changes we made in response to the first expert review, we conducted a second expert review with one of our experts. We regard our tool as effective for the intended purpose

if it is able to fulfill the tasks described in Section 3, in a convenient manner and to the satisfaction of our expert. The revision of the software shown to the expert contained all features tested in the first review round, complemented with all visualizations presented in Section 7: the *TopicTrends*, the *TopicCalendar*, the miniature *TopicGenetics* and, additionally, the barycentric revision of the *TopicAssembly*. The only noticeable difference was the glyph design used in the *TopicCalendar*. Our expert saw the symmetrical pie charts () instead of the stacked bar-chart approach () we opted for later on due to a higher resemblance to our overall design.

Getting familiar with Brock's work was already appreciated during the first review, but even more so during the second. Already in the first round, all experts were able to recognize Brock's main works at a glance (see section 5). Our expert was excited about how the new *TopicGenetics* miniatures improved the recognition of which topic has been important to Brock during a certain period as well as providing a kind of topic overview to begin with.

Regarding the *TopicAssembly* she was impressed how the expert curated topics improved the visualization. The barycentric revision appeared much clearer to her in assigning documents towards potential topics than the older MDS-based one, especially, while looking at both versions showing the same older LDA-extracted topics (as in Figure 6). Furthermore the orientation of the glyphs' spikes was easier to recognize for her in the new revision. However, more important to her was being able to select certain topics (in the miniature *TopicGenetics*) she want to investigate in the *TopicAssembly* together, which improves the overall lucidity in her eyes.

Our expert was also very fond of the *TopicCalendar* as visual introduction to his topics. She stated that, the *TopicCalendar* condenses the topical information even more than the *TopicGenetics* when it is initially showing only the aggregated topic distributions of the decades. She assessed it as an easy entry point for less experienced readers that can be hierarchically explored later on. Although the version of *TopicCalendar* we presented to her still used pie glyphs at that time and she was not sure about the meanings of the angles of the pie, she stated that, she was a priori confident about the meaning of the black and gray colors used in the pies, which gave her strong advice what to explor further regarding particular topics. As mentioned above, we shared this concern about the angles, which is why we later on changed the glyphs to a stacked appearance.

The *TopicGenetics* also facilitates in examining his work life and different work periods. A quick look at the *TopicGenetics*—without reading the topic labels—sufficed for our expert to recognize certain work periods, such as the "Documenta" period in the 80s. She told us about how Brock conducted the "Besucherschulen" at the "Documenta" exhibition, which is highly reflected in his writings at the time and therefore visually prominent in the *TopicGenetics*. Later on the matter pops up now and then in certain texts, however, this period was over and is somehow distinct from his previous and later work life.

For further surveying and comparing Brock's topics we first looked at the *TopicTrends* with our expert. Although she was involved in defining and extracting Brock's topics, she was surprised by some of the temporal patterns, whereas others were quite familiar to her (see "Documenta" above). For instance, the topics "Individium/Subjekt" and "Philosophie/Wissenschaft" had the most similar temporal appearance. However, comparing both directly using the overlay functionality of the *TopicTrends* reveals that this does not necessarily mean that they often occur in the same documents. According to the expert, comparing multiple topics like that is best supported by the *TopicTrends*, while surveying a single topic is more effective using the *TopicGenetics*. The *TopicTrends* is also apt for constructing and verifying hypotheses. The surprising patterns revealed by the *TopicTrends* suggested the hypothesis that topics can be divided in early topics and late topics which interact differently with each other. While early topics tend to be self-contained and showing only little interrelation among them, late topics tend to interact more strongly with each other co-occurring more often in the same documents.

Although the *ConceptCircuit* features' – expressing persons and locations – was mostly intended for readers less familiar with his work, it revealed some interesting findings which were even new to our expert. For instance, she was very aware that Brock wrote about God in multiple texts, yet she was surprised how often the entity really occurred in his work. One of the most exciting features to her was the instant filtering of documents in all visualizations by the different search methods. She stated that being provided with snippets containing the search phrases would have saved her a lot of time when working on her thesis. Moreover, her effort in searching, reading and collecting would have been further reduced by using the search in combination with the *TextGenetics* and the *TopicGenetics*, as they enable her to identify related documents along with their assigned topics and their impact in Brock's working periods at a single glance. Also, the importance of the last point raised during our initial interviews with the experts about having immediate and fluent access to the text by simply clicking on the glyphs became obvious when our expert talked about particular documents and wanted to show us an interesting sections. By incident, we even found an incorrect paragraph ordering in the digital version of one of the documents.

9 Conclusions and Future Work

In this paper we presented a visualization framework that is suited to illustrate and analyze the contents of a medium-sized text collection. Several visualizations were developed; each of them reflects another aspect of the text collection: the *TextGenetics* in combination with wildcard-based textual search focuses on the temporal structure while the *ConceptCircuit* relates the facets "persons" and "locations" to the documents shown in the *TextGenetics*. The *TopicGenetics*, *TopicCalendar* and *TopicTrends* provide insight into expert-curated topic occurrences during different decades. The *TopicAssembly* provides an overview of automatically extracted or manually curated topics and their related documents.

The framework received positive feedback from our experts and Bazon Brock himself, especially the *TextGenetics* including the advanced search capabilities and the combination with the *ConceptCircuit* were much appreciated. Also,

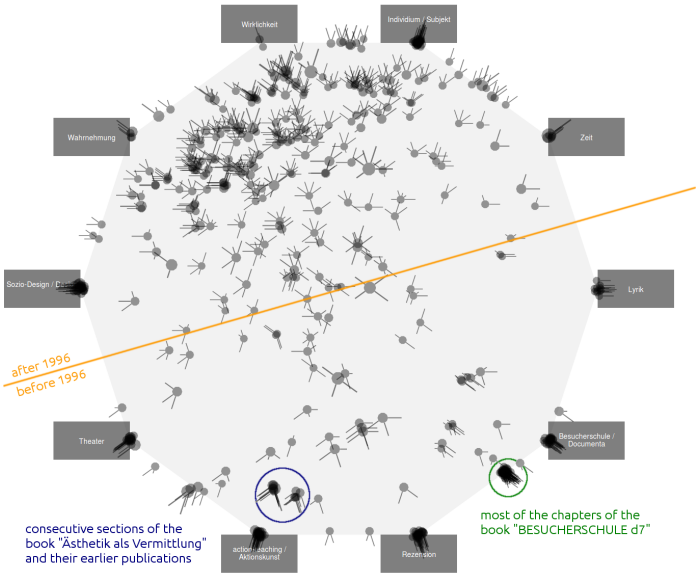


Fig. 11. The *TopicAssembly* based on generalized barycentric coordinates provides interesting insights into particular expert-curated topics that have been previously chosen with the help of the *TopicTrends*. Most of the topics before 1996 (the lower half of the visualization) tend to be more self-contained; showing only little interrelation among them. The tightly packed clusters (marked in blue and green) form consecutive chapters in "Ästhetik als Vermittlung (1977)" and most of the chapters of "BESUCHERSCHULE d 7 (1982)". Topics after 1996, however, tend to interact more strongly with each other, forming several overlapping clusters. It seems that Brock shifted from focusing on one topic at a time to addressing the greater picture, thus relating several topics to each other; possibly a matter of sagacity and seniority. In order to reduce overplotting and getting a better impression of the number of documents, jittering was applied for documents sharing the same position. This was particularly helpful for the manually curated topics since they had discrete ratings of only 5 levels.

the idea of the *TopicAssembly* was considered as being potentially useful; however, the results of automatic topic extraction via LDA were not satisfactory and were controversially discussed, ultimately resulting in the involvement of domain experts for topic compilation and topic assignment to all documents. The performed comparisons of the modeled topics against the expert-curated ones indicate that none of over 90 tested LDA configurations results in topics that lead to similar clusters as the expert curated topics. The sheer number of 38 expert-defined topics necessitated designing new topic-related visualizations such as the already mentioned: *TopicGenetics*, *TopicCalendar* and *TopicTrends*. In one of our next steps we will ask our experts to hierarchically structure the topics with the goal to reduce the overviews and to interactively explore the detailed topics.

Figure 11 illustrates how well the *TopicAssembly* also works for expert-curated topics. It shows one of the most interesting findings about Brock's work which was revealed upon closer inspection of the *TopicAssembly* in a generalized barycentric layout. Moreover, the fundamental concepts of the TopicAssembly such as topic-oriented placement of documents and the topic-directing glyphs, can be very well used with other multi-dimensional projection techniques such as least squares projection. In order to reduce overplotting we plan to explore grid-based layouts in combination with summarizing glyphs to depict the set of topic distributions in each grid cell.

Of course, the search backend itself can be extended in multiple ways. Besides the already implemented phrase search, a wildcard search for syllables or inflections was a feature that our experts were looking forward to. At the moment, only the results of a single search query can be visualized. In some cases, however, the user could be interested in narrowing the results further by submitting a refinement query. Therefore the system must be capable of depicting the results of multiple queries at once. Ideally, the user should be able to distinguish between the results of these queries and quickly determine overlaps between them. Furthermore, filters for different aspects of the text collection such as time, books and text lengths were also of interest to our experts.

Based on our experiences we want to provide some take-away messages regarding highly specialized corpora with a very elaborate vocabulary.

- It's all about domain expertise. Approaching experts at the earliest stage possible is key for exploring, structuring and understanding the nature of the corpus and deriving the different facets that are crucial to visualize. However, those discussions can also be difficult at the beginning if there is no functionality available.
- Expert reviews were immensely helpful (even with the small group of 3 experts and the author himself) and showed the shortcomings of the reviewed version. Additional visualizations and interactions were designed to overcome these issues later on.
- Probabilistic topic models are worth being tried, albeit we believe that the more elaborate language a corpus uses, the poorer the results will be. Analyzing the coherence of the automatic results and having them assessed by domain experts might help to find sensible parameters of the automatic approaches but it might also lead to a rejection of the automatically generated results.
- Using a consistent design vocabulary aids in ending up with a visually pleasing visualization. While not being always necessary for gaining insight, it might be quite important for acceptance by the general public, students, scholars as well as experts or the authors such as Brock, who was very fond of the visual ideas.

The reviews showed that the system provides useful means to fulfill the intended tasks described in Section 3 and hence is able to aid with most of the questions the experts posed. Of the many questions towards Brock's opus stated in Section 3 only the question about changes in his writing style remained somehow unresolved. While our visualizations allow the exploration of changes in topics over the course of Brock's creative period, the relationship to changes in writing style remains future work.

We plan to integrate the most promising components into the current website of Bazon Brock in order to provide his readers with a better survey of his work and the visual tools to explore it in detail. Albeit, the proposed visual concept is very focused on the work and person of Bazon Brock, every visualization technique can be easily adapted

to corpora of other authors e.g. the publicly available works of Nietzsche [57] or Kant [58]. Their huge impact in philosophy and the correspondingly large number of experts could lead to a dedicated and customizable visualization system for the work of individual authors, a small group of authors or a thematically or otherwise constrained set of documents. Our work contributes some initial components for such a visualization system.

ACKNOWLEDGMENTS

Foremost, the authors wish to thank Bazon Brock for consenting to our proposition to analyze and visualize his entire opus as well as providing feedback to our results. We also thank Bianca Girbinger, Andrea Seyfarth and Stefan Wilke for participating in the expert reviews. We are particular grateful to both, Bianca Girbinger and Andrea Seyfarth, for curating the topics and document assignments.
This work was supported in part by the German Federal Ministry of Education and Research (BMBF) under grant 03IP704 (project Intelligentes Lernen) and grant 03IPT704X (project Big Data Analytics).

REFERENCES

[1] C. Collins, F. B. Viégas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST '09, 2009, pp. 91–98.
[2] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "Tiara: A visual exploratory text analytic system," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10, 2010, pp. 153–162.
[3] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization." in *Proceedings of the 2010 IEEE Pacific Visualization Symposium*, ser. PACIFICVIS '10, Apr. 2010, pp. 121–128.
[4] B. Brock, "Bazon Brock Homepage," http://bazonbrock.de/, Accessed: 06/2017.
[5] C. Lodder, *Russian Constructivism*. Yale University Press, 1983.
[6] S. Baier and B. Dümpelmann, *Kazimir Malevich: The World as Objectlessness*. Hatje Cantz, 2014.
[7] L. Boersma, B. Rutten, and A. Shatskikh, *Kazimir Malevich and the Russian Avant-Garde: Featuring Selections from the Khardziev and Costakis Collections*. Walther König, Köln/Stedelijk Museum Amsterdam, 2014.
[8] G. Lupi and M. Buttignol, "Kandinsky," https://i1.wp.com/www.brainpickings.org/wp-content/uploads/2013/05/accurat_kandinsky.jpg, A project by Accurat (www.accurat.it). Accessed: 06/2017.
[9] V. Pellegrini, "The atlas of kant's legacy," http://www.densitydesign.org/wp-content/uploads/2013/08/100-parole-definitivo-b-thumbnail-visualizing.jpg, Accessed: 06/2017.
[10] P. Ciuccarelli, "Minerva - data visualization to support the interpretation of kant's work," http://www.densitydesign.org/2013/08/minerva-data-visualization-to-support-the-interpretation-of-kants-work/, Accessed: 06/2017.
[11] F. B. Viégas, S. Golder, and J. Donath, "Visualizing email content: Portraying relationships from conversational histories," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ser. CHI '06, 2006, pp. 979–988.
[12] E. R. Gansner, Y. Hu, and S. North, "Visualizing streaming text data with dynamic graphs and maps," in *Proceedings of the 20th International Conference on Graph Drawing*, ser. GD'12, 2013, pp. 439–450.
[13] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann, "The infosky visual explorer: Exploiting hierarchical structure and document similarities," *Information Visualization*, vol. 1, no. 3/4, pp. 166–181, Dec. 2002.
[14] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu, "Facetatlas: Multifaceted visualization for rich text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1172–1181, Nov 2010.
[15] J. Singh, S. Zerr, and S. Siersdorfer, "Structure-aware visualization of text corpora," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, ser. CHIIR '17, 2017, pp. 107–116.
[16] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09, 2009, pp. 497–506.
[17] 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," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, Dec. 2011.
[18] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2012–2021, Dec 2013.
[19] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, Jan. 2002.
[20] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher, "Serendip: Topic model-driven visual exploration of text corpora," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2014, pp. 173–182.
[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research.*, vol. 3, pp. 993–1022, Mar. 2003.
[22] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.
[23] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, 2006, pp. 113–120.
[24] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, pp. 7:1–7:30, Feb. 2010.
[25] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Neural Information Processing Systems (NIPS)*, 2009.
[26] D. Blei and J. Lafferty, "A correlated topic model of science," *Annals of Applied Statistics*, vol. 1, pp. 17–35., 2007.
[27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, 2004.
[28] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013.
[29] C. Pich, "Applications of multidimensional scaling to graph drawing," Ph.D. dissertation, University of Konstanz, 2009. [Online]. Available: http://kops.ub.uni-konstanz.de/volltexte/2009/8399/
[30] S. Cheng and K. Mueller, "Improving the fidelity of contextual data layouts using a generalized barycentric coordinates framework," in *Proceedings of the 2015 IEEE Pacific Visualization Symposium*, ser. PACIFICVIS '15. IEEE, 2015, pp. 295–302.
[31] K. Strimmer and A. Von Haeseler, "Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment," *Proceedings of the National Academy of Sciences*, vol. 94, no. 13, pp. 6815–6819, 1997.
[32] L. Hamel, O. Zhaxybayeva, and J. P. Gogarten, "Pentaplot: A software tool for the illustration of genome mosaicism," *BMC bioinformatics*, vol. 6, no. 1, p. 139, 2005.
[33] R. Chasin, D. Woodward, J. Witmer, and J. Kalita, "Extracting and displaying temporal and geospatial entities from articles on historical events," *The Computer Journal*, vol. 57, no. 3, pp. 403–426, 2013.
[34] Y.-F. R. Chen, G. Di Fabbrizio, D. Gibbon, S. Jora, B. Renger, and B. Wei, "Geotracker: Geospatial and temporal rss navigation," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07, 2007, pp. 41–50.
[35] C.-C. Pan and P. Mitra, "Femarepviz: Automatic extraction and geo-temporal visualization of fema national situation updates," in

Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology, ser. VAST '07, Oct 2007, pp. 11–18.

[36] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, "Leadline: Interactive visual analysis of text data through event identification and exploration," in *2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14-19, 2012*, 2012, pp. 93–102.

[37] C. Harrison, "Biblical social network (people and places)," http://www.chrisharrison.net/index.php/Visualizations/BibleViz, Accessed: 06/2017.

[38] P. Venetis, G. Koutrika, and H. Garcia-Molina, "On the selection of tags for tag clouds," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11, 2011, pp. 835–844.

[39] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale, "Sparkclouds: Visualizing trends in tag clouds," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1182–1189, Nov. 2010.

[40] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *2014 47th Hawaii International Conference on System Sciences*, Jan 2014, pp. 1833–1842.

[41] A. Endert, R. Burtner, N. Cramer, R. Perko, S. Hampton, and K. Cook, "Typograph: Multiscale spatial exploration of text documents." in *Proceedings of the 2013 IEEE International Conference on Big Data*, ser. BigData '13, 2013, pp. 17–24.

[42] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, and K. Hofl, "Compare clouds: Visualizing text corpora to compare media frames," in *In Proceedings of the IUI Workshop on Visual Text Analytics*, 2015.

[43] P. DeCamp, A. Frid-Jimenez, J. Guiness, and D. Roy, "Gist icons: Seeing meaning in large bodies of literature." in *Proceedings of the 2005 IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST '05, 2005.

[44] D. A. Keim, "Designing pixel-oriented visualization techniques: Theory and applications," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 59–78, Jan. 2000.

[45] D. A. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, ser. VAST '07, 2007, pp. 115–122.

[46] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1221–1228, Nov. 2008.

[47] C. Harrison, "Web trigrams: Visualizing google's tri-gram data," http://www.chrisharrison.net/index.php/Visualizations/WebTrigrams, Accessed: 06/2017.

[48] S. Luz and S. Sheehan, "A graph based abstraction of textual concordances and two renderings for their interactive visualisation," in *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, ser. AVI '14, 2014, pp. 293–296.

[49] P. Riehmann, M. Potthast, H. Gruendl, , M. Trenkmann, B. Stein, and B. Froehlich, "Wordgraph: Keyword-in-context visualization for netspeak's wildcard search," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1411–1423, 2012.

[50] D. W. Galenson, "Two paths to abstract art: Kandinsky and malevich," *Russian History*, vol. 35, no. 1-2, pp. 236–250, 2008.

[51] K. Malevich, *Suprematist Composition (Oil on canvas)*. Wilhelm Hacke Museum, Ludwigshafen, 1915/16.

[52] M. A. Hearst, "Tilebars: Visualization of term distribution information in full text information access," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '95, 1995, pp. 59–66.

[53] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05, 2005, pp. 363–370.

[54] D. Ramage and E. Rosen, "Stanford topic modeling toolbox," http://www-nlp.stanford.edu/software/tmt/tmt-0.4/, Accessed: 06/2017.

[55] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Advanced Visual Interfaces*, 2012. [Online]. Available: http://vis.stanford.edu/papers/termite

[56] M. Streit and N. Gehlenborg, "Points of view: Bar charts and box plots," *Nat Meth*, vol. 11, no. 2, pp. 117–117, Feb 2014, this Month. [Online]. Available: http://dx.doi.org/10.1038/nmeth.2807

[57] N. Source, "Digital critical edition of nietzsche's works and letters (ekgwb)," http://www.nietzschesource.org/documentation/en/eKGWB.html, Accessed: 06/2017.

[58] Universität Duisburg Essen, "Akademieausgabe von Immanuel Kants Gesammelten Werken," http://korpora.zim.uni-duisburg-essen.de/kant/verzeichnisse-gesamt.html, Accessed: 06/2017.

Patrick Riehmann is a PostDoc with the Virtual Reality and Visualization Research Group at Bauhaus-Universität Weimar. His research interests include visualization of time-oriented data and textual data, graph drawing, and multi-touch interfaces.

Dora Kiesel is a PhD candidate with the Virtual Reality and Visualization Research Group at the Bauhaus-Universität Weimar. Her research interests include data visualization with focus on argumentation and text analytics.

Martin Kohlhaas is with Kohlhaas&Kohlhaas Design Agency focused on design services for screen and paper, web sites and web apps. He is member of the igroup.org project consortium known for the igroup presence questionnaire. He is frequently faced with Information Visualization issues in the projects of his Agency.

Bernd Froehlich is a full professor with the Computer Science Department at Bauhaus-Universität Weimar and head of the Virtual Reality and Visualization Research Group (www.uni-weimar.de/medien/vr). His research interests include real-time rendering, visualization, 3D user interfaces, 3D display technology, immersive telepresence, and support for collaboration in colocated and distributed virtual environments.