Visual Analytics for the Exploratory Analysis and Labeling of Cultural Data

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet

Informatik

Vorgelegt
von M.Sc. Christofer Meinecke

geboren am 13. November 1994 in Gifhorn

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Gerik Scheuermann, Universität Leipzig &
   Prof. Dr. Stefan Jänicke, University of Southern Denmark
2. Prof. Dr. Andreas Kerren, Linköping University & Linnaeus University

.

Die Verleihung des akademischen Grades erfolgte mit Bestehen der Verteidigung am 13.10.2023 mit dem Gesamtprädikat *magna 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

Christofer Meinecke graduated in Computer Science at Leipzig University, Germany in 2019. During his master's studies, he gained experience in working on interdisciplinary digital humanities projects in the TASTEN project at the Musical Instrument Museum Leipzig. Since his graduation, he has worked as a research assistant in the Image and Signal Processing Group at Leipzig University, where he was involved in the projects Data Mining and Value Creation (DMW) and Smart Regional Development Infrastructure (SARDINE). Furthermore, he engaged in a number of interdisciplinary research collaborations with humanities scholars and people working in the GLAM sector. His research interest includes Information Visualization, Digital Humanities, Visual Analytics, and Machine Learning for cultural data with a strong focus on human-in-the-loop processes and participatory design.

Acknowledgments

Although the process of finishing a Ph.D. thesis is done by only one person, many people contributed in different ways to fill the following pages. First and foremost, I thank my supervisor Stefan Jänicke for guidance in my academic way since my bachelor thesis in 2017, for his support in the years of preparing this dissertation, and for teaching me how to engage in meaningful interdisciplinary research.

Another big thank you goes to David Joseph Wrisley for our long cooperation going back to my master's thesis in 2019 and to Estelle Guéville. I am very grateful to both of you for your (digital) humanities perspective and the discussions on our previous and ongoing projects. I look forward to our next projects! I also want to thank Chris Hall for the discussions and his perspective on exhibition design for the virtual museum project. A big thank you to my other Co-Authors, Daniel Wiegreffe, Jeremias Schebera, Jakob Eschrich, and Ahmad Dawar Hakimi.

I thank Gerik Scheuermann for allowing me to be part of the BSV group and to finish this thesis in Leipzig. I also want to thank Daniel Wiegreffe again for co-supervising me in Leipzig and letting me participate in several research projects. Next to Daniel, I thank Christian Heine, Dirk Zeckzer, Christina Gillmann, and Andreas Niekler for their advice on research problems and the academic world.

I want to thank my (former) peers in the VIS4DH group, Richard Khulusi, Jakob Kusnick, and Martin Reckziegel, for the discussions on information visualization and digital humanities research problems and from time to time discussions on some fun web development problems. I also want to thank my peers in the VDA group, Jeremias Schebera, Yves Annanias, Michelle Philipp, and Nico Graebling, for the joint cooperation, the input on research questions, the office talks, and from time to time some welcomed distractions. I also thank all of my former and current colleagues at the BSV group, who are not mentioned by name for numerous conversations on computer science and other topics of interest, and for sharing the struggle of wrapping up a Ph.D. thesis. For their bureaucratic support, I would like to thank Karin Wenzel and Heike

Zschoche. I thank my bachelor's and master's students for allowing me to work with them on some of my research ideas.

I also want to thank all of my friends from home for their long friendship and ongoing support over the past 10 to 15 years. I really appreciate you. In particular, big thanks and love to Dennis, Leo, Lars, Björn, Malte, Alex, Jan, Michi, Kim, and Lara. I also want to thank my former study peers, in particular, Dominik, Enno, Konstantin, and Leo. Another big thank you goes to my Leipzig Family for emotional support and free time activities. In particular, big thanks to Marlo, Lisa D., Konny, Lukas, Lisa W., Fenja, Lisa M., Lennart, JJ, Basti, Sophie, Otto, Miriam, Thea, Marie, Lena, Paul, Alina, Mathias, and Manuel.

The last thank you goes to my family. Especially to my brother Daniel and my parents, Karin and Gerhard. Daniel, thank you for always being there and taking care of my upbringing when needed. To my parents, thank you for my upbringing and for allowing me to leave home at the age of 16 in order to move to the city and grow up. I appreciate your support over the past years and without your help, it would not have been possible to study and write this thesis.

Contents

If we are to preserve culture, we must continue to create it.

Johan Huizinga

1 Motivation

Cultural data can come in various forms and modalities, such as text traditions, artworks, music, crafted objects, or even as intangible heritage such as biographies of people, performing arts, cultural customs and rites. The assignment of metadata to such cultural heritage objects is an important task that people working in galleries, libraries, archives, and museums (GLAM) do on a daily basis. These rich metadata collections are used to categorize, structure, and study collections, but can also be used to apply computational methods. Such computational methods are in the focus of Computational and Digital Humanities projects and research. For the longest time, the digital humanities community has focused on textual corpora, including text mining, and other natural language processing techniques. Although some disciplines of the humanities, such as art history [WD16] and archaeology [Opg21] have a long history of using visualizations. In recent years, the digital humanities community has started to shift the focus to include other modalities, such as audio-visual data [AT19]. In turn, methods in machine learning and computer vision have been proposed for the specificities of such corpora [vN22].

Over the last decade, the visualization community has engaged in several collaborations with the digital humanities, often with a focus on exploratory or comparative analysis of the data at hand [WFS+18, JFCS16]. This includes both methods and systems that support classical Close Reading of the material and Distant Reading [Mor13]

methods that give an overview of larger collections, as well as methods in between, such as Meso Reading [JW17a]. Furthermore, a wider application of machine learning methods can be observed on cultural heritage collections [FKP+20]. But they are rarely applied together with visualizations to allow for further perspectives on the collections in a visual analytics or human-in-the-loop setting. Visual analytics can help in the decision-making process by guiding domain experts through the collection of interest. However, state-of-the-art supervised machine learning methods are often not applicable to the collection of interest due to missing ground truth [MZ18]. One form of ground truth are class labels, e.g., of entities depicted in an image collection, assigned to the individual images. Labeling all objects in a collection is an arduous task when performed manually, because cultural heritage collections contain a wide variety of different objects with plenty of details. A problem that arises with these collections curated in different institutions is that not always a specific standard is followed, so the vocabulary used can drift apart from another, making it difficult to combine the data from these institutions for large-scale analysis [RMD+22].

This thesis presents a series of projects that combine machine learning methods with interactive visualizations for the exploratory analysis and labeling of cultural data. First, we define cultural data with regard to heritage and contemporary data, then we look at the state-of-the-art of existing visualization, computer vision, and visual analytics methods and projects focusing on cultural data collections. After this, we present the problems addressed in this thesis and their solutions, starting with a series of visualizations to explore different facets of rap lyrics and rap artists with a focus on text reuse. Next, we engage in a more complex case of text reuse, the collation of medieval vernacular text editions. For this, a human-in-the-loop process is presented that applies word embeddings and interactive visualizations to perform textual alignments on under-resourced languages supported by labeling of the relations between lines and the relations between words. We then switch the focus from textual data to another modality of cultural data by presenting a Virtual Museum that combines interactive visualizations and computer vision in order to explore a collection of artworks. With the lessons learned from the previous projects, we engage in the labeling and analysis of medieval illuminated manuscripts and so combine some of the machine learning methods and visualizations that were used for textual data with computer vision methods. Finally, we give reflections on the interdisciplinary projects and the lessons learned, before we discuss existing challenges when working with cultural heritage data from the computer science perspective to outline potential research directions for machine learning and visual analytics of cultural heritage data.

Overview of Publications

This dissertation is based on the following publications by the author and reuses text, figures, and other results of them:

Chapter 1 & 7:

- Labeling of Cultural Heritage Collections on the Intersection of Visual Analytics and Digital Humanities [Mei22]

Chapter 2 & 3:

- Detecting Text Reuse and Similarities between Artists in Rap Music through Visualization [MJ21]
- Explorative Visual Analysis of Rap Music [MHJ22a]
- Visualizing Similarities between American Rap-Artists based on Text Reuse [MSEW22]

Chapter 2 & 4:

- Automated Alignment of Medieval Text Versions based on Word Embeddings [MWJ19]
- Explaining Semi-Supervised Text Alignment through Visualization [MWJ21]

Chapter 2 & 5:

- Towards Enhancing Virtual Museums by Contextualizing Art through Interactive Visualizations [MHJ22b]

Chapter 2 & 6:

- From Modern to Medieval: Detecting and Visualizing Entities in Manuscripts of Marco Polo's Devisement du Monde [MWJ22]
- A Visual Analytics Framework for Composing a Hierarchical Classification for Medieval Illuminations [MGWJss]

Remark

Although this dissertation is the work of a single 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 academic researchers are familiar with.

Only those who know the past can understand the present and shape the future.

Ferdinand August Bebel

2

At the Intersection of Visual Analytics and Digital Humanities

Visual analytics was initially defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [CT05]. In the context of this thesis, we assume that a visual analytics system combines methods from information visualization, human-computer interaction, and semi-automatic data processing, such as machine learning, to help a domain expert in the sense-making, reasoning, and decision-making process [KAF+08]. A schematic overview is given in Figure 2.1. Visual analytics of cultural data is also closely related to *Cultural Analytics* and the digital humanities, with the former being a term coined to emphasize the focus on studying all cultural data rather than sole historical materials [Man16]. In the following, we discuss the state-of-the-art of text reuse and computer vision methods for cultural data, as well as visualization projects focusing on visual cultural heritage collections. We will then look at projects for interactive data labeling and have a deeper look at projects that combine information visualization, machine learning, and human-computer interaction in a visual analytics setting for cultural data.

According to the United Nations Educational, Scientific and Cultural Organization (UNESCO), "culture should be regarded as the set of distinctive spiritual, material, intellectual and emotional features of society or a social group, and that it encompasses,

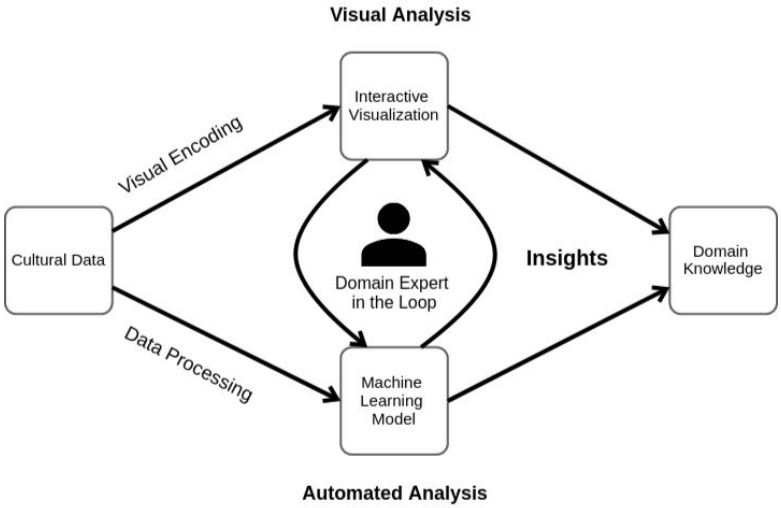


Figure 2.1: The Visual Analytics process of cultural data.

in addition to art and literature, lifestyles, ways of living together, value systems, traditions, and beliefs" [UNEO01]. Based on this definition, cultural data can come in various forms and modalities. Different concepts of culture are also discussed in the state-of-the-art report for visualization of cultural heritage by Windhager et al. [WFS+18]. In general, cultural data can be divided into historical data (cultural heritage data) and contemporary data [Man16]. An overview can be seen in Figure 2.2. Cultural heritage includes tangible culture, intangible culture, and natural heritage [Ahm06]. While natural heritage includes landscapes and biodiversity, tangible and intangible heritage refers to the combination of tangible and intangible cultural assets that contribute to the knowledge and culture of a society and have an outstanding universal value from the point of view of history, art, or science. An overview of tangible and intangible heritage can be found in Figure 2.3. These assets were preserved by the groups or societies they belong to, but in the last centuries, this was mainly done by institutions in the GLAM sector. Contemporary culture can be included in cultural heritage, but this is not guaranteed, as heritage is part of a selection process [Log07]. In the case of tangible assets, digitization and the process of creating metadata is an important part of preservation, as it allows the creation of a *Digital Twin* [GV17, HNBG22] that can ease access to the object of interest and enables the application of computational methods for exploratory and comparative analysis of a collection. For intangible assets, creating a tangible asset can be the first step in preserving an excerpt as a digital representation, but it is also possible to directly create a digital excerpt. For example, the biography of a person can be part of intangible heritage; by writing it down or filming

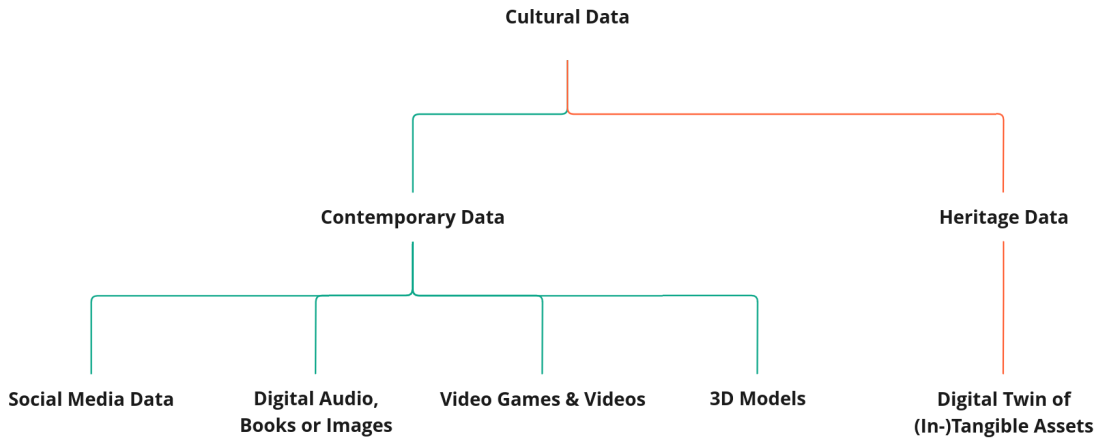


Figure 2.2: Cultural data taxonomy distinguishing between contemporary and heritage data. Both contemporary and heritage data can take the form of several modalities including 3D models, text, image, video, or audio data.

a video of the person talking about their life, a tangible asset or digital representation can be created, but this only includes some aspects of the person's life and not all details. Similarly, for the performance of a traditional dance, it is possible to write down the instructions, but it is also possible to take pictures or create a video of the performance as a digital representation. These digital representations also provide new challenges and research directions for the visualization community [MWL+22] on how to facilitate access to cultural heritage for researchers and the general public. In contrast to historical data, contemporary data sets often do not need to be digitized, since they usually have a digital representation or are already digital. This includes, for example, data and content from social networks or digital platforms collected by crowd-sourcing.

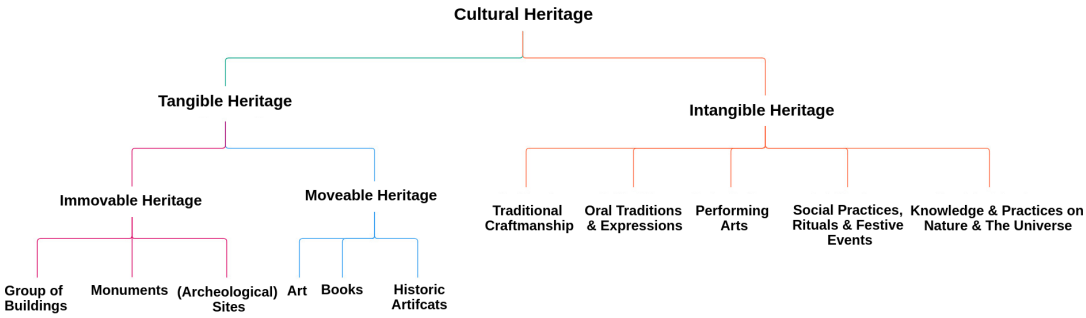


Figure 2.3: Tangible and intangible heritage taxonomy. The taxonomy is partially based on the UNESCO cultural heritage classification [UNEU03, Ahm06]. The UNESCO adjusts their classification over time, so this taxonomy could not be in line with the currently used and should only give an overview to better map cultural assets to data types.

2.1 Visualization of Textual Variance & Text Alignment

Text variants are important elements of different domain-related tasks. The general goal is to find similar and divergent patterns within two or more texts, i.e., a text alignment. Text alignment application scenarios can be divided into three areas [YJ20], first collation, which examines and records similarities and differences between variant text editions, second detection of text reuse, such as fragments, allusions, or paraphrases, and third translation alignments, where cross-lingual connections are focused. For example, methods have been developed to support the analysis of text reuse patterns [ARRO+17, JGS15, JGBS14], and, more specifically, plagiarized text fragments [RPSF15]. Furthermore, text variants appear in different languages, and visualizing automatically aligned fragments can help translators manually adjust them [You19]. However, most applications focus on different versions of a base text. Asokarajan et al. [AEA+16] tailored their system to support the analysis of lemma-level similarity for classical Latin texts. Other systems focus on directly comparing two different versions of a text [BKSK12, Sch17, WJ13]. Some systems do not apply text similarity measurements and use manually collected annotations to compare two critical editions [BJP+19]. In order to compare different translations of Shakespeare's Othello, ShakerVis [GCL+15] uses a vector space model and applies parallel coordinates and scatter plots to analyze the occurring patterns, while Alharbi et al. [ACL20] visualize alignments in parallel translations through stream graphs. Hazem et al. [HDS+19] align medieval devotional text editions using different methods, including pre-trained word embeddings, and visualize text similarity in a heat map. A detailed overview of text alignment visualizations can be found in the survey by Yousef and Jänicke [YJ20].

2.2 Computer Vision for Cultural Heritage Collections

In 2019, Arnold and Tilton proposed the *Distant Viewing* framework for working with large collections of visual material from cultural heritage in the digital humanities [AT19]. Before that, many works outside the digital humanities community proposed computer vision methods for visual cultural material. For example, various works detect objects in artworks using convolutional neural networks [WCH14, WCH16] including weak supervision methods [GGLB18, IFYA18] where only image-level annotations are available as ground truth and no instance-level annotations of specific objects. Previous work by Crowley et al. [CZ14b, CZ14a, CPZ15] showed that classifiers trained on natural images can retrieve paintings containing the selected category. Garcia and Vogiatzis [GV18] presented SemArt, an art training set for semantic understanding, while Strezoski and Worring [SW18] presented baselines for multi-

ple art classification tasks on the Omniart data set. Garcia et al. [GRN19] computed context-aware embeddings of images using image features and metadata presented in a knowledge graph. Pre-trained neural networks were also applied to historical photographs from the twentieth century to find similarities and observe trends [WS20]. Such photographs are old but still quite close to contemporary image data sets compared to fine-art paintings. For medieval images, computer vision algorithms were used to spot patterns [ÚSN+20], to classify crowns [YMCO10] or gestures [SCO11]. Lang and Ommer [LO18] applied computer vision methods to find general recurrences and organize medieval manuscripts to assist iconographic research. The survey of van Noord [vN22] gives an overview of computer vision projects that focus on iconic image analysis and discusses the challenge of a cultural gap between the model and human interpretation. Other works focused on aligning illustrations in manuscripts using image collation methods [KSD+21] or aligning texts from medieval manuscripts with their illuminations through visual-semantic embeddings [BCGC18, CSB+20]. However, none of these works applied advanced interactive visualization methods to explore the results, let alone to adjust and enrich the metadata from the collections.

2.3 Visualization of Cultural Heritage Collections

Images are the most widely used data type to visualize collections of cultural heritage [WFS+18] as cultural assets are commonly digitized as images with associated metadata. Most of the time, a collection is presented in an exploratory way by giving an overview of an image or a collection of documents [Bed01, WRZ+15, BTC10]. Furthermore, the use of timelines [GPD17, GAFM11] is a common method for showing the historical contexts of artifacts, while maps are used to show the geographical context [MHR12, DMTS14]. Other approaches combine multiple visualizations showing different facets of cultural heritage collections to give manifold perspectives on objects of interest [DPC17, GvTGD18, BDH21]. Crissaff et al. [CRD+17] created a system inspired by traditionally used lightboxes to explore, compare, organize, and annotate art image collections. Junginger et al. [JOV+20] displayed objects contained in photographic plates in an image cloud where object categories can be further investigated. The Bohemian Bookshelf by Thudt et al. [THC12] follows the serendipity principle to explore a collection of digital books and gives multiple entry points to the underlying data.

When visualization and machine learning methods are applied to a collection, they are mostly used to plot images onto a two-dimensional space by dimensional reduction. Pflüger and Ertl [PE16] proposed a visualization system for large image sets based on clustering and projection methods. Crockett [Cro16] plotted high-dimensional clusters

of images through dimensional reduction and arranged slices of images in histograms based on visual and non-visual features. Hochmann and Manovich [HM13] plotted social media images based on features such as color, brightness, and time. Similarly, Hristova [Hri16] compared art from Aby Warburg's Mnemosyne Atlas, while Yamaoka et al. [YMDK11] compared Time Magazine covers, mangas, and paintings. Strezoski et al. [SFM+20, SGBW18, SSB+19] presented several visualization systems for art collections using machine learning methods on the Omniart data set [SW18]. Including an art recommendation system in which users are presented with art that they can like or dislike [SFM+20], and exploratory interfaces for artworks based on color, sentiment, or time [SGBW18, SSB+19].

2.4 Interactive Data Labeling

Machine learning methods are found in many domains, often without incorporating domain expert feedback or visualizations that explain their functionality. As a consequence, a large number of systems and architectures are designed as black boxes, which is particularly acute for deep learning models [CL18]. Visualization can help the machine learning process by increasing trust [CMJ+20] and helping the sense-making process [ERT+17]. In a human-in-the-loop process, user interactions, such as data labeling, are used as feedback to a model to iteratively refine it. This can help to overcome training data limitations [CBC+20, BHZ+17], while inducing domain knowledge into the model. Labeling can be seen as the assignment of numerical values to an object of interest [WDC+17], the assignment of categorical labels to an instance [BGC10], or the definition of relations between objects or label categories [SSKEA19]. All these cases can be seen as single- or multi-label problems. Especially when working with visual material from the GLAM sector, multi-label methods are needed in order to assess the variety of depictions in an image of interest.

In general, visual annotation systems support manual labeling tasks through visualizations. Zhao et al. [ZGB+16] proposed a graph visualization to explore annotations within an annotation system. Some systems support only a small predefined vocabulary [WHHA11], or data properties, such as outliers [CBY10], while others support textual annotations without a defined vocabulary [EB12]. Other systems focus more on the collaborative aspect between multiple users through monitoring functionalities such as inter-annotator agreement [QMSM17], common ground construction in asynchronous collaborations [CAB+11], or organizing findings, hypotheses, and evidence in a collaborative visual analytic system [MT14]. Furthermore, labeling can be supported by visualizations to explore data and machine learning methods to recommend points of interest [FDB18, KKZE19]. Other visualization systems close the loop between ma-

chine learning algorithms and users by supporting different labeling tasks through interactive visualizations.

Active learning [WSZ+20] is a popular method of reducing the amount of manual labeling. In an active learning process, a user labels data samples that are queried by an algorithm based on different strategies to improve the underlying model while minimizing the amount of work. Some initial works [FOJ03, WFH+01, HKBE12, HNH+12] extended the concept of user interaction to interactive learning, in which a user interacts with visualizations to create a classifier. Modern visual analytics systems combine interactive visualizations with active learning strategies to better understand the classifier and support several tasks and domains such as correcting mislabeled training data [XYX+19], text data annotation [KPSK17, SLT17], labeling documents [CPY+19], constructing sentiment lexicons [MBM14] or classifying the relevance of tweets in real-time [SLK+19]. Bernard et al. introduced visual interactive labeling (VIAL) [BZSA18] as a concept to combine active learning with visualization systems to explore and select data points for labeling. Furthermore, Chegini et al. [CBC+20] showed that VIAL can outperform active learning under specific conditions and can help solve the cold start problem. Therefore, combining active learning with visualization using visual encodings that expose the internal state of the learning model and the use of knowledge from visual perception theory can help the labeling process [LLL+18].

2.5 Visual Analytics for Cultural Data

The definition of visual analytics changed over time; in the beginning, Cook and Thomas proposed visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces" [CT05]. In the following years, another definition was coined by Keim et al. with a stronger focus on automatic algorithms: "Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning, and decision-making on the basis of very large and complex data sets" [KAF+08]. The slight difference in definition results in several publications on cultural data that fall under the initial definition of visual analytics, but without using machine learning or similar automatic algorithm procedures, which are outside of the scope of this thesis [KBD15, BCS+16, XEJJ14, CRT+22].

There are several works on visual text analytics that close the loop between the model and the domain expert, i.e., putting the human in the loop. Including word vector space manipulation [PKL+17], stance classification [KPSK17], labeling and classification of question types [SJS+21], extracting social network from newspapers [KLB14], constructing lexicon-based concepts or refine topic models [EASS+17, EAKC+19,

CLRP13], but they focus mainly on contemporary data and do not include cultural heritage data. Similarly, VIAL processes have already been applied to suggest keywords for articles in a literature collection [ABB18] and the labeling and classification of music [RAZ+18]. At the intersection between digital humanities and visualization research, several works support the analysis of historical material through Close and Distant Reading methods with named entity recognition and topic modeling, but do not close the interaction loop between the domain expert and the applied model [CYCD18, MDG16, JLK+16, VCPK09, HPK+21]. In contrast to that, the VarifocalReader [KJW+14] presents use cases from German poetics and English classic literature in a human-in-the-loop setting, allowing topic segmentation, named entity recognition, automatic summarization of text segments, and active learning to create automatic annotations.

The field of archaeology has a long history of using visualizations [Opg21] but only a few works use semi-automatic or automatic methods. Examples are automatic methods for analyzing rock carvings [DPT+12], semi-automatic tools for deterioration risk analysis of ancient frescoes [LZS16], and decision-making on flood risk in cultural heritage [LZSW17]. Semi-automatic approaches were also proposed for ornamentation on painted pottery. In particular, for the creation of archaeological drawings by extracting ornamentals [LKK+20], and for the classification and recommendation of surface patterns by interactively segmenting and labeling parts of the surface [LHP+22].

In recent years, visualization methods have often been applied in projects dealing with cultural heritage data [WFS+18, JFCS16], but there is a lack of works that combine state-of-the-art machine learning and visual analytics methods on historical material, especially works on cultural heritage that close the interaction loop between the domain expert and the model to help domain experts solve humanities research questions. The reasons for this are probably related to the young age of the field and the limited amount of training data. Of course, not all visual analytics projects focusing on cultural data need to close the interaction loop, but this can still help answer domain-specific research questions and, in the case of applied machine learning methods, it can reduce skepticism and reservedness in using them from the humanities side. Especially for low- and under-resourced languages and historical visual material, a human-in-the-loop setting combining machine learning with visualization can help to tackle the limited amount of resources.

2.6 Addressed Challenges

We address multiple problems at the intersection of digital humanities and visual analytics and provide solutions for them. A common theme of the projects is that state-of-the-art machine learning algorithms were required together with visualizations to

answer domain-specific research questions. Especially, the works on medieval vernacular poetry in Chapter 4 and the work on medieval illumination in Pars Bibles in Chapter 6 close the loop between the machine learning model and the domain expert to study data related to cultural heritage. In the following, we will briefly state the domain-specific problems and our proposed solution, which will be presented in detail in the following chapters of this thesis.

Problem 1: Since the early days of rap music, references to pop culture, as well as other rap artists, have been an integral part of the artistry of the lyrics. Rappers may use them to introduce their shared personal background, such as where they grew up. In addition, rap musicians refer to each other by adopting fragments of lyrics, for example, to give credit. Listeners may be interested in finding artists similar to the artists they already know, or in finding patterns and references on a song- or line-level. Crowd-sourced knowledge platforms like Genius.com [Inc14] can help in this process through user-annotated information about the artist and the song, but do not include visualizations to help users find patterns and structures across several songs or artists. Also, due to the large number of lyrics, finding patterns can be an arduous task without automated methods to detect text reuse.

Solution 1: In Chapter 3 we present two visualization systems to analyze text reuse in rap lyrics from Genius.com. The systems support the user to detect text reuse and allusions between songs and to explore connections between artists. Furthermore, artists and their lyrics can be analyzed using multiple exploratory visualization methods to support domain-specific tasks. We also trained a neural network specifically tailored for rap lyrics to compute similarities.

Problem 2: The analysis of variance in complex text traditions is an arduous task when carried out manually. Text alignment algorithms provide domain experts with a robust alternative to such repetitive tasks. Existing white-box approaches allow to establish syntax-based metrics taking into account spelling, morphology, and order of words. However, they produce limited results because semantic meanings are typically not taken into account. On the other hand, methods based on natural language models include semantic meanings but tend to fail on low-resource and under-resourced languages because of limited training data. This gap creates the need for semantic methods that deal with limited data while including domain knowledge.

Solution 2: Our interdisciplinary collaboration between visualization and digital humanities scholars combined a semi-supervised text alignment approach based on word embeddings that take into account not only syntactic but also semantic text features, which is presented in Chapter 4. In our collaboration, we developed different visual interfaces that communicate the word distribution in the high-dimensional vector space generated by the underlying neural network for increased transparency, assessment of

the tool's reliability, and overall improved hypothesis generation. We further offer visual means to enable the expert reader to feed domain knowledge into the system at multiple levels with the aim of improving both the product and the process of text alignment. This ultimately illustrates how visualization can engage with and augment complex modes of reading in the humanities.

Problem 3: In response to the COVID-19 pandemic, public spaces, such as museums and art galleries, are experiencing increasing demands to offer virtual online access. While current solutions seek to replace or augment a real visit, online tours often suffer from being too passive and lack interactivity beyond recreating the physical space [VKCA20, Hof20] to keep virtual visitors meaningfully engaged with an exhibition. Museums and art galleries seeking to broaden and engage their audience more deeply should offer intriguing experiences that invite the visitor to explore, be entertained, and learn by interacting with the content.

Solution 3: In collaboration with a museum exhibition designer, we propose a novel virtual museum experience in Chapter 5 that utilizes multiple visualizations to contextualize a gallery's digitized artworks with related artworks from large image archives. We use the WikiArt data set that includes more than 200,000 images and offers diverse metadata used for comparative visual exploration. In addition, we apply machine learning methods to extract multifaceted information about objects detected in images and to compute similarities between them. Visitors to our virtual museum can interactively explore the artworks using different search filters, such as artists, styles, or object classes detected within an image. The results are displayed through interactive visualizations that offer different perspectives on artwork collections, leading to serendipitous discoveries and stimulating new insights. The utility of our concept was confirmed by an evaluation with virtual museum visitors, including people from the general public and humanities scholars.

Problem 4: Annotated data is a requirement for the application of supervised machine learning methods, and the quality of annotations is crucial for the result. Especially when working with cultural heritage collections that comprise a manifold of uncertainties, annotating data remains a manual and arduous task that needs to be carried out by domain experts. Furthermore, contemporary hierarchies for image classification and object detection like ImageNet and Open Images contain a manifold of classes that are not present in medieval illuminations and only partial of the classes a medieval scholar could be interested in. Our project started with two already annotated sets of medieval manuscript images, which, however, were incomplete and comprised conflicting metadata based on scholarly and linguistic differences. Our aim was to create (1) a uniform set of descriptive labels for the combined data set and (2) a high-quality

hierarchical classification that can be used as a valuable input for supervised machine learning.

Solution 4: To reach these goals, we developed a visual analytics system to enable medievalists to combine, regularize, and extend the vocabulary used to describe these data sets. Visual interfaces for word and image embeddings as well as co-occurrences of the annotations across the data sets enable annotating multiple images at the same time, recommend annotation label candidates, and support composing a hierarchical classification of labels. Our system itself implements a semi-supervised method, as it updates visual representations based on the medievalists' feedback, and a series of usage scenarios document its value for the target community in Chapter 6.

I seek Sun, deceive none, for each one must teach one.
Keith Edward Elam

3

Exploratory Visual Analysis of Rap Lyrics and Rap Artists

Rap music emerged from a long history and tradition as a rhetoric of resistance [Kop02] for marginalized groups to express their social and economic struggles rhythmically and poetically into a standalone music genre. Initially, the genre remained primarily within the boundaries of its corresponding subculture. But in the 1980s, with the emergence of *gangsta rap* through groups such as N.W.A. and artists such as Snoop Dogg or Dr. Dre, rap music made its way into the mainstream [AS05, Lig99]. Today, it is one of the most popular music genres with great influence around the world [McK]. Rap music as part of hip-hop culture combines "creative use of language and rhetorical styles and strategies" [Kop02]. This characteristic of rap music creates similarities to literature in regard to using poetic language or referencing other artists like the rephrasing of famous quotes or from a musical standpoint through sampling. In particular, intertextuality can enhance the enjoyment of music through emotions such as nostalgia. Since the early days of rap music, references to pop culture but also to other rap artists have been an integral part of its lyrical craftsmanship. Rappers can share personal connections through their background, such as the city or neighborhood where they grew up or even gang affiliation. Because of these relations, they often reference similar themes, places, or culturally specific phrases. Rivalries also play

an important role. Controversies between formerly affiliated rappers such as members of the group N.W.A, rappers being affiliated with different gangs, or rivalries spanning the whole genre such as the East Coast vs. West Coast clash in the 1990s often result in *diss tracks*. In these, musicians mock each other, often reusing or referencing their adversary's lyrics to use against them. More positively, artists sometimes reuse phrases from other musicians to pay homage to them and their lyrical craftsmanship, be it out of mutual respect, or in the effort of a younger artist to refer to those who inspired them [Lig99]. Detecting all these references can be a difficult task, because the listener needs a lot of knowledge of the genre and its history. User-crafted annotations from platforms like Genius.com [Inc14] can help in this process.

However, an issue that arises with anything related to commercial success is plagiarism [Cha11]. Websites like Genius.com [Inc14] offer annotated song lyrics, while services like Spotify [AB08] and SoundCloud [Lim07] provide access to millions of songs on demand. With tools like these, discovering music has never been easier. This easy access combined with the promise of financial success achievable through rap music may lead aspiring artists to plagiarize successful ones in the hopes of garnering attention. Due to the sheer amount of lyrical content, automated means of detecting text reuse can help find cases of plagiarism. Furthermore, these automated procedures can also be used to identify similar artists on the basis of their lyrical content. These data may then be utilized to help fans of the genre find new artists similar to those they already enjoy. References to other artists and commonly used phrases could be traced back to their origin, allowing those interested in rap music to deepen their knowledge. Visualizations can be applied to communicate such similarities and to further ease the process of detecting them.

We combine natural language processing techniques with visualizations to communicate similarities in lyrics to domain experts and casual users who are interested in music. References can result in similar lines, these cases can be found by similarity searches based on word and sentence embeddings, as the embedding space preserves semantic relations. The similarities found can give starting points to further search for cases beyond rephrasing like plagiarism or can just increase the knowledge about the genre and its history. In particular, the domain problem of detecting similar lines can be seen as a text alignment problem [YJ20]. We visualize the text alignments starting with an edge in a graph as an aggregate over two artists, followed by streamlines representing the songs and showing dependencies between them, and finally the side-by-side inspection of two lyrics in a collation manner. For this, we apply word embeddings to the lyrics of Genius.com [Inc14], which are enriched with metadata about the songs and the artists and additional annotations about the lyrics. The data is used to compute edges between artists with weights depending on line similarities in the lyrics. Fur-

thermore, we extended these methods [MJ21] with visualizations for an exploratory multi-faceted analysis of the data. For this, we designed and applied visualizations to communicate the sentiment of the lyrics of an artist, to compare the vocabulary of different artists, and to compare the development of rap genres. Visualizations can help to give a better understanding of the music genre and the relations between different artists. Thus, supporting multiple visual text analysis tasks in the digital humanitiess [JFCS16], such as corpus, sentiment, and text reuse analysis using Distant Reading methods [Mor13]. This approach is generalizable to the lyrics of all genres of music and different languages.

In addition, we present a second visualization system built on top of the first that uses state-of-the-art technologies like RoBERTa [LOG+19] in order to compute similarities between lyrics and a Neo4J database to store all artists, songs, and lines as nodes in a graph. The system also allows the user to explore similarities between artists, detect cases of plagiarism and allusions between songs, and discover new artists or songs.

3.1 Related Works

3.1.1 Similarity of Musicians

Similarity Analysis of Musicians is one of the use cases in the state-of-the-art report on visualizations of musical data by Khulusi et al. [KKM+20]. Although the text of a song is not directly musical data, it is still connected to the music and the musician.

Similarity measurements for musicians can be divided into multiple categories [KTT09]. These categories include collaborative filtering of user data [SH12], computation of co-occurrences of words on web pages [SKW05b], biographical information [JFS15], and content-based methods to focus on audio or textual data from the songs themselves. In our work, we focus on the lyrics of musicians; therefore, approaches that focus on user data, biographical information, web content, and sound features are not applicable. Works using user data to measure the similarity of musicians include Amazon sale statistics [Vav17] or Spotify listing histories [Spo18]. Similar works are done by Gibney [Gib11], Cano and Koppenberger [CK04] and Gleich [GRZL05] based on user data and web co-occurrences. All these methods visualize, similar to us, the data through graphs either focusing on a given artist or the whole database, but they do not include additional visualizations to inspect a more detailed level of the data. Similarly to the former works, platforms like Genius.com are crowd-sourced and include rich annotated metadata about musicians and, more importantly, the transcribed lyrics of the artist. These text collections can be analyzed in terms of text reuse and overall similarity. Some works compared the vocabulary of rap artists extracted from Genius.com for American [Dan14] and German artists [Sch15]. Another work used the vocabulary to

define the similarity between the artists [FD17], which is also in the focus of our work. Still, these works focus only on the vocabulary and do not include other aspects of the underlying data.

Another way to find and visualize the relations between artists would be to observe the influence of the musicians of the past on the currently active musicians. For example, MuzLink [LH21] allows exploring collaborative and influential relationships between musical artists using connected timelines. Other works tried to observe this influence through graph visualizations showing the history of rock [LA18] or by finding artists that are prototypical to a genre [SKW05a]. Similar approaches can be of interest to rap music because of the long-established culture of referencing and collaboration, where new upcoming artists are referencing previous artists or are supported by established artists. There are also works in the field of music information retrieval that focus on lyrics to compute similarities between artists, but without visualizing them [LKM04, BH03]. In contrast to the prior works, Jänicke et al. [JFS15] designed a visual analytics system that supports the profiling of musicians based solely on biographical characteristics. Similarly, Oramas et al. [OSEAS15] compute artist similarities based on biographical information and word embeddings.

We use Genius data for an automated semantic analysis of songs to generate similarities between artists and explore their lyrics with several visualizations. The first prototype uses fastText [GBG+18] in order to compute similarities between song lines, which only provides word vectors trained on Wikipedia and Urban Dictionary [Pec99] that are not fine-tuned on lyrics. For the second system, we use two transformer models, which are fine-tuned on the task of semantic textual similarity and natively produce sentence vectors. Additionally, we train a model specifically on the collected corpus of rap lyrics to include domain knowledge of this specific task.

3.1.2 Song Similarity

The similarities between songs are often addressed in the music information retrieval community and can be divided into context-based methods [KS13] and content-based methods [DSK21]. Content-based methods focus on the audio signal, while context-based methods can include all information that is not part of the audio signal itself, e.g., metadata or lyrics. We follow a visual text analysis process and disregard content-based methods, as this information is also not available in the Genius data.

In contrast to our approach, many music information retrieval systems focus on sound features, but often combine them with the lyrics [RASJ14]. Yu et al. [YTRC19] combined textual and audio features using deep cross-modal learning to retrieve lyrics by audio and audio by lyrics, but did not include visualization. LyricsRadar [SYN+14]

allows users to browse song lyrics while visualizing their topics in a two-dimensional space. Furthermore, graph-based visualizations to tackle plagiarism detection based on sound features are designed by Ono et al. [OCF+15] and De Prisco et al. [DPLM+16].

3.1.3 Text Reuse

Our focus lies on textual data and has similarities with works based on textual alignment and text reuse presented in Section 2.1. Common methods to visualize text reuse patterns are Grid based [ARRO+17, JGBS14], Sequence-aligned [AEA+17] or Text-oriented Heat Maps [DPDT18]. More popular are side-by-side views supported by stream graphs and aligned barcodes [RPSF15, JW17a, MWJ21]. On a line-level, variant graphs [JGF+15, RGP+12] and tabular views [DM11] can help visualize similarities and differences. Our prototype application aims to find similar artists by detecting possible occurrences of text reuse in their song lines. From a text alignment perspective, we visualize text reuse scenarios at the song- and line-level with collation methods, where we treat similar lyrics as textual variations [JW17a]. To visualize these occurrences we utilize a combination of Side-by-side Views and Aligned Barcodes that aid in pairwise collation as well. For the collation of more than two similar lines, which can be seen as text variants, we use Variant Graphs [JGF+15].

3.2 Data Processing

3.2.1 Data

We collect song lyrics and metadata about rap artists from Genius.com [*], which is a website where casual users and even artists themselves can transcribe song lyrics and annotate them with additional information. Genius started as *Rap Genius* in 2009 but changed its name in 2014 to include knowledge of other genres of music and other types of media, such as literature. Annotations can be added by any user, but must be accepted and reviewed by a moderator. These annotations can include references to other songs or artists, possible interpretations of certain lyrics, an explanation of specific words or phrases, e.g. slang or wordplay, or connection to pop culture, the artists' personal life, or historical and current events. On top of that, Genius provides metadata such as featured artists, release dates, record labels under which a song was released, and many more; all this information can be extracted through the Genius API.

For our first prototype, we used a subcorpus from the Genius Expertise data set [LB21]. The entire data set includes 223.257 songs crawled from Genius.com in the time frame

[*] https://genius.com

between September 2019 and January 2020. We filtered the data set for English songs associated with the rap and hip-hop genre, resulting in a subcorpus with 37.993 songs by around 2300 different artists. Furthermore, Genius provides access to additional metadata, annotations, user information, and information about the artists. We crawled the missing artist metadata so that these can be used for the artist profile, which can be seen in Subsection 3.3.2. Since crowd-sourced data has the property of having troll entries, the subcorpus had to be cleaned off them, resulting in a corpus with 35.783 entries. Additionally, we crawled lyrics and metadata from Genius.com of 28.969 songs by around 600 German rap artists and groups.

For the second system, we compile a list of 219 American hip-hop artists based on popularity and personal interest. We collect Genius data on the most popular songs by artists up to a maximum of 200 songs per artist. Thus, our database includes a total of 25,654 songs with 1,598,466 lines. The data contain information about the artists, such as their names, a short description, and the artists' songs including their lyrics. The lyrics were lowercased, and the punctuation and special characters were removed. Since relationships and similarities are our main focus, we used a Neo4j [†] graph database to store artists, songs, and lines as nodes. We focus on textual alignments between individual lines to establish connections between songs and artists. For this, the text was split into parts, so that each line in a song is represented by its own node in the database. Beyond the lyrics, the aforementioned annotations were used to enrich the line nodes with information about which part of a song they belong to and who they were performed by. To preserve order, each line node also gets an index according to its position within the song. We connect these line nodes with similarity relationships based on the findings of our search for textual alignments. Each song is represented by a node as well, containing information about the title, release date, associated album, featured artists, etc. The line nodes are connected to their respective song nodes via a *part-of* relationship. Thanks to this, it is later possible to compute song-level similarities and explicitly connect songs through similarity relations. Finally, the same is done for the artists, as those too are represented by their own nodes containing their name, description, alternate names, etc.

3.2.2 Textual Alignment

For the first system, we apply fastText word vectors [BGJM17] to compute similarity values between artists based on their lyrics. We choose fastText vectors because they include out-of-vocabulary words, which are a common phenomenon for rap lyrics due to slang, adlibs, or neologisms. Furthermore, a model trained on the Urban Dictio-

[†] https://neo4j.com

nary [WMM+20] is available, for which we hoped to gain a better contextualization of word vectors for slang or adlibs. For the German corpus, we used fastText word vectors trained on Wikipedia. We treat each line in the lyrics as a sentence, for which a sentence vector is computed by unsupervised smooth inverse frequency [Eth18]. Therefore, the sentence vector is a weighted average of the word vectors. The weight depends on the frequency of words, the size of the vocabulary, and the average length of sentences in the corpus.

For the second system, we use RoBERTa [LOG+19] to find semantically similar lines. The model takes a string of words as input and produces an embedding vector that represents the semantic meaning. We use two versions of RoBERTa, one ready-to-use version specifically fine-tuned on the task of semantic textual similarity called 'stsb-roberta-base' [‡], and the same model fine-tuned on our corpus of rap lyrics, which we give the name 'rapBERTa'. The reason for this additional training is the widespread use of slang, neologisms, and pop culture references in hip-hop. The hypothesis is that, in learning rap-specific language, rapBERTa may also perform better in finding meaningful semantic similarities in a corpus of rap lyrics. The model produces sentence embeddings for each line of the corpus.

For both systems, the sentence vectors are added to a faiss index structure [JDJ21] to query the lines that are the nearest neighbors for an efficient similarity search. The index is used to find the 15 nearest neighbors for each line, i.e. the most similar lines within the corpus based on cosine similarity. We focus on lines instead of sentences because rap artists write their lyrics line by line, and lines are often sentences. For the second system, the resulting relations between song lines are added to a Neo4j graph database with the corresponding similarity value as *neighbor of* relationships.

3.2.3 Artist Similarity

The user should be able to discover similarities between artists in a graph. Therefore, it is necessary to compute a similarity score between artists based on the relationships between their respective song lines. There are different viable approaches to compute an artist-to-artist similarity. One way is to use a rank-based metric s_{ab} using the cosine similarity between the target lines and its nearest neighbors of an artist a to an artist b.

$$s_{ab} = \sum_{i=0}^{n} \sum_{r=1}^{k} cos(l_i, l_r) \cdot ((k+1) - r)$$

[‡] https://huggingface.co/cross-encoder/stsb-roberta-base

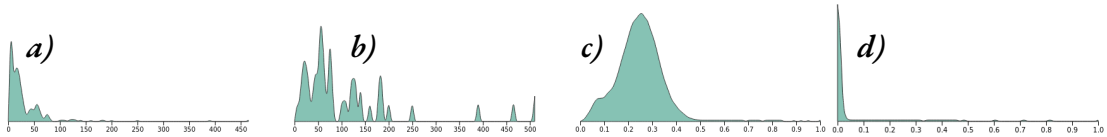


Figure 3.1: Kernel Density Estimate plots of the German graph. Showing the minimum number of songs (a), the maximum number of songs (b) the edge weights after min-max normalization with Box-Cox Transformation (c) and without (d).

For two artists, we use all their lines that are nearest neighbors, i.e. the most similar ones in the corpus. For each of these pairs l_i and l_r, we take the cosine similarity $cos(l_i, l_r)$ multiplied by the number of nearest neighbors $k + 1$ minus the rank r of the neighbor l_r. Through this, we obtain a rank-based weighting, which favors lines with a lower rank in the nearest neighbors list. We then take the sum of all such pairs for two artists. This value is further normalized using the total number of lines of all the songs of the artist. For two artists a and b this results in two similarities s_{ab} and s_{ba} because the nearest neighbor relationship between the sentence vectors is not symmetric. Both values are summed together to get a single-edge weight for the graph. To allow better filtering, we apply a Box-Cox transformation [BC64] and a min-max normalization to the edge weights. This gives similarity values that are easier to interpret by humans between 0 and 1. We apply a Box-Cox transformation because it transforms a skewed distribution into one that is close to the normal distribution. The resulting distribution can be seen in Figure 3.1c, and the original skew distribution of the edge weights can be seen in Figure 3.1d. This approach is used in the first prototype.

Another approach to compute an undirected artist-to-artist similarity score is to use the number of *neighbor of* relationships found between the song lines and the cosine similarity in those relationships. Not all artists have the same number of songs stored in the database, and therefore the number of lines for each artist varies as well. Especially, newer musicians may not have recorded that many songs or have not yet had all of their songs added to Genius by users. To account for this, the relative number of *neighbor of* relations from artist a to b (p_{ab}) can be used to compute the similarity between artists. Of course, there is also the counterpart from the artist b to a where the relative number of *neighbor of* relations is defined by p_{ba}. Another component is the average line-to-line similarity between two artists a and b $sim_{avg_{ab}}$.

$$sim_{avg_{ab}} = \frac{\sum_{i=0}^{n} \sum_{r=1}^{k} cos(l_i, l_r)}{n \cdot k} \tag{3.1}$$

Intuitively, if there are two pairs of artists with the same relative number of similar lines between them, the artist pair with the higher average line-to-line similarity is

closer. The difference from the computation of the similarity value for the first prototype is that an average is used instead of a rank-based metric. Finally, the minimum of p_{ab} and p_{ba} is combined with the average line-to-line similarity.

$$s_{2_{ab}} = min(p_{ab}, p_{ba}) \cdot sim_{avg_{ab}} \tag{3.2}$$

The rationale behind using the minimum of p_{ab} and p_{ba} is that a high p_{ab} indicates that a large number of lines of artist a are similar to lines of artist b. However, this is not enough to indicate that both artists closely resemble each other. It could simply mean that artist a reuses themes often that artist b only features in some of their songs. If a p_{ba} that is smaller than p_{ab}, however, results in a similarity score between artist a and artist b that is higher than those between artist a and any other artist or artist b and any other artist, it makes a strong case for the close relationship between them. This approach is used in the second application. Furthermore, we decided against a numerical similarity filter and therefore a normalization for the second prototype, as the similarity value was hard to understand for casual users.

3.3 First Visual Interface

In the process of creating the first prototype, we formulated seven research questions. These questions were derived on the basis of the information available in the Genius data and through discussions with people interested in rap music.

- **Q1**: Which artists have collaborated, were part of the same record label or group, and are similar based on their lyrics? (Graph Subsection 3.3.1)
- **Q2**: What are the most similar artists or songs for a specific artist? (Artist Profile Subsection 3.3.2)
- **Q3**: Which songs of two artists have similar lines, are they remixes, covers, or interpolations? (Side-by-Side Alignments and Variant Graphs Subsection 3.3.3)
- **Q4**: When were songs released that are associated with a specific genre? (Genre Timeline Subsection 3.3.4)
- **Q5**: What vocabulary is used in a genre, by an artist, or in a song, and how does the vocabulary differ, are there dominant words? (TagCloud and TagPie Subsection 3.3.5)
- **Q6**: What is the sentiment for a specific song? Are there artists who, on average, have a negative or positive sentiment? (Sentiment Barcodes Subsection 3.3.6)

For each question, a visualization can be used, and different tasks are performed. We create a graph with a force-directed layout where the edges between the nodes are based on the similarity of the lyrics to identify similar artists while exploring the graph. For the song and line comparison tasks, we apply a line-level alignment approach based on side-by-side views to allow comparison of lyrics. The vocabulary of a song, an artist, or a genre can be inspected with TagClouds and even compared with other songs, artists, or genres with TagPies [JBR+18]. Furthermore, we visualize the sentiment of a song as a colored barcode, and the genre tags as a boxplot-inspired timeline.

The biggest challenge in the design process was to present the low-level line similarities in a way that a user could quickly get an overview of the corpus. Due to the corpus size, it is not possible to give a detailed overview of all the line similarities. Therefore, we decided to aggregate the line similarities into a single value that can be used as an edge weight and, therefore, to show the relation between artists. This allows one to bridge from a line-level to a song-level or artist-level and also to encode other information, like the social relation, into the edge.

Following Brehmer and Munzner's task abstraction [BM13, Mun14] the domain-specific tasks are to *derive* references between songs, *identify* similar musicians and songs based on their lyrics, *explore* a network of musicians, to *compare* the lyrics, the sentiment used and the vocabulary of different artists and songs, and to give an overview (*summarize*) of the different facets of the data set.

3.3.1 Artist Graph

Following the Information Seeking Mantra [Shn96], we started by giving an overview by visualizing the similarities between artists as a node-link diagram. For this, we represent each artist as a node and use the artist similarity as the edge weight, so an edge indicates that two artists are similar based on their lyrics.

Design. For the graph, we chose a force-directed layout, as they are easy to understand, flexible to use with graph aesthetic criteria, and easy to interact with to change the positions of the nodes. To reduce visual clutter, the user can filter the edges shown with sliders. For this, the similarity values and the minimum and maximum number of songs of an artist can be used. The distribution of the edge weights and the distributions of the minimum and maximum number of songs of the artists connected by the edges are displayed as Kernel Density Estimate Plots, which can be seen in Figure 3.1a-c. This allows a user to visually assess the impact on the edges shown in the graph. We use a bandwidth of 1 to create a smooth estimation of the distribution.

After filtering, all nodes without an edge that satisfies the conditions are removed. Furthermore, we color-coded the edges to show different relations. A blue edge indi-

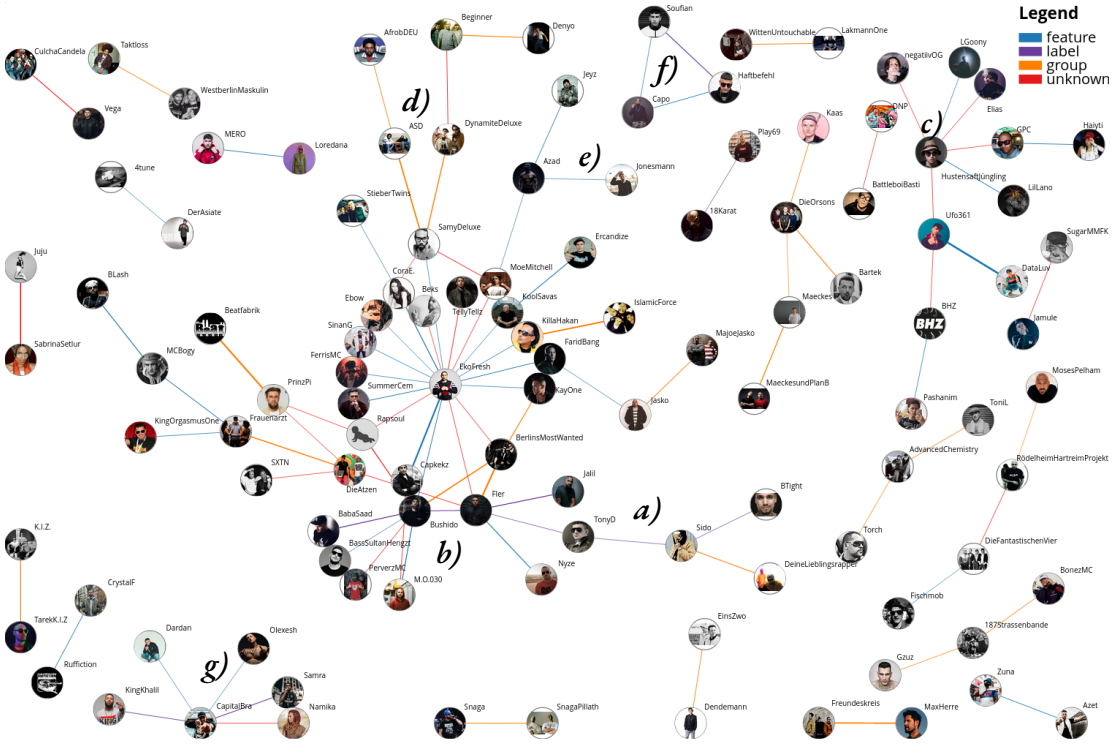


Figure 3.2: An excerpt of the similarity network of German rap artists based on the most similar lines in their lyrics. Label and collaboration partners tend to be connected.

cates that two artists have at least one song together, a purple edge indicates that the artists are or were signed by the same record label, an orange edge is a *part of* relation for group members, while a red edge shows an unknown relation. We choose red for the unknown relations to better highlight them, as they represent an unknown or missing social relation. The different types of relationships show social connections beyond the lyrics, which can give hints about why the lyrics of the two artists are similar. We extracted the relation types from the lyrics and the Genius.com metadata, and for some cases, we added the record label or *part of* relation with domain knowledge. Furthermore, we map the similarity value of two artists to the thickness of the edge to highlight relations with a higher similarity value.

Next to the graph is a list of the most similar song pairs. The song pairs are color-coded from white to red on a linear scale depending on the number of nearest neighbors. An example of the English corpus is shown in Figure 3.4a. When clicking on a pair of songs in the list, the side-by-side alignment view is displayed (Subsection 3.3.3).

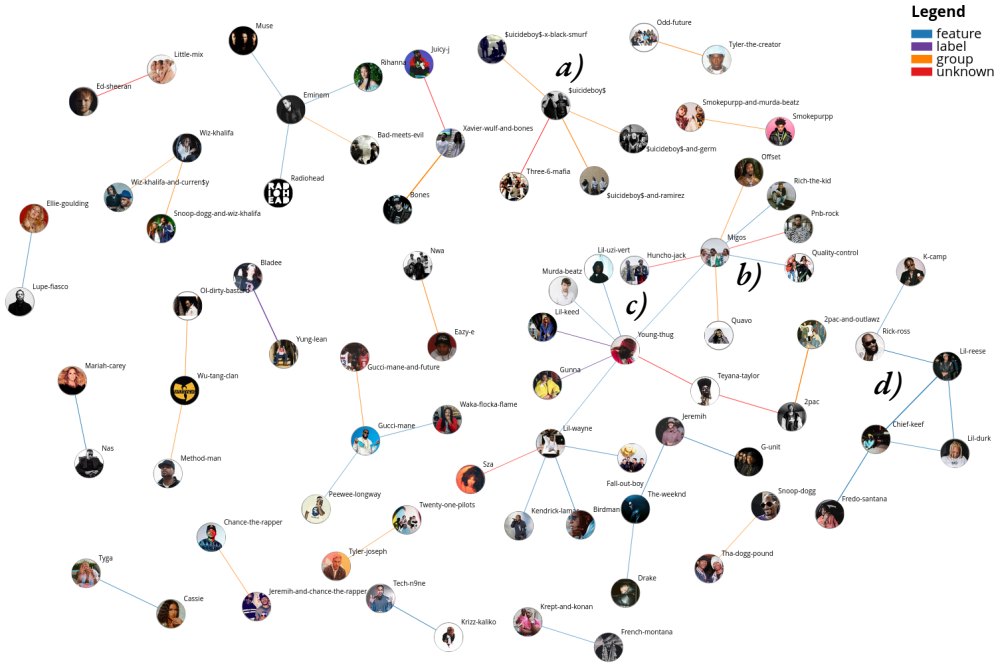


Figure 3.3: An excerpt of the similarity network of English rap artists based on the most similar lines in their lyrics. Collaboration partners tend to be connected.

Additionally, the user can search for two specific artists of interest or click on an edge in the graph to investigate the songs of the artists.

Use Case. Taking into account knowledge about individual artists, their style, and history, it becomes apparent that the graphs in Figure 3.2 and 3.3 show meaningful connections. We can observe not only subgraphs of artists that share thematic and even stylistic similarities, but sometimes even clusters within those subgraphs that point to a deeper connection between artists.

The subgraph of the German corpus (Figure 3.2) shows different clusters. The nodes in position (a) show previous members of the German label *Aggro Berlin* and the rap crew *Die Sekte* such as *Sido*, *Tony D*, and *BTight*. At position (b) previous members of *Ersguterjunge* and *Berlins Most Wanted* can be seen, such as *Bushido*, *Fler*, *Kay One*, *Eko Fresh*, *Nyze*, *M.O.030*, and *Baba Saad*. Above these artists, more Berlin-based artists like *Prinz Pi* can be seen. Another interesting thing to note is that *Eko Fresh* is connected with a large number of artists, showing his influence on the German rap scene through collaborations and the support of new artists. Position (c) shows *Hustensaft Jüngling* and other artists with whom he collaborated. Some of the edges are created because of the exhaustive use of brand names like Gucci or Fendi, or

drugs like Lean. At position (d) the Hamburg-based group's *Beginner* and *ASD* with some of their members can be seen. Positions (e) and (f) show Frankfurt-based artists like *AZAD*, *Jonesmann*, *Jeyz*, and *Haftbefehl* with his brother *Capo* and label member *Soufian*. The label members and feature partners of *Capital Bra* can be seen in position (g). In addition to these examples, multiple *part-of* and feature relations can be found.

Another subgraph showing the English corpus can be seen in Figure 3.3. Position (a) shows *$uicidebo$* and *Three 6 Mafia*, where *$uicidebo$* reused multiple lines from *Three 6 Mafia* songs, which can also be seen by multiple entries in the list of the most similar songs in Figure 3.4a. At position (b), *The Migos* and two of their members, *Offset* and *Quavo*, can be seen together with multiple feature partners. Around *Young Thug* (c) are artists with whom he collaborated, such as *Lil Uzi Vert*, who he influenced, and *Lil Keed* and *Gunna*, who are signed by his record label *YSL Records*. Position (d) shows multiple artists from Chicago that are associated with the Drill genre, such as *Lil Durk*, *Lil Reese*, *Chief Keef*, and his cousin *Fredo Santana*.

3.3.2 Artist View

An interesting property of the Genius.com data is the rich annotated metadata, including references and information about the artists. We give an overview of some of the metadata from Genius.com and display a list of the most similar artists based on their lyrics and all of the songs of the artists in the artist profile view. This view is accessed by clicking on a node in the graph or the artist's name in the side-by-side view.

Design. The list of most similar artists is color-coded in the same way as the graph, but instead of the edge thickness, saturation is used. Through this list, the user can further explore other artists. The list of songs includes for each song the ten nearest neighbors color-coded in the same way as the list of the most similar songs. Furthermore, Genius.com metadata is used to display relationships with other songs. These relation types are: samples, sampled in, interpolates, interpolated by, cover of, covered by, remix of, remixed by, live version of, and performed live as. By clicking on a color-coded nearest neighbor, the alignment view pops up. Therefore, a user can explore the network and find different points of interest to further investigate alignments.

Use Case. Figure 3.4b shows the profile of the Berlin-based rap group *BHZ*. The list of similar artists shows mostly artists who are either from Berlin or have at least one song with them. Below, we can see that the song *LSD* is an interpolation of *Yung Kafa & Kücük Efendi - Saphir*.

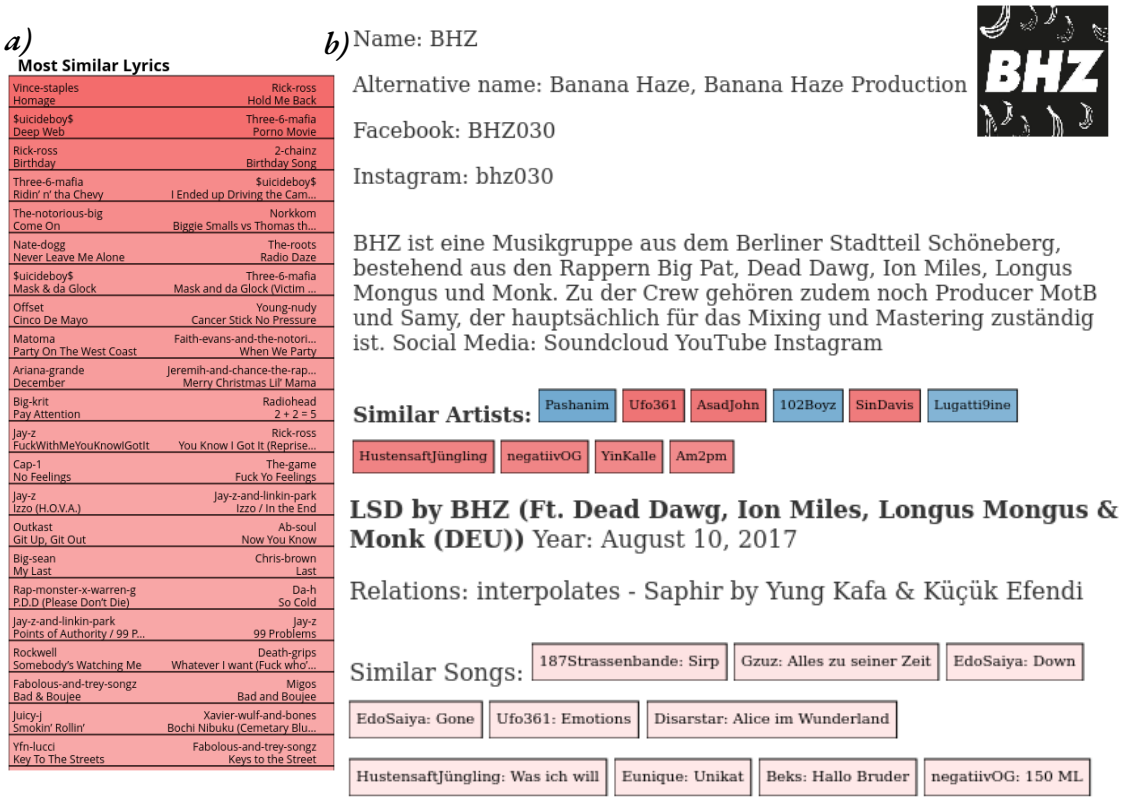


Figure 3.4: The most similar English songs (a) and the artist profile of the German rap group *BHZ* (b).

3.3.3 Exploring and Comparing Artists' Lyrics

The nearest neighbor relationship can be used to compare the songs of two artists of interest on different levels. On a song-level, it shows all relations between two artists and on a line-level the exact nearest neighbor relations of two songs can be seen.

Design. We use stream graphs to visualize the nearest neighbor relations between songs. For this, the number of nearest neighbors is mapped to the saturation of the edge between two songs. To reduce visual clutter, the filter mechanism can be applied. A user can filter according to the number of nearest neighbors and the release date of the songs. The lyrics of the songs can be read by clicking on a streamline. Both song lyrics are placed side-by-side, while the nearest neighbors of each line are shown. This allows the user to read the lyrics side-by-side while investigating the alignments. Each alignment is visualized as a streamline that connects the lyrics. Furthermore, the user can filter the alignments based on a slider. The filter values correspond to the cosine

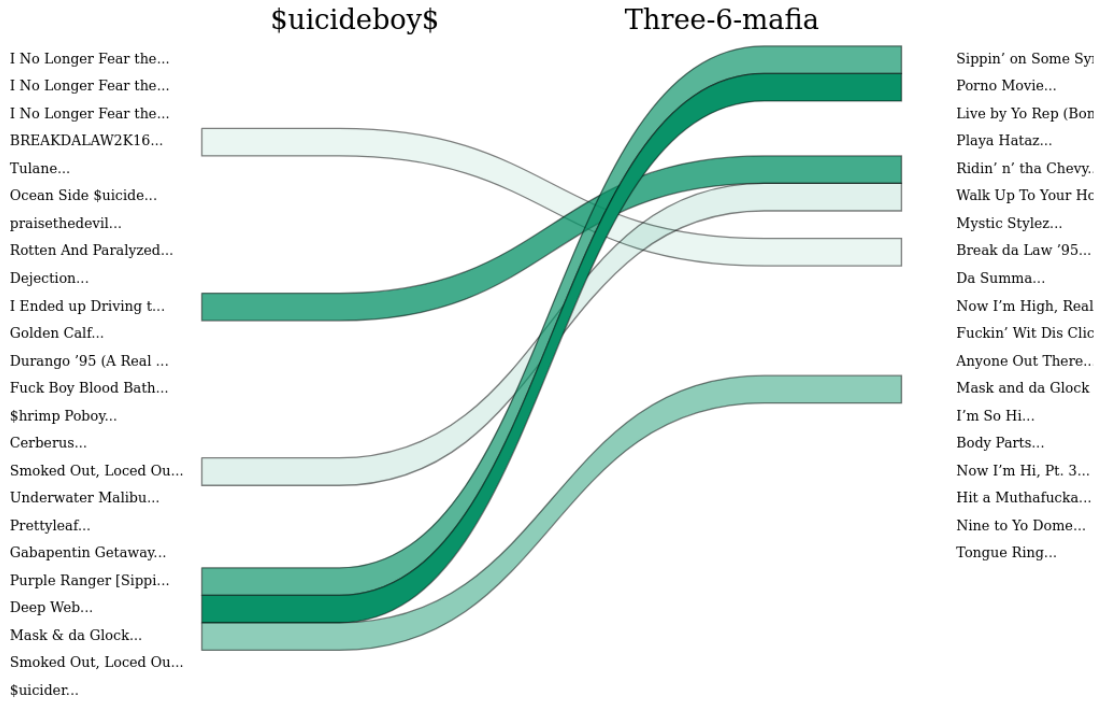


Figure 3.5: Excerpt of the song-level of *$uicideboy$* and *Three 6 Mafia*. All connected songs show cases where the *$uicideboy$* reused lines from *Three 6 Mafia*.

similarity between the lines in the alignment. This allows to further investigate the nearest neighbors of two songs of interest. When clicking on a streamline of interest, the alignment is visualized as a variant graph using TraViz [JGF+15]. Furthermore, all lines that are the nearest neighbors of both lines are shown with TraViz, as seen in Figure 3.7. This highlights words that have been reused or shared between lines. These nearest neighbors can be used to move to another pair of songs of interest where the alignment occurred.

Use Case. Figure 3.5 shows an example of the similarities between songs of two artists. In this case, the songs of the *$uicideboy$* and *Three 6 Mafia* are displayed side-by-side. Some of these pairs can be found in Figure 3.4a and are part of a lawsuit filed by *Three 6 Mafia* against *$uicideboy$* [Dar20]. For these songs, samples or lines that are part of the hook were reused. Examples of monolingual alignments of two songs can be seen in Figure 3.6. Figure 3.6a shows *Kool Savas - Komm mit mir* and *Alligatoah - Komm mit uns* where *Alligatoah* parodies the original song by *Kool Savas*. The excerpt in Figure 3.6b shows *Sido - Du bist Scheiße* and *Tic Tac Toe - Ich find dich scheiße* where the song by *Sido* sampled the one by *Tic Tac Toe*.

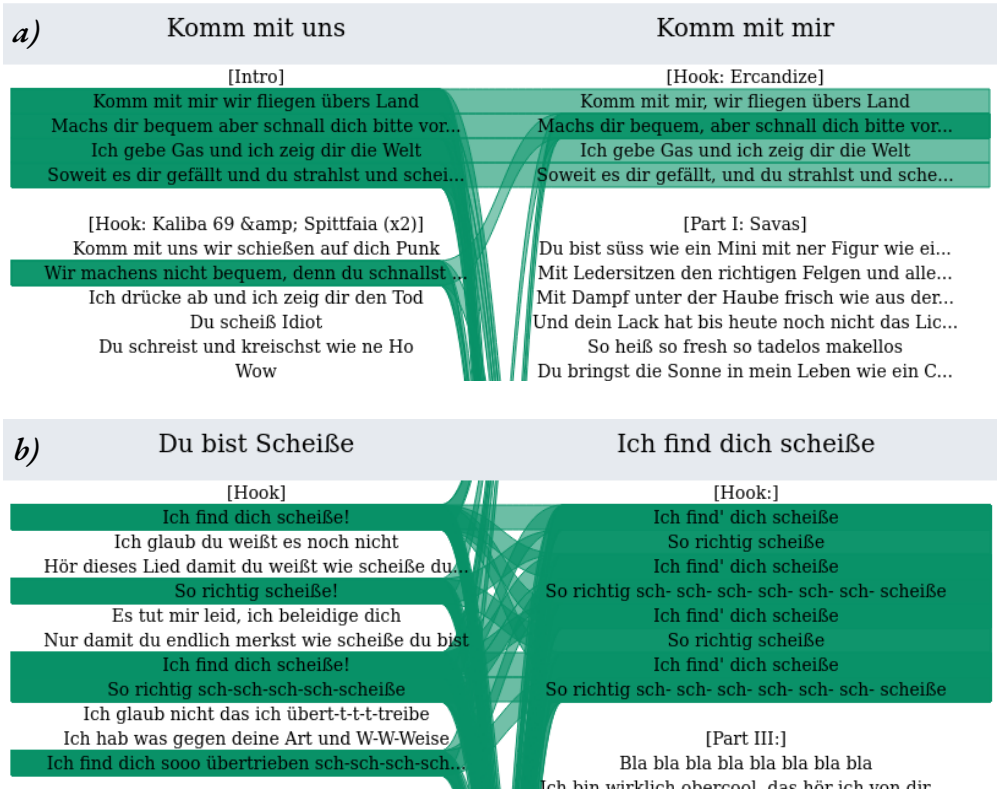


Figure 3.6: Excerpt of multiple monolingual alignments on the line-level (a) *Kool Savas - Komm mit mir* and *Alligatoah - Komm mit uns* and (b) *Sido - Du bist Scheiße* and *Tic Tac Toe - Ich find dich scheiße* .

3.3.4 Compare temporal distributions of genre

Rap music can be divided into multiple different subgenres. In order to see the development of these subgenres, we display them on a timeline based on the annotations of the Genius Expertise data set [LB21]. First, we filtered the genre tags to exclude non-rap tags and non-English tags, resulting in around 40 genres.

Design. For each genre, we computed a boxplot representation with the lower and upper whisker, the lower and upper quartile, and the median of the release dates of the songs annotated for a genre. The oldest release date of a genre is used as the endpoint of the lower whisker, whereas the newest release date is used as the endpoint of the upper whisker. The median is encoded as a colored circle. The range between the lower whisker and the lower quartile shows where the first 25% of the release dates are located, the range between the lower quartile and the median is the next 25%. The same applies to the range between the median and upper quartile and the upper whisker and

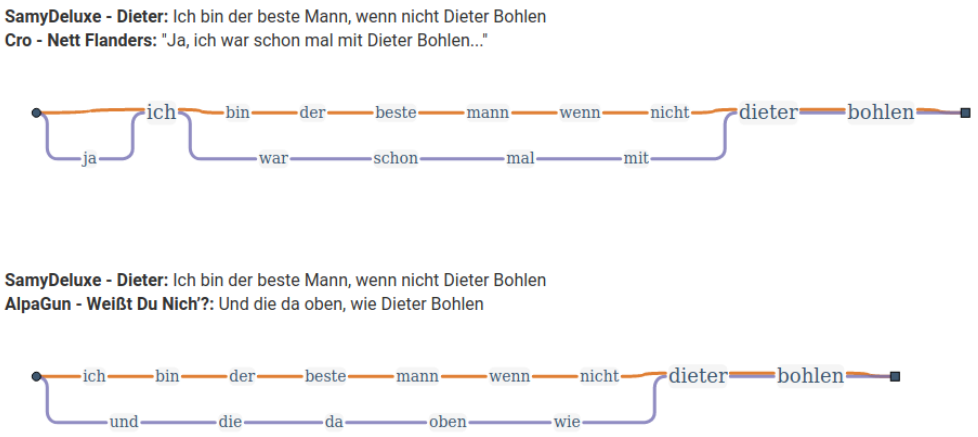


Figure 3.7: Two of the nearest neighbors of a line by *Samy Deluxe* displayed with TraViz.

upper quartile. In the visualization, we sorted the genre tags by the earliest release date of a song. We use colors to better differentiate between genres, but they do not convey similarity between genres.

Use Case. For example, we can see in Figure 3.8 that for the *Dipset* genre, the lower whisker and the lower quartile, and the upper whisker and the upper quartile are the same. Furthermore, the distance between the lower quartile and the median (2004 - 2006) is less than the distance between the median and the third upper quartile (2006 - 2016), showing that 50% of the songs were released in a short time period between 2004 and 2006. The *Dipset* movement goes back to *The Diplomates*, a hip-hop group that released their first studio album in 2003, which influenced many international artists in the following years. Furthermore, for many genres, the distance between the lower whisker and the lower quartile is significantly greater than the distance between the upper quartile and the upper whisker. This suggests that the data set contains more newer songs than older ones, possibly due to missing older songs and/or an increase in the publication of rap songs in recent years.

3.3.5 Compare Vocabulary

In order to compare the vocabulary of the artists, we allow a user to select multiple artists to visualize the vocabulary in a TagPie [JBR+18]. Before visualization, stopwords are removed from the vocabulary, and all words are lemmatized.

Design. The word font size is assigned on a logarithmic scale so that the smallest value is assigned to 10 and the largest value is assigned to 50. Furthermore, the user can change the number of tags that are shown, and the measurement that is used for

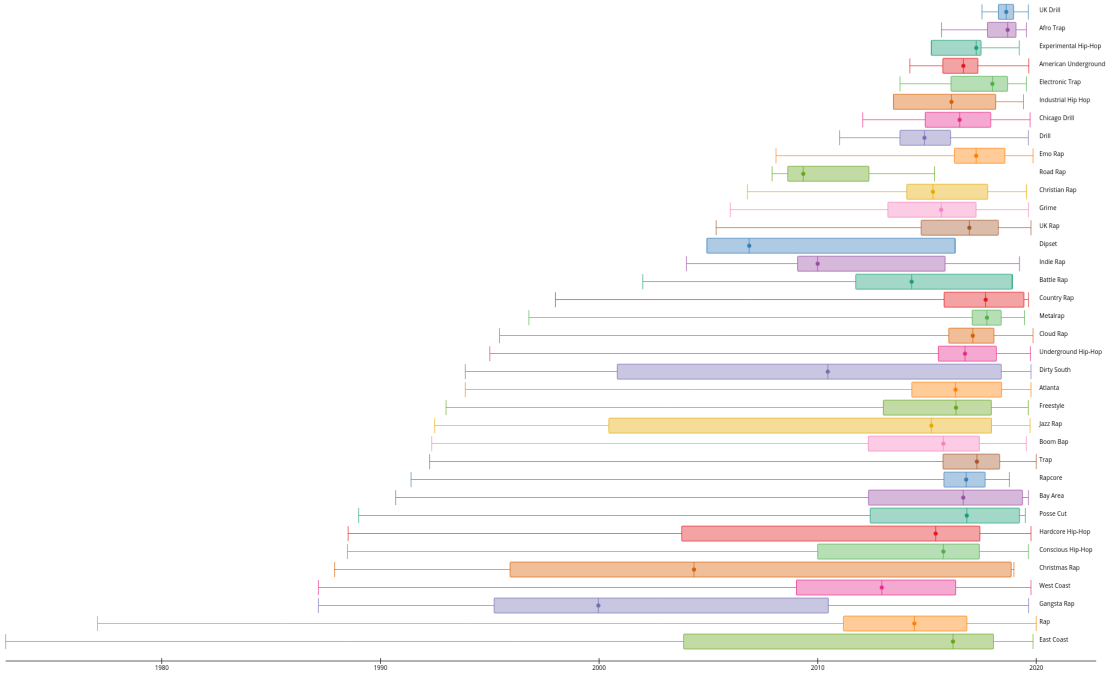


Figure 3.8: A timeline visualizing all rap associated genres in the Genius Expertise data set. Each genre is shown as a boxplot.

the font size in the visualization. The measurement for a word w and an artist a can be either $f_w(a)$, $y_w(a)$ or the z-Score i.e. $z_w(a)$

$$y_w(a) = f_w(a) - \sum_{a_i \in A} f_w(a_i)$$

$$z_w(a) = \frac{f_w(a) - \mu_w}{\sigma_w}$$

$f_w(a)$ is the number of times a word w is used by an artist a. $y_w(a)$ is defined by subtracting the number of occurrences by all artists $a_i \in A$ from $f_w(a)$ and the z-Score denotes the number of standard deviations $f_w(a)$ is below or above the mean value of the word w throughout the corpus. While $f_w(a)$ only shows the number of occurrences, $y_w(a)$ can be used to highlight words that are unique to an artist in the corpus or rarely used by other artists. Similarly, the z-Score allows one to detect words that are common to a group of artists, but more rarely used by others. An example for $f_w(a)$ and the z-Score can be seen in Figure 3.9.

Use Case. Taking only high-frequency words without normalization as in Figure 3.9a, we see many generic words that are often used in old-school hip-hop tracks. These do

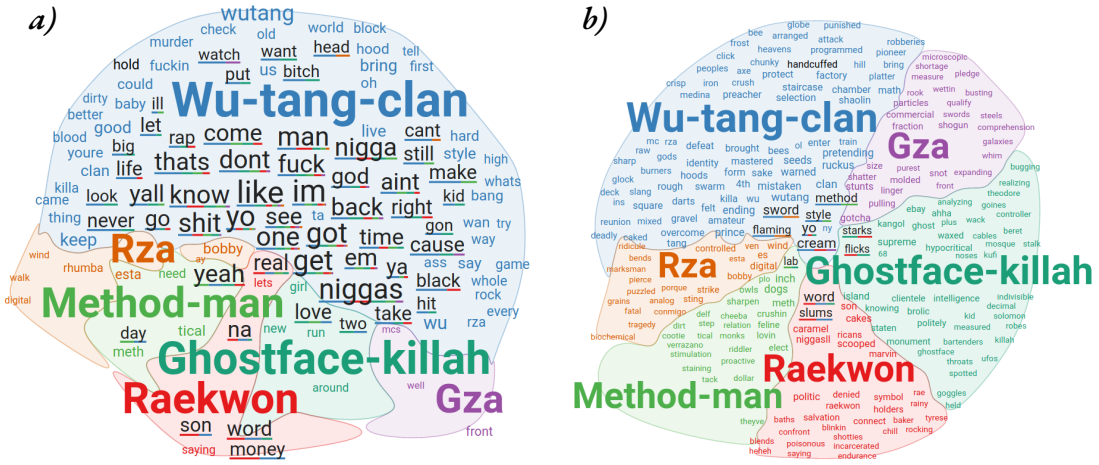


Figure 3.9: TagPie showing the most frequent used words by the *Wu Tang Clan* and various members (a). TagPie (b) shows words by z-Score e.g. words that are more frequent to the rest of the corpus.

not show the wordiness of individual artists of the *Wu-Tang Clan*. Using the z-Score in Figure 3.9b, results in getting words that are more descriptive for the individual artists. For example, *cream* (C.R.E.A.M. - Cash Rules Everything Around Me) is the most streamed and best-recognized song by the group. *Starks* stands for Tony Stark, which describes *Ghostface Killah's* alter ego. The words *sword*, *flaming*, and *style* are all related to their debut album *Enter the Wu-Tang (36-Chamber)* which has a Shaolin theme.

3.3.6 Sentiment Analysis

Another facet of textual data is the sentiment it conveys. To communicate the sentiment, we computed a sentiment score for each line in the corpus between 1 (negative) and 5 (positive). For this, we use *Huggingface* [WDS+20] and the BERT model of *NLPTown* [Tow20] for Multilingual Sentiment Analysis. With the sentiment score for each line, we computed an average sentiment score for each song and each artist in the corpus.

Design. In the visualization system, a user can view a list of German or English artists ordered by sentiment score in ascending or descending order. Next to each artist is a colored rectangle based on the average score. When clicking on an artist of interest, an ordered list of the artists' songs is displayed. For each song, a colored rectangle shows the average value of the song and a colored barcode shows the sentiment throughout the song for each line. Sentiment scores are mapped on a divergent color scale between red (negative) and blue (positive), with white as the neutral value.

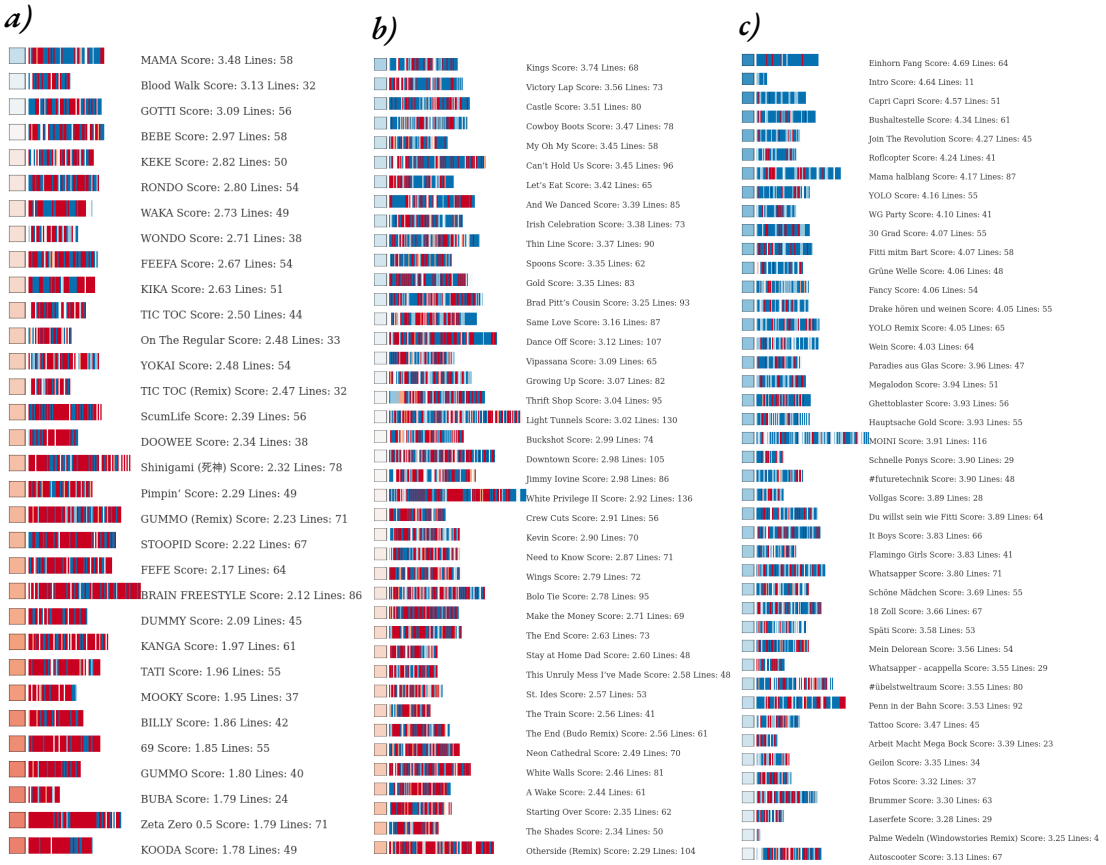


Figure 3.10: Sentiment Barcodes for the songs of *6ix9ine* (a), *Macklemore & Ryan Lewis* (b), and *MCFitti* (c). A red bar indicates a negative sentiment and blue a positive sentiment for a line.

Use Case. Figure 3.10 shows the songs of the American rapper *6ix9ine* (a), *Macklemore & Ryan Lewis* (b), and the German rapper *MCFitti* (c) ordered by average sentiment. *MCFitti* songs have a high sentiment on average, reflecting his more cheerful music, most of *6ix9ine* songs have a low sentiment on average reflecting his aggressive style of music, while the songs of *Macklemore & Ryan Lewis* range from positive to negative sentiment, showing diversity from party songs like *Can't Hold Us* and *And We Danced* and more serious themes such as drug addiction in *Otherside* and "Black Lives Matters" as well as white privilege in *White Privilege II*.

3.4 Second Visual Interface

The second prototype is designed for users of the general public interested in rap music, with a stronger focus on the text reuse aspect and less on the other facets of the

data set. The aim is to offer a tool that supports an exploratory analysis of selected American rap musicians and their lyrics. Therefore, the following three levels of tasks (with corresponding sub-tasks) were developed:

1. **Analyze Artists:**
    1.1. **Find similar artists:** As someone generally interested in hip-hop, a user could reasonably want to discover artists similar to those they are already familiar with or even fond of.
    1.2. **Explore an artist:** Knowledge about the artist's background may give the user context for similarities between artists or potential references.
    1.3. **Compare different artists:** A user familiar with hip-hop might want to explore groups or pairs of artists they already consider similar by listening to their music and inferring which songs and lines are the closest thematically. By doing this, it is possible to find artists directly referencing each other. On top of that, looking at artists that emerged in the same time period, the user could discover certain trend words or phrases from that time period and even if the meaning of the phrase or word has evolved over time.
2. **Analyze Songs:**
    2.1. **Find similar songs:** Users may also be interested in finding songs that are lyrically similar to their favorite song or a song of interest.
    2.2. **Explore a song:** A user with prior knowledge of influential artists and songs could explore the influence of those songs by searching for other songs that reference specific lines. However, commonly used phrases could be traced back to their origins within hip-hop.
    2.3. **Compare different songs:** When a user finds a song of interest, they could be interested in comparing the song with other songs by different artists.
3. **Find similar lines:** A user could be interested in finding lines that are similar to a line of interest and also finding all occurrences of a line throughout the whole song corpus. Thus, looking at different contexts a line has been used in, possibly even discovering whether that context and therefore the meaning has evolved over time.

The described tasks follow a granular order, from artists to songs to lines. Thereby, each task and its sub-tasks can serve as an entry point for each other, but can also be treated separately in the tool.

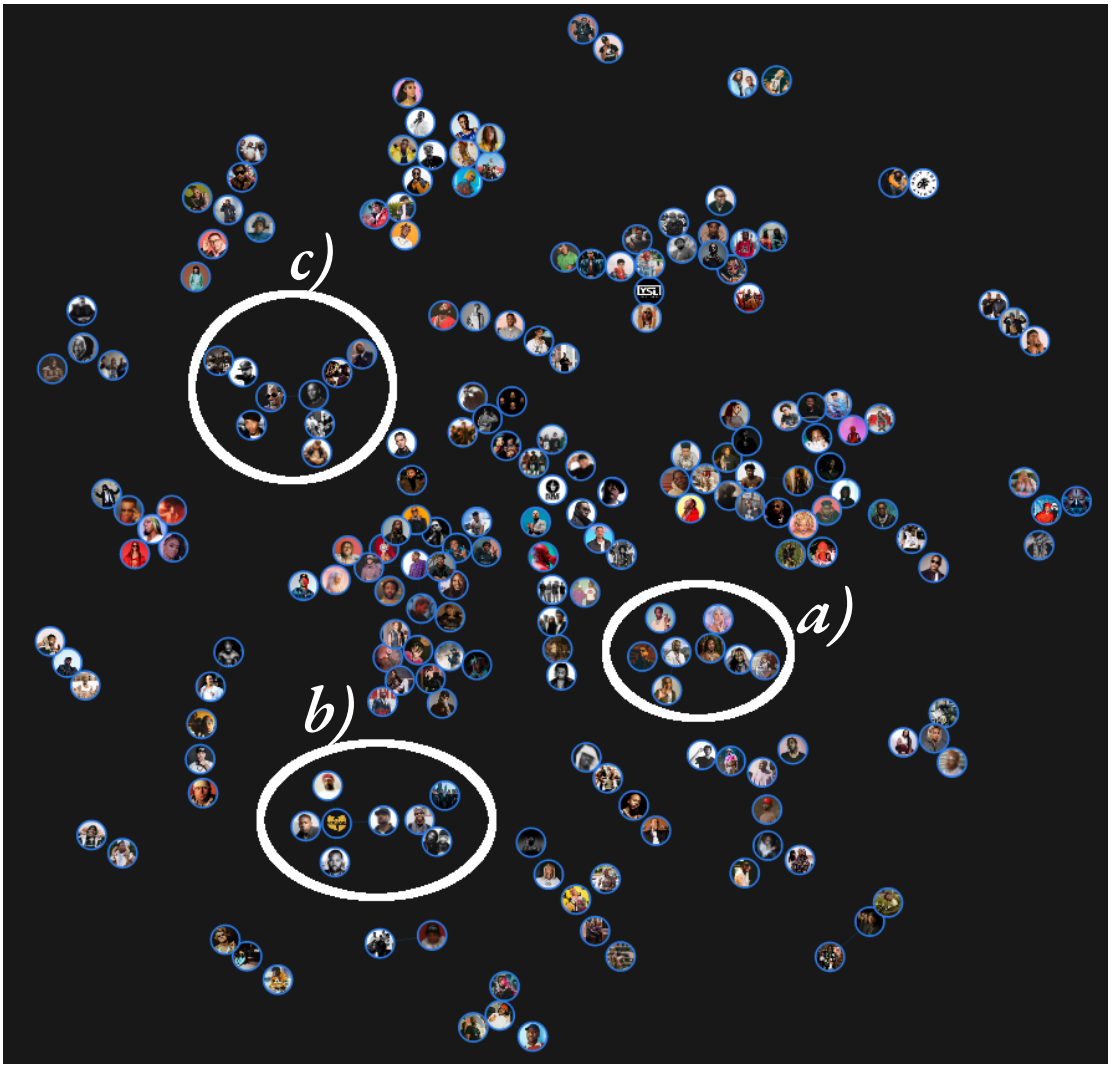


Figure 3.11: The artist graph of the second prototype, artists that are similar based on their lyrics are connected. Different kinds of clusters can be observed. (a) Shows one subgraph with a cluster containing Atlanta-based rappers *Offset*, *Quavo* and *Take-off*. (b) Shows a subgraph containing the artists *Raekwon*, *Ghostface Killah*, *Method Man*, *Redman*, and *GZA*, all part of the *Wu-Tang Clan*. (c) Shows *N.W.A* members *Dr. Dre* and *Ice Cube* together with several artists connected to them.

3.4.1 Artist Graph

Similarly to the first prototype, access to the data set is given through a force-directed graph layout.

Design. For the graph layout in the second system, the library "vis.js" [§] is applied, which uses the Kamada Kawai algorithm [KK89] for the initial layout and the Force Atlas 2 algorithm by Jacomy et al. [JVHB14] for the final layout. Each artist is represented by a circle that contains an image of the artist. An edge between two artists indicates that they are the most similar based on their lyrics. This leads to the formation of subgraphs consisting of lyrically related artists. Additionally, the length of an edge represents the value of the similarity score. Through this, denser clusters within those subgraphs manifest, indicating even more closely connected artists. The connections between the artists within the subgraphs and the spatial proximity of the artists within the clusters help the user quickly identify groups of similar artists. With this baseline of information, the user can then explore the lyrical connections of artists within these groups (Task 1. & 2.).

Use Case. For the second prototype Figure 3.11a shows a cluster containing Atlanta-based rappers *Offset*, *Quavo*, and *Take-off*. As the graph shows, these three are quite closely related lyrically. This makes sense because they are also related in the literal sense and form the rap trio known as *The Migos*. We can also see a close connection between *Offset* and *Cardi B*, who are married in real life and therefore regularly feature in each other's songs. Three of the other rappers in this subgraph are also based or at least born in Atlanta. Similarly, Figure 3.11b shows a subgraph containing the artists *Raekwon*, *Ghostface Killah*, *Method Man*, *Redman*, and *GZA*, all part of the *Wu-Tang Clan*, which is also part of the subgraph. The additional artists featured in the subgraph; *Cypress Hill* and *Heltah Skeltah*, also emerged in the same time period as the *Wu-Tang Clan*, around 1990. Furthermore, apart from *Cypress Hill*, they all come from New York, influencing and being influenced by the 1990s era East Coast Hip Hop. Figure 3.11c shows *N.W.A* members *Dr. Dre* and *Ice Cube* together with several artists connected to them. Including *Snoop Dog* and *Warren G*, two artists who collaborated with *Dr. Dre*, and groups where *Ice Cube* was a member.

3.4.2 Artist View

In the second system, information about the artists and a list of the songs can be accessed by double clicking on their image in the graph. This opens a pop-up artist view

[§] https://visjs.org

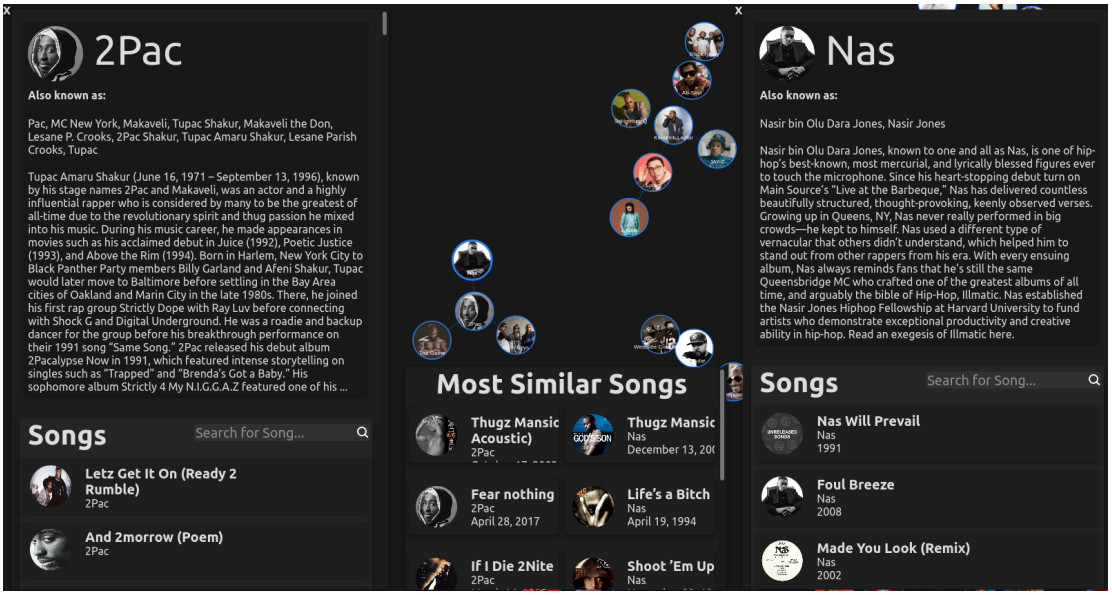


Figure 3.12: The artist view of *2Pac* and *Nas* shows biographical information, a list of songs, and the most similar songs from both artists.

that contains information about the artist and a list of their songs, which supports task 1.2. ("Explore an artist").

Design. When selecting the first artist, the corresponding artist view opens on the left side. Selecting an additional artist will open a second artist view on the right side. As both of these artist views are shown together with the artist graph, the user never loses context, as they can still see the area of the graph they were exploring. Any subsequent selections of an artist will change the right-hand artist view to display the newly selected artist, while the left-hand artist view stays the same. At the top of the artist view, the user can find a short text about the artist that was collected from Genius along with the other data. These descriptions offer knowledge about the artist's background, giving the user context for similarities between artists or potential references.

Use Case. An example of the second system showing the rappers *2Pac* and *Nas* can be seen in Figure 3.12. The most similar songs include *2Pac and Ice Cube - Fear Nothing* and *Nas and AZ - Life's a Bitch*. Examining both songs shows that the former song reused the chorus of the latter song.

3.4.3 Exploring and Comparing Artists' Lyrics

In the second system, opening two artist views offers the first direct comparison method (also gives an entry point to Task 1.3. "Compare different artists"). An additional pop-up will appear in the middle between the two artist views, showing pairs of the artists' most similar songs. This also allows to support Task 2.1. ("Find similar songs"). Selecting one of these pairs will open a song view, in place of their corresponding artist view. If the user wants to compare two specific songs (Task 2.3.), the artist view also allows them to search and select any of the songs of the corresponding artists in the database.

Design. Whenever two song views are open at the same time, all their textual alignments are shown in a side-by-side view. All pairs of similar lines are marked and connected by colored streamlines (Task 3.). This visualization can be thought of as a graph in which song lines are vertices with edges connecting them to similar song lines. A group of lines that are all similar to each other form a connected subgraph. Each of these connected components has its own color, so the user can easily distinguish between the groups. The colors are equidistant in regard to their hue, but the same in saturation and brightness. This color scheme was chosen to highlight that the groups are qualitatively different.

Each song remains individually scrollable, so different parts of both songs can be compared and explored (Task 2.2. & 2.3.). Each song view offers additional options to explore the data. By clicking on the artist name at the top of the song view, the user can go back to the artist view to compare another of the artists' songs with the one still open on the other side. It is also possible to compare not only two artists, but also to use one specific song as a starting point to traverse the data (Task 2.2.). If the user wants to find references to a song, opening only one song view makes a list of similar songs appear in the middle of the screen. To be even more specific, each line of a song view is clickable. Selecting one line opens a list of all similar lines from other songs on the opposite side of the screen (Task 3.). This enables the user to explore the usage of certain phrases between different artists and possibly trace who is referencing whom. Additionally, a Text Variant Graph is provided that helps to compare all similar lines and also supports task 3. ("Find similar lines"). Having found a particularly interesting line similar to the one originally selected, the user has the option to click on it in the list. Thus, the list of similar lines is replaced by the song view corresponding to the clicked on song line, once again enabling the comparison of the two songs. Additionally, a user can perform a full-text search for a specific song name or occurrences of a specific line.

Use Case. For the second system, Figure 3.13 shows the previously mentioned example in which one song reused the chorus of another song. Figure 3.14 shows a search

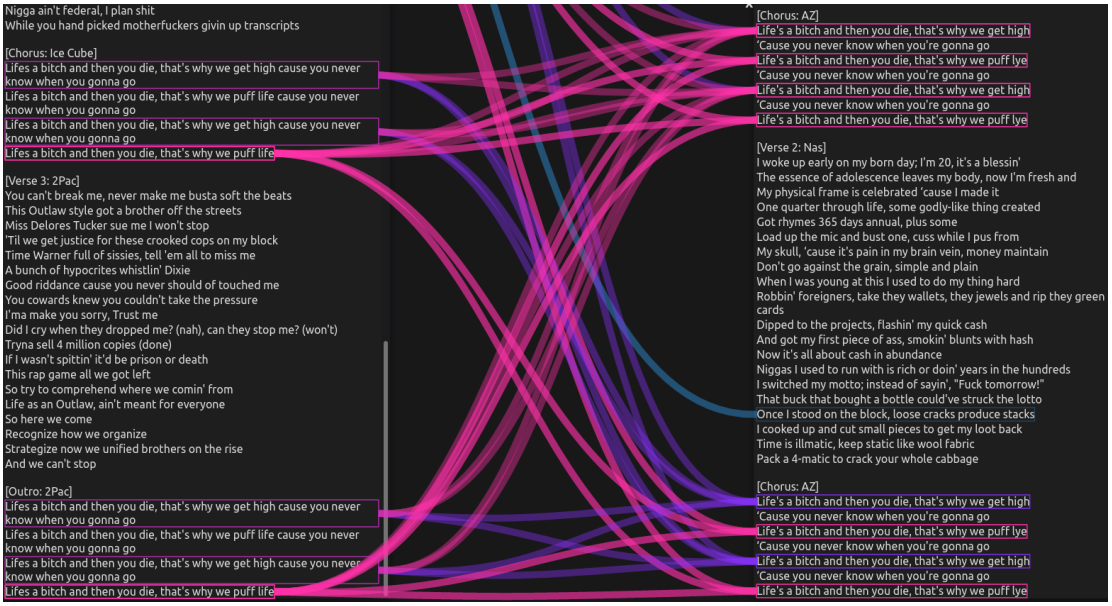


Figure 3.13: Side-by-side view of the songs *2Pac and Ice Cube - Fear Nothing* and *Nas & AZ - Life's a Bitch*. The former song reused the chorus of the latter song. Each group of lines that are similar to each other is assigned a unique color so that the user can easily distinguish them.

example for the line *Each one, teach one*, which is an African-American proverb that originated in the United States during slavery when black people were denied education. When someone learned to read or write, it became their responsibility to teach someone else. The phrase is still used today in many songs.

3.5 User Feedback

We conducted an informal evaluation with six fans of rap music who have general and scene-specific knowledge of the German and US rap scene. They used the system for approximately half an hour to one hour to explore the graph and the relations between the artists and the lyrics. One user suggested adding filtering by year for the song-level side-by-side view to focus on specific parts of the artist's career, e.g., when two artists were part of the same group or if only early works or new works are similar. For example, he noticed a greater similarity in the lyrics of *Tony D* and *Sido* when they were both part of the rap group *Die Sekte*. A user noted that the list of similar songs in the profile view is helpful for detecting songs about the same or similar topics, for example, love, cars, or drugs. Multiple users noted that the TagPies created by z-Score help confirm hypotheses about the vocabulary of two or more artists. For example, a

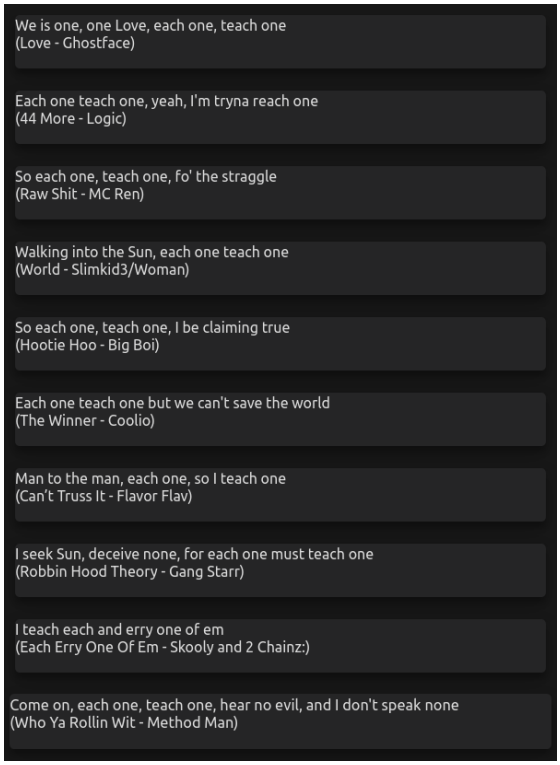


Figure 3.14: Top search results for the phrase *Each one, teach one*.

user thought that the vocabulary of *Flatbush Zombies* and *The Underachivers* is similar, which he then confirmed with the visualization. Users also noted that the relations in the graph make sense as long as the similarity value does not decrease too much.

3.6 Discussion

3.6.1 Imprecision & Incompleteness

A limitation of our approach is the data itself, as the data from Genius.com comprise different facets of inconsistencies [KKM+20] i.e. imprecision and incompleteness. Although Genius.com always had a strong focus on rap music, there are probably always songs or artists that are not included, and therefore, resulting in an incomplete data set. Furthermore, missing metadata information about artists or songs also leads to incompleteness. To increase the knowledge base, other sources of information could be crawled and linked to the data. The data is also imprecise for several reasons. Some "songs" collected through the Genius API were, in fact, not songs. Such oddities in-

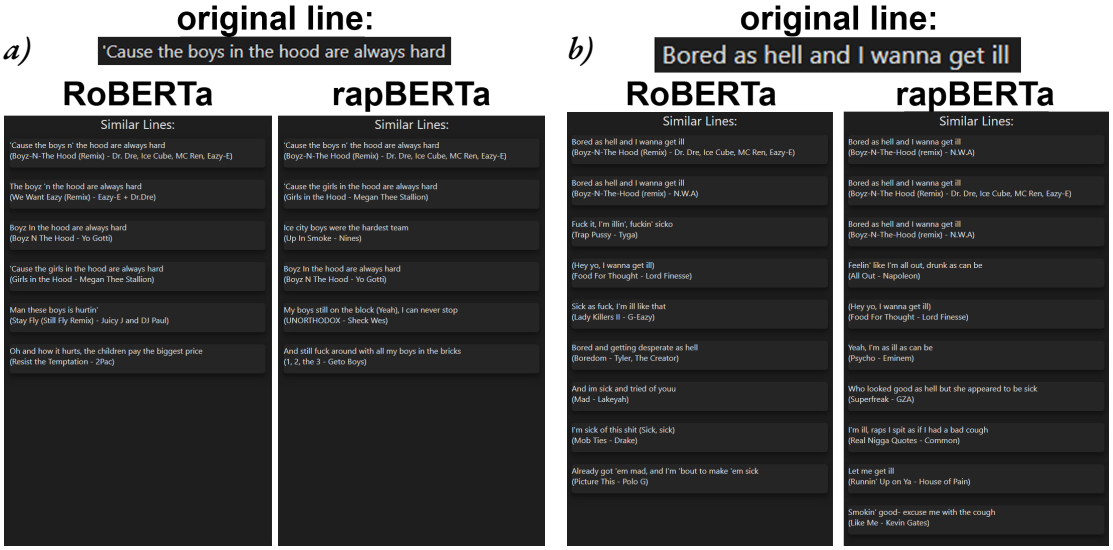


Figure 3.15: The most similar lines according to the two models in the second system for the lines (a) "Cause the boys in the hood are always hard" and (b) "Bored as hell and I wanna get ill".

clude body maps of artists' tattoos and recipes. For example, in 2013, rapper 2Chainz released a cookbook, which is listed on Genius as his album. Thus, some recipes from his book exist in the Genius database and are treated just like regular songs. All of this is a byproduct of the crowd-sourced nature of Genius. Another inconsistency is given by music genres, which are ambiguous terms without a clear starting point. Genre definitions or the association of a song or artist with a genre can change over time, as new genres emerge. The visualization of the temporal information of the genre tags can give an overview of the different types of rap genres and also about new emerging genres, but it is not precise.

Imprecision is also given by machine learning methods such as sentiment analysis and word and sentence embeddings. These methods are biased based on the data on which they were trained. Another imprecision is given by the artist's similarity. The use of cosine similarity and also the inclusion of a fixed number of nearest neighbors influence the text alignments. Currently, the alignments are often occurring because of the use of the same proper names, like the artist's names or cities, and the usage of the same adlibs. Therefore, including a threshold and other metrics could be helpful. Unfortunately, the Genius.com data set has no ground truth, so it is not possible to evaluate the quality of the alignments and the similarity metric. Nevertheless, the exploratory analysis with the visualizations allows for an informal evaluation through domain knowledge about artists' relations.

Also, alignments often occur between the hook or the refrain of two songs, so for future works, it would be better to treat them differently to focus more on less clear similarities. Exploring the data has also made clear that the lyrical nature of rap sometimes poses a problem for the models' understanding of song lines, especially references created through metaphors or rhyme structures. While handling lines that contain words found in the dictionary well, text passages that make use of neologisms and slang are prone to misinterpretation by the models. Even including word vectors trained on the Urban Dictionary does not tackle this issue in the first prototype. For the second prototype, it can be observed that RoBERTa often succeeds in finding lines with meaning similar to those of the first few nearest neighbors. However, not all neighbors always match well. Additionally, it seems that rapBERTa has fewer problems in understanding words that are not meant literally based on their context. Examples can be seen in Figure 3.15. Figure 3.15a, shows how *hood*, *block* and *bricks* can be used interchangeable and in Figure 3.15b, *ill* means *drunk* which becomes apparent when reading the lines. Based on the similar lines found by RoBERTa and rapBERTa, it appears that rapBERTa has at least partially learned this meaning, while RoBERTa only knows the literal meaning of the word. However, it still finds semantic similarities where there are none. In many cases, this can be attributed to the fact that the surrounding lines must be taken into account to understand the meaning of one line. Furthermore, considering the amount of data on which the standard RoBERTa model is trained to achieve high scores on the Semantic Textual Similarity Benchmark [CDA+17], the corpus on which rapBERTa was fine-tuned is very small.

3.6.2 Future Works

For future works, training or fine-tuning a model on a much larger corpus of rap lyrics may produce better results in dealing with specific slang, neologisms, and pop culture references utilized in rap music. Moreover, employing an approach where the context that the model can use to learn the meaning of words is not limited to the one line that contains the word, but expanded to its surrounding lines could improve the performance as well. To improve the performance on the task of detecting similar lines, a manually assembled data set of similar and dissimilar lines could be used to fine-tune the model. This could be supported by visual interactive labeling or active learning [BHZ+17] for example, in a crowd-sourced environment. Despite their shortcomings in regard to the domain-specific language of rap, the data generated by both models often found similar artists who share real-life connections, pointing to the solidity of the approach. Building on the prototype and taking advantage of the expandability of the graph database, the application could be expanded to include a much larger

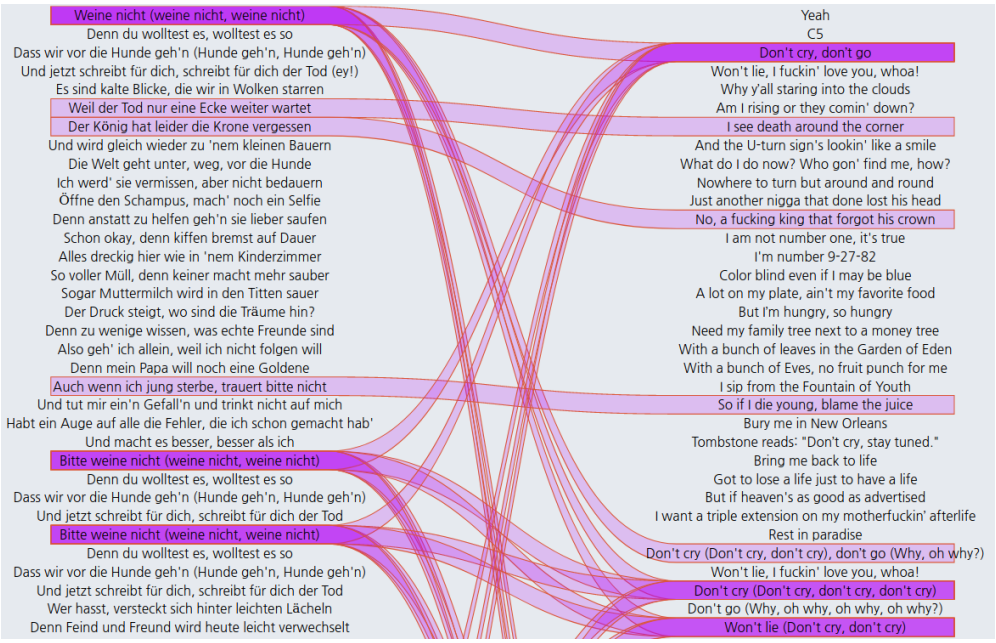


Figure 3.16: Side-by-side view of *Kontra K - Weine nicht* and *Lil Wayne & XXXTENTACION - Don't cry*. The former sampled the melody of the hook of the latter and also translated some of the lines as a homage.

number of artists from different genres. Thus, users would be able to discover more artists, especially with the inclusion of lesser-known ones.

It is also possible to extend this approach from monolingual to cross-lingual lyrics to detect cases where, for example, German artists reused passages from American artists. For this, as a proof of concept, we used the lyrics of around 20 international artists to find cross-lingual alignments between their lyrics and the lyrics of the German artists. We applied the LASER [AS19] model pre-trained for 93 different languages to create multilingual sentence embeddings. The LASER encoder maps similar sentences from different languages to similar vectors and can be used without any additional fine-tuning. An alignment in this case can be seen as a translation. We found some initial results in which the German artists communicated that they reused parts of English songs. An example of this can be seen in Figure 3.16. Furthermore, the approach is expandable to all genres of music and the entire Genius.com database with more than 12 million lyrics. Therefore, a possible future work would be to use more data to detect cross-lingual references and to compare the similarity of songs based on their lyrics on a large-scale through new Distant Reading methods. For example, with the visualization of alignments beyond the line-level, one can inspect multiple texts at the same time or cross-line connections.

To extend the similarity analysis, the combination of lyrics and sound features is of interest. Similarly to Yu et al. [YTRC19] sound features can be included next to the lyrics to create a multi-modal approach that includes similarities, for example, in mood, melody, tempo, or rhythm. For this, sampling information from crowd-sourced websites such as WhoSampled.com [Lim08] can be used to show more relationships between songs and artists. Another interesting approach would be to include other facets of cultural data, such as literature, to display the development of famous quotes or proverbs such as *Each one teach one* over time, cross music genres, and beyond lyrics. Additionally, a temporal visualization that includes historical events could provide insight into how these events impacted music.

The application of stylometry methods could be of interest in using frequencies of uncommon words such as the Burrows Delta [Bur02] to find lyrics that are unusual for a given artist and that are more similar to the lyrics of another artist, and thus can serve as an indicator for ghostwriting. In the past, such ghostwriters were often not communicated to the audience: "the silent pens might sign confidentiality clauses, appear obliquely in the liner notes, or discuss their participation freely" [Cam16].

3.7 Summary

We propose two prototypes to compute the similarities of rap artists and to find intertextuality between monolingual song lyrics based on word embeddings and transformer models. The analysis is supported by visualizations to explore similarities between the lyrics of rap artists. The investigation of the lyrics is further supported by different views showing the metadata from Genius.com and visualizing similar songs or lyrics through stream graphs to find similar songs and investigate monolingual alignments in their lyrics. Furthermore, we allow for a multifaceted exploratory analysis of the lyrics that includes the sentiment of the songs, the vocabulary of the artists, and the development of rap genres. Thus, supporting multiple visual text analysis tasks on the Genius data. We explain the current limitations of the systems that we observed through user studies. Furthermore, we outline possible directions to focus on, like finding cross-lingual alignments on a large corpus of song lyrics and cross-modality.

What is a true reading, if not an activity involving both the reader and the culture to which he belongs, and ... the author and his own universe.

Paul Zumthor

4

Explaining Semi-Supervised Text Alignment through Visualization

In contrast to contemporary textual data such as lyrics, text editions of cultural heritage pose new challenges for machine learning and visualization. A type of such cultural heritage text editions is medieval vernacular literary text, which often exists in multiple versions that are characterized by significant differences in length and structure. This textual instability is known as *mouvance* [Zum92] and takes on a wide variety of forms: differences in regional or scribal dialect, influences of an oral tradition, as well as the poetic modification of wording, rewriting, even omission or rearrangement of large parts of the text. These unique properties of literature pose a challenge when analyzing different versions of a text manually. The principle aim of the visual analysis of *mouvance* is to generate new perspectives that allow expert readers to draw conclusions on dependencies across the different versions and dialects of the language, as well as to track, compare, and assess the use of language, its meanings and time-dependent changes. The precondition for such an analysis is to find similar text fragments across different text versions, a technique known as *text alignment* [YJ20]. To perform the alignment of complex texts marked by *mouvance*, we determine the similarity at the level of lines (verses or sentences). Figure 4.1 illustrates an example of alignment of two

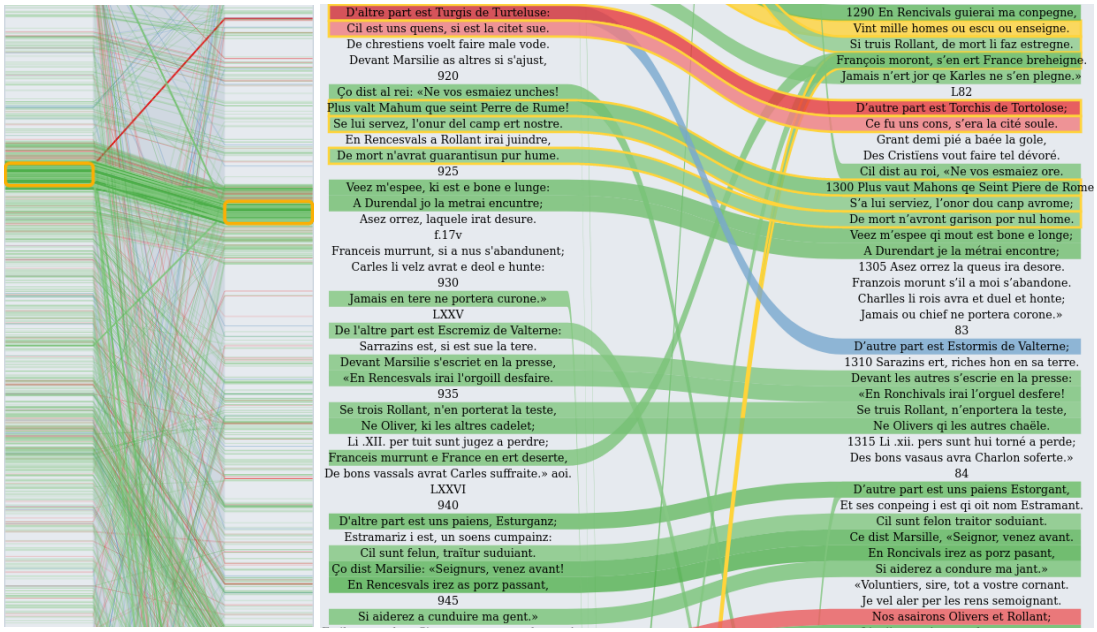


Figure 4.1: A barcode and a side-by-side view of two versions of the *Song of Roland* show different types of alignments.

versions of the medieval French epic poem, the *Song of Roland*. Colored streams connect lines of the two versions that share a certain degree of similarity.

Straightforward solutions to determine accurate text alignments do not exist for medieval vernacular literary texts. As a first solution, the visual analytics system *iteal* [JW17a] was proposed for interactive visual comparison of complex text versions to support professional reading [SLB+09], for researchers we call here "expert readers". This user-driven, parameter-based, white-box system—henceforth *iteal-V1*—uses string similarity and word n-grams in order to align and visualize different versions of a text. Although *iteal-V1* provides expert readers with a transparent framework for studying differences and similarities across different text versions, its major shortcoming lies in the neglect of semantic text features such as words with inflectional endings, synonyms at the word level, and stylistic features formed by the combination of words such as paraphrases or analogies. As taking into account such features that determine alignments across text versions is crucial to producing an optimal result, we replaced the white-box text alignment computation backend with an unsupervised word embedding method [MWJ19] to accommodate semantic alignments. This second version of *iteal*—henceforth *iteal-V2*—is a fully automated text alignment approach. However, since expert readers are typically not specialists in machine learning or advanced natural language processing, implementing such pipelines for domain-specific problems without providing a means

to understand or interact with the results can be problematic [AB18]. Furthermore, expert readers would like to be able to observe, evaluate, and critique such automated processes and are increasingly interested in peeking into computational black boxes to understand their assumptions and inner logic [Sam19, VZAA20].

With such a critical and interpretative expert reader in mind, we created a series of extensions in a version that we call *iteal-V3*. Both *iteal-V1* and *iteal-V2* are limited in that they do not incorporate the expert reader's domain knowledge into the calculation of textual alignments. Given the strict rules of *iteal-V1*, instead, we directed our attention toward generating a method to allow the expert reader to adjust the word embeddings, with the effect that the text alignment results also change in an iterative manner. We offer feedback mechanisms and novel manifold perspectives on alignment and word relations with a particular eye for their legibility. *iteal-V3* can be used to label line and word relations by exploring the neighborhood of lines and words in the vector space, while simultaneously providing insights on step-wise generated results. We needed to develop new visualizations in order to match semantically closer concepts with the word embeddings and to explain their behavior to the expert reader/collaborator, whom we refer to below as *DJW*. Over multiple iterations, the expert reader cannot only observe the changes in the alignments until a satisfactory result is obtained, but can also assess the impact of human input on two levels: the alignment of the poetry as well as the adjustment of the words in the vector space. Our system makes the argument that the alignment of complex poetic traditions is not a linear process, but rather an iterative one based on cooperation between the model and the expert reader.

Continuing our long-standing interdisciplinary collaboration [JW17b, JW17a, MWJ19], we adopted a participatory design process, proven to be valuable in the design of visualizations to be understood and used by domain experts [JKKS20]. In summary, the contributions of this process to the community are as follows.

- **Semi-automated Visual Text Alignment:** We provide an interactive, semi-automated text alignment approach, which combines visual analytics, word embeddings, and an iterative refinement process.

- **Visualizations for Word Transportation:** We designed visualizations to explain the computation of the Word Movers Distance [KSKW15], the distance measurement of our word vector approach.

- **Visualizations for Word Vector Neighborhood:** We introduce new visual means to observe, interact with, and manipulate word vectors. Manipulations of word vectors affect the alignment results in our system, which makes it important to

allow the expert reader to trace changes in both word vectors and their neighborhoods across different iterations.

- **Reflection on the Participatory Design Process:** We document our design process that includes iteration-dependent reflections on how the underlying word embedding and potential changes were perceived and what visual cues were required to better understand alignment computation.

4.1 Related Work

Our work combines three different lines of inquiry. First, we focus primarily on the visualization of text variations and text reuse on a line-level alignment, the different methods and application scenarios were already highlighted in Section 2.1. Second, we design multiple visualizations, which focus on the relation of the k-nearest neighbors of a word vector of interest. Third, through the interactions with the model, we engage with research focusing on active learning, and related works applying a human-in-the-loop scenario to include domain knowledge for textual analysis. The following subsections are dedicated to these aspects.

4.1.1 Mouvance & Critical Editions

Before the invention of printing, texts were copied by hand in manuscripts and the language in them bears the marks of elements of an oral culture. There exist multiple problems when a scribe would copy a work, for example, common writing errors such as orthographic errors or "eye skip" and missing words. More often, authors added or removed parts, exercising their poetic license, changing the meaning of passages, or simplifying parts of the text. Textual criticism is a specific application area of text reuse that deals with the comparison of the wide variation of such texts. Some scholars of textual criticism attempt to find or to construct an archetype of multiple text versions; others prefer to analyze the variance across the whole tradition as evidence of the text's reception in different contexts. For the construction of an archetype, Lachmann [Tim05] proposed a workflow. The first step is to gather as many editions as possible to create a corpus. The second step is the collation, i.e., the comparison of all editions to find differences between them. Then a family tree is constructed to detect the archetype. In the last step, an archetype is reconstructed considering the similarities and differences of the editions. For medieval text versions, especially for vernacular literature, Lachmann's archetype method can prove to be quite troublesome. Vernacular medieval literary texts are often authorless with uncertain dating, and the biggest

problem for textual criticism is the *mouvance* of these texts. *Mouvance* is a term introduced by the medievalist Zumthor [Zum92] and addresses the instability of medieval text variations that emerge through the above-mentioned elements of an oral vernacular culture. This can lead to changes in word order, as well as poetic modification of wording, rewriting, or even omission or rearrangement of large parts of the text. The problems of *mouvance* and the methods for dealing with them were also outlined by Jänicke and Wrisley [JW17b]. Regardless of the philological approach, our system is useful inasmuch as it allows a corpus of similar text versions to be explored, allowing a user to find similarities and differences between them.

4.1.2 Visualizing the Nearest Neighbors of Word Vectors

Modern natural language processing pipelines often apply dense word vectors as a representation of words. This shift from sparse one-hot encoding to dense word vectors was brought upon by Mikolov et al.'s word2vec [MCCD13, MYZ13, MSC+13]. Despite the wide application of word embedding models, only a few works visualize the vectors and their relations. Most of them apply dimensional reduction through PCA, t-SNE, or UMAP to project the vectors to a two- or three-dimensional space and then visualize them as a scatterplot like the Embedding Projector [STN+16], WebVectors [KK17], UTOPIAN [CLRP13], DataDebugger [XYX+19], or ConceptVector [PKL+17]. In contrast to these works, we do not primarily focus on dimensional reduction. Instead, we simplify nearest neighbor graphs [KYL+19, Kas18] by offering a one-dimensional representation of word neighborhoods to make the constitution of the vector space comprehensible to the expert reader. Similarly to most of the above-mentioned related works, we allow the inspection of the nearest neighbors of a word of interest, but we go beyond inspection, also allowing the original vector space to be changed through interaction. Changes in the neighborhood relation are further visualized after such interactions. Similar interactive methods based on word vectors were applied to interactively construct lexicon-based concepts [PKL+17] or to refine topic models [EAKC+19, CLRP13].

4.1.3 Human-in-the-Loop for Text Analysis

In recent years, various methods and concepts have been introduced to tackle the opacity of black box systems [ERT+17, JLC19] in order to give users of these systems ways to understand them, interact with them, and even critique their performance. The application of user interaction as feedback to a model is indicative of a human-in-the-loop process in which a model is iteratively refined. However, the question remains as to how users of a system can assess the stepwise refinement. A popular concept for model

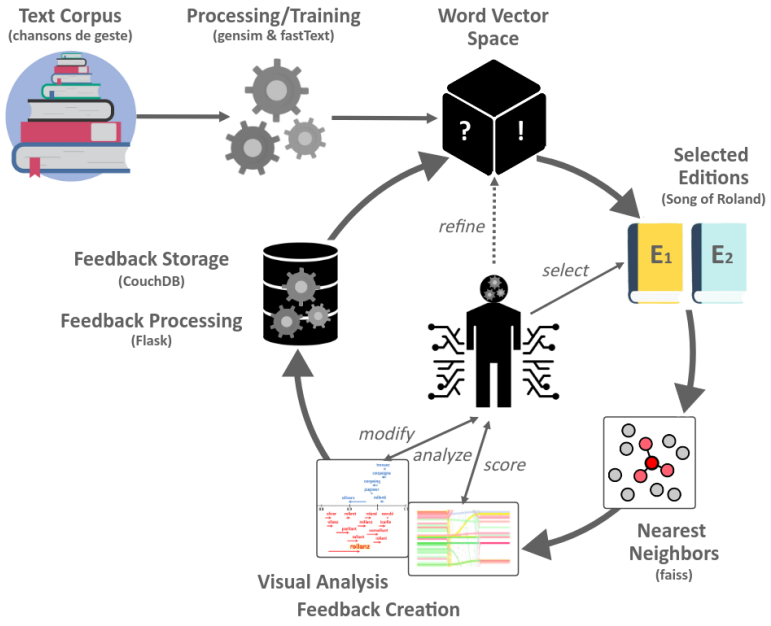


Figure 4.2: Our human-in-the-loop process applying word embeddings and visualization to perform textual alignments on low-resource and under-resourced languages.

refinement is active learning, which is applied when manual data labeling is impractical. The user labels data samples that are chosen by the system to maximize feedback and minimize labeling time. In some cases, this process is combined with interactive visualizations to better understand the classifier [KPSK17, HKBE12, SLK+19]. In contrast to these approaches, we let the user solely explore the poems side-by-side while labeling the relations between the lines and words in both poems, thus modifying the underlying word embeddings. These changes are then communicated through several visualizations.

4.2 Project Overview

The interdisciplinary collaboration of this project began in 2015. With a corpus of medieval poetry at hand, the goal was to develop a visual analytics framework capable of discovering aligned text fragments taking into account the expert reader's domain knowledge about the phenomenon of *mouvance* [JW17b]. In 2017, *itealV1* [JW17a] was published as a result, determining alignments based on string similarity and shared word n-grams. Although string similarity can disambiguate many of the medieval French words, the limitation of this approach is its inability to take into

account semantic features characteristic of vernacular, orally-influenced poetry, such as synonymic replacement, formulaic intertextuality, word reorganization, or significant orthographic difference. To address these problems, we proposed *iteal-V2* in 2019, an automated approach based on word embeddings [MWJ19].

This work extends the *iteal* portfolio with *iteal-V3*, introducing novel visual metaphors to communicate the shape of the vector space, the word neighborhoods, and the iterative changes introduced into the vector space. Following Munzner's guidelines for task abstraction [Mun14] the domain-specific tasks for all *iteal* versions are to *derive* alignments of lines in the poem that can then be *explored* by the expert reader, who then *identifies* alignments of interest and *annotates* them as true or false alignments. *iteal-V3*'s visualizations serve the need to understand and adjust the word embeddings used for alignment computation. What follows is a description of the text corpus and details of how the word embeddings are computed. The whole process is summarized in Figure 4.2. We tested our system for medieval French epic poetry, but the pipeline is applicable to other languages and generalizable to other corpora with a high degree of intertextuality [KMJ20].

Text Corpus and Pre-Processing. Our historical text corpus consists of multiple medieval French poetic works, in the epic genre known as *chansons de geste* along with some texts of the *Romance of Alexander* legend which share epic-like characteristics. The corpus varies in terms of language variety, epic cycle, and century and consists of around 30 different works. Our alignment here focuses on the oldest of the epic legends and arguably the most complex, the *Song of Roland*. The Roland tradition was chosen for its significant intertextuality and variance. For example, different versions of the *Song of Roland* can vary from 2000 to 8000 lines long. The shared narrative aspects of the Roland tradition across the different versions make the exercise of comparing them a compelling task. We focus on the alignment of texts taken from single-manuscript editions of the *Song of Roland*: the Oxford manuscript (approximately 4,000 lines) and the Venice 7 manuscript (approximately 8,000 lines). The manuscripts are written in major regional varieties of medieval French, and this variety adds another layer of complexity to the alignment. The entire corpus was cleaned from diacritics (editorial emendations not present in medieval language), unnecessary white spaces, and artifacts created through Optical Character Recognition.

Word Embeddings and Post-Processing. A pre-trained model for modern French was available, and so we began by carrying out an alignment of a structurally conservative modern translation of the Oxford version (Petit de Julleville) [PdJ78] with the original medieval text. *DJW* evaluated the results as arbitrary due to the wide gap between the medieval and modern French languages. Consequently, we had to use a model for twelfth-century French for which no solutions exist. Therefore, we trained

a model based on the text corpus described above using the *gensim fastText Skip-Gram* implementation [ŘS10], introduced in *iteal-V2*. We applied *fastText* [BGJM17], due to its ability to grasp orthographic variance of different dialects and word modifications over time, thus addressing the issue of highly variant spellings. An evaluation of the different approaches can be found in Section 4.4. Due to the small corpus size, we decided to use a 100-dimensional vector space to compute word embeddings. After the training phase, the word vectors were normalized and post-processed. Subsequent normalization ensured unit length and improved the quality of word vectors, since the vector length is known to correlate with the frequency of the word [TMdS19], that is, in our case, when dealing with rare orthographic variants, not important for the meaning of a word. For the post-processing step, the method of Mu and Viswanath [MBV17] is applied to eliminate the common mean vector and the top dominating direction of the word vectors. This leads to more uniformly distributed vectors, which can help to better express word similarities and further reduce the influence of word frequencies on the vectors [TMdS19].

Compute Alignment Candidates. When comparing two text versions, a sentence vector is computed for each line using unsupervised Smooth Inverse Frequency [Eth18]. With *faiss* [JDJ21] these sentence vectors are added to an index structure to query the nearest neighbors of each sentence based on cosine similarity. For each text version, an index is constructed, and the other version is queried. This process results in a list of potential alignment candidates for each line in both versions, thereby reducing the computation time for the following steps. For each candidate, two sentences X and Y, the Word Movers Distance (WMD) [KSKW15] is computed by solving an optimization problem to find the minimum cumulative Euclidean distance between the word vectors of the sentences, that is, the minimal cost required to transport the sentence X to the sentence Y. We denote the set of transportation pairs of two sentences X and Y as $T_{X,Y}$. A word transport pair is a tuple $(w_1, w_2) \in T_{X,Y} : w_1 \in X \wedge w_2 \in Y$, while a sentence X is a bag of words $X = \{w_0, ...w_n\}$. To give a better explanation of the Word Movers Distance (WMD), we added an example in Figure 4.3 with different sentence lengths and multiple occurrences of the same word. In the case of different sentence lengths, a word can be moved to multiple words or multiple words can be moved to the position of the same word. We applied the WMD because it performs well for the nearest neighbor classification [KSKW15]. Furthermore, the underlying metaphor of the transport of the word can be easily visualized and interpreted. The resulting list of nearest neighbors for each line is used as input for the visualization system.

Need for Refinement. A visual analytics system that facilitates the study of variant text traditions must address multiple usage scenarios as well as the means of visualization and interaction for a user-driven process of gaining insight. Such a process could

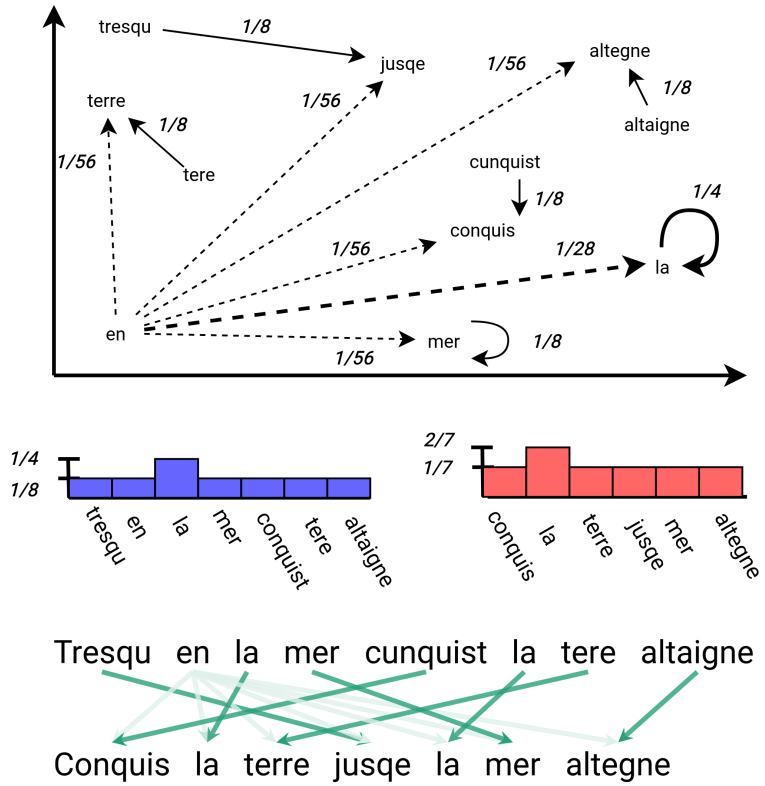


Figure 4.3: Word transportation when computing the WMD. The histograms summarize the word movements at the top, the blue bins are transported to the red ones. At the bottom, our visualization of the WMD can be seen.

include automatically detecting alignments, assessing the quality of these alignments, removing false positives, and adding new undetected alignments. When the alignment detection process was switched to a word embedding model in *iteal-V2*, new scenarios appeared for the expert reader, but the elimination of the parameter adjustment opportunity of *iteal-V1* made it difficult to interpret the results. In particular, it was not traceable why the system aligns specific lines drawing on the underlying word embedding model. Traceability had been granted in *iteal-V1* by the string similarity approach, but the change to a word embedding changed the workflow. Thus, for *iteal-V3* we conducted an iterative process that allowed user-led refinement of the model to proceed in iterations (stages). To communicate changes at the line and word level after each iteration, a variety of visualizations and interaction techniques were developed to evaluate line alignments and word vector relations, as well as to observe stage-dependent modifications.

Participatory Visualization Design. Bearing on the authors' experiences gained in a variety of interdisciplinary digital research, we followed a participatory visual design process to carry out this project [JKKS20]. This process is based on, but also extends, task-based development models [Mun09], as most design considerations and adaptations were debated in depth among all project members [Wri18]. Our stage-based visualization development led to vibrant reflections on required adjustments on the one hand, but, more importantly, to entirely new visual perspectives on data and alignments on the other.

4.3 Iterative Design of *iteal-V3*

Our participatory design process started with a prototype, which allowed *DJW* to compare the results generated by *iteal-V1* [JW17a] with those of *iteal-V2* [MWJ19]. The prototype offered an alignment view that allowed the introduction of user-generated input into the semi-automated system by labeling line-level alignments according to their reliability. In later stages, we added visualizations to explore the neighborhood of word vectors and to allow for word-level modifications to the vector space. After each session, user feedback on line and word level is used to adjust line-level alignments among the text versions. Changes to the word embeddings can be inspected by the expert reader in the subsequent session. *iteal-V3* can be also applied to other alignment scenarios of two text versions provided that both a list of potential alignment candidates and the embedding model used to compute them are available.

Through multiple iterations, we developed a series of visualizations to inspect stage-dependent alignment changes and word embedding features. In what follows, we describe the visual encoding and means of interaction that we designed to engage with complex questions in the human textual record and the workflows of the expert reader. An overview of the interface can be seen in Figure 4.4.

4.3.1 Alignment View

DJW wants to inspect the results of the alignment of the Oxford and Venice manuscript using the unsupervised word embedding method of iteal-V2. In the beginning, he sees the barcode view showing a zoomed-out version of the poem and the alignments, which he can use to jump to a specific area of interest in the poem. Next to it, he can see the side-by-side view, which allows him to read the editions while exploring alignments. For each line in the Oxford manuscript, the first nearest neighbor in the Venice manuscript is used for the alignment. Currently, the first stage is selected, which is the model after training and without user interactions. DJW can later switch to a higher stage to see the influence of his interactions with the

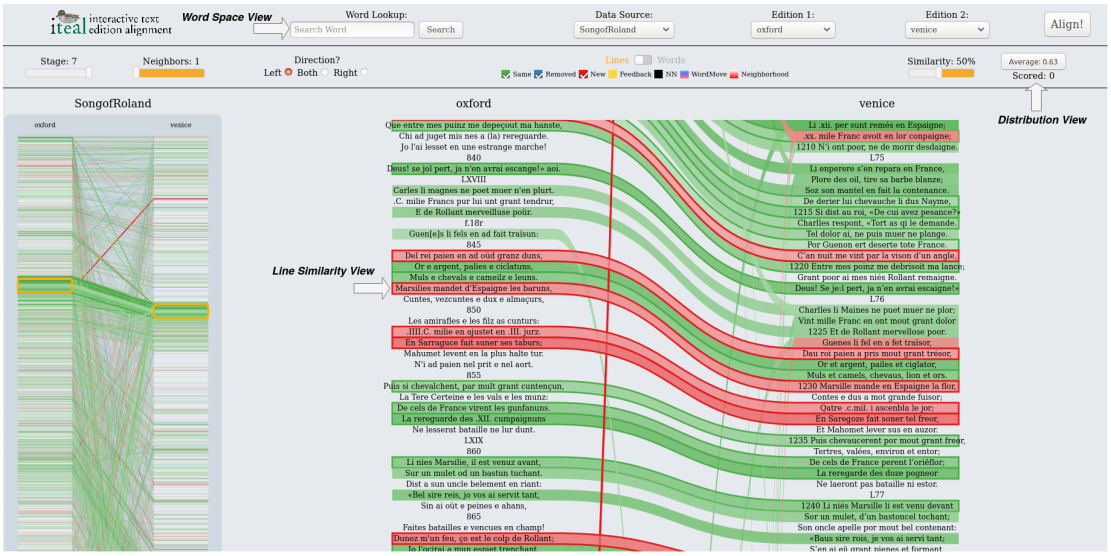


Figure 4.4: An overview of the iteal interface and how to access the different views.

system. To reduce visual clutter and to focus on highly similar alignments, DJW can increase the similarity threshold, which is by default the average similarity value.

Tasks. The alignment view is designed to support exploring the alignments computed using the word embedding approach. It makes alignment patterns in a barcode and a side-by-side view visible, and it aids in identifying particular alignment tuples that attract the expert reader's attention. A coloring scheme is implemented to facilitate the identification of stage-dependent changes in alignment patterns.

Design. The parameter-driven *iteal-V1* system offered various visual means to inspect alignments and show changes after parameter changes. For *iteal-V3*, the original alignment view was extended to communicate stage-dependent modifications and feedback information in a more dynamic way. A sample output of the barcodes and side-by-side views that display alignments as colored lines can be seen in Figure 4.1.

The expert reader can interactively change the set of displayed alignments using different sliders. The stage slider allows for inspection of different iterations (after feedback) of the model and the differences between them. The nearest neighbor slider determines the number of nearest neighbors that are displayed for each line. Since the neighbor relationship is not symmetric, the expert reader has the option to change if the neighbors of the first text, the second text, or both are displayed. Furthermore, the similarity threshold can be increased to allow only inspecting high-quality alignments. We denote the alignments of stage i and the selected number of nearest neighbors k of two editions E_1 and E_2 as $A_{i,k} = \{(X, Y) : X \in E_1 \wedge Y \in E_2\}$. We additionally de-

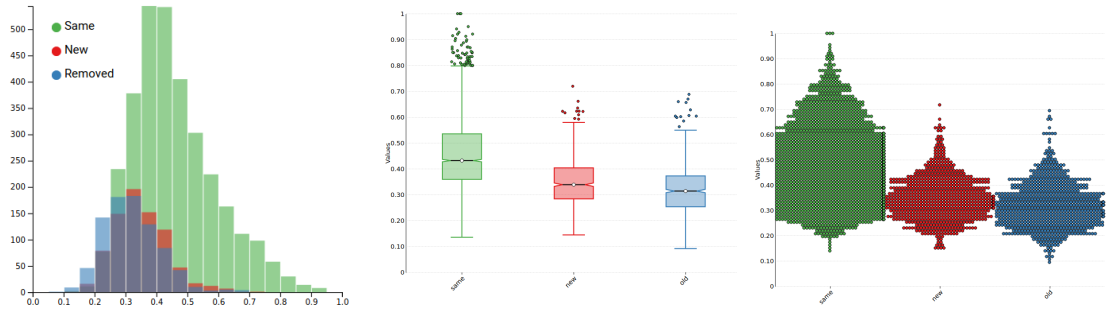


Figure 4.5: The distribution view displays the distribution of the similarity value of the alignments in A_g, A_r, and A_b.

note the alignments found in the current stage as A_c and the alignments found in the previously selected stage as A_p.

In the first stage, all alignments are displayed as green streams connecting two lines, one for each text version. For higher stages, alignments are grouped in different sets and color-coded. The set $A_g = A_c \cap A_p$ includes green-colored alignments found in both the current and the previously selected stage. The set $A_r = A_c \setminus A_p$ represents new red-colored alignments that were not found in the previous stage. Finally, the set $A_b = A_p \setminus A_c$ represents blue-colored alignments that were found in the previous stage, but not in the current stage. The system allows enabling or disabling the different alignment sets to focus on nearest neighbor relations of interest, and the quality of the alignment is communicated through the saturation of the line. If the feedback option is enabled, all alignments labeled in the previous stages are visualized as yellow lines in the barcode view. In the side-by-side view, alignments already contained in any of the sets A_g, A_r or A_b receive a yellow border; otherwise, since they have been manually annotated, they appear in yellow as well.

For reference, the previously selected stage is used. If no stage has been selected, the first stage is used as a fallback to show the total changes from the beginning of the feedback process. If the similarity threshold is increased, alignments that no longer match the new value appear gray, but keep a colored border. To obtain a statistical view of automated alignments, we added a distribution view to the system, which is composed of a histogram and a plot view (see Figure 4.5). The histogram shows the distribution of the similarity of the green, red, and blue alignments, respectively. The histogram acts as a starting point for the exploration of alignment. It shows, at which similarity value, the number of alignments changes to what extent, conveying a feeling on how the alignments are distributed. The plot view displays the distribution of these alignments as notched box-plots or bee-plots. While box-plots are an effective way

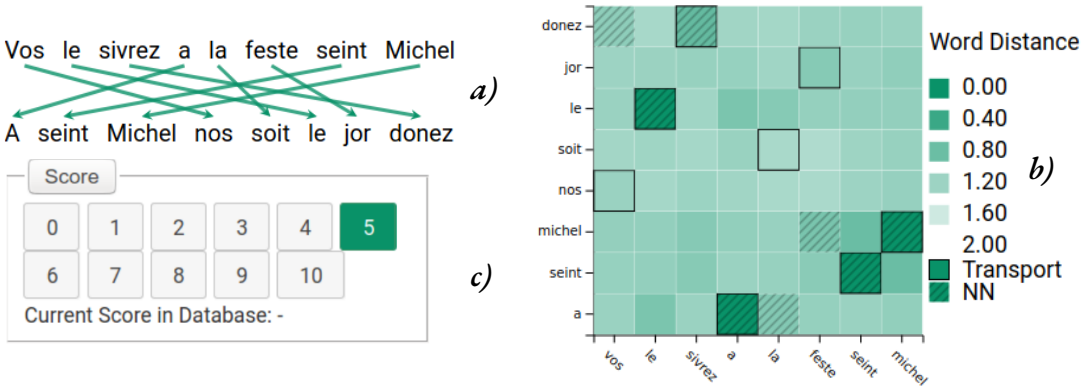


Figure 4.6: The Line Similarity view allows to analyze the word transportation (a) and the word similarities (b) of an alignment of interest and to rate it (c).

to display the distribution and the statistical properties for an experienced user, a less experienced user can benefit from the bee-plot.

Both the barcode and the side-by-side views are interactive, allowing for a flexible exploration of the alignment space through scrolling or clicking on a text section of interest. If an alignment is selected, the line similarity view pops up.

Usage Scenario. At first glance, it seemed to *DJW* that *iteal-V3* offered less control of the visualization than previous iterations, but in reality, it changed both the process and the types of possible alignment and foregrounded the idea of alignment as a gradual process. With the changes in functionality and the color-coding of stepwise reading, the new system was actually more effective in arriving at high-quality, nuanced alignments. Furthermore, the kinds of alignments that were automatically found were of a different nature. They resembled the broken n-grams and orthographic variance that had been identified previously, in addition to new kinds: lines of structural similarity, even lines sharing repeated formulaic speech or synonymous meanings. The alignment example depicted in Figure 4.6a illustrates how the shared string *seint Michel* referring to the feast day of Saint Michael is identified in the lines, but also the copresence of synonyms *feste* (feast day) and *jor* (day) contribute to the alignment of the lines. Although an expert reader might not initially recognize this phenomenon as a strong alignment, the semantic and structural relations that emerged on account of proximity in vector space provided unexpected, yet positive, suggestions for expanding the notion of intertextuality.

4.3.2 Line Similarity View

Now, DJW wants to inspect one alignment of interest in the side-by-side view. When he clicks on the corresponding stream, the line similarity view pops up. At the top, he can see the word transportation that is used to compute the WMD, the system's underlying similarity measurement. At the bottom, he sees a heat map showing the distance between the word vectors in both lines, their nearest neighbors, and, again, the word transportation. In a higher stage, the new and previous similarity values of the alignment could also be seen, as well as the score (label) saved in the database for the alignment. He decides to label the alignment and save the score to the database to include this feedback when computing the next iteration.

Tasks. The major purpose of the line similarity view is to provide a visual explanation of the functioning of the WMD in the automated alignment computation that leads to the showing of this particular alignment tuple. If desired, the expert reader can include domain knowledge into the model by annotating the chosen alignment with a qualitative score.

Design. The two feedback visualizations to inspect an alignment of interest can be seen in Figure 4.6. The first visual depiction (Figure 4.6a) conveys the word transport of the WMD by saturated green arrows that originate from the words of the first sentence to the words of the second sentence. The color scale ranges from white to green to show how much of a word has been moved to the connected word. Additionally, a heat map shows the distance of the word vectors for the words in both sentences (Figure 4.6b). A high saturation indicates a lower distance following a linear color scale from green to white. In this view, word transportation is communicated through a solid border, whereas the nearest neighbors of each word are displayed as striped squares. The heat map gives a quick overview of the distances among the observed words in the vector space. Both visualizations help to get a feeling for word transportation and therefore to explain why the corresponding lines are considered similar in the vector space. Lastly, the line similarity view provides a scoring scale (Figure 4.6c) to be used to rate the alignment on a scale of 0 to 10; this feedback is one of two possible means for the expert reader to refine the vector space for the next stage. Another way to label the nearest neighbors is to click on a particular line of interest in the side-by-side view. The expert reader is then presented with a list of nearest neighbors with the associated similarity values and scoring scales. The pop-up also lists the nearest neighbor of the chosen line from the previous stage to observe changes induced by the expert reader's feedback. To further investigate word vector neighborhoods, the Word Vector Space View can be opened.

Usage Scenario. Assessing the alignment of poetry using a numerical scale is not a straightforward task, so *DJW* judged the relative quality of the alignment of two lines using different features that can occur together in the same line. Generally speaking,

an alignment received a 10 if the linguistic information was exactly the same, even if the words in the lines were spelled differently, close synonyms were used, or verbs were found in different tenses. A score of 0 indicated that there were no shared words or semantics between the lines. Using the ten possible intervals, *DJW* would label based on the amount of linguistic information shared in the line relative to the number of words or syllables. For example, a score of 5 was assigned to the sentence in Figure 4.6a since a little more than half of the words in the line are similar, with two important differences, one was a synonym and the other the definite article of another gender.

4.3.3 Word Vector Space View

DJW decides to investigate the neighborhood of the words included in the aligned lines in the word vector space. He can select a particular word and a number of nearest neighbors that are shown in the word vector space view. He can move the words closer or farther away from the target word in order to change the underlying vectors, which are used for the computation of the alignments in the next iteration. If DJW wonders why two words $\vec{a}$ and $\vec{b}$ do not match as pairs for the computation of the WMD, he can use the neighborhood intersection view to explore the situation. There he sees the common neighborhood of $\vec{a}$ and $\vec{b}$ and all words that are closer to $\vec{a}$ and, respectively, $\vec{b}$, than the target words are to each other. On the basis of the results, he can decide to change the distance between some of the word vectors. After labeling alignments in the line similarity view and moving words in the word vector space view, he can submit the collective feedback and trigger a re-computation of the word embedding displayed in the next stage.

Tasks. Our word space visualizations provide a simplified depiction of the neighborhood of words for the expert reader to understand line-level alignment decisions more easily. Whereas the word space view makes the neighborhood of a single word explorable, the neighborhood intersection view makes the neighborhoods of two words comparable. The expert reader is also able to make word-level adjustments by decreasing or increasing the distance between words.

Word Space View. In the word space view, the k-nearest neighbors of a word of interest are displayed on the x-axis, as seen in Figure 4.7a. To prevent the overlap of words, a collision detection is used to adjust their y-coordinates. This view gives an intuitive overview of the neighborhood of a word vector and allows the expert reader to change the distance d between two word vectors $\vec{a}$ and $\vec{b}$. One or multiple words can be selected and moved to a new position via drag and drop, especially if the expert reader concludes that their distance is inaccurate and that two words should be either closer to each other (or more distant) in the vector space. To support this task, sample sentences giving the word in context can be observed in a popup. For each adjusted

word, the new x-coordinate is used in the next iteration to adjust the word vector corresponding to the new distance d'.

$$\vec{a}' = \frac{\vec{a} + (\vec{b} - \vec{a}) \ \cdot (1 - \frac{d'}{d})}{\|\vec{a} + (\vec{b} - \vec{a}) \ \cdot (1 - \frac{d'}{d})\|}$$

The vector $\vec{a}$ of the moved word is moved closer to or farther away from the target vector $\vec{b}$ on the line between them. Finally, the vector is normalized to ensure unit length, resulting in a small inaccuracy in the distance d'. Through this visualization, the expert reader can adjust the distances between words of interest, thereby changing their vectors. This approach could prove to be helpful when applying word embeddings to under-resourced languages in tackling training limitations.

Neighborhood Intersection View. In order to enable a more in-depth analysis of how the vector space is composed and to illustrate the relation among a set of word vectors, we provide a means for visual comparison of the neighborhoods of two word vectors. This is especially helpful if the scholar wants to investigate why synonymous words are placed far from each other or if they are transported to less related words. The visualization can be seen as a one-dimensional projection of a high-dimensional sphere with diameter d being the Euclidean distance between the selected words $\vec{a}$ and $\vec{b}$. The two neighborhoods of $\vec{a}$ and $\vec{b}$, which are placed along two vertical axes as outlined in Figure 4.7b, are laid out in three sections of the screen. The words inside the sphere have a distance to both $\vec{a}$ and $\vec{b}$ smaller than d, and are placed between the vertical axes reflecting their orientation inside the sphere. Words outside the sphere with a distance smaller than d to either $\vec{a}$ or $\vec{b}$ are placed to the left and right of the vertical axes. Since the neighbor relation of the vectors is not symmetric, the words are color-coded to highlight which neighbors belong to which word of interest. For a specific alignment, all combinations of words in both sentences can be investigated. Two words can also be arbitrarily chosen for comparison.

Usage Scenario. In addition to synonyms for any given word, we also found what could be called synonymous collocates. Taking the example of the word *paien* (pagan), *DJW* found in its immediate word space several prevailing racialized stereotypes of the genre and the medieval period, words such as *Turc* (Turk) and *Sarrazin* (Saracen), as can be seen in Figure 4.7. Word embedding models are particularly efficient ways to expose cohesive discursive patterns of genres, "complete with [their] biases" [Liu20, LGLE18]. Notable in the word space was the appearance of other derived forms of the same words (nominal, adjectival, or adverbial forms) that in the word space reflected a semantic cluster significant to medieval French epic. In Figure 4.7a, we find the nominal declension *paiens*, three derived nouns suggesting a place or state of being of the

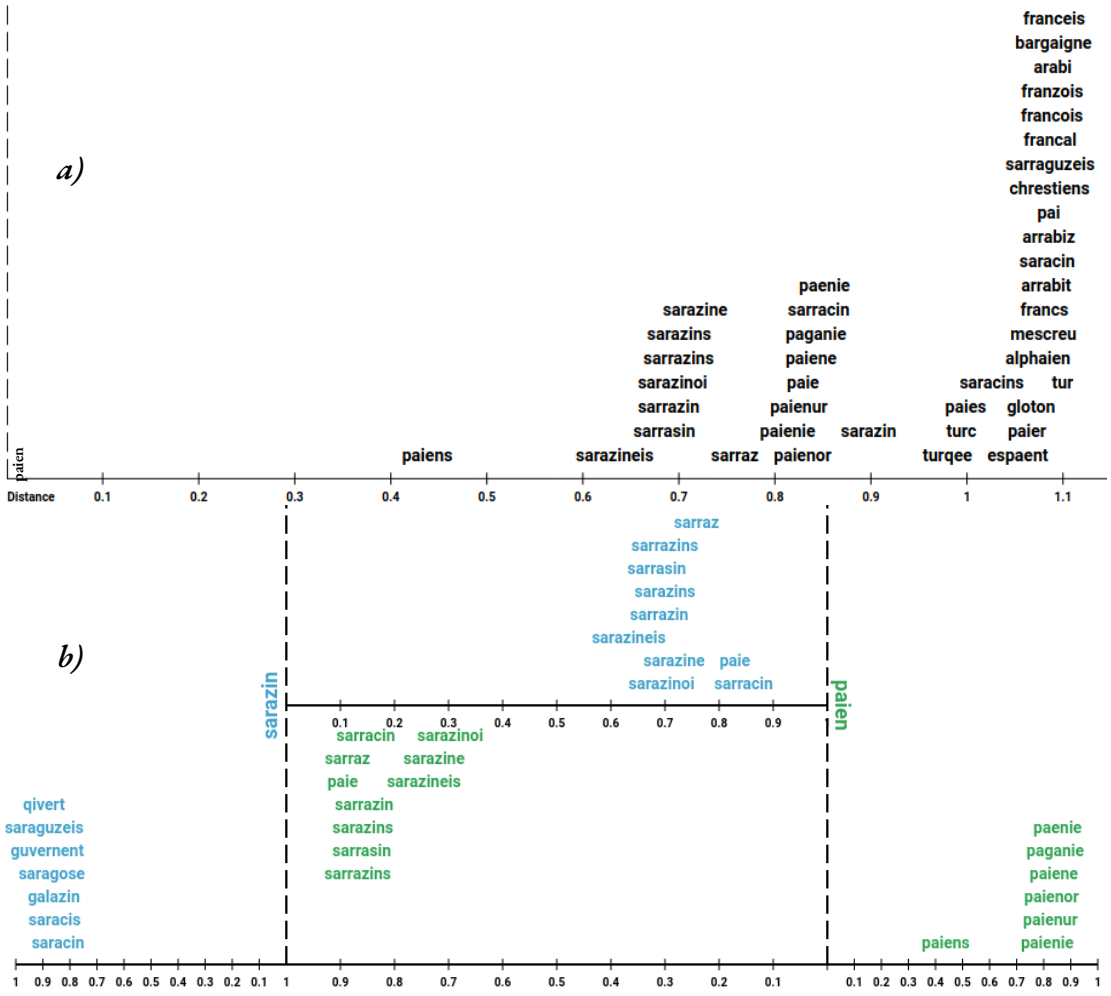


Figure 4.7: The word space (a) and the neighborhood intersection view (b) visualize the nearest neighbor relation of words of interest.

pagans, *paienie*, *paenie*, and *paganie*, as well as an analogous group of words related to *sarrazin*.

Since the word space view in Figure 4.7a was designed for *DJW* to move words based on their similarity, a similar approach was adopted as the one described above in Section 4.3.2; The same word, but with different spellings or inflections, was moved fully, while the derived forms or synonyms were moved only a partial distance. Often we would find verbs of totally different meanings but in the same inflection (infinitives or third-person plural), usually from the end of the poetic line in rhyme-word position. In addition, proper names, place names, and ethnonyms tended to occur within the

nearest neighbors, not on account of synonymity, but perhaps due to their common position within the prosody or syntax. As *DJW* became familiar with the neighborhood view for the word space, it became obvious how useful it could be as a cross-dialect workaround to group words from a common lemma or to build cross-dialect synonym lists, perhaps to replace the problematic ones we have today [vW59].

4.3.4 Compare Stages

DJW can now investigate the second stage. In the alignment view, the alignments are now color-coded to show which alignments stayed the same and which were either removed or added as a result of his interaction. He can also focus on the alignments that he labeled in the previous stage. In addition, he can enable and disable the different alignment types. Furthermore, he can apply the word-level view to focus on the word vectors to find places in the poem where the vectors changed either significantly or not at all. He can then search for a specific word of interest and investigate its neighborhood. In the word space difference view, DJW can see the change of the nearest neighbors of a word of interest compared to the last stage. He can see which words were moved farther away and which moved closer. In a higher stage, he can investigate the changes over all iterations. With this input, he can repeat the labeling of alignments and words to further improve the model.

Tasks. All visualizations in this category are visual feedback mechanisms to reflect on how the adjustments of the expert reader have affected the model. They serve to compare word-level changes across stages, to identify words that have been impacted but have not been directly modified by the expert reader, and to open avenues for continued labeling activities.

Word-level View. In order to observe word vector changes, the expert reader can toggle from the alignment to a word-level view. Instead of alignments, words are now exposed using different colors. The word-level view serves two different purposes. Either the difference between a word's vectors in two different stages or the change of a word's neighborhood in the vector space is displayed. These changes are indicated by colored word backgrounds. The hue of the color depends on the amount of word vector change. To facilitate easy visual differentiation, colors are assigned to five sets of word moves based on the maximum observed change $d_{max} = max\{d(newV_w, oldV_w) : w \in V\}$, with d being the Euclidean distance between the new and old vector w, and V being the vocabulary of the text versions of interest. The colors assigned to the bins are: *none* for $d = 0$, i.e. if the vector has not changed at all, *blue* for d in $(0, \frac{d_{max}}{4}]$, *purple* for d in $(\frac{d_{max}}{4}, \frac{d_{max}}{2}]$, *pink* for d in $(\frac{d_{max}}{2}, \frac{3\ d_{max}}{4}]$ and *red* for $d > \frac{3\ d_{max}}{4}$. An example can be seen in Figure 4.8a. For the word neighborhood, saturation is used to encode the amount of neighborhood change similarly using a linear scale between white and

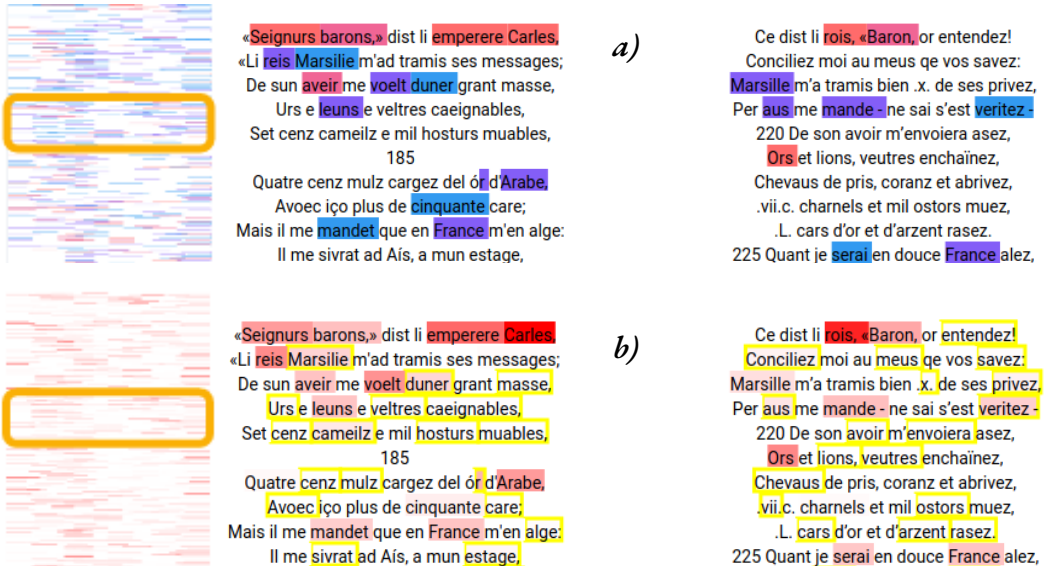


Figure 4.8: The word-level view, which allows spotting places of interest, shows either the change of (a) a word vector or (b) a word's neighborhood.

red. An example can be seen in Figure 4.8b. We compute the change in the neighborhood for each word w as $\sum_{i=1}^{k} d(newV_{w_i}, oldV_{w_i})$ with k set to 50 and w_i being the i-th nearest neighbor of w. To prevent interferences among both channels, word moves and neighborhood changes are observed separately. As in the alignment view, feedback information can be superimposed using yellow color. This includes words manually moved in the previous iterations, which receive a yellow border in the side-by-side view. This helps to spot feedback interactions that may have had an impact on the vector space. When focusing on neighborhood changes, we use the metaphor of yellow borders to highlight words without vector changes. This supports finding words, not touched by any feedback interaction, with changed neighborhoods.

Word Space Difference. In order to observe how the neighborhood of a word has changed across two stages, either through directly moving words using the word space view or by alignment labeling, we designed the word space difference view. Similar to the word space view, the words are displayed based on their similarity to a target word. The difference is the inclusion of information of a previous state of the model. The changes are encoded as arrows starting at the old value and pointing to the new value as outlined in Figure 4.9a. In the case of minimal or no changes, circles are used instead. The words are separated into three groups, decreased distance, no change and increased distance. These three groups are stacked atop each other. Blue color encodes decreased distance, red color encodes increased distance and grey encodes no change.

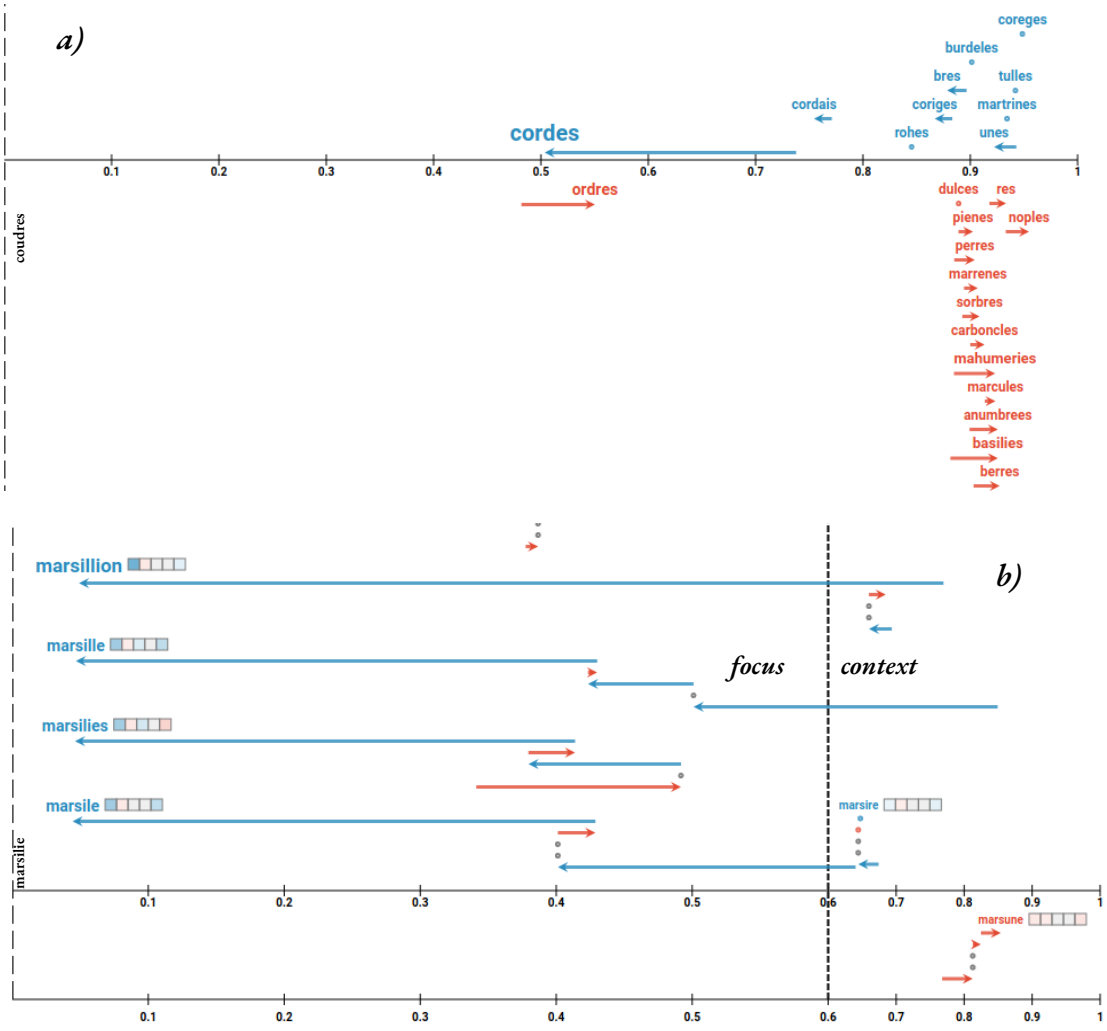


Figure 4.9: The word space difference view gives an overview of the changes in the vector space of a word of interest for (a) one or (b) multiple iterations.

As per user preference, the font size of the word can either encode the absolute distance or the change in the distance. Although both are encoded also by the position of the word and by arrow length, this function is useful for guidance through the neighborhood.

Word Space Difference Sequence. We extended the word space difference view to communicate changes in the neighborhood of a word after multiple iterations, an example is shown in Figure 4.9b. For a reference word, multiple glyphs indicating distance change per stage are stacked. This information is further encoded in a heat map

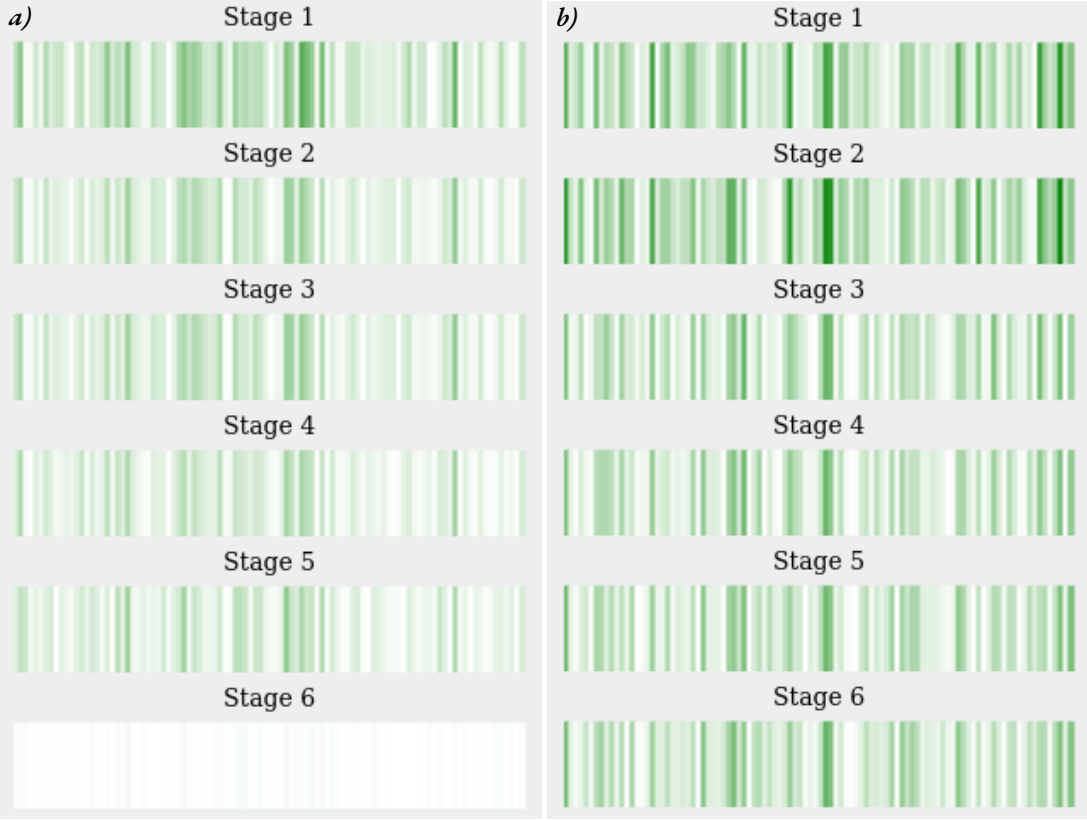


Figure 4.10: The dimension heat maps for each stage for the words (a) *marsilie* and *marsille* and for (b) *paien* and *sarazins*. A higher saturation encodes a larger difference in the dimension. The Stage 6 heat map in (a) is almost completely white because of the low difference, which can also be seen in Figure 4.9b.

placed next to a word. The visualization gives a quick overview of how the neighborhood of a word of interest has changed for each iteration and throughout the whole process. We also used this view to observe the changes in the vector space after normalization and post-processing. To prevent scaling problems, we applied a focus+context metaphor. The focused part on the left-hand side shows the close neighborhood of a word while the context part on the right-hand side provides screen space for the remaining vector space. The expert reader can change the ratio of focus and context by dragging the separating vertical axis.

Dimension Heat Maps. Inspired by a barcode visualization for the comparison of word embeddings [LCHJ16], we designed dimension heat maps to allow the expert reader to see the changes between two word vectors. An example of the words *marsilie* and *marsille*, which are different variants of the same word, can be seen in Fig-

ure 4.10a. Two words that are prevailing stereotypes of the medieval genre *paien* and *sarazins* can be seen in Figure 4.10b. Each dimension is encoded as a line ranging from white to green. Saturated green highlights larger differences, and white means that there are no differences. Across five iterations, the difference between the word vectors is visibly reduced. The large change between Stages 5 and 6 is related to a word move in the last feedback session, in which *DJW* moved *marsille* very close to *marsilie*, resulting in almost the same vector for both words. Something similar happened to the words *paien* and *sarazins*. The visible difference between Stages 2 and 3 occurs because both words were moved by *DJW* closer to the word *sarrazins*. For both pairs of words, the changes in the other stages correspond to the more fine-granular alignment labeling process.

Usage Scenario. Although the process of aligning the poems is not complete, with iterative labeling and word movement the system allows for a gradual convergence on strong patterns of intertextuality that were not identified by *iteal-V1* [JW17a]. The challenge of such a complex multistage task was to trace the impact of the changes *DJW* made. The color-coding scheme applied in Figure 4.8a was helpful not only for keeping track of the many words that were changed but also, for purposes of coverage, to be able to redirect attention to sections of the poem or even sections of the poetic line (beginning, middle, or end) for drawing *DJW's* attention to elements of alignment that may have been neglected in previous iterations. The same can be said of the saturation, where *DJW* would pay more attention to words that were not yet affected by neighborhood changes. For example, in the list of animals given in the Venice poem in Figure 4.8b in lines 6 and 7, the neighborhoods of the names of the animals *ors et lions* (bears and lions), *veutres* (hunting dogs) and *chevaus* (horses) had moved but the various action verbs at the end of the line had not. Furthermore, for a more precise indication of the word space difference discussed in Section 4.3.3, the word space difference and the word space difference sequence are particularly helpful for a more precise separation of words that often occur in rhyme position, for example *cordes* (the Spanish city, Cordoba) and *ordres* (order or position), yet without a strong semantic similarity. Visualizing the stepwise progression of words that were affected by word moves with the dimension heat map ultimately provides the expert reader interested in forms of complex intertextuality with the ability to focus not on a perfect alignment, but rather to self-pace and self-monitor while exploring complex textual scenarios with companion tools, exploring the relations between different phenomena at hand, and assessing the evidence on display [SLB+09].

4.3.5 Alignment Labeling

The most important feedback opportunity for DJW is labeling an alignment, dependent on its feasibility on a scale from 0 (entirely unreasonable) to 10 (perfect match). In general, the scale indicates how similar the words in the alignment are, including syntactic as well as semantic features. Since it can be difficult to accept or reject alignments given the nature of the poetry at hand, DJW requested a means to be able to interpret the lines in more depth.

After labeling several alignments, they are used as feedback to the word vectors. Our feedback approach is inspired by the Rocchio Algorithm [Roc71], which moves a query vector closer to relevant documents and farther away from irrelevant documents. Our adjusted Rocchio formula results in:

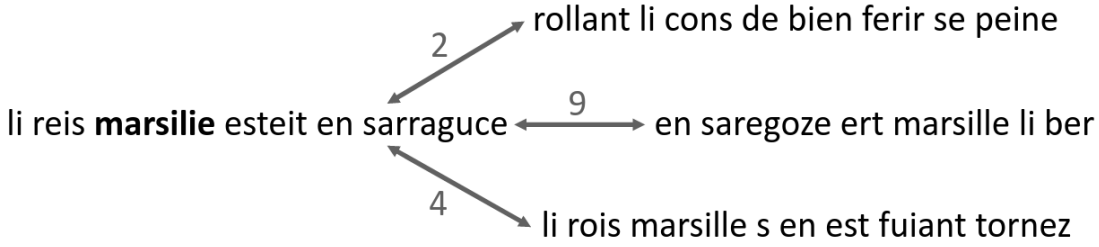
$$\vec{v}_w = \alpha \cdot \vec{v}_w + \beta \cdot \frac{1}{|D_{p_w}|} \cdot \vec{v}_{p_w} - \gamma \cdot \frac{1}{|D_{n_w}|} \cdot \vec{v}_{n_w}.$$

In the classical formulation, D_p and D_n are sets of relevant and irrelevant documents. In our case, they correspond to bags of words, which should be closer to the target word w or farther away. In contrast to the Rocchio Algorithm, we do not focus on a query vector, instead, we apply an update for all word vectors in the labeled alignments, which we separate into three bins: positive feedback (alignments scored higher than 6), negative feedback (scored lower than 4), and mixed-case (scored 4 to 6). We decided to apply this approach because of *hemistiche* (half-line) alignments. This is important across versions of a medieval poem since sometimes the information of one line is transposed into a single line in the target poem and, at other times, it is split in half across two separate lines. This can also be an issue when a poem is recast in a different meter and recombination of syllables or words is necessary. For a given word w, D_{n_w} includes, for all alignments (X,Y) with $w \in X$ or $w \in Y$, all words of the other sentence in the alignment if the score is lower than 4. Alignments with a low score are typically generated through overlapping function words or misplacement of rare words in the vector space, so it can be beneficial for the following iteration to move all words appearing in this false alignment slightly away from each other. Similarly, D_{p_w} includes all matches of the word transportation problem $T_{X,Y}$ for the word $w \in X$ or $w \in Y$ in all labeled alignments (X, Y) with a score higher than 6. The case of a score between 4 and 6 corresponds to half-line alignments. Because the sentences are not totally dissimilar, the transport pairs are added to D_{p_w}, while all the other combinations of word pairs are added to D_{n_w}. This combines positive with negative feedback to reduce the risk of moving similar words away from each other. To explain the feedback procedure in more detail, we added formal equations for the bags of feedback. This results for the

positive feedback bag of a word w in:

$D_{p_w} = \{(t, s(X,Y)) : \exists (X,Y) \in \mathcal{A} : ((t,w) \in T_{X,Y} \vee (w,t) \in T_{X,Y}) \wedge s(X,Y) \geq 4\}$
and for the bag of negative feedback in:
$D_{n_w} = \{(t, s(X,Y)) : \exists (X,Y) \in \mathcal{A} : ((t \in X \wedge w \in Y) \vee (w \in X \wedge t \in Y))$
$\wedge (s(X,Y) \leq 3 \vee (4 \leq s(X,Y) \leq 6 \wedge (t,w) \notin T_{X,Y} \wedge (w,t) \notin T_{X,Y}))\}.$

D_{p_w} includes all matches of the word transportation problem $T_{X,Y}$ for the word $w \in X$ or $w \in Y$ in all labeled alignments (X,Y) with a score of 4 or higher. D_{n_w} includes, for all alignments (X,Y) with $w \in X$ or $w \in Y$, all words of the other sentence in the alignment if the score is lower than 4. Additionally, for sentences with a score between 4 and 6 (half-line alignments), all word pairs except transportation matches are used. For the half-line alignments, we also tested not using the transportation pairs as positive feedback, so only excluding them from the negative feedback, but this resulted in lower similarities for almost all of them. For both bags, multiple occurrences of the same word t are possible from different alignments with different scores. Consider the following three alignments:



For the line *li reis marsilie esteit en sarraguce*, we have a high-rated alignment, a half-line alignment, and a low-rated alignment. The bags for the word *marsilie* would be:

$D_p = \{(marsille, 9), (marsille, 4), (fuiant, 4)\}$ and
$D_n = \{(li, 4), (rois, 4), (s, 4), (en, 4), (est, 4), (tornez, 4), (rollant, 2), (li, 2), (cons, 2), (de, 2), (bien, 2), (ferir, 2), (se, 2), (peine, 2)\}$.

Another difference from the original formula is the computation of $\vec{v}_{p_w}$ and $\vec{v}_{n_w}$.

$$\vec{v}_{p_w} = \sum_{(t,s) \in D_{p_w}} \frac{s}{10} \cdot \vec{v}_t, \quad \vec{v}_{n_w} = \sum_{(t,s) \in D_{n_w}} (1 - \frac{s}{10}) \cdot \vec{v}_t$$

Instead of a simple centroid, we compute a weighted centroid based on the score s of the alignment (X,Y), in which the words w and t co-occurred. α, β and γ are weighting values to further control the influence of the original vector, the positive centroid,

and the negative centroid on the new vector. Modern information retrieval systems set $\alpha = 1, \beta = 0.8$ and $\gamma = 0.1$. We deviate from these default values and set $\beta = 0.5$ and $\gamma = 0.5$ to treat interactions of the expert reader in an equal way. The labeled alignments are stored together with the new distances of the moved words in the Word Space View. Both types of feedback are applied to adjust the word vectors. Labeled alignments are first processed, followed by new distances registered after word moves.

4.4 Evaluation

Our project setup profited from our long-standing collaboration and trust as a team, which allowed us to avoid the far-reaching misunderstandings typical of projects at the intersection of visualization and digital humanities [Jän16, HEAB+17, BEC+18, SRF+19]. Our close collaboration also made possible the opportunity for an implicit evaluation of the co-designed iterative visual design process. Such evaluations have been previously documented in various publications focusing on applications in digital humanities [ARLC+13, JFS15].

Iterative Labeling & Qualitative Evaluation. The design of our visualization system was iterative, where we met once a week to reflect the results of a stage and plan the steps for the next. To begin, we focus on *Song of Roland* as outlined in Sections 3 and 4. In five sessions, *DJW* labeled alignments and then moved words in the word space view. We met once a week to reflect on the results of each stage and to plan the steps for the next, revisiting the visualization design and incorporating *DJW's* feedback into the model. In *stage 1*, 100 alignments were labeled and 40 words were moved manually by *DJW*. After this stage, we added the word-level view to better communicate the changes in the next stage. Additionally, we highlight manually moved words and an option to filter for previously labeled alignments, allowing DJW to keep track of the feedback interactions. In *stage 2*, 40 alignments were labeled, and 110 words were moved by *DJW*. To highlight the interesting cases in the word view, a list of stopwords was removed. This list consisted of 300 of the most frequent words from the training corpus, retaining topical words specific to epic. Additionally, the neighborhood intersection view was added to allow for a comparison of the neighborhood of two words of interest. In *stage 3*, 100 alignments were labeled and 250 words were moved by *DJW*. We added a word space difference view for sequences to communicate a summary view of the change in the neighborhood up to this point over multiple iterations, facilitating the discovery of significant clusters in word space and their concomitant moves. In *stage 4*, 110 alignments were labeled, and 290 words were moved by *DJW*. No additional changes to the system were made, but the combination of the visual system (the nearest neighbor alignment possibilities, the distribution view, and the color-

coding of labeled alignments and alignments found) allowed for more coverage of labeling and word moves across the poem. In the final *stage 5*, 180 alignments were labeled and 70 words were moved by *DJW*.

The results of our extensive work with the *Song of Roland* have confirmed the extent of intertextuality in this tradition commented on by generations of scholars but never systematically and visually demonstrated. For an extended evaluation of our human-in-the-loop process, we carried out a sixth iteration on three different text traditions, each of which has multiple versions. First, we work with two of the four "branches" of the *Romance of Alexander*, a term used to indicate different segments of the life of Alexander the Great compiled in medieval French [Mey82]. These branches were part of the training corpus for the initial model. Furthermore, we included two decasyllabic versions of the *Life of Saint Alexis* and two octosyllabic versions of the *Life of Mary the Egyptian*, which were not part of the training corpus and therefore contained out-of-vocabulary words that were added to the model by *DJW* through labeling.

Quantitative Evaluation. To carry out a neutral assessment of whether our model suits its intended purpose, we sampled up to 60 sentences from each of the four text sources where the nearest neighbor had changed after the sixth scoring session. For each of these sentences, *DJW* was presented with a blind choice between the two nearest neighbors found before and after the sixth scoring session, and he had to rate which nearest neighbor was more accurate or if neither one was. We chose the sixth session because it was the first session during which *DJW* worked with all text sources. We have chosen to sample 60 sentences, since, for some of the text sources, there are no more samples for which a sufficient change in the vector space occurs. The results, summarized in Figure 4.11, document the gradual improvement of the model with our suggested methodology. For all text sources, more of the nearest neighbors determined after the sixth scoring session (green bars) were picked as the more accurate ones. The red bars show the lower number of cases in which *DJW* picked the nearest neighbor determined before the sixth scoring session as more accurate. The orange bars are cases where neither of the two neighbors has been rated more accurate, which can mean either both are good candidates, or both are equally bad. Both cases can be partially explained by the difference in length in the versions, which resulted for some lines in no suitable alignment or multiple candidates for others.

Analysis. *DJW* further selected a reason for his evaluation based on the various alignment features discussed above in Section 4.3. Because of the subjective nature of textual alignments between medieval text sources, we did not apply significance measurements for this process; instead, we focused on emphasizing their characteristics while measuring them against an assessment based on content-specific knowledge. Since the branches correspond roughly to temporal segments of Alexander's life instead of

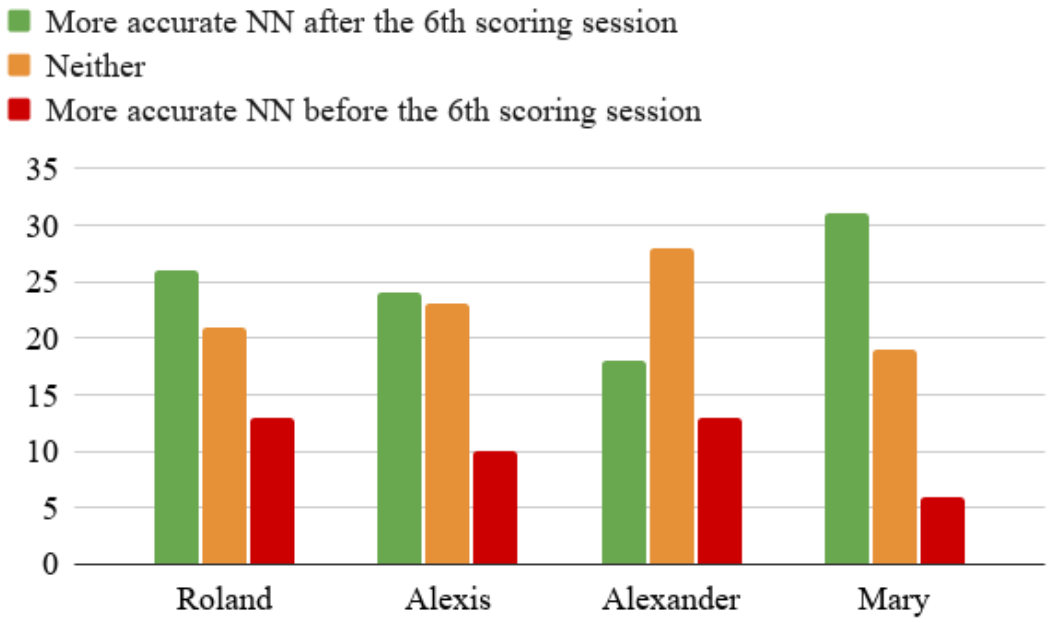


Figure 4.11: Quantitative Evaluation Results

rewritten versions of the same epic cycle, the lowest scores from the group were expected from these branches, where *iteal-V3* was able to find only examples of similar poetic lines across the tradition. Most of the improvement of the model stemmed from the other three text traditions. It is important to note that the reasons for alignment were also not equally distributed across the various text sources but seem to correspond to the nature of the orally-inflected texts in question. It was the choice of a differently inflected verb or noun that led to the choice of a new nearest neighbor in the case of the *Life of Saint Alexis*. In the case of the *Romance of Alexander*, it was the choice of synonymic features that led to the choice of a previous nearest neighbor, while in *Life of Mary the Egyptian* it was the deciding factor for a new nearest neighbor. In the end, it was the orthographic difference that was the most dominant factor in the choice of new and old nearest neighbors in the case of the *Song of Roland*. These data reflect not only the relative narrative similarity of parts of the versions of the Roland but also their significant dialectal differences. In particular, although the *Song of Roland* was not touched in the sixth scoring session, it showed 26 out of 39 new alignments that were better than the previous model.

	R@1	R@5	R@10	Total
Pre-Trained	0.13	0.23	0.27	
iteal-V2	0.37	0.55	0.60	
iteal-V1				0.32

Table 4.1: The recall@k for the different approaches.

Comparison with Modern French and *iteal-V1*. We tried a pre-trained French fastText model on a ground truth case where we used a more modern translation of the Oxford version and the original medieval text. Both versions have the same length, and the more modern version is a line-by-line translation of the original manuscript. For both versions, we computed the recall@k with the pre-trained model and our initial model. Furthermore, we also tested *iteal-V1* with the parameter setting that was used in the original paper [JW17a], that is, including stopwords, string similarity of 0.5, 3 shared broken n-grams, and coverage of 0.4. For methods based on word embeddings, we consider an alignment to be found if it is part of the k-nearest neighbors of one of the two lines. As *iteal-V1* does not work based on nearest neighbor relations, we used all pairs found by the parameter approach, which can be less than or more than the k-nearest neighbors' case. The results can be seen in Table 4.1. It is important to note that our model is not trained on any text written in modern French, while the pre-trained model has no information about the dialect in the original Oxford manuscript, and *iteal-V1* works purely on string similarity and shared n-grams. Even in this case, the pre-trained model gave us bad results. The new *iteal-V2* model outperformed the pre-trained model on modern French and the parameter approach. Furthermore, even the parameter approach using string similarity and word n-grams outperformed the pre-trained vectors, which showed us that a model for modern French is not suited for this corpus. Still, because of our corpus size, the pre-trained vectors are likely to have a better geometric property but are nevertheless worse for our use-case scenario. This evaluation is not based on a real-world use case, as both Oxford versions have the same length and are easy to align by hand, but it demonstrates the disadvantage of the pre-trained model even in a case where it could have an advantage compared to our base model because of the modern language of one of the two text versions. For a real-world use case, e.g. the Oxford and Venice manuscripts, it is not feasible to compute measurements like precision and recall because an alignment of these versions can highly deviate due to different interpretations by different expert readers.

4.5 Discussion

Our development of a human-in-the-loop process was based on intense interdisciplinary exchange, from which we gained valuable insights for our respective scholarly backgrounds. Additionally, it led us to assess the limitations of our approach and discover directions for future research.

VIS Reflections. During the participatory design process and the scoring sessions, we recorded the interaction of *DJW* with the different visualizations. In the beginning, he mostly focused on labeling line alignments. The initial cautious interactions with the Word Space View changed throughout the scoring sessions, so that in the end the word interactions captured his attention more than labeling lines. Over time, the interaction with the different options to enable or disable the different sets of alignments also increased. Throughout the process, 525 different alignments were labeled and around 770 word vectors were manually moved. The rating of the alignments involved alignment interactions, observing an alignment of interest and its associated word transportation visualization, as well as direct interaction with the line to inspect and rate a list of nearest neighbors of the sentence. The Word Space View was typically accessed from the Line Similarity View, and the direct search for a word of interest was seldom used. A reason for this behavior can be observed when looking at the feedback in the database: in later stages, the labeling of an alignment was often combined with a Word Space interaction for the words appearing in these alignments as *DJW* was moving through the poem. The way *DJW* worked through the poem also relates to findings in "slow analytics" where literary analysts interacted with a poem through multiple iterations that build on top of each other to build a larger sense of meaning [BMHC16].

DJW Reflections. From the point of view of the researcher using the system, it is a rather complex environment, and its complexity has both benefits and drawbacks. To begin with the latter, the learning curve with such a system can be steep, not because its visual semantics are unclear, but rather because they are so precise and interconnected. Learning to read efficiently within such an environment can take time and, in particular, learning to integrate high-level observations from the vector space into more granular reading practices. The researcher must become used to navigating the various decision-making and presentational views of *iteal-V3* and to assess how or if they can integrate such data into decision-making. On the positive side, it is possible to have rather complex interactions with poetic texts, to compare them in novel ways, and in so doing, refine the word space of the genre in question. Debates about multi-scalar reading in recent years have uncovered rich examples of Distant Reading practices in "a long view of disciplinary history" [Und17]. Scholars have also stated that Close and Distant Reading are not opposites [Bod17] and a loose consensus has emerged in the

critical literature that not only interpretation at different scales is a valuable contribution to contemporary literary studies, but also that visualization has a key role to play in facilitating such innovative reading practices [JFCS16]. *iteal-V3*'s visual system does not blend all aspects of Close or Distant Reading–it would be absurd to claim that it did–but it does combine a very specific task of professional reading, the process of synoptic comparison of texts in view of understanding textual genetics, with additional views of more abstract representations of the word space of the genre. In *iteal-V3*, visualization is not only the static output of computational analysis of texts, as unfortunately is the case in much literary historical criticism, but instead forms an environment in which active interaction between close, distant, and other acts of reading that fall in between might take place.

Limitations. The approach we have presented here has limitations. The computation of the new vector of a word might exhibit inaccuracies. An example is the inaccuracy in the movement of words through normalization, although the magnitude of this effect seems negligible. Another problem is that the feedback in a new stage can overwrite the feedback of a previous stage. For example, in the second stage, the word *ociz* was moved to the word *ocis*, and in the third stage to the word *morz*, which results in overwriting the previous feedback. A similar problem arises when a word is moved to two different words during the same feedback session. A solution for both cases would be to move a word to a position where it can satisfy both conditions, but such a position does not exist for many cases. Moving two words closer to each other in the same iteration leads to the same problem. For these cases, the current solution changes the vectors of these words before all other words to prevent indirect changes. Another limitation of our system is that the re-computation of the alignments is done stage-wise after multiple interactions, leading to feedback delays for a single interaction. Real-time computation is currently not feasible because computing the WMD for two sentences is too complex. The resulting system was designed and developed in close collaboration with one expert reader (*DJW*) to address his needs. Consulting other users could guide us to other feedback visualizations that meet different information needs. Furthermore, the resulting model could be fitted to the needs and interests of *DJW*, and this genre, which is one of the most complex, if not the most complex, in medieval French from the perspective of *mouvance*. Different expert readers could be interested in different relations between words and lines, resulting in different interactions with the system and different word vectors. Furthermore, the main features our system processes consist of sentences, words, and character n-grams, although our focus lies in the alignment of epic poetry. For the comparison of poetry, focusing on syllables might further improve, achieving new results when comparing related texts.

Future Work. Our stage-based interdisciplinary exchange led to many ideas being implemented throughout the project and identified room for future improvements. For example, visualizations could provide more granular information on how user feedback changed the vector space. The current solutions do not convey how the labeling of one particular alignment during an entire feedback session affected the entire word vector space. Moreover, a comparative view on the influence of different labels could provide valuable information on how feedback is processed and on optimal use of the system. An active learning approach could support the scholar in generating feedback; for example, an algorithm based on string similarity and vector similarity could select pairs of likely synonyms and variants to be presented to and approved by the scholar. This could, on the one hand, ease the scoring session and support the generation of dictionaries valuable to domain scholars, on the other. A setup with multiple scholars refining a single vector space could also be of interest because of its social and collaborative potential in the humanities. Visualizing similarities and differences in their interactions with the system could provide valuable information for future developments in visualization. Another interesting aspect would be to focus on a larger collection of vernacular literature, although the corpus of texts exhibiting such a high level of *mouvance* is somewhat limited. A visualization system to compare more than two texts in the same traditions could direct the scholar to hitherto unknown alignment patterns.

4.6 Summary

Our interdisciplinary collaboration began several years ago with the goal of establishing a system that supports the semi-automated alignment of unstable medieval text versions. Our journey led us first to a parameter-based white-box approach (*iteal-V1*) to compute alignments based on syntactic text features [JW17a], and second to a black-box approach based on word embeddings (*iteal-V2*) that also considers semantic text features and thematically related concepts [MWJ19].

The work on *iteal-V3* documents our efforts to develop a series of visualizations capable of intuitively conveying the complex structure of the word vector space for professional reading. Furthermore, we created means to integrate scholarly feedback back to the vector space model and visual cues to observe how user-driven modifications affect local neighborhoods in the vector space. Our stage-based development process attests to the fact that explainable visualizations like the ones presented in this thesis are capable of building bridges between computer science and other domains, thereby expediting gradual trust building in complex algorithmic processes. Benefiting from the expert reader's feedback, the word embedding approach finally led to a better performing alignment computation transferable to related text-based scenarios.

God hides in the details.

Aby Moritz Warburg

5

Towards Enhancing Virtual Museums by Contextualizing Art through Interactive Visualizations

In recent years, the digital humanities community started to apply large quantitative analysis on visual material. For this, the term *Distant Viewing* [AT19] was coined based on the established term *Distant Reading* [Mor13] as the use of quantitative methods such as machine learning and visualization can provide new insights into collections of cultural heritage stored by public institutions and can give new access points to cultural assets. Public institutions seek to attract as broad an audience as possible, which results in an imperative to cater to the growing virtual online audience in conjunction with regular visitors. This has been magnified by the closure of public institutions across the world during the COVID-19 pandemic, meaning that virtual online access has remained the only option for domestic and international visitors [VKCA20, Bie21]. Traditional art gallery exhibitions are typically limited to a finite number of works, arranged in a fixed order, and accessible by visitors walking around within a physical space. The range of artworks shown can be further limited by considerations such as the fragility of the items, visitor fatigue, and the availability of items from existing col-

lections or on loan from other institutions. These limits often prevent an institution from displaying the ideal content to support a given theme.

A virtual museum can be seen as an extension or a complement to a physical museum which removes some of these constraints by providing a solution to the space limitations of a physical exhibition and by giving a safe alternative to displaying and investigating fragile objects. Ideally, the design of a virtual museum should go beyond presenting only the content and information that the museum has available digitally from its own collections [AS98, Sch04]. "The term virtual museum has been defined as a logically related collection of digital objects composed in a variety of media which, because of its capacity to provide connectedness and various points of access, lends itself to transcending traditional methods of communicating and interacting with visitors; it has no real place or space, its objects and the related information can be disseminated all over the world" [Sch04]. This core idea of the virtual museum goes back to André Malraux's project the Imaginary Museum [Mal54, SFKP09] which he described as a museum without walls, which was a montage of photos from all over the world and from different time ranges.

While existing types of virtual experiences seek to replace or complement a real visit, online tours often suffer from being too passive and lack in-depth interactivity to keep virtual visitors meaningfully engaged with the content for more than a short time. We propose a virtual museum tour enhanced by direct access to various selected visualizations that contextualize the artworks within the gallery. Our approach allows the virtual visitor to explore and compare paintings using machine learning methods and visual interfaces that arrange related artworks as an array of icons. Following the concept of generous interfaces using multiple representations to reveal the complexity of the cultural collection [Whi15].

This serves the visitor with various entry points to discover the underlying art collection and thus to promote serendipitous discoveries [Ric88]. In the context of our virtual exhibition, we offer different arrangements of art and visualization to give visitors a deeper and more fulfilling museum tour experience. Our intention is to encourage visitors to explore more of the virtual museum exhibits. Therefore, visitors can explore collections or exhibitions with individually targeted activities, such as searching, selecting, and comparing artworks.

We conducted an informal evaluation with 61 participants from different backgrounds to evaluate the concept of a virtual museum in a three-dimensional environment combined with information visualization principles that contextualize the artworks. Logging all activities during virtual museum visits gave us interesting insights concerning the visitors' adaptability to the visual interfaces, how long they observed

the diverse visual depictions, their movement patterns inside the virtual space, and, in general, their interest in art. In summary, our contributions are as follows.

- **A Virtual Museum Model & Tour** that connects the virtual version of a 'real' museum gallery with an adjoining exploration room, where the visitor is enabled to regard a painting in the context of a large image archive. Visitors are engaged in exploration bearing on machine learning methods to generate diverse perspectives on art collections through interactive visualizations, which also allow serendipitous discoveries.
- **An Evaluation** that informs on the behavioral aspects of museum visitors and their characteristics in a three-dimensional space, providing valuable insights regarding the adaptability of the virtual museum concept and for future improvements of virtual museums.

Our evaluation revealed that not only did it appeal to the general public, but our approach also seems to serve the increasing desires of humanities scholars to quantitatively analyze art collections. Although the use of visualization methods in art history can be traced back to Aby Warburg's Mnemosyne Atlas in the 19th century, digital methods are rarely available [WD16]. Or, as Drucker stated it "To date no research breakthrough has made the field of art history feel its fundamental approaches, tenets of belief, or methods are altered by digital work" [Dru13]. Nevertheless, computational methods, e.g., machine learning and visualization, can support working practices and can give new insights into the objects of interest and therefore help answer research questions. Furthermore, virtual museums could serve as an interdisciplinary investigation object and extend the conventional museum space by providing enhanced visitor experiences in terms of engagement and attraction [Bie21].

5.1 Related Works

Our work focuses on virtual museums and, more generally, visualizations of cultural heritage collections (Section 2.3). In particular, visualization of images and the objects contained within them, facilitated by neural networks and other computer vision methods (Section 2.2). We follow the serendipity principle to explore a collection of cultural heritage. Similarly, Thudt et al. [THC12] followed the same principle to explore a collection of digital books. Junginger et al. [JOV+20] displayed objects contained in photographic plates in a Close-Up cloud similar to our Picture Clouds. Furthermore, other approaches also combine multiple visualizations showing different facets of cultural heritage collections to give manifold perspectives on the objects of interest [GvTGD18, BDH21, DPC17].

5.1.1 Virtual Museums

In recent years, different types of virtual museums have been proposed to display artifacts of cultural heritage. Kabassi [Kab17] evaluated the state-of-the-art museum websites, including three-dimensional environments and mobile apps, while Bekele et al. [BPF+18] surveyed augmented, virtual, and mixed reality technologies for cultural heritage collections. Walczak et al. [WCW06] presented the ARCO system to explore a virtual museum with AR and VR. Huang et al. [HCC05] presented an augmented panorama approach. Lugrin et al. [LKS+18] presented a location-based virtual museum for more than 100 users. For three-dimensional artifacts, Loscos et al. [LTF+04] present a virtual museum where users can touch statues with haptic feedback, and the Atalaya3D project [MRB18] created three-dimensional scans of sculptures and historical sites that can be visited in a three-dimensional environment. Carvajal et al. [CMB20] created a virtual museum based on three-dimensional image acquisition and modeling.

Some works focus on the creation of a personalized virtual museum. For this, Hayashi et al. [HBN16] present an approach to automatically generate a virtual museum based on user's bookmarks, while VIRTUE [GSP+19] allows users to create and navigate within their own exhibition of two- and three-dimensional objects. DynaMus [KKP16] is a three-dimensional virtual museum framework to create virtual exhibitions, and Liarokapis et al. [LSB+04] present a visualization framework to visualize artifacts from cultural heritage in a virtual museum. All these approaches focus on the recreation of a museum in a virtual environment to present objects of interest, but without visualizing additional information about them. Therefore, we see an opportunity to enhance virtual museums using visualization methods that contextualize objects of interest and give new perspectives on cultural heritage collections by allowing visual exploration.

5.2 Virtual Museum Project

The project started with the idea of applying the concept of quantitative text analysis to paintings, e.g. *Cultural Analytics* [Man16] and *Distant Viewing* [AT19]. Two visualization scholars started prototyping a series of two-dimensional visualizations, each of which generated a new quantitative perspective on art collections. Our team was later complemented by an expert in museum exhibition design, who, in collaboration, designed a virtual museum concept that made it possible to make the multiple exploration abilities of our visualizations accessible to the general public.

Our Vision. Current virtual museum solutions mostly aim to recreate digital representations of existing exhibitions. The major shortcoming is that they seldom take advantage of interactive visualization principles to give multi-perspective views on the

objects of interest. In order to make virtual museums more interactive, we regarded them as an ideal environment in which to integrate our prototypical two-dimensional visualizations for quantitative image analysis. Visualizations can help contextualize the paintings and provide new information that is currently not supported by other virtual museums. Consequently, we wanted to find out if quantitative views are of interest to museum visitors and if they are seen as a valuable complement to real museums. Therefore, we formulate multiple abstract tasks and design visualizations to satisfy the underlying information need.

Tasks. The exploration of the virtual museum can spark different questions in the visitor's mind. We formulate the following abstract tasks based on Munzner's task abstraction model [Mun14]. First and foremost, the function of a virtual museum is to *present* artifacts to an audience in an *enjoyable* way that empowers the visitor to *explore* the collection and therefore to *discover* new insights into art. In particular, visualizations should allow the visitor to *discover* new images, styles, genres, or artists. Furthermore, a virtual museum should allow the visitor to *query* for particular paintings or artists. Referencing the large data set allows the visitor to easily *discover* and *identify* similar images. When focusing on an artist of interest, a *summary* of the artist's development over time, their oeuvre, can be observed. The ability to *compare* images and artists with each other is also easier on a large-scale through digital tools like a virtual museum than it would be in a real museum. We also allow interactions to perform different *search* tasks. Sometimes visitors want to *lookup* a specific image or an artist of interest. Other times, they might want to interact with the collection freely and *explore* it, resulting also in serendipitous findings. The concrete tasks for each wall are explained together with the design in Section 5.3.2.

Data. We combined the paintings from Wikimedia and the WikiArt corpus. The WikiArt corpus consists of more than 180,000 paintings from more than 3,600 different artists ranging over 199 different styles. It is one of the largest publicly available digital art data sets and includes information about the year of creation of the painting, the genre, the media used, the location, and the series of which it is part. There are also some manually annotated tags for the content of the images and sometimes descriptions. The Wikimedia painting set consists of around 20,000 images with over 1,400 different artists. Our total data set without duplicates contains around 200,000 images from more than 4,800 artists. We applied machine learning methods to contextualize this data set further.

Object Detection and Pre-Processing. We applied the Faster R-CNN [RHGS15] trained on the Open ImagesV4 corpus [KRA+20] to the data. The Open Images data set consists of around 9 million natural images with different image-level annotations. For around 1.74 million images, annotations about object bounding boxes exist that

were manually annotated. An image contains on average 8.4 bounding boxes, therefore, resulting in 14.6 million bounding boxes for the whole data set. For object classes, a hierarchy consisting of 600 different classes is used, including parent-child class relations. All these properties make the Open Images data set a good choice for training a model to detect a wide range of objects. The image feature extractor (backbone) of the network is an Inception Resnet V2 [SIVA17] trained on the ImageNet data set [DDS+09]. The ImageNet data set consists of more than 14 million images and is therefore an appropriate training data set to create a generalized feature extractor. For each image in our data set, Faster R-CNN predicts 100 bounding boxes with a confidence score between 0 and 1. For this, the region proposal part of the network proposes rectangular regions of interest, and the detector part classifies these regions based on the object classes. This results in around 20 million bounding boxes, for which we only included bounding boxes with a confidence score of 0.5 or higher to reduce errors. Our final set includes around 530,000 object bounding boxes. Furthermore, we use the top layer of the image feature extractor to compute image embeddings for each image and each bounding box in our data set. These image embeddings are vectors with a dimension of 1536 and are used to compute the nearest neighbors for the images and the bounding boxes. All image embeddings are added to a faiss [JDJ21] index structure, where the similarity is based on the Euclidean distance between them. For each image, the k most similar images and, for the bounding boxes, the k most similar bounding boxes are queried by searching the index based on the vector of the image or bounding box. We also apply the nearest neighbor search to find duplicates in the two data sets. For this, we first find the nearest neighbors for each image, and then we disregard all images by different artists. All neighbors with a Euclidean distance of 0 are treated as duplicates. For the remaining images, the titles are compared with string similarity to prevent removing a similar image by the same artist. The titles are cleaned from special characters and lowercased. When the titles are identical, we remove one of the images. For removed images, we aggregated the metadata of the two sources.

5.3 Virtual Museum Design

We started with the creation of two-dimensional visualizations of the paintings and their metadata. When considering the complexity of in-depth visual exploration of paintings, there is an advantage in presenting and organizing the results of the different interactions in an engaging way within a three-dimensional exploration room. The objective is to provide an intuitive virtual experience that takes advantage of the existing knowledge of visitors about how to navigate and explore three-dimensional spaces, and also to make it easy to learn by exploring the space. It is proposed that spa-

tial awareness of the virtual visitor makes it easier for them to visualize, comprehend, and navigate a complex array of visual results in "the round" compared to displaying them on an infinite and more abstract two-dimensional web page. We also see an advantage in the visitors' ability to conduct their exploration seamlessly back and forth between the gallery and a virtual exploration space, resulting in longer visitation times with less chance of leaving the site.

As virtual museum tours are typically situated in authentic representations of real visitable places such as museums or art galleries, we seek to extend the online experience by adding a virtual exploration space. Once a real building has been digitized, the resulting model can be easily extended. In this way, a virtual portal to another space can be added as a new piece of the building, either static or hidden, and revealed by clicking on a target.

5.3.1 Exhibition Design

In order to evaluate this enhanced virtual museum tour, we created a working prototype consisting of a realistic simulation of a gallery space with an additional exploration room accessed through a portal (Figure 5.1). We created a virtual exhibition for evaluation using 12 paintings selected from the WikiArt data set, from the period 1887-1939, and included the styles: Art Nouveau (Modern), Fauvism, Impressionism, Realism, Pointillism, Social Realism, Ukiyo-e, and Symbolism. When sizing the two spaces, we took into account the number of paintings to be displayed, a natural angle of view, and the navigation by the visitor. The Gallery measures 12mx12m with a ceiling height of 4m, with space for 3 paintings on each wall and space on one wall for the portal to the exploration room. Each painting is set in a three-dimensional photorealistic frame with the centers at the visitor's eye-level and on the left side of each painting is a label that shows the artist, the title and the year. The Exploration Room measures 18mx18m with a ceiling height of 5m. We found that a larger space is required so that the walls can accommodate the visualizations within the visitor's field of view.

The two rooms were rendered using photographic textures sampled from real buildings to give an impression of a gallery with realistic lighting, in the same way as other virtual tours represent real-world environments. The floor and ceiling in both spaces are the same to provide a feeling of continuity. The walls in the Gallery are sampled from a real well-lit gallery interior; however, those of the exploration room are stylized in plain grayscale colors to give prominence to the visualizations. To facilitate navigation between the two spaces, we added a way-finding sign above the portal. The final model is imported as a Collada file with the three.js library [Dir13] to render it with WebGL in the browser. Furthermore, we add a CSS3D renderer to add interactive vi-

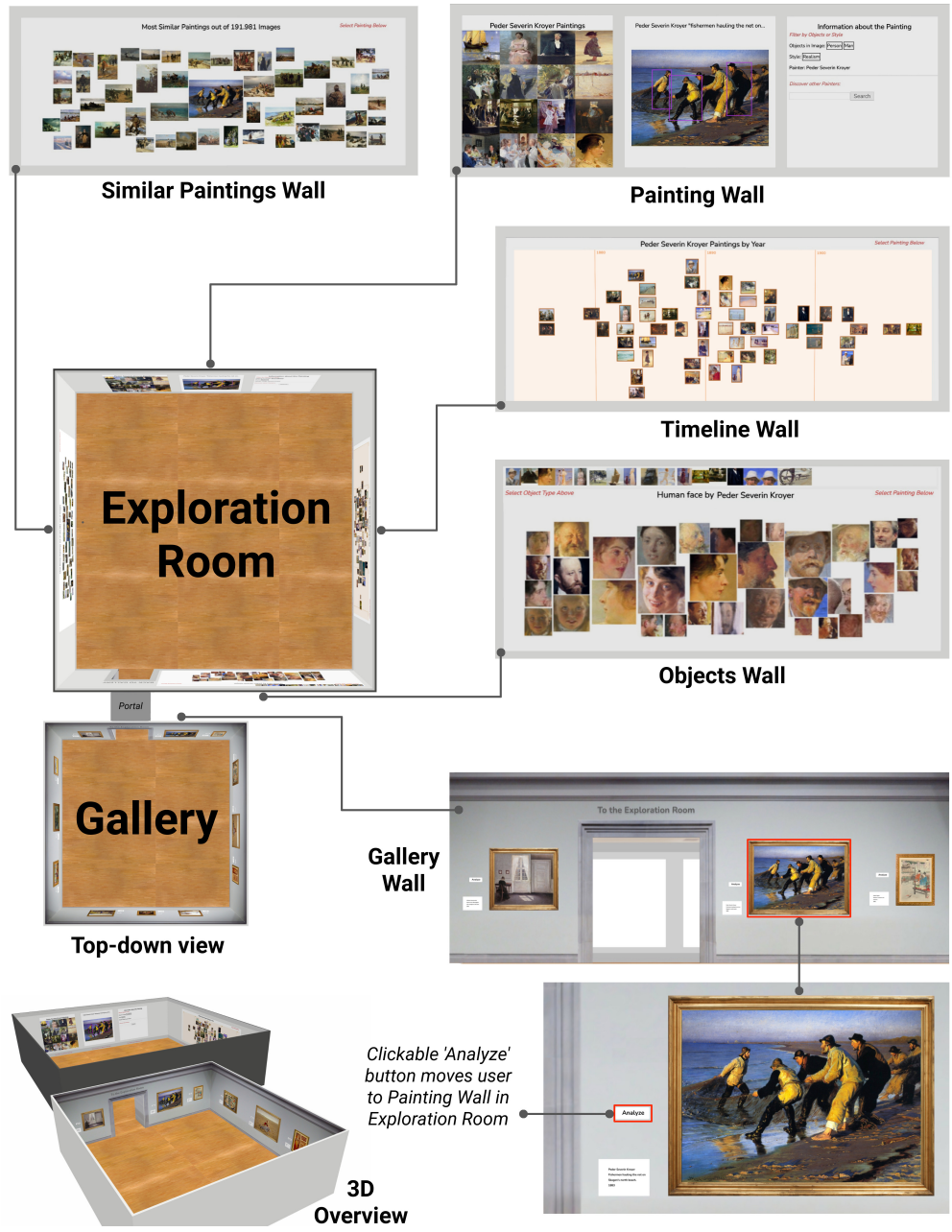


Figure 5.1: The design of the virtual museum. Showing the visualizations on the different walls of the top-down view.

sualizations with d3 [BOH11] to the three-dimensional scene. This renderer applies three-dimensional transformations based on the CSS transform property on the DOM elements. Each visualization is mapped to a wall of the exhibition. Visitors can move around the virtual museum using the arrow keys or WASD keys. Using a mouse or a touchpad, they can rotate the camera by holding the left mouse button. Using the mouse wheel or the touchpad scroll functionality, they can zoom in and out. To prevent visitors from walking through the floor or ceiling, we restrict movements in the y-direction. Furthermore, we apply ray casting to prevent movements through walls. Visitors can interact with the paintings and input components by clicking on them.

5.3.2 Virtual Museum Tour

The virtual museum displays artworks in different contexts based on specific attributes, such as objects contained in the image, the artist, the style or the year. This approach intends to extend the virtual experience by exploring objects of interest in a more faceted way, rather than purely presenting them to the audience. A visitor starts in the Gallery surrounded by artworks from different artists and various styles. The Gallery includes the following 12 paintings placed on four walls depicted in Figure 5.2:

- Peder Severin Kroyer "Fishermen hauling the net on skagen s north beach" (1883)
- Ogata Gekkoprint "From series women s customs and manners" (1895)
- Camille Pissarrothe "Harvest of hay in eragny" (1887)
- Albert Blochthe "Garden of asses II" (1939)
- Anna Ancher "Harvesters" (1905)
- Hans Andersen Brendekilde "Worn out" (1889)
- Alphonse Mucha "Austria" (1899)
- Vilhelm Hammershoi "Interior from strandgade with sunlight on the floor" (1901)
- Andre Derain "The basin of london" (1906)
- Georges Seurat "Harbour at port en bessin at high tide" (1888)
- Homer Watson "The flood gate" (1901)
- Henri Fantin Latour "Peaches" (1903)

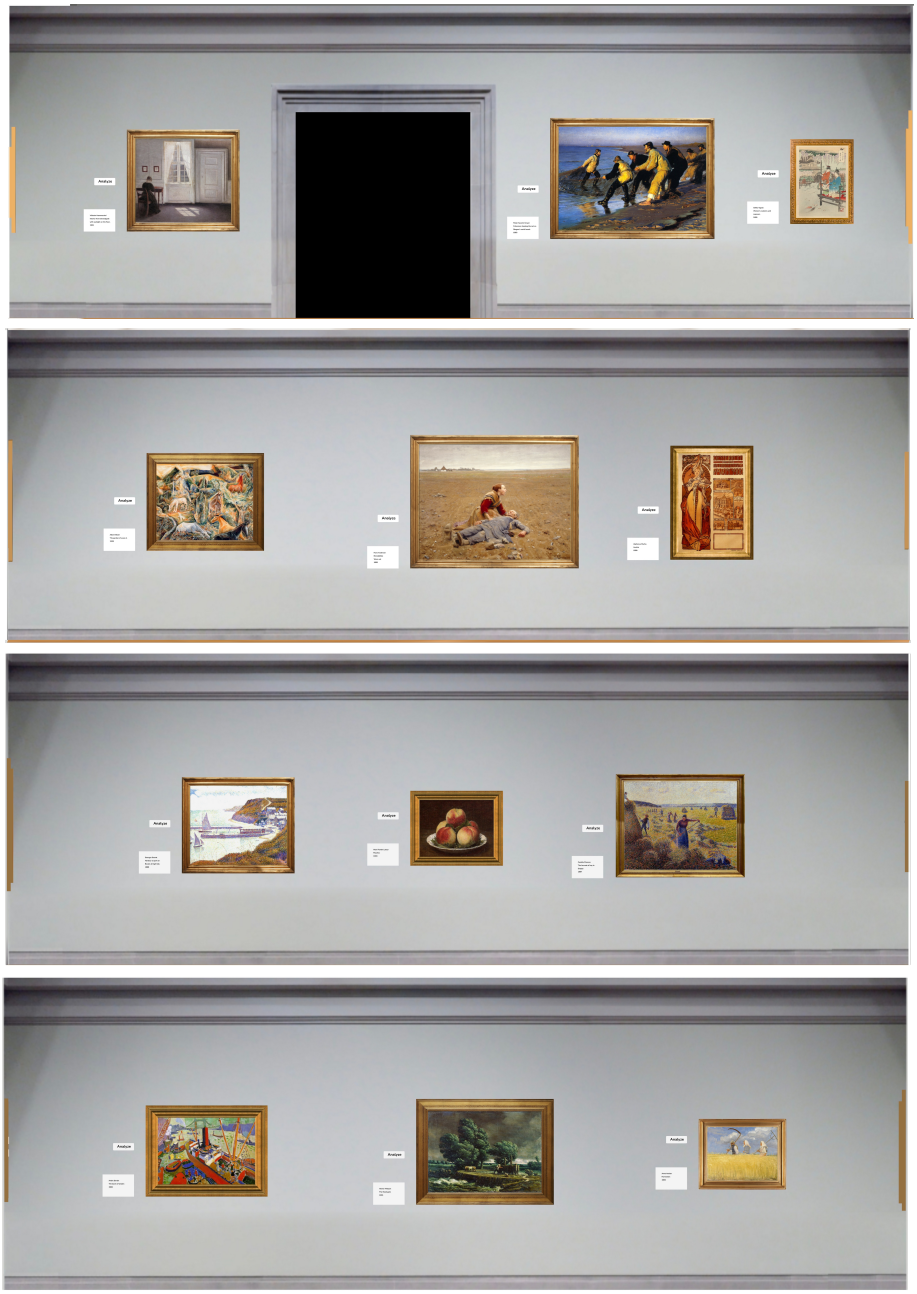


Figure 5.2: The walls of the gallery room.

When clicking on an image of interest, the visitor can choose to further analyze the image. For this, visitors are moved to the Exploration Room. The room consists of four walls, each displaying interactive components that contain visualizations to contextualize the artist's artwork, either by similarity, creation year, depicted objects, or other metadata. Figure 5.3 shows the composition of visualizations in the Exploration Room when clicking on Henri Fantin Latour's painting "Peaches"; the large amount of fruits and flowers is notable.

In the following, we describe a visit to the virtual museum based on the interactions of a participant in our evaluation. A video to follow the narrative is available: https://www.youtube.com/watch?v=EY1ucEb9llQ. The visitor's interactions are followed by the concrete tasks that can be performed and the rationale for the design of the visualizations.

5.3.3 Painting Wall

Silke visits a virtual museum for the first time and has no prior experience with virtual three-dimensional environments. After observing the painting "Austria" by Alphonse Mucha, she decides to analyze it further. In the Exploration Room, the painting is shown together with some metadata and other paintings by Alphonse Mucha on the Painting Wall. After observing different paintings by the artist for a while, she searches for "Gustav Klimt". Now she is presented with "Adam and Eva (unfinished)". Next, she clicks on a bounding box in the image with the label "Woman".

Tasks. The main purpose of the Painting Wall is to display the painting of interest for enjoyment, but it also serves as a control panel to change the context, that is, the artist, the style or the depicted objects. The center of the wall enriches the image with information about the objects in it with their bounding boxes. The right-hand side of the wall can be used to lookup or search for a specific image, artist, or style. Furthermore, filters can be applied to find images that contain a specific class, multiple specific classes, or have a specific style. The left-hand part of the wall gives an overview of some of the other images in the selected context, and therefore a suggestion of other potential paintings of interest.

Design. The Painting Wall of the exhibition displays the currently selected image in the correct aspect ratio together with all WikiArt metadata about the image, e.g., style, artist, and year (Figure 5.1). When hovering over the image, the bounding boxes are shown with their class labels. Similarly, when hovering over the object filter, the bounding boxes for the specific objects are also displayed. In the metadata field, a visitor can also search for another artist to change the context of the room, or select objects contained in the image, or the style as a filter. The other paintings are also up-

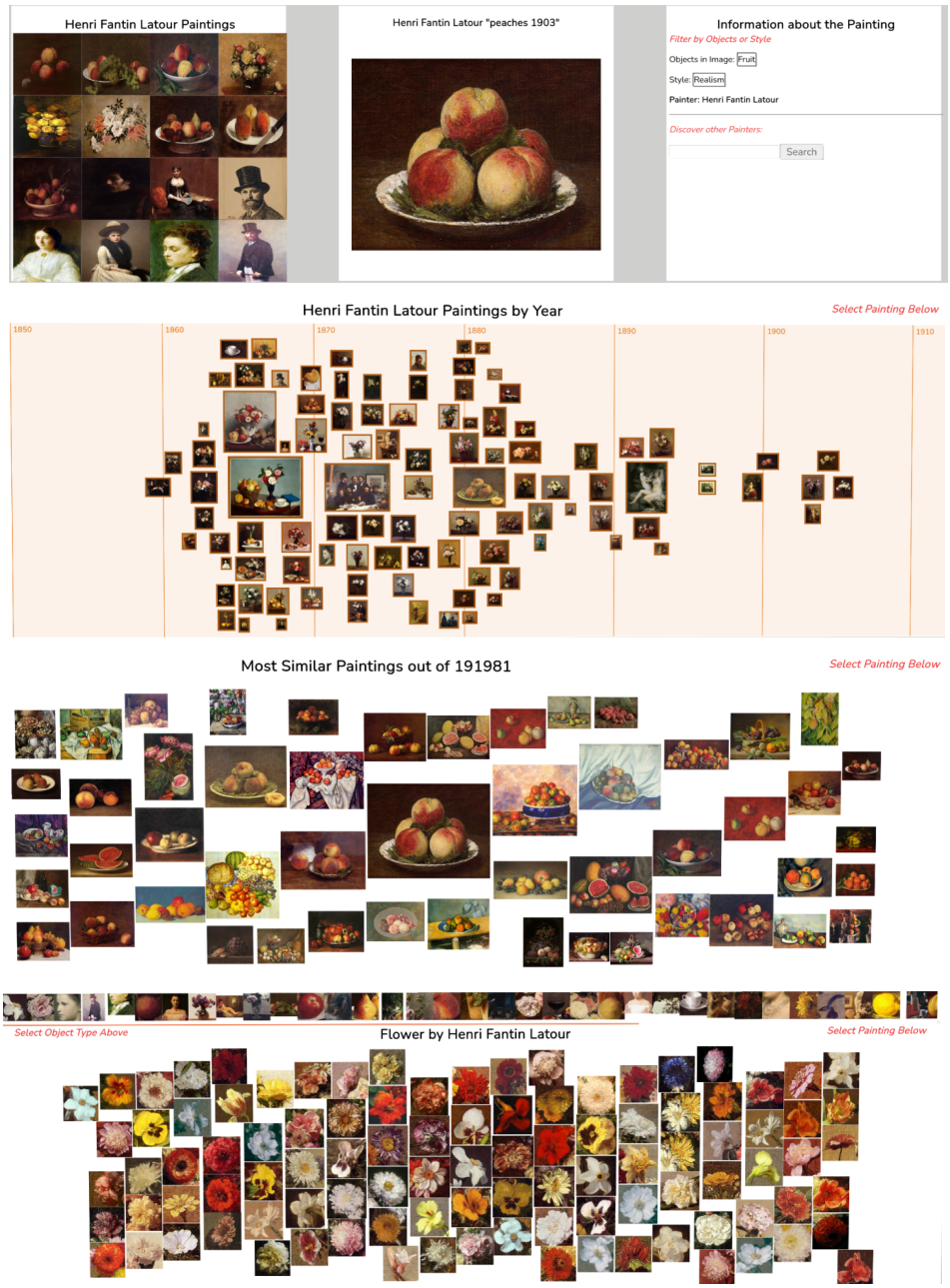


Figure 5.3: The walls in the exploration room focusing on Henri Fantin Latour - Peaches.

dated when the visitor interacts with the room, e.g., by searching, filtering, or clicking on an image. By clicking on another image, that image is displayed on the Painting Wall, and the context is changed if the painting is by a different artist. If a visitor is more interested in a particular style than an artist, a style can be used as the context. Further contexts are possible by using the objects contained in the image, the genre, the series the painting is part of, or manually annotated tags. For the evaluation prototype, we excluded the search for these additional metadata, to prevent the experience from becoming too complex [MMF19].

5.3.4 Objects Wall

After selecting the bounding box, Silke sees the Picture Cloud of women painted by Gustav Klimt on the Objects Wall. She can now click on one bounding box of interest in the cloud or change the selected label to show another object class painted by Gustav Klimt.

Tasks. The Objects Wall gives insight into what classes of objects an artist painted and which are the most frequent. Furthermore, it gives an overview of the variety of different depictions in each of these object classes.

Design. On the Objects Wall, the objects contained in the paintings are displayed in a Picture Cloud (Figure 5.1). At the top, the bounding box with the highest confidence is shown for each class. These are sorted by frequency in descending order and can be selected to change the content of the cloud to a specific class. The cloud shows all objects of the class for the given artist. This can give different perspectives on the paintings of the selected artist. The sizes of the images are based on the confidence of the bounding box, and they are placed in descending order on the basis of this value on an Archimedean spiral. When a style or an object class is selected on the Painting Wall, they are also applied to this Picture Cloud. Through this, objects of a specific style can be analyzed to see the broad range in a specific style. Furthermore, objects that are co-occurring with other objects can be found.

5.3.5 Similar Paintings Wall

Silke goes back to the Gallery to further look at the painting "Fishermen Hauling the net on Skagens north beach" by Peder Severin Krøyer. She decides to analyze the painting. In the Exploration Room, she chooses to look for similar paintings after looking at the image for a few minutes. On the Similar Paintings Wall, the most similar paintings are displayed in a Picture Cloud. There she finds "Clam Diggers" by T.C. Steele and decides to further analyze this image. With the interactive option to look at similar images more closely, she finds artists with whom she is not familiar.

Tasks. The Similar Paintings Wall supports serendipitous discoveries while exploring the data set of the virtual museum. Paintings similar to a painting of interest can be found, and through this other similar artists that are not known by the visitor can also be found.

Design. The Similar Paintings Wall shows the most similar paintings of the currently selected painting (Figure 5.1), for example, the k-nearest neighbors based on the Euclidean distance of the image embeddings of the paintings. The target painting is displayed at the center, and the other paintings are scaled in size by similarity (while preserving the aspect ratio) and placed on an Archimedean spiral around the painting. With a mouse click, one of the nearest neighbors can be further analyzed on the Painting Wall. In contrast to the other visualizations, the nearest neighbor cloud does not focus just on one specific artist or style. This supports the serendipitous discoveries of other paintings and new artists.

5.3.6 Timeline Wall

Later, Silke searches again for Gustav Klimt and decides to focus on his paintings on the timeline. In the timeline, she now focuses on some of his other works like "Girl with Long Hair, with a sketch for 'Nude Veritas'" and "The Virgin". It gives her an impression of the development and changes in Gustav Klimt's style and themes.

Tasks. The Timeline Wall gives visitors an overview of the development of an artist over time (Figure 5.1). It also helps to place the painting of interest in the career of the artist. Furthermore, the development of different styles or genres can be observed, and even the depiction of specific objects over time. In addition to selecting a specific attribute, it is also possible to analyze a specific time range.

Design. The Timeline Wall displays a set of images on a timeline using Timages [Jän18]. For this, the currently selected context is used. This can be the artist, the style, or the objects that appear in the images. Through this, paintings of a specific style or a specific artist can be observed over time. Paintings are placed as thumbnails on a horizontal timeline by filling in a polygonal region. While preserving the aspect ratios of the image sizes, the images are scaled and ordered decreasingly according to a custom relevance metric. The thumbnails are then horizontally placed as close as possible to the vertical center and the x-position, which corresponds to the painting's year of origin. Varying sizes help to observe images with higher relevance to the topic of interest; e.g., by default they are scaled to a similar size, but the original image size or other measurements are possible.

5.3.7 Additional Visualization Designs

We also computed a two-dimensional embedding with UMAP [MHM18, MHSG18] for each image based on the embeddings of Inception Resnet V2 [SIVA17]. UMAP embeddings are used to compute clusters for each class of objects with HDBSCAN [MHA17] and outliers with the Local Outlier Factor [BKNS00]. To allow for uncertainty in cluster assignment, we apply a soft-clustering strategy. For each image, a probability distribution for the cluster assignment is computed, which allows us to assign an image to multiple clusters, and therefore to assign outliers that are not part of a specific cluster to multiple clusters. To prevent large and heterogeneous clusters, we apply a leaf selection method for cluster selection, which is more likely to result in smaller and homogeneous clusters. To visualize the clusters later, we computed the centroid for each cluster in the two-dimensional space. In addition to dimension reduction and clustering based on the different classes, we compute the same for artists and styles, as they are the most frequent metadata. To give a better overview of a specific object class, artist, or style, we presented each cluster by the image closest to the centroid of the cluster. To prevent overlaps, we applied a collision detection that adds a small offset to the x- and y-coordinates. When a cluster is clicked, the images of the clusters move to the outer part of the space, while the cluster results appear in the center. If the cluster is too large to fit all results on the canvas, an additional soft-clustering is applied to the cluster, presenting only the centroids of each new cluster. Images can be part of multiple clusters, and by mouse-over the other clusters are highlighted. In addition, a tooltip shows the class, style, and artist distribution of the cluster. The timeline visualization can be linked to the cluster map to also allow the user to investigate the time distribution of a cluster of interest.

The outliers can be displayed in a Picture Cloud similar to the nearest neighbors of an image of interest with the distance as a scaling factor to see the most unusual paintings of an artist, a style, or a specific object class. An example of the class Elephant can be seen in Figure 5.4.

We excluded visual depictions resulting from the above-described technique due to their complex composition and the limited intuitiveness of the results. Dimensional projections of a large number of high-dimensional vectors result in images with a lot of clutter, which can be seen in Figure 5.5. In this figure, images of the Ukiyo-e style are presented, including images depicting nature and water on the upper right and humans in different situations on the left side. We also applied a cluster mechanism to reduce clutter, but the results were not satisfactory.

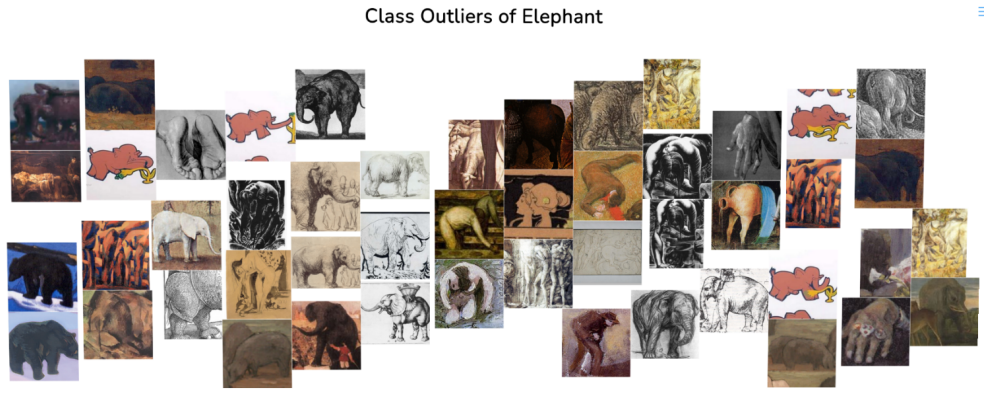


Figure 5.4: Picture Cloud showing outliers of the class Elephant. Revealing multiple wrong classified instances.

5.4 Evaluation

In order to assess if our virtual museum concept would be accepted by the general public, we conducted an informal evaluation [Twi93] suitable for the intended purpose. The evaluation took the form of an online experiment, which has been shown to deliver valuable results for visualization studies [HB10]. As we designed our virtual museum in a participatory design process to ensure the adaptability of museum visitors, we did not conduct a pilot evaluation earlier. To mitigate the disadvantages compared to controlled laboratory studies [WVL+15], we asked participants to visit the virtual museum only once. To ensure this condition, each participant had to register for a preferred time slot of two hours, in which they could freely decide when to enter and leave the museum, providing a suitable time frame for casual exploration. We asked potential participants only if they would like to visit a virtual art museum. To observe how an enhanced three-dimensional virtual art museum is perceived by visitors from the general public, we did not include any introduction to the project or any preview of our virtual museum concept or its visual interfaces. After finishing the virtual museum tour, visitors were asked to complete a questionnaire to reflect on the experience. Along with the logging data, we look at the feedback provided from different angles. The questionnaire can be found in Section 5.7.

Participants. The setup as an online evaluation helped us reach a large number of interested participants. To obtain a heterogeneous group of participants for our virtual museum evaluation to be conducted during one week, we invited scientific staff from different universities and research institutes, students, and acquaintances through mailing lists. 61 of the invited confirmed their willingness to participate. Virtual museum visitors, of whom 24 were women and 37 were men, have varying backgrounds and be-

Figure 5.5: Two-dimensional representation of the images with the style Ukiyo-e.

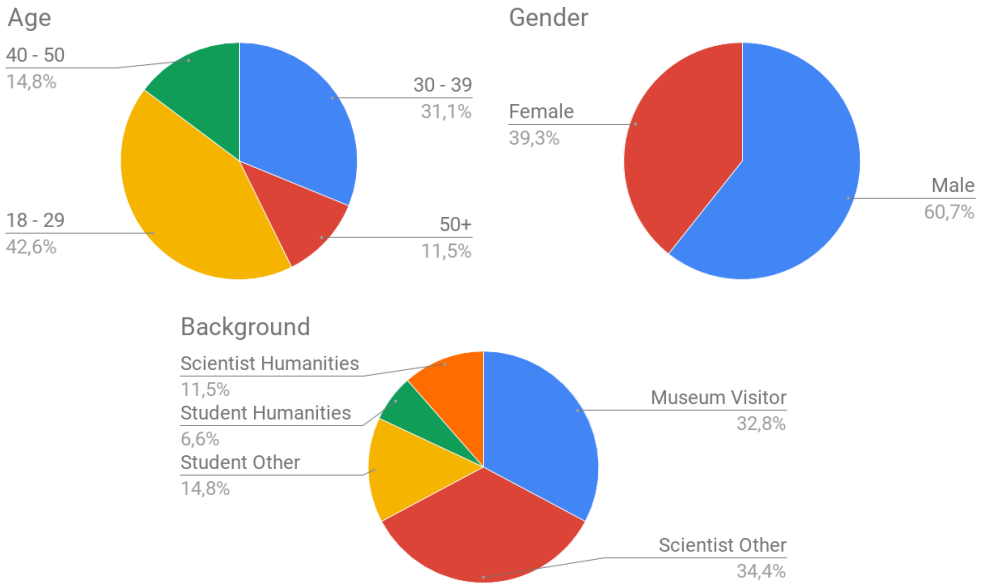


Figure 5.6: The age, gender, and background distribution of the participants.

long to different age groups. An overview is given in Figure 5.6. The age group of 18 to 29 years included the majority of participants, followed by 19 between 30 and 39 years old, 9 between 40 and 50 years old, and 7 virtual museum visitors older than 50 years. 28 visitors had a scientific background (7 with a background in humanities), 13 were students (4 with a background in humanities), and 20 were non-scientific, general museum visitors.

Logging. For each participant, we monitored all movements in the virtual museum, all interactions, and the actual duration of the virtual museum visit. For the movements, we saved the position after moving and the new rotation of the camera. Furthermore, we applied ray casting to compute the duration the participants looked at a specific image in the gallery or a wall with visualizations. For each interaction, we save the type of interaction, e.g., searching for an artist, applying a filter, clicking on a bounding box, or an image.

5.4.1 Movement & Focus

The gathered logging information of the participants gave us valuable insight into how our virtual museum is used as a whole and the individual interfaces in particular.

Movement Patterns. In order to get an overview of how museum visitors move in the virtual space, we computed a trajectory for each visitor in the form of a line string

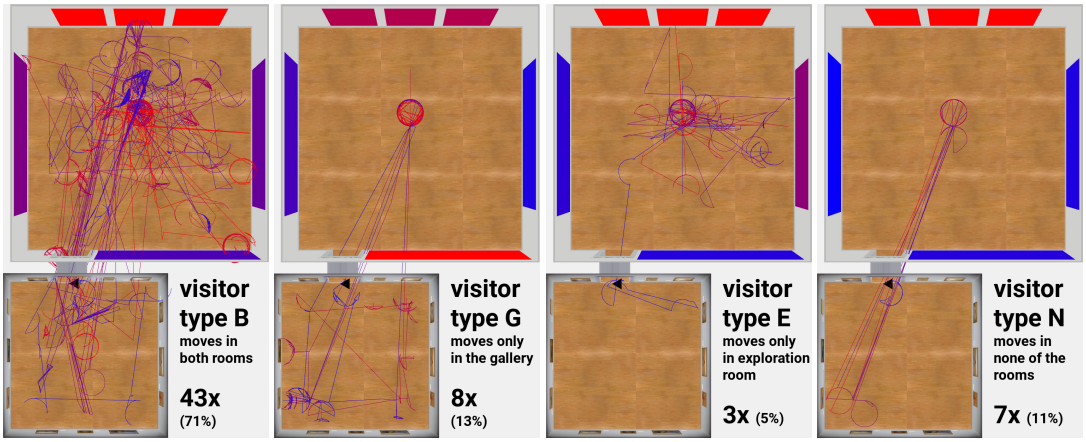


Figure 5.7: Different patterns of movements of museum visitors. The entry point is depicted with a black triangle. The color gradient of the trajectory flows from blue to red over time. The amount of time a visitor spent focused on particular walls in the exploration room is also shown on a color gradient from blue (less time) to red (more time).

that graphically depicts movements throughout the museum stay. We color-coded individual lines according to the time of movement using a color gradient from blue (entering the museum) to red (leaving the museum). We observed four different movement patterns, which are illustrated in Figure 5.7. Circular artifacts occurring in all patterns denote rotations of a museum visitor. The straight lines through the walls are the result of a direct transportation from the gallery to the exploration room.

- **Visitor type B:** 43 visitors used the entire space in **B**oth rooms to move, turn around, and explore the different visual interfaces. They spent an average of 39 minutes in the museum, which is equal to the mean duration of all visitors. The majority of the 71% of visitors showed this movement pattern; in other words, they used the degrees of freedom we provided them with, confirming our decision to offer a three-dimensional space for both rooms.

- **Visitor type G:** Eight visitors moved almost entirely in the **G**allery showing typical movements that one would expect from a real museum visit. Inside the exploration room, visitors typically remained in the initial position and only rotated to observe all walls from a distant perspective. However, the color gradient of the movement trajectory (rather blue in the gallery and red in the exploration room) and the relatively high mean of 47 minutes for a museum visit indicate that these visitors required some time to adapt to the virtual space.

- **Visitor type E:** Three visitors disregarded the gallery and went directly into the **E**xploration room. They spent less than 30 minutes each (mean: 18 minutes)

and in their qualitative reflections focused more on the technological concept of the museum.

- **Visitor type N:** Seven visitors made limited use of the three-dimensional environment, indicating unfamiliarity with three-dimensional technology. They only rotated to observe and interact with the walls, resulting in **N**o movement. The mean time spent in the museum was 25 minutes. However, all visitors of this type confirmed their interest in returning to the virtual museum.

It should be noted that movement patterns do not show a significant influence on museum visit ratings, information content ratings, or the intuitiveness of the visualizations provided. Thus, patterns tend to reflect the heterogeneous group of visitors and different behaviors in virtual spaces, as observed in previous studies [SOR04].

Focused Art. From our Gallery, 12 visitors started their tour by selecting the Danish painting "Fishermen hauling the net on Skagens north beach" by Peder Severin Krøyer (see Figure 5.8), 9 chose the Japanese painting "Print from series womens customs and manners" by Ogata Gekko. These paintings were also analyzed by most people, but none of the paintings were ignored. Except for the gallery paintings, the most analyzed paintings were "Interiør med syende pige ved vinduet" by Carl Holsøe (21), "Camille Monet in the Garden at Argenteuil" by Claude Monet (16), "Girl on the Beach" by Peder Severin Krøyer (13), "Besog hos bedstemor" by Hans Andersen Brendekilde (13) and "Sunlight in the blue room" by Anna Ancher (10). Except for the French artist Claude Monet, most of the mentioned artists have Danish nationality, like Carl Holsøe who was related to Krøyer and Hammershøi, closely connected in themes and style. Thus, these results reflect the virtual museum's capacity to "get deeper and deeper into different artists and styles", as a visitor remarked. More than 700 different images, painted by 202 different artists, have been observed on the Painting Wall. The most frequent artists who were not part of the Gallery were Carl Holsøe (48), Claude Monet (43), Salvador Dali (34), and Pablo Picasso (30). In total, the visitors searched for 105 different artists; the most frequent ones were Claude Monet (22), Vincent Van Gogh (17), Leonardo Da Vinci (9), Pablo Picasso (8) and Edvard Munch (6). The object filters applied most frequently were *Human face* (24), *Person* (23), *Clothing* (19), *Woman* (15), *Tree* (14), *Man* (11) and *Boat* (10).

Engagement with Visual Interfaces. We recorded the number of visitor interactions with each wall and the time they spent observing each wall, in other words, their overall engagement with the visual interfaces offered. Therefore, we color-coded the walls from blue (less engagement) to red (more engagement) on a linear scale; the results are shown in Figure 5.8. Based on click interactions, most activity has been registered for the Painting Wall (763 clicks), followed by the Objects Wall (325), the Timeline Wall

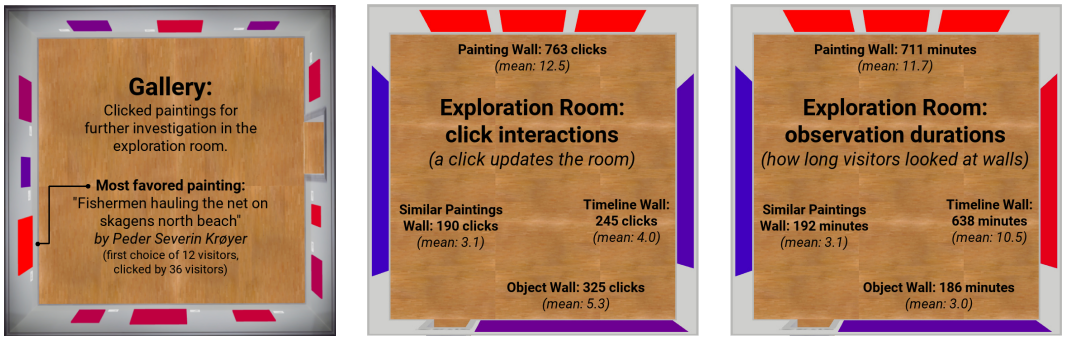


Figure 5.8: Engagement, color-coded on a linear scale from blue (less engagement) to red (more engagement), with paintings and walls in the virtual museum and the mean per participant.

(245), and the Similar Painting Wall (190). Each of these interactions changed the context of the Exploration Room, either by focusing on a new painting by the same artist on the Painting Wall, also affecting the Similar Paintings Wall, or by selecting a painting by a new artist, thereby updating the whole Exploration Room. Visitors looked at the Painting Wall for 711 minutes in total, followed by the Timeline Wall for 638 minutes, the Similar Paintings Wall for 192 minutes, and 135 minutes for the Objects Wall. The Painting Wall reached the highest values for engagement, which is not surprising as visitors initially look towards this wall when entering the room or clicking a painting elsewhere. Surprisingly, the Objects Wall registered more click interactions than the other two walls, which can be seen as an indicator of the visitors' interest in focusing on the individual elements of paintings. Furthermore, the Timeline Wall was looked at for the second longest time, which is explicable by its capacity to tell stories about the oeuvre of artists. Finally, we compare the time visitors spend in each of the two rooms. The Gallery was visited for 682 minutes (mean: 11.2), while the Exploration Room was visited for 1,727 minutes (mean: 28.3). Considering the circumstance that our invitation neutrally asked participants to "visit a virtual museum", and that the visitors started in the Gallery, those numbers underpin the value of our solution to be an important complement to real museums.

5.4.2 The Value of our Virtual Museum

We asked museum visitors to evaluate the utility and importance of our virtual museum. The results are shown in Figure 5.9. First, we asked them to rate their visit on a 5-point Likert scale from *boring* (1) to *exciting* (5). Although only four visitors rated their visit as rather boring (7%), a majority of 39 visitors found it to be a rather exciting experience (64%). The remaining 18 visitors did not express a trend. With a me-

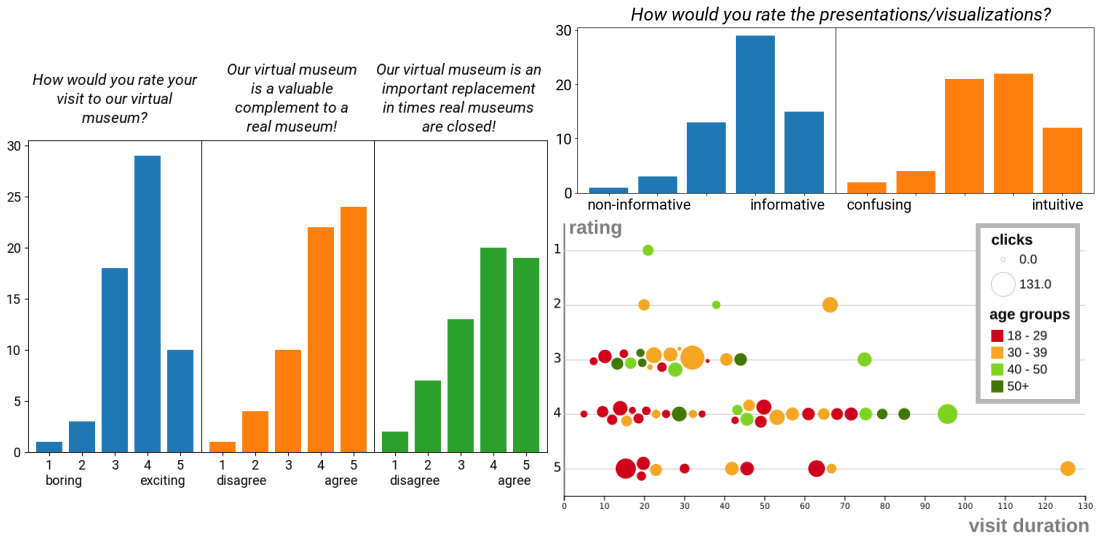


Figure 5.9: Participants' ratings on the value of the virtual museum and information content and intuitiveness of the visualizations, also excitement relates to age group and engagement.

dian of 4 (rather exciting) and a mean of 3.72, most visitors enjoyed their stay at our virtual museum. Second, we ask if our virtual museum is seen as a valuable complement to real museums. 46 visitors (75%) supported this capacity of our solution, reaching a median of 4 (rather agree) and a mean of 4.05. According to qualitative feedback, this is first and foremost related to exploratory functionality that provides visitors with new means of engaging with art. Third, accounting for the closure of real museums in times of the pandemic, we wanted to know the visitors' opinions on whether our virtual museum would be an important replacement. Although 9 visitors rather disagreed, the majority of 39 visitors (64%) expressed their gratitude for having a museum-like space in which they can appreciate and discover art. A representative comment from an excited visitor aligns with the objectives of our solution: *"The advantage of digital interaction possibilities in virtual museums compared to real classical exhibitions is immense from my perspective and increases the attraction for me to participate in exhibitions."* To obtain a more detailed overview of the positive and negative aspects of our virtual museum prototype, we asked visitors if they would return. Both positive and negative feedback are discussed below.

Positive Feedback. A majority of participants (47 out of 62, 76%) declared interest in visiting our virtual museum again, highlighting the benefits of the exploratory environment. A representative comment was *"It is exciting that you can 'find your way'deeper and deeper into different artists and styles."* One participant valued this richness of acces-

sible information and remarked that it is *"difficult in a real museum"* to link different artists and their works. Another participant pointed out more clearly that the means of exploring art are limited in real museums (*"From my perspective, this makes it more appealing than many of my previous classic visits to museums."*), which confirms the capacity of the virtual museum as a valuable complement. Furthermore, the vastness of the data set was seen as an attractive reason for visitors to return: *"There was far too little time to even come close to looking at the whole treasure trove of pictures."*

Negative Feedback. 15 out of 62 participants (24 %) indicated that they would not revisit the virtual museum. The reasons given relate to aspects that real museums provide and those that virtual ones do not. One comment addresses two limitations of our current solution: *"Compared to a real museum I miss the atmosphere and the contemplative - and on the other hand in-depth information."* First, our virtual museum does not include a curated collection. It only offers one room, mimicking a real museum, in which we arrange paintings that generate various entry points to the analysis room. On the other hand, the vast image collection suffers from incomplete metadata, a remnant of duplicates, and an error-prone object detector. Second, it does not include social aspects that emulate the atmosphere of real museums at all. Lastly, some participants expressed their general disinterest in art (*"I'm not really into paintings."*)

5.4.3 Acceptance of our Virtual Museum Concept

We analyze the information provided by the participants to discover whether any characteristics correlate with the acceptance of a virtual museum solution. In addition to the general assessment of the value of the virtual museum, we asked visitors to rate on a 5-point Likert scale if the provided visualizations were informative and intuitive (see Figure 5.9). Both aspects were positively evaluated, reaching mean values of 3.89 for information content and 3.62 for intuitiveness.

Concept Addresses Needs in Humanities Research. Museum visitors with a humanities background (seven scientists and four students) rated the experience more favorably compared to other study groups. With a mean value of 4.45 (eight times rating 5), the assessment of the museum as a valuable complement to real museums is higher compared to the entire group. This group also gave the highest scores for information content (mean: 4.18) and intuitiveness (mean: 4.45) of the visualizations presented. This might be related to the concept of *Distant Viewing* [AT19] that transfers the idea of quantitative text analysis (Distant Reading) to image collections. Using temporal and similarity analysis and extracting objects from paintings on a large-scale and making the results explorable, our virtual museum provides novel *Distant Viewing* avenues for humanities research.

Acceptance Correlates with Age and Engagement. Figure 5.9 provides a multi-faceted picture of the factors influencing the museum visit ratings from *boring* (1) to *exciting* (5). It is evident that all visitors who gave the highest rating were younger than 40 years. None of the visitors of the group with the youngest participants (aged 18 to 29 years) gave a rating below 3. This age group also reaches the highest mean values for rating the visit (4.03), the virtual museum being a complement (4.38) or a replacement during times when real museums are closed (4.11). These results may be related to the generation of digital natives feeling more at home in a virtual environment, having more interaction skills and being more technologically adept [Jar19]. However, the chart also shows that the longer the duration time of the older visitors, the better their rating. The size of the circle reflects the number of clicks, in other words, the number of images selected in the Exploration Room. As there is no clear tendency visible, one can conclude that visitors may be of different types: those who aim to gather more information and those who spend more time observing the paintings. However, what the distribution indicates is that ratings correlate with actual visit durations. Although the actual mean visit duration of visitors who rated the virtual museum as (rather) exciting (with 4 or 5) is 43.5 minutes, the other visitors spent only 28.5 minutes on average.

Perceived vs. Actual Duration of Museum Visit. Due to consistent logging of all visitor activities within the virtual museum, we were able to determine the actual duration of museum visits for all participants. Although the mean actual duration of the visits was 39 minutes with an average entry delay of 25 minutes, 13 visitors spent more than an hour in the museum. In addition to logging the time, we also asked visitors in the final questionnaire how much time they spent in the museum. 13 visitors underestimated, while only four visitors overestimated their stay. Visitors tended to underestimate the duration of their stay, especially when they spent more time in the museum. According to studies in psychology [Fra63, Fra84], active participation and higher levels of motivation lead to perceiving time as shorter than it appears to last. Therefore, we consider this to be an indicator of casual entertainment.

Does it Have to Be 3D? Two visitors remarked that the three-dimensional environment in its current form is superfluous (*"The 3D-stuff has no value for me."*), or suggested taking advantage of immersive technologies (*"The use in VR glasses would make the 3D element more important."*). In our design phase, we evaluated opportunities and drawbacks of various possible implementations of our prototype visualizations. We decided on a three-dimensional representation of the gallery to emulate the real space as best as possible. On the one hand, this consideration serves the desire of partaking visitors of our evaluation to reproduce the atmosphere of real museums. On the other hand, it is in line with existing virtual museum implementations. As these are often only available as desktop applications to reach a large audience, we did not target a vir-

tual reality solution for our prototype implementation. However, we added an option to deactivate the three-dimensional control for visitors who prefer to explore in a two-dimensional environment.

5.5 Limitations & Future Work

The evaluation results underpin the utility of our virtual museum as a complement to real museums that offer new avenues for the general public to engage with art collections. We registered some limitations of our current solution, some of which we were aware prior to the evaluation. However, our main focus was to connect quantitative analysis of paintings with virtual museums and to evaluate how visitors adopt and value such *Distant Viewing* concepts. In the following, we discuss how our concept might be improved to make virtual museums even more engaging for the general public.

A Curated Virtual Museum. Some of the evaluation participants remarked that they were missing a curated exhibition where a story is told based on a selection of objects and artworks to target a particular set of goals. For example, a goal could be to reveal new information about an artist's life and work. In this context, the search menu for the exploration room may feature custom filters and preset search parameters written by the curator, which result in visualizations that support a particular curatorial goal. As the story of an artist is inherent in the Timeline Wall, this could explain why it was observed longer than the other visualizations. Some visitors also addressed the need for "a tour guide's insight on the artwork". These results are not surprising, as public audiences prefer curated collections that have a story to tell [TGBvdB14, EE16], while the inclusion of information retrieval principles, for example, search options to fulfill different information needs, can lead to designing for an expert audience. Nevertheless, our system allows for serendipitous discoveries without expert knowledge. The general intention of our concept is to connect the Exploration Room to real museum exhibitions that are already curated. Also, we could improve the storytelling in the Exploration Room by including external sources such as Wikipedia. It should be noted that the digital versions of the paintings provide the correct aspect ratio of the real painting, the resolution is lower, and the real-world size is missing for many of the paintings from the data set.

The Social Virtual Museum. Some participants pointed out that much of the atmosphere of a real museum (the architecture of the building, background noise, or even smell) is missing, which leads to discomfort being "alone in the virtual space". Currently, our solution does not support shared visits; in other words, it does not generate or strengthen social and emotional connections between visitors. The participating vis-

itors expressed their wish to interact with other people "to discuss the art and share impressions". This lack of social aspects is common for most current virtual and online cultural experiences [VKCA20]. A solution would be to allow multiple people to visit the virtual museum together, for example, by introducing avatars to the virtual space and, therefore, allowing synchronous interactions. There is also potential for increasing degrees of "gamification" as well as the integration of social media interactions. To further include more of the actual atmosphere, background noise from real museums, including sounds of footsteps, could be included.

Personalization of Virtual Museum Visit. Currently, our interface offers a search option for artists in which visitors are interested and then filters the results by painting styles. Although we developed a series of other means to explore the art collection (e.g., searching for particular objects and comparing those on the timeline), we did not include those in the evaluation prototype in order to keep the interface as intuitive as possible. The problem of additional representations and metadata was also reported by Ma et al. [MMF19] "when designing visualizations for museums, additional representations should be carefully considered, and secondary data may need to be left out". However, some visitors wished for more personalized search functionality. A further possibility to motivate visitors to interact with the provided interfaces would be to allow them to input their own data, for example by uploading images that they are interested in comparing with our art collection [MCS17].

Moving Beyond Paintings. Our museum only focuses on paintings. Visitors expressed the desire to expand our collection with three-dimensional artifacts, such as statues, musical instruments, or tools. However, three-dimensional object reconstructions are more expensive, more complex, and more error-prone [SFKP09]. In addition, our methods are tailored for processing two-dimensional image data, which we would have to adapt to three-dimensional sources.

Cross-Depiction Problem & Incompleteness. One limitation of our virtual museum is related to the machine learning approach and the data used to train the methods. The neural networks that we applied were trained on ImageNet and OpenImages. Both data sets contain only real photographs, instead of fine-art paintings. This leads to the cross-depiction problem [HCWC15]. For example, neither an abstract cat nor a cubist cat resembles a real cat. Due to this and the general error proneness of automatic approaches, the bounding boxes of the objects in the images can be wrongly labeled, and not all objects have been detected. Also, hierarchies such as OpenImages suffer from incompleteness because they do not contain all kinds of objects. To improve automatic object detection, context-aware approaches [GRN19] that combine image features and metadata can increase accuracy. Furthermore, a crowd-sourcing approach that combines manual annotation of museum visitors with machine learning

approaches such as few-shot learning can extend hierarchies like OpenImages and can correct wrong objects. Based on the feedback from the evaluation, we started to add means of annotating images with bounding boxes, and so creating new labels. Incompleteness is given not only in the class hierarchies, but also in the data. Although we included around 200,000 paintings, there are still many missing paintings from different artists, and some artists are not yet part of the data set.

Profiling of Artists. Another future direction is to put more emphasis on artists rather than paintings. A visitor suggested including more "information about the artist and how they relate to artists from a similar group/time". For this purpose, we could include biographical information from external sources and apply profiling techniques, proven to deliver valuable results for prosopographic data sets [JFS15], with the aim of discovering artists similar to an artist of interest. Similarity can be computed based on metadata such as activity period, style, genre, or social relationships with other artists. This can even be expanded by including similarity metrics based on depicted objects, themes, used colors, and image embedding similarity.

5.6 Summary

Our work contributes to a virtual museum model that connects a virtual version of a Gallery with an Exploration Room that contextualizes artworks with a large image archive based on WikiArt. For this purpose, we take advantage of machine learning techniques to extract object information from artworks and determine similarities among them. The results are depicted as interactive visualizations that provide a novel virtual museum experience to visitors, allowing them access to paintings beyond those exhibited in a real museum. From our evaluation, we can conclude that a virtual museum is not a replacement for a real museum, but during times when real museums cannot be visited, they are a viable alternative. Our solution was further regarded as a valuable complement to a real museum, exemplified by how the interactive visualizations, composed for objects of interest, can intrigue new visitors, give them new thoughts, and address the information needs of general visitors and humanities scholars. Lastly, it is important to note that designing a virtual museum is not about replicating a real museum in a virtual space, but more about extending the notion of a museum taking advantage of digital methods.

Virtual museum visitor survey

1. **What is your age?**
   - ○ 18 - 29
   - ○ 30 - 39
   - ○ 40 - 50
   - ○ 50+
2. **What is your gender?**
   - ○ Male
   - ○ Female
   - ○ Non-binary
3. **You are a ... ?**
   - ○ Scientist (Humanities)
   - ○ Scientist (Other)
   - ○ Student (Humanities)
   - ○ Student (Other)
   - ○ Museum Visitor
4. **How often did you visit a museum in a year before the COVID-19 restrictions?**
   - ○ 0
   - ○ 1 - 2
   - ○ 3 - 4
   - ○ 5+
5. **Have you ever visited a virtual museum before?**
   - ○ Yes
   - ○ No
6. **Have you used any of the following technologies before?**
   - ☐ Augmented Reality

☐ Virtual Reality

☐ 3D Games

☐ 3D Movies

☐ Other 3D Environments

☐ None of them

7. **How long was your visit to our virtual museum?**

   ○ less than 10 minutes

   ○ 10 - 30 minutes

   ○ 30 - 60 minutes

   ○ over 60 minutes

8. **How would you rate your visit to our virtual museum?**

   Boring ○—○—○—○—○ Exciting

9. **Our virtual museum is a valuable complement to a real museum!**

   Disagree ○—○—○—○—○ Agree

10. **Our virtual museum is an important replacement in times real museums are closed.**

    Disagree ○—○—○—○—○ Agree

11. **What did you miss compared to a real museum?**

    _____

    _____

    _____

12a. **Would you visit this virtual museum again?**

    ○ Yes

    ○ No

12b. **Why or why not?**

    _____

    _____

    _____

13. **How would you rate the presentations/visualizations?**

Non-informative ○—○—○—○—○ Informative
Confusing ○—○—○—○—○ Intuitive

14. **How would you rate the navigation/orientation?**

Confusing ○—○—○—○—○ Intuitive

15. **What did you learn from this experience?**

_____

_____

_____

16. **What information about the paintings were missing?**

_____

_____

_____

17. **Do you have any suggestions to extend and/or improve the virtual museum experience?**

_____

_____

_____

Iconology ... must ultimately do for the image what linguistics has done for the word.

Ernst Hans Josef Gombrich

6

A Visual Analytics Framework for Composing a Hierarchical Classification for Medieval Illuminations

Visual material stored in public institutions is not always in the shape to be used for state-of-the-art computer vision methods. Reasons for this include limited metadata, the preservation status of the cultural object, or the quality of the digital twin. This is more acute for images in ancient manuscripts than for contemporary artworks. Historical library collections contain hundreds of thousands of ancient manuscripts that have been described over the years in different print publications and cataloging systems. In today's world of digital libraries, manuscripts are often fully digitized, but researchers in art history and book history who want to access these collections in a more granular fashion often turn to earlier iconographic databases created in the era of partial digitization of the 1980s and 1990s on account of the rich metadata they contain. These annotations in earlier cultural heritage collections provide a propitious opportunity for applying supervised machine learning methods. However, despite originating in common thesauri and controlled vocabularies and growing in size over the years, it is not a straightforward task to use these annotated images for machine learning. Not only have the original vocabularies of various collections "drifted apart", but

the collections also contain manifold uncertainties, such as imprecision, incompleteness, and non-homogeneity in both the annotated metadata and in the data of the collection itself [BEM+19]. One interest lies in organizing and structuring similar data sets with overlapping metadata that allow researchers to gain deeper insight into specific materials or phenomena, which for artificial reasons have been siloed in different cultural institutions. Additionally, more robust general connections between divergent image collections should allow greater retrieval and discoverability in the cultural heritage sector between varied vocabularies or even across multilingual metadata schemas.

Another problem with annotations is that of missing labels or inappropriate hierarchies that do not match those usually used for machine learning. The assumption that labels are independent of each other often does not represent real-world scenarios, therefore, taking into account the relations between labels can improve classification [DMG+20] or retrieval tasks [BD19]. The most common label hierarchy in computer vision is the ImageNet [DDS+09] hierarchy, which uses a subset of WordNet [Mil95] labels. This hierarchy can be problematic for multiple reasons, in particular, for historical image domains, where its natural images and hierarchies associated with them are not applicable to the kinds of images found therein. Also, relations in hierarchies like WordNet can over time be seen as no longer appropriate, e.g., abusive language, or they can include non-visual concepts that can be problematic for image annotation [YQFF+20].

As a starting point for *Distant Viewing* of medieval illumination, we applied computer vision methods to a data set of images from manuscripts of the French Marco Polo textual tradition, images that demonstrate strong visual coherence. Present in multiple manuscripts, the "Devisement du monde" is famous for descriptions of extra-European travel and the depiction of Asian cities [Cru19]. We set out to see whether repeated visual features in this image corpus are detectable by object detection, and what visualization would allow us to better understand Polo's depiction and how modern image hierarchies might be adapted to the specificities of medieval manuscripts.

We then shifted our focus to two large image databases from medieval manuscripts digitized from French libraries, Mandragore [ndF03] and Initiale [dreddtdCndlrsSdme12]. The images come from the "Paris Bible" tradition, a genre of Latin manuscript that includes the Old and New Testaments of the Christian Bible with widespread diffusion in Europe in the thirteenth and fourteenth centuries [Lig12]. Although the genre exhibits significant repetition in the themes and forms of the images across the biblical books, the two databases in question contain conflicting metadata fields that prevent us from automatically combining different labels. The process of connecting the data sets can be achieved through different approaches, including extending existing metadata using a significant amount of domain-specific knowledge, some initial normalization of

already existing metadata, as well as the creation of a logical hierarchy for downstream tasks with the images.

We designed a visual analytics system that supports combining two (partially-)annotated image data sets from the same genre but from different sources with differences and inconsistencies in the used vocabulary. In particular, we apply a semi-automatic process to unify the annotations of a subset of the Mandragore and Initiale data sets by creating a shared, high-quality label hierarchy. In our case, labeling or annotating can be seen as assigning multiple categorical labels to an image and defining relations between labels. The system allows annotating multiple images at the same time while suggesting existing annotations from the different data sets, recommendations made using word embeddings, co-occurrences of the annotations across the data sets, and image embeddings of the images in the collection. Furthermore, it allows the creation of label hierarchies appropriate for the corpus by using the given metadata and additional concepts. We decided to construct the label hierarchy from both the annotations already present in the data sets and entirely new terms, since a specific vocabulary related to the cultural horizons of the period was required, and existing external hierarchies do not include all of the domain-specific vocabulary like people and objects linked to religious practices.

Continuing our long-standing interdisciplinary collaboration [JW17a, MWJ21], we adopted a participatory design process [JKKS20] to address the problem described above. In addition to the creation of a combined image data set that is in itself a valuable resource for medievalists, the main contributions to our community are as follows:

- A **multi-layered visual analytics framework** that tailors Shneiderman's Information Seeking Mantra [Shn96] to navigate large sets of images from image embeddings to detailed annotation views.
- A **multi-view image annotation environment** that provides various visual interfaces to explore various aspects of the data to help evaluate the similarity and relatedness of images.
- A description of **user pathways** documenting various strategies on how domain experts can use such annotation environments, allowing valuable insight to be gained for related scenarios.
- A **label hierarchy for medieval illuminations** as a direct result produced by content specialists using the system. It can be straightforwardly applied to scenarios in which object detection is performed on specific historical sets of images with related themes.

Although the question of the specific medieval manuscripts represented in our corpus might seem quite specialized, the situation of divergent versions of common vocabularies and the desire to resolve and combine labels across knowledge bases is common to many research areas, especially in the humanities. Our system is designed to support various ways that subject specialists from different backgrounds look at sources (here different kinds of specialists in pre-modern culture, such as paleographers, art historians, codicologists, and philologists), and by extension, different publics in visual cultural studies, each of which has very different academic training and looks for different details in groups of images. Consequently, our solution is adaptable to related image annotation scenarios in which it is desirable to revise, create, and/or organize domain-specific labels in a hierarchical structure.

6.1 Related Works

We visualize images of cultural heritage [WFS+18] but in contrast to other works, we do not focus on image exploration [THC12, DMTS14, DPC17]. In order to include exploratory methods for multiple image data sets and their labels, set visualizations [DVKSW12, LGS+14] can be combined with other visualizations, because labeling a collection with categorical labels can be seen as defining multiple sets over the collection [AMA+16]. The focus of our work lies on image labeling, which is similar to visual annotation systems, interactive labeling, and other human-in-the-loop processes discussed in Section 2.4. Furthermore, our work shares similarities with those of profiling [JFS15] and recommender systems [PBT14].

The ability to annotate spatial regions in images is required for the application of localization methods, such as object detection. Annotation tools such as the VGG Image Annotator [DZ19] support this, but do not apply visualizations to communicate the data distribution and other features of the data. Although allowing to annotate spatial regions like bounding boxes is often important, it is not in the focus of our work, as we first want to regularize the vocabulary of the already existing annotations. Some works only support a predefined vocabulary [WHHA11], while others support textual annotations without a defined vocabulary [EB12]. In contrast to them, we use an already existing vocabulary and allow to extend and regularize it. Other work focuses more on the collaboration aspect between multiple users [QMSM17, CAB+11, MT14]. As we currently designed the system for a small number of experts who communicate with each other, we did not focus on the collaboration aspect in detail but still included some methods. Our approach shares similarities with the work on visual-interactive labeling [BZSA18]. In contrast to them, we do not include active learning strategies, but our exploratory approach can be extended to include recommendations

based on active learning. Currently, the identification of labeling candidates is solely dependent on domain experts. However, the applied visualizations help in this process by showing similarities based on embeddings and metadata. The reason to exclude active learning in the current state is the large size of the used vocabulary with over 1700 labels, the skewness of the distribution of the already existing labels and the multi-label classification setting.

6.2 Detecting and Visualizing Entities in Manuscripts of Marco Polo's Devisement du Monde

For image classification and object detection, there are large data sets with class hierarchies, such as ImageNet [DDS+09] and Open Images [KRA+20]. These data sets and their underlying hierarchies are neither particularly effective at identifying the wide variety of entities depicted in medieval manuscripts nor detecting entities well given the representational density of medieval illumination. In our work, we argue that networks trained on natural image data sets can provide both a first impression [CZ14a], and a convenient starting point for building new classes and hierarchies and can even be used to extract some initial training samples from small to medium-sized image corpora. We applied computer vision methods to a data set of some 700 medieval illuminations from seven manuscripts and built a visual interface to explore and annotate the results. We were interested in the possibility of editing the classes of contemporary hierarchies, replacing them with categories more appropriate for the period and the corpus.

6.2.1 Data & Image Processing

Each image shows a page with a visual scene that depicts different aspects of Polo's description. We applied object detection using Faster R-CNN [RHGS15] trained on Open Images [KRA+20]. The label hierarchy of Open Images consists of 600 different classes, including parent-child relations. Object detection extracts 100 bounding boxes for each image with a confidence score and a label for the detected entity. The result was 71,400 bounding boxes. Furthermore, we extracted image embeddings for each bounding box detected with an EfficientNet [TL19] trained on ImageNet [DDS+09]. For image embeddings, we use faiss [JDJ21] to query the most similar bounding boxes for each example based on the Euclidean distance between embeddings. This allows us to see parts that are more similar to that of another image to an image of interest.

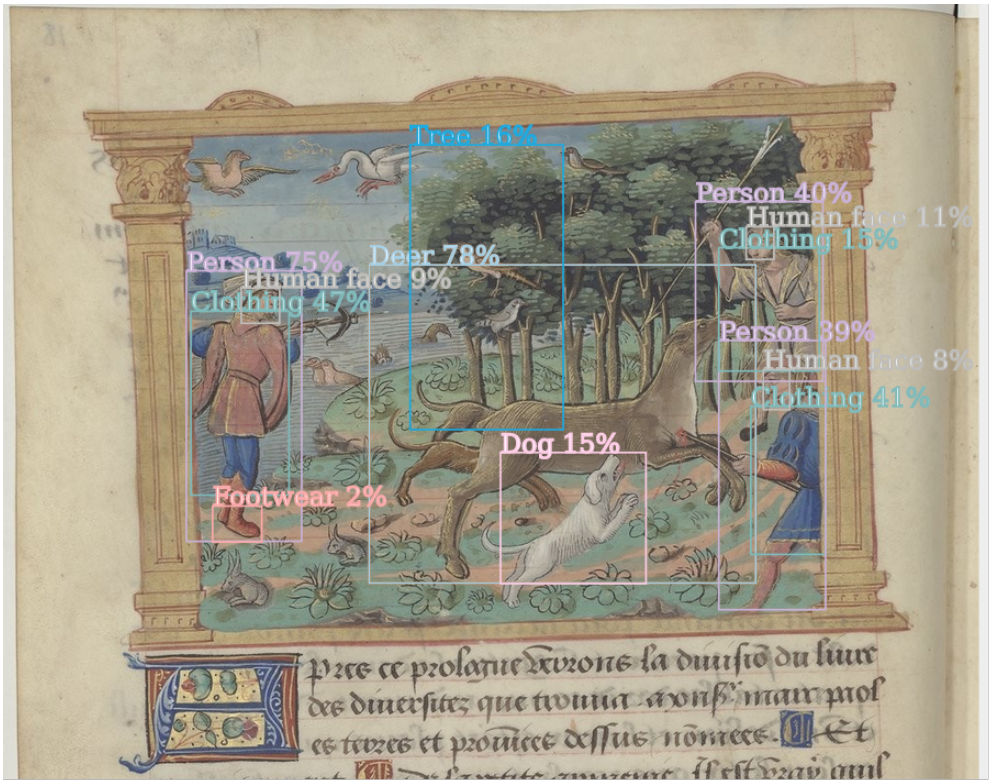


Figure 6.1: A page of the data set with entities found by a neural network in an illumination from BnF Arsenal ms 5219.

6.2.2 Visual Interface

The design of the visual interface facilitates the exploration of the image data set and comparison with different representations of specific entities. For this, the object classes can be accessed through a Tag Cloud where frequency is encoded by font size, or through a tree that visualizes the Open Images hierarchy with all classes found in the Marco Polo data set. Such interfaces for visual exploration and annotation allow the professional viewer/reader to focus on a given interest to annotate new areas or investigate objects found inside the image, delete them, or even edit their labels [SLB+09]. To prevent visual clutter, they can filter by confidence value and select or deselect specific classes. When focusing on one specific bounding box, it is also possible to display the bounding boxes that intersect, that are inside or outside the box of interest. Figure 6.1 shows a page of the data set with the entities found by the neural network.

For a given object class, all depictions are displayed in a two-dimensional grid ordered by the confidence score assigned by the neural network. Examples can be seen

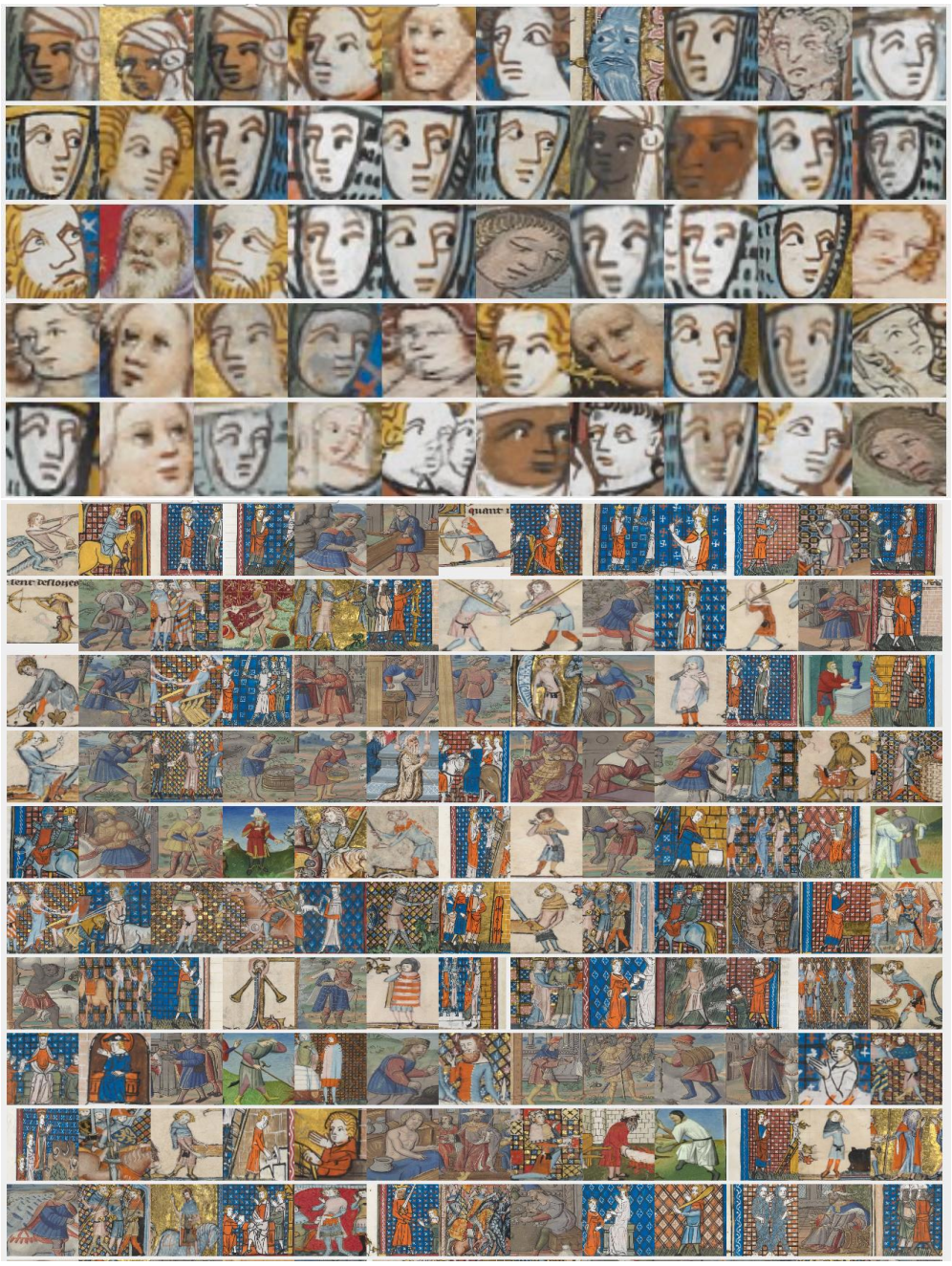


Figure 6.2: An overview of samples of faces and human figures marked with the highest confidence scores.

Figure 6.3: The TagPie gives an overview of the classes found by the neural network (green) and those from human annotations (purple).

in Figure 6.2. The interface is designed for both discovery and revision by clicking on a bounding box of interest, which leads us to see the most similar bounding boxes. It is also possible to select multiple bounding boxes and delete or re-label them in case of an imprecise classification. Furthermore, the expert viewer can annotate areas in the image with new classes, thus contributing to a new category in the Tag Cloud (Annotated) and transforming it into a TagPie [JBR+18], seen in Figure 6.3.

6.2.3 Discussion

Whereas some anachronistic categories remained throughout the output of the initial system, other objects such as those mentioned above in Figure 6.2 led to quite convinc-

ing recognition. Furthermore, summary views of the visual interface proved particularly effective in demonstrating the tensions found between the codified visual languages of medieval French manuscripts and the diachronic innovative attempts at representing "unprecedented images of the world beyond Europe's borders", as well as domains in which patterns in those tensions were particularly pronounced [Cru19]. On the other hand, the interface created to explore, revise, and manipulate features in the Marco Polo visual corpus provides us with a stepping stone for working with larger visual corpora built from across the global Middle Ages. As our inquiry evolves, finding ways to guide the viewer from the extracted objects and their computed confidence levels back to full images and relevant metadata will be crucial to allow sufficient contextualization to facilitate interpretation. Furthermore, our current method for revision and addition of labels is open-ended, but in future work, we intend to lead the annotation toward established art historical vocabularies to ensure future discoverability. Future work will also focus on ways to achieve the "best of both worlds", allowing research to move from the modern to the medieval, that is, for contemporary hierarchies to be adjusted and augmented by domain- and period-specific terminology with the support of expert knowledge.

Creating this visual pathway for visual exploration and hypothesis generation using computer vision techniques is not a trivial task, since the metadata of legacy databases of manuscript illumination (Mandragore, Initiales, Digital Scriptorium, etc.) also vary in both size and granularity. Furthermore, methodologies are needed to combine or unify the vocabulary of different data sets, bridge the gap between general and domain-specific vocabularies, and create expert hierarchies of entities found in manuscript illumination to create appropriate training data sets to deal with cross-depiction issues [HCWC15]. This leads us to focus on legacy databases such as Mandragore and Initial. We then started applying computer vision and visual analytics methods to the Paris Bible data set.

6.3 Paris Bible Project

Up until the early thirteenth century, manuscripts were mostly produced in workshops attached to courts or by monks in monasteries, but the creation of universities in medieval Europe greatly influenced this form of written, cultural production. "Paris Bibles" emerged in the thirteenth century Europe as a mass-produced written object in response to new forms of literacy, namely teaching and preaching. After 1220, these hand-copied Bibles contained a corrected text and followed a standard order, introduced by prologues and divided into chapters, usually including related series of illuminations (that is, hand-painted images) and decoration particularly collocated with

the prologues and chapter beginnings. At first glance, the images found in Paris Bibles seem similar from one manuscript to another, yet a closer look shows how different these illuminations are, from the use of colors and the details of representation, presence or absence of particular objects or people, differences which can be attributed to the origin of the manuscript, to the individual illuminator or workshop. Examples of the first book of kings can be seen in Figure 6.4.

Our data set of images from Paris Bibles is made up of a small subset of the Mandragore [ndF03] and Initiale [dreddtdCndlrsSdme12] databases corresponding to the examples of Paris Bibles; it contains respectively 1.633 images from 53 manuscripts and 11.472 images from 241 manuscripts. Each digital image illustrates one or two pages of a specific manuscript and can contain one or several illuminations depicting various scenes and objects. In addition to the images, the data set also includes general geospatial and temporal information on the manuscripts, a topical description of the images, and tags indicating the book of the Bible depicted.

Mandragore is an iconographic database of medieval manuscripts created in 1989 at the Bibliothèque nationale de France. It describes the decoration of more than 200.000 manuscript descriptions. It is based on a controlled vocabulary of 20.349 labels, 530 of which are used in this subset. This vocabulary was originally based on "Thésaurus Garnier" [Gar84]. It continues to be enriched on a daily basis by curators, librarians, and researchers. In a later stage, it was enriched with the aim of identifying, in each illumination, all the objects, places, people, and iconographic subjects represented. There is both descriptive and interpretive vocabulary included.

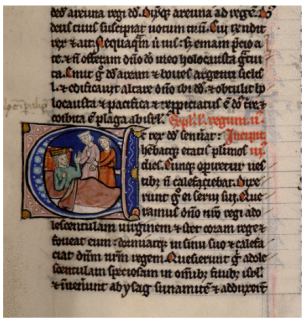
Although both databases contain samples of Paris Bibles, they were, in fact, created at very different times and with different priorities. Created in the early 1990s **Initiale** is an online catalog of medieval manuscripts belonging to the public libraries of France, beyond the Bibliothèque nationale de France. Initiale includes about 10.000 manuscripts and more than 90.000 illuminations from these manuscripts. Developed with research in mind, it uses a refined iconographic index with a controlled vocabulary and offers art-historical analyses of the decoration. This vocabulary was also originally based on the "Thésaurus Garnier" [Gar84] but evolved in a different direction over the last 30 years [Lal01]. Our subset uses 1734 words from this controlled vocabulary, only 279 of which are shared with Mandragore.

While deep, the divide between Mandragore and Initiale is an artificial one, owing to the history of institutions and collections, rather than the original historical material. This is far from optimal for scholars who would like to see relationships in the larger picture of medieval Bibles. Moreover, similar images of the same unit of text have completely different attention paid to them, which can also be seen in Figure 6.4. Since the two databases use a different vocabulary, a desirable endpoint is not so much

Latin 18

abishag, bed, bethsabee, david, king, natân, prophet, sick

Latin 17947



abishag, bed, david, king, sick

Paris, Bibl. Mazarine, 0038

abishag, arms forward, bedridden, biblical scene, crown, david, hand covered, hand on chest, hand on shoulder, historiated initial, leaning, lying down, open hand, pillow, raised hand, servant

Tours, BM, 0008 (tomes I-II)

bed cover, bedding, biblical scene, cap, crown, david, garment, hand to his cheek, historiated initial, look, pillow, raised hand, servant

Figure 6.4: Examples of the book Kings 1 in the data set with their annotations. The upper ones show images from Mandragore, while the lower ones show images from Initiale. These examples show the variation of the images in the data set. Even in these cases where the same scene is depicted, preservation status and the background colors vary. They also, show how Mandragore and Initiale focused on different concepts in the images. For example, Initiale includes positions and gestures.

a fuller description of the scenes depicted like *creation* or *death*, but a more granular depiction of details found in the images that "make up" the scene like a cross, a certain species of bird, a blue background, or a desk. What has been required in the process of connecting the data sets is a significant amount of subject knowledge and some initial

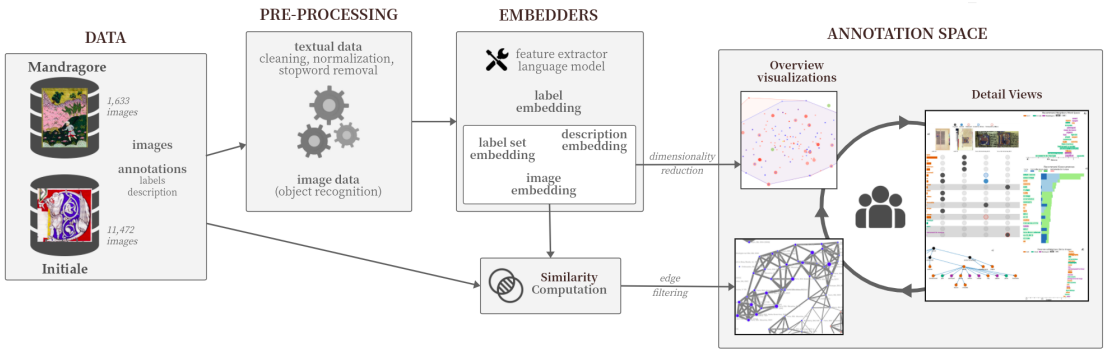


Figure 6.5: Systematic overview of our semi-automated image annotation workflow.

normalization of some of the metadata. The resolution of metadata led to an expanded code for books of the Bible, based on the OSIS Book abbreviations following the SBL Handbook of Style [Ale99].

6.4 Visual Analytics Framework

The goal of our visual analytics system–composed of multiple parts–is to allow annotating medieval illuminations and composing a label hierarchy of the objects depicted in the images. A systematic of our annotation framework is shown in Figure 6.5. The first part concerns pre-processing of textual data. Pre-processing of image data in the form of object recognition would be valuable; however, high-quality label hierarchies describing medieval illuminations do not exist, and object detection or image classification based on modern hierarchies like Open Images [KRA+20] or ImageNet [DDS+09] fail. Only a few objects that are depicted inside our image data set are found, since many classes of these hierarchies consist of objects that did not exist in the given time period. The pre-processing is followed by applying machine learning methods to generate vector representations in order to compute image similarities. After these processing steps, the annotator requires an entry point to the data. In our case, we aggregate the different facets of the data according to the manuscripts to which they belong. The relations between the manuscripts are then visualized. In addition to the manuscripts, we also needed to present and filter the actual images in the data set. For this purpose, we apply dimensionality reduction methods based on embeddings of the images and their metadata. The next step is to enable the inspection of a specific subset of the image data. The subset can be selected on the basis of the same or similar metadata or relations in a vector space. This is needed to allow one to annotate images by adding missing concepts and objects that are depicted, but not yet labeled as such. This task

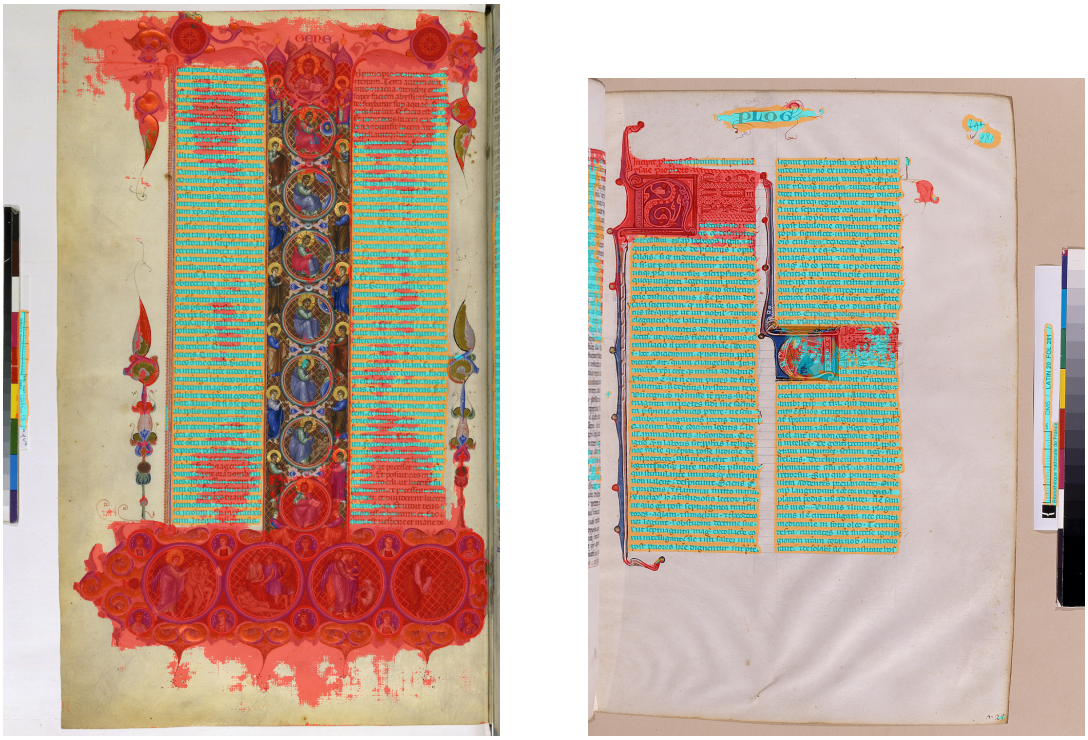


Figure 6.6: Two images of the Madragore data set with image segmentation results of the docExtractor. The red areas are detected as images, the blue areas are detected as text paragraphs, and the orange areas are detected as text borders. The results are not bad but too error-prone to extract all images automatically and study the illumination in detail.

can be supported by different recommendation methods like word and image similarities, co-occurrences or even active learning methods. The annotation space also includes an interactive graph to support creating a high-quality label hierarchy for medieval illuminations to generate a valuable resource for medievalists.

6.4.1 Image and Text Processing

The first step of our framework consists of processing images, textual metadata, and annotations. For image pre-processing, we tested multiple methods to extract the illuminations from the images. Although the method in Grana et al. [GBC11] based on the Otsu algorithm [Ots79] showed good results on subsets of the data set, it was not adaptable to the entire corpus due to the varying preservation status and background colors of the manuscripts. Furthermore, we tested pre-trained models of the docExtractor [MA20] but the results were not satisfactory for the entire corpus. Two examples

showing some problems can be seen in Figure 6.6. Due to this, we use the entire image for the next processing steps.

For the next image processing steps, we apply the EfficientNet B7 [TL19] that was pre-trained on ImageNet [DDS+09]. We use the top layer of the network to compute the image embeddings for each image in the data set. These embeddings are vectors with a dimension of 2560 and are used to compute similarities between the images, i.e., the nearest neighbors. Although the network was trained on natural images, the features can still be used to compute similarities between images in the corpus [CZ14a]. All embeddings are added to a *faiss* [JDJ21] index structure, where the similarity is based on the Euclidean distance between them. For each image, the k most similar images can be queried by searching the index based on the vector of the image. We also use the nearest neighbor search to find duplicates in the data sets that could be a result of the structure of the data set or the crawling process. All neighbors with a Euclidean distance of 0 are treated as duplicates. In the cases where the same image has different entries for the same metadata attribute, the entries were combined.

For the pre-processing of the textual annotations, we lowercased all words, and removed diacritics and special characters. Additionally, for the computation of embeddings, we did not include stopwords from annotations with more than one word to prevent high similarity in cases where stopwords overlap. We apply two pre-trained models for modern French *fastText* [BGJM17, GBG+18] and *CamemBERT* [MMOS+20] to embed the annotations. For the *fastText* model, we use word vectors, and in cases where an annotation is composed of multiple words, we compute an average vector. For the *CamemBERT* model, we apply a mean pooling to the hidden state embeddings of the neural network. We also add the vectors to a *faiss* index. Some of the images are also annotated with a sentence that describes the images for which we also compute embeddings. The *fastText* vectors seemed to better grasp the word relations based on the nearest neighbors of the annotations. Because most annotations are single words, we disregarded the *CamemBERT* embeddings, although they could be better in other application scenarios where the annotations consist of multiple sentences.

We also compute for each image in the data set three types of two-dimensional embeddings with UMAP [MHSG18, MHM18], which we use to visualize the images in a two-dimensional space. The different UMAP embeddings are based on the image embeddings, the word embeddings of the annotations of an image, and the embeddings of the descriptions. In order to ensure that the same data results in the same embeddings, we use a fixed random state for the computation. This is important because the annotation embeddings of the images change with successive annotations by the domain expert. It is also important to consider the time it takes for the different methods to

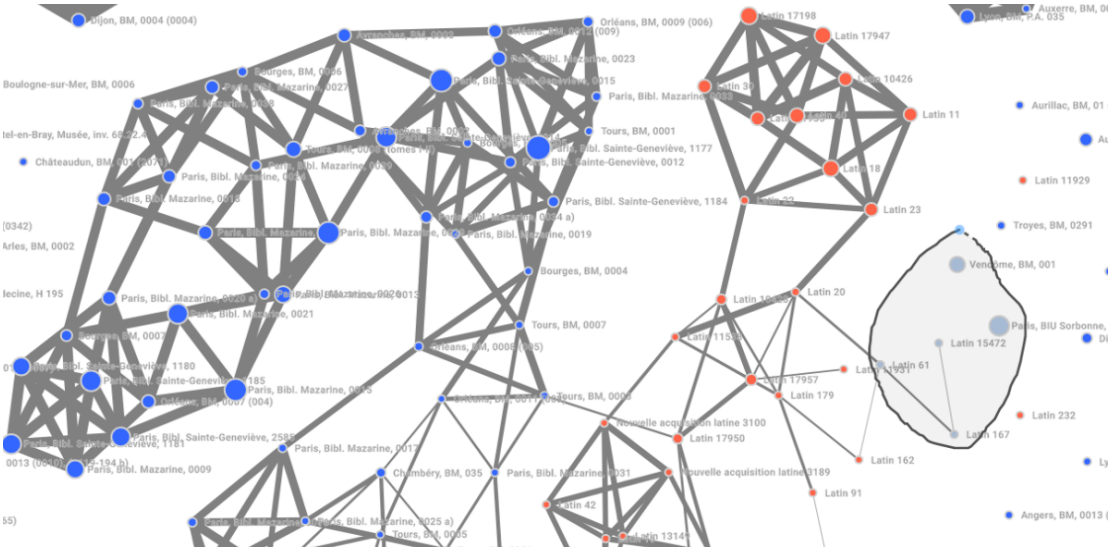


Figure 6.7: An excerpt of the manuscript graph based on annotation similarity. Blue nodes are part of Initiale and red nodes are part of Mandragore. Showing the separation of both data sets and the similarity between the manuscripts in the respective data set. The grey area shows the selected manuscripts, some of these are connected and some without a connection.

compute. It would also be possible to include other types of embedding, such as color or other textual metadata.

To compute the similarities between the manuscripts in the data set, we apply multiple similarity measurements. To define an image similarity, we compute an average vector for each manuscript based on its image embeddings and then compute the Euclidean distance between all manuscripts. The annotation similarity is defined in a similar way, where we use the average of all word vectors of annotations that are associated with a manuscript. We do the same for the textual descriptions. For each type of measurement d we save the maximum distance $maxDis_d$ in order to transform the distances into similarities in the range 0 and 1 by computing $\frac{maxDis_d - dis_{m_i,m_j}}{maxDis_d}$ for each distance between two manuscripts m_i and m_j. Combining image similarity with annotation similarity or description similarity can be helpful for multiple reasons; similar images depict similar scenes, and similar scenes are likely to have similar annotations or similar descriptions. Also, not all images in the data set are annotated, so the image similarity allows us to include them.

6.4.2 Manuscript Graph

The domain expert wants to inspect the annotation status of the data set. For this, they selected the annotation similarity to compute the layout of the manuscript graph. They can see multiple small clusters of manuscripts of the Initiale data set and one big connected component of Initiale manuscripts that are connected through a few edges with a large number of manuscripts from the Mandragore data set. They also noticed that several manuscripts are not connected to any other manuscripts in the graph. For now, they decide to focus on a subset of these manuscripts by selecting them.

Tasks. The manuscript graph is the entry point into the visual collection that allows the selection of manuscripts of interest for which images can be accessed. The graph shows similarities between the manuscripts in the data set based on several features such as image similarity, annotation similarity, or other metadata.

Design. The manuscript graph (Figure 6.7) displays the manuscripts as nodes and connects them by edges based on user-selected similarities. The nodes are color-coded according to the data set to which they belong, and the size indicates the number of images that belong to this manuscript. Next to each node, a text label shows the name of the manuscript. The graph layout is a force-directed layout in which the nodes repel each other and the edges pull the nodes together. The edge thickness shows the selected similarity value, on a scale from the selected threshold to 1.

The domain expert can select one or multiple similarity metrics for the graph, such as image similarity, annotation similarity, or description similarity. When two or more similarities are combined, the edge value corresponds to the average of the values. To avoid visual clutter, it is possible to select the maximum number of edges a node can have and filter the displayed edges based on a similarity threshold, which filters all edges below the threshold. If more than the selected number of edges satisfies the threshold, only those with the highest similarity value are displayed. Before selecting an additional similarity metric, the metric can be added as a graph overlay to see how the metric would impact the graph. Overlay edges have no impact on the currently displayed layout; new edges are colored blue, and edges that would disappear are colored red. It is also possible to drag nodes to another position and zoom in and pan the graph. In order to explore a specific subset of manuscripts, the domain expert can use a lasso selection to draw a polygon around the nodes. The selected nodes are then increased in size and colored steel-blue. A drawer on the left side that can be toggled shows a bar chart, a timeline, and a tag cloud with information on the metadata of the selected manuscripts. It is also possible to show places in the graph where images do not have annotations by clicking the recommended button.

Usage Scenario. In the manuscript network based on annotation similarity, we would avoid drawing a perimeter around several already connected manuscripts, since

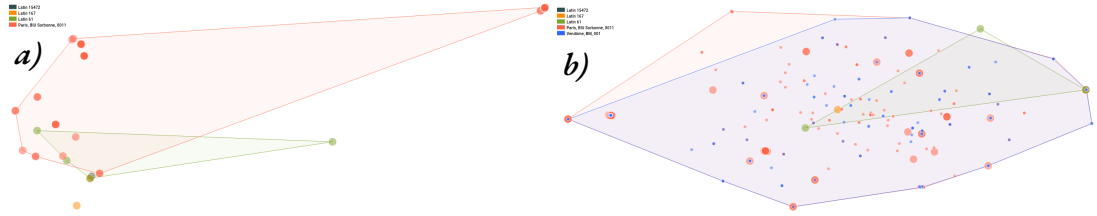


Figure 6.8: Point cloud of images based on the embeddings of the annotations (a). Images without annotations are not visible. After selecting some of the images and changing the used embeddings to the textual description (b) several images without annotations are displayed next to or on top of already annotated images. This allows to find sets of images with the same description i.e. the same or similar content where some images are already annotated and others are not.

this would only reinforce the existing connections in the metadata. Instead, we would select a group with a trade-off of connected and unconnected points. Such an example can be seen in Figure 6.7. If the number of selected manuscripts were large, we would filter them in the next step by the book of the Bible. This provides us with a wide range of possibilities, and also the risk of not finding many connections.

6.4.3 Image Point Cloud

The domain expert first also selects annotation similarity for the point cloud. It is now visible that not all selected manuscripts contain images with annotations. The domain expert selects multiple images, thus increasing the size of the circles (Figure 6.8a). Then they add the image similarity to show all of the images of the selected manuscripts, which reveals more than 100 images that are not annotated. In order to find a good starting point, they change the similarity to the description similarity, which shows that some of the images of the manuscript Vendôme, BM, 001 that are not annotated have almost the same description as some of the annotated ones. They select one of the clusters of similar images to start annotating (Figure 6.8b).

Tasks. The purpose of the Image Point Cloud is to give an entry point to the annotation process by showing two-dimensional representations of the images of the selected manuscripts, so that similar images are presented close to each other in the space. Filtering mechanisms based on metadata and coordinates allow one to focus on a specific subset of images that can be selected for annotation purposes.

Design. The images of the selected manuscripts can be explored in a point cloud below the graph (Figure 6.8), where each circle represents an image. For the two-dimensional representation of an image, UMAP is applied to multiple embeddings of the data sets. The domain expert can select and combine embeddings based on the images, the word

vectors of the annotations, and the description of the images. When combining multiple embeddings, the average is used.

Each selected manuscript has a different color, which is displayed in a legend together with the manuscript name. To highlight the positions of the images of a specific manuscript, the convex hull of the points is drawn as a contour. For this, other designs based on reduced convex hulls are possible, such as a butterfly plot [SSZW08] to reduce visual clutter in the case of convex hulls that overlap. The convex hull can be toggled by clicking on the manuscript in the legend, and it is possible to zoom in and pan the point cloud. In the case of similar or the same two-dimensional representation for multiple images, or when multiple manuscripts with a large number of images are selected, this can result in overplotting. In order to solve this, it is possible to filter the displayed points based on metadata such as text units and annotations, or by drawing a rectangle around them, which recomputes the layout. On mouseover, the image is shown. A lasso selection can be used at the points to select a set of images for the annotation process. The selected points are increased in size to better highlight them. When the drawer on the left side is opened, the domain expert can click on a button to switch to the annotation space. The current state of the graph and the point cloud are saved, so when the domain expert is done with the annotation of a subset, they can go back to the same place they worked on before.

Usage Scenario. A possibility of knowing where to start would be to target images in the same units of text. For this, we would filter by book of the Bible. If our goal is to work progressively by a category, say one book, we choose more images at this point. Another option is to select a group of similar images.

6.4.4 Annotation Space

When inspecting the selected images of both manuscripts the domain expert noticed that both sets consist of several depictions of people with swords. The only annotation presented is "épée" (a type of sword). After annotating some of the images with missing information, they decided to go back and focus on another part of the manuscripts. When inspecting the now selected images, they notice that one of the images is annotated with "hirondelle" (swallow) and another with "paon" (peacock). One of the recommended words is "oiseau" (bird). Furthermore, one of the other selected images has the annotation "harpe," (harp) with multiple recommendations on musical instruments. They decide to create an edge between these annotations in the label hierarchy.

Tasks. The annotation space allows the domain expert to see the current annotation status of a number of images and to add and remove annotations. The domain expert can zoom in on the details of the images in order to annotate them. It is also

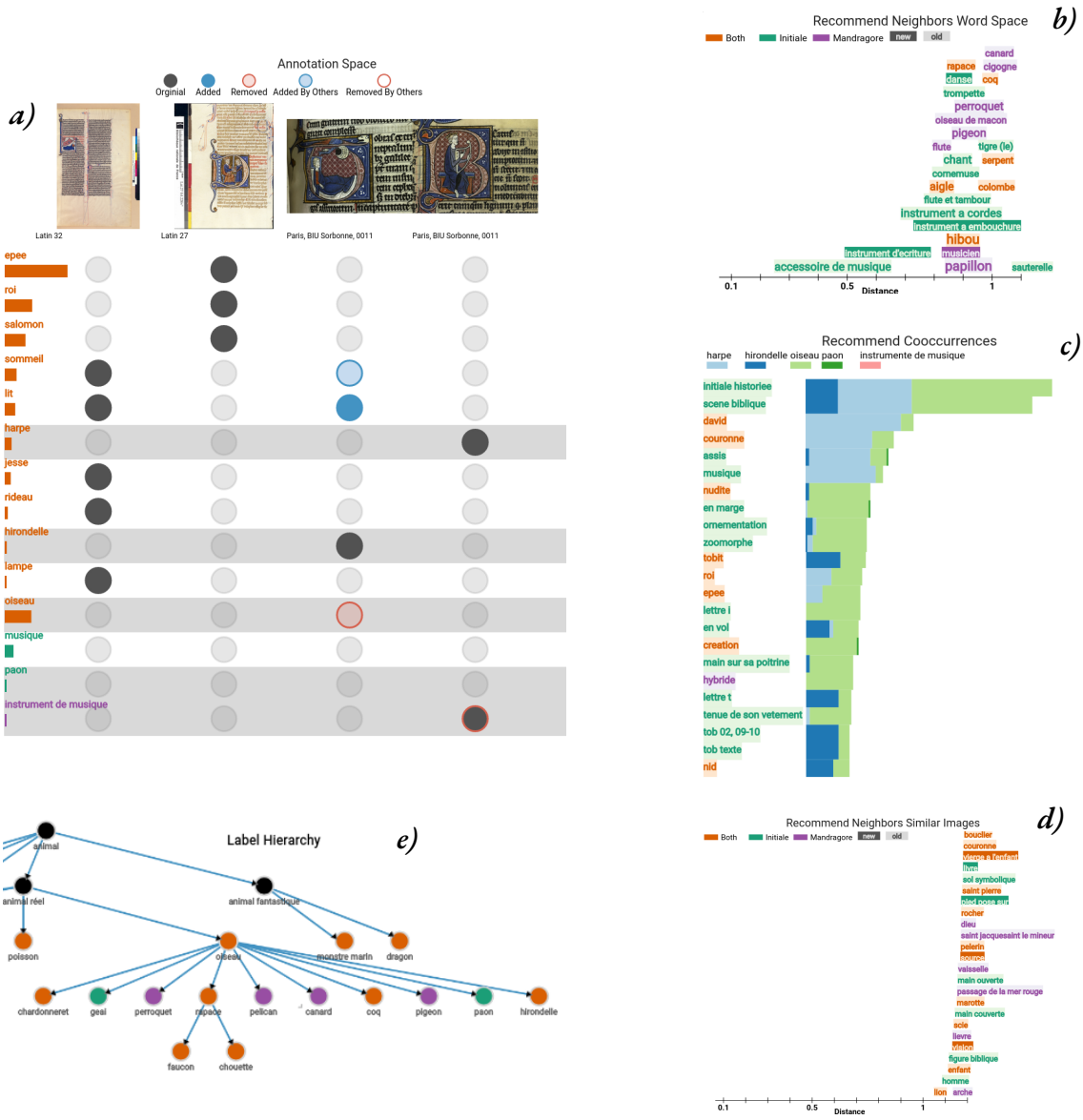


Figure 6.9: The annotation space (a) shows four manuscripts and their annotations. Some annotations were added by different users and some were removed for more specific ones. The word space (b) shows words that are similar to the ones currently selected in the annotation space. It is visible that after the selection of "instrument de musique" multiple words related to music were added. The recommended co-occurrences (c) show for example related terms like "couronne" (crown), or "musique". The similar words from the most similar images (d) contain some entries about different animals, but a lot of the terms seem rather general and are not that similar to the selected words. A reason for this could be that not images with a similar scene are found as nearest neighbors but images from the same manuscript i.e. with a similar background color. The excerpt of the label hierarchy (e) shows multiple types of birds (oiseau) that were connected by the user.

possible to add new labels that are not part of the data set and to filter images based on metadata like the same unit of text. Furthermore, the annotation space recommends possible labels, communicates why they are recommended, and to which data set they belong to. Furthermore, it helps to construct a label hierarchy and to inspect the current status of the hierarchy.

Design. In the annotation space, the domain expert can filter the currently selected images, based on metadata, to select a subset to focus on for annotation. The selected subset of images is placed on the top and their position is fixed to allow them to be seen when scrolling through the list of annotations. On the left side (Figure 6.9a), the current annotations of the images are presented together with a bar chart to show the number of appearances inside the corpus, color-coded based on the data set. Annotations are sorted on the basis of the number of selected images they belong to in descending order. When annotations belong to the same number of images, the number of occurrences in the whole data set is used, also in descending order. When clicking on the bar, the domain expert can see other images that are annotated with this annotation in a pop-up. For each word and image pair, a circle is shown that is either black to show annotations or gray and less saturated in the case where no annotation exists.

When clicking on a circle, an annotation can be added or removed. An added annotation is shown as a blue circle, and a removed annotation is shown as a red circle with less saturation. Annotations that were added by other users are shown in a less saturated blue and with a red border if an existing annotation was removed. The legend of the colors is provided at the top. When the domain expert is done, they can save the annotations to the database. All saved changes can be inspected in a history pop-up showing the timestamp, the user, and the changes. This allows one to keep track of the own interactions and the interactions of other users. To better examine an image for potential annotations, the image can be viewed in high resolution in a pop-up by clicking on it. If a specific word is not of interest to the annotation process, it can be removed from the annotation space by clicking on it. The other visualizations are linked to the annotation space, so that an update of the annotation space also updates the other views. To select one or multiple words for the other views, a row in the annotation space can be clicked, which is then highlighted in a dark gray. By default, the content of the other views is based on all the words in the annotation space.

On the right side, words are recommended based on multiple criteria (Figure 6.9b). The first visualization displays the words most similar to the currently selected words in the annotation space, similar to the word space view in the iteal system in Subsection 4.3.3. Each word is placed on the x-axis on the basis of its minimum distance from the target words. We apply a collision detection to adjust the y-coordinate to prevent words from overlapping. The words are colored according to the data set to

which they belong. To highlight which words are added based on the last selection in the annotation space, we use a less saturated color to present either the background for old words or the font color for new words. When hovering over a word, the words in the annotation space that are the most similar are shown in a tooltip together with the distance to give a reason why a word is recommended. We use the same visualization to show the most similar annotations of the most similar images of the currently selected images (Figure 6.9d). This can be helpful in cases where the selected images do not have annotations, but similar images do.

In Figure 6.9c the words that co-occur the most frequently with the currently selected words are displayed. The words are ordered based on their total number of co-occurrences. To distinguish between the co-occurrences with different words, we visualize the occurrences with a stacked bar chart for each word. The bars are ordered based on the ordering in the annotation space. The legend of the colors is given at the top. Because the number of annotations can grow really fast, the colors repeat themselves every 12 steps. Although this is not a problem when selecting a small number of annotations in the annotation space to get recommendations. On hovering over a rectangle, the number of co-occurrences is displayed in a tooltip. Similarly to the previous visualization, we use a less saturated color to either present the background for old words or the font color for new words.

Usage Scenario. Putting multiple images together raises interesting questions about how annotation could be improved (or simply changed) when noticing differences across multiple examples simultaneously. Due to screen space limitations, the annotation space accommodates about five images at a time for labeling without scrolling. We choose the desired number of images one at a time according to different criteria: the same book of the Bible from different manuscripts or images coming from the same manuscript, both of which are assumed to have common metadata, although proceeding by book seems most efficient for this stage of the process of connecting the data sets. First, we make sure that the selection of images does not include doubles; then we build our set of images for annotation. Then we temporarily blacklist any extraneous annotations. Choosing one of the possible annotations, we open each of the images to see if there are missing labels, and if so, we add them, referring to the recommended word space and co-occurrences.

6.4.5 Label Hierarchy

The domain expert creates an edge between the annotations "hirondelle" (swallow) and "oiseau" (bird) and also between "paon" (peacock) and "oiseau" (bird). While swallow and bird are annotations that appear in both data sets, peacock only appears in the Initiale data set. Fur-

thermore, they add the parent concept "animal", which was not present in both data sets. They also added a relation between the "instrument de musique" (musical instrument) and the recommended musical instruments. As "instrument de musique" only appears in the Mandragore data set, they decide to draw an edge to "musique", which only appears in Initiale. Therefore, we connect the annotations that use different vocabulary from both data sets with each other.

Tasks. The label hierarchy view shows the current state of the underlying label hierarchy of the data set. Furthermore, it is possible to modify the hierarchy by adding new nodes and creating edges between nodes. This allows one to classify metadata into categories like themes or objects and also to connect metadata from different data sets.

Design. To draw a label hierarchy, we use the Sugiyama framework [STT81] to draw a directed acyclic graph. In the first step, it is checked with a depth-first search for each node if the graph contains cycles. If there is a cycle, i.e. an edge that leads to an already visited edge, the edge is removed from the hierarchy but added back after the layout is computed. For more complex hierarchies, methods to include domain knowledge when removing cycles [SAN+17] would be helpful. In the next step, nodes are assigned a layer for which we use the Network Simplex method [GKNV93] to minimize the length of edges in the graph. Then the nodes are ordered on the layer they are assigned to reduce edge crossings, for this additional dummy nodes are added to the internal computation to replace edges that span more than one layer. For this, we use a top-down one-layer crossing minimization approach that orders the nodes based on the aggregation of their parents' indices. In the last step, the nodes are assigned a specific coordinate using the quadratic programming approach, while minimizing the distance between the connected nodes, the curvature of the edges, and the distance between the disconnected components. For each step, other methods of the Sugiyama framework are possible.

After the layout is computed, the nodes are color-coded based on the data set to which they belong. An excerpt of the hierarchy can be seen in Figure 6.9e. The edges removed from the cycle detection are drawn in red to indicate to the domain expert that there is a conflict that should be resolved to preserve the acyclic structure of the label hierarchy. At the beginning, all annotations that belong to the currently selected images are shown together with their ancestors and descendants in the hierarchy. It is possible to add new nodes to the hierarchy by selecting words from the recommendations, searching for a specific word in the data set, and adding a new word. The nodes of words that are not part of the data sets are colored black. To modify the hierarchy, the domain expert can draw a yellow edge from one node to another and click on an edge to remove it; after each operation, the layout is recomputed. After saving, the

changes are also added to the history, allowing other users to investigate them. Edges that were drawn by other users are presented by a dashed line in the label hierarchy.

Usage Scenario. In examining the annotations, if there are related ones, then we use the label hierarchy creation view to create child and parent relationships between them, or by adding new parent categories that are not presented in the data set to link metadata from both data sets.

6.4.6 Feedback Computation

New annotations are added in real-time to the data set, which also updates the annotation co-occurrences and can result in new recommendations in all views. Because the graph similarities and the UMAP embeddings are computationally more complex, they can take up to a few minutes. In order to avoid unnecessary computations, the update of the graph similarities and the UMAP embeddings is done asynchronously in the background. If a domain expert is interested in inspecting the state of the graph after a specific session, they can select a state of the graph and the embeddings using a slider.

The resulting label hierarchy (Figure 6.10) is used to adjust word embeddings through retrofitting [FDJ+15]. This results in moving words closer together in the vector space that share an edge in the graph. For this, a second vector space is created and saved. This changes the nearest neighbor search of annotations as now the union over the nearest neighbors of both vector spaces is presented in the word space.

6.5 Complementary User Pathways

Carrying out a systematic relabeling of images from two legacy iconographic databases would be almost impossible by hand. Not only has it taken public institutions many decades to get to the point where the data are now, but the two sets are artificially siloed, as we have mentioned above. Both tasks–adjusting the annotations based on recommendations and creating the hierarchy of annotations–as organized in our visual analytics system provide a framework for understanding the logic of previous annotators, but also to rethink and expand the two data sets of common cultural artifacts into a more unified data set. The process of annotating can be undertaken simultaneously from several perspectives and following different objectives. This is not due to any fault on the part of the domain specialists involved, but is quite typical of different training and points of view with respect to humanistic research.

For User 1, choosing a "trade-off" set in the manuscript graph, split between ones that had been found to be similar and others that had not, was one way of beginning to understand the processes by which previous annotators had passed through the ma-

Figure 6.10: The current label hierarchy of medieval Latin Bible illuminations. Leaves are colored grey, while inner nodes are black. The level in the tree is given by the indentation. The pseudo root of the tree is not displayed.

terial and to identify missing annotations. The selection of a small set of four to six images for annotation inevitably led to many false pathways and re-selection of new images, but in combination with the recommendations in the word space and the co-occurrences, in particular, guided by the color coding, led me to find numerous annotations not included in either the Initiale or Mandragore data sets. In fact, when a missing annotation was identified, the recommended images provided an effective way to cycle through a number of other possible candidates for annotation. Important in this process in the initial rounds were the indications of frequency, which allowed for commonly occurring labels to be explored with priority. In the cases of synonymous or near synonymous labels, the hierarchy was useful in the beginning as a way of linking and establishing an order between them.

On the other hand, User 2 started the process not by labeling the images with missing annotations but by working on the hierarchy. Starting with the hierarchy helped to understand the variety of labels from a holistic point of view and to understand how they relate to each other. The hierarchy also helped to organize and group the labels by themes, either well-known and widely used (e.g., animals and furniture) or specific to this corpus (positions of the body, steps of creation or any other biblical scene, various descriptions of god, etc.) Furthermore, adding and organizing labels in the hierarchy also helped detect potentially missing labels in the system. Annotating the illuminations with a good idea of the existing vocabulary acquired by working on the hierarchy improved the quality of the annotation system. Sometimes, several labels belonging to the same hierarchy exist when only the most precise one would be enough. Because the words appearing higher in the hierarchy implicitly belong to it, they do not need to be added to the annotation system. Knowing the hierarchy system also helps to be more precise in the description and the labels used; for example, a lower label, more detailed, encompasses more information.

6.6 Discussion

Following a participatory design process [JKKS20], we extend Munzner's nested model [Mun09] with frequent interdisciplinary exchange on a multitude of design aspects of our system. Thus, the visual interfaces that make up our system were subjected to regular implicit evaluations by target users. This process revealed limitations, some of which were addressed during our iterative process, on the one hand, and future directions, on the other.

6.6.1 Limitations

Our visual analytics framework to build hierarchies is currently only tested with two visual data sets, but should be extensible to more than two. However, scaling issues can be encountered in the color coding for more than two data sets, which can be addressed. Similar problems would occur if more than two users were using the system.

Creating the hierarchy can be difficult, as the images can contain a larger number of annotations. The currently applied graph-drawing algorithm does not include domain knowledge about annotations. Related words that are not connected are often positioned far from each other. This problem could be addressed by including methods that automatically group similar words, thus recommending possible connections to the user through spatial proximity. A growing hierarchy also presents some scaling issues. These make navigating tasks cumbersome and it is hard to keep track of the overview even with methods like zooming and panning. This leads, on a larger scale, to the resulting label hierarchy suffering from incompleteness, as all hierarchies suffer from this problem. The current state of the label hierarchy includes 169 terms in the data set and can be used for image classification tasks, for example, for a specific subtree of the hierarchy like "positions" [ABSMJ23]. To further include object localization tasks, such as object detection, weak supervision methods [IFYA18], or an additional annotation process would be needed to add bounding boxes. In addition, including other types of relationships outside of parent-child relationships, such as synonyms, could be of interest. Furthermore, including more means to show the data distribution could help in the process of creating the label hierarchy.

We did not extract the illuminations from the scanned manuscripts prior to processing because the state-of-the-art methods for extracting illuminations were not robust. The complexity of page structures, backgrounds, and preservation statuses led to insufficient and unusable results. Instead, we use the raw image presented in the data set. Depending on the content of the image, this leads to a smaller depiction of the illuminations and also in background noise through the text and the background. Another limitation lies in the cross-depiction problem [HCWC15], as the applied neural network to compute the image embeddings [TL19] is pre-trained on natural images and not on medieval illuminations.

6.6.2 Qualitative Evaluation

The nature of the data set makes a quantitative evaluation of the system at this stage of our research near impossible. However, we are able to offer a qualitative analysis of the implementation of the system and a comparison of the observations of the two users in Section 6.5. We have created a system that allows for annotation and hierarchy

creation among the full data set of some 13.000 images and 2.000 labels. In this first step of research with the combined data sets of Initiale and Mandragore, the two users working on the system were able to generate with relative ease an initial labeling hierarchy that included more than 50 parent labels with many more children (Figure 6.10). In most cases, parents were created imminently from existing labels and, in a minority of cases, they were based on abstraction and created anew. With more time, users are enthusiastic about progressing much further, as the process allows the inherent relationships of the labels to emerge in an organized manner. Although they found this process to be much faster than labeling images one by one in the annotation space, they nonetheless expect that the creation of new labels in the annotation space will facilitate the long-term expansion of this hierarchy.

Furthermore, the two approaches they took to the materials in the system were neither contradictory nor inconsistent with each other's work. Instead, both users wanted the annotation space and the hierarchy builder to be linked so that their work was complementary. The success of the system in our eyes is its interconnected and non-contradictory qualities, which allow multiple users with different experiences of analytics systems, metadata, and humanities databases to work in a cohesive and favorable manner compatible with their experiences. Moreover, the users were eager to continue the work and advance the process to the next step. It was commented that because the data sets already contain annotations, the hierarchy will probably expand faster than the number of annotations, which may actually be beneficial for passing to a new stage of the work including downstream tasks using it, such as automated object recognition.

6.6.3 Future Works

Next to the limitations that give multiple directions for future work, there is a multitude of additional future work possible. To align the matching illuminations in the data sets, for example depiction of the same person, image collation methods could be applied [KSD+21]. Aligning annotations, descriptions, and images with visual-semantic embeddings could also help the recommendation process and automatic image annotation [BCGC18, CSB+20]. Another future work would be to communicate the changes in the vector space. The two-dimensional representation of an image changes with new annotations and re-computation of UMAP. Changes over several iterations could be communicated in the point cloud and in the convex hull.

For both the hierarchy and annotations, instead of the user cycling through the images manual to complete the tasks, the system can be combined with incremental machine learning methods that combine automatic recommendation with user interaction. This would help create robust and generalizable models. One possibility would

be to include an active learning component to combine it with the current labeling process. Also, including methods to focus on the uncertainties of other metadata, like spatio-temporal annotation, can be a next step. We currently disregard this information because spatio-temporal metadata is annotated to the manuscripts, but is often uncertain. Such a system moves from annotating and training to prediction, illustrating new relationships that the new metadata can suggest.

It would be interesting to extend the system to a larger number of users who potentially do not know each other. An image annotation and exploration system that is conceived for multiple actors, that takes input from them and allows us to see disagreement and debate amongst scholars about how we label and classify, but also that attempts to provide multiple classifications. Given that our system is designed to create a label hierarchy as we annotate and there is potential disagreement between annotators, we will need to explore visual modes for representing (and resolving) ambiguities and disagreements. For annotations, methods based on Fleiss' Kappa [Fle71] could be used to display inter-annotator agreement. Visualization should help to keep the human in the system, by showing the different paths users took and the different decisions they made, and so making the collective contribution of knowledge and labor visible.

Cultural collections of images are siloed, and metadata can be inconsistent or nonexistent. The knowledge of the images found in this single genre (the Latin Bible of the thirteenth and fourteenth centuries) should obviously be expanded to encompass many more periods and genres of Christian art. The larger picture of the project is to leverage what is known about the images in one place with some expert input to expand what is known in others. The hierarchy co-constructed in our visual analytics system would ideally be used to expand to other genres. In the long view, learning how to predict classification for the entire data set or any other digitized collections of medieval art could be fascinating repeated themes; style transfer to stained glass, other cross-genre depictions found in other legacy art historical collections. Such a system would be of benefit to the larger community of medievalists and other scholars interested in images.

6.7 Summary

We have presented a visual analytics framework to create annotations and label hierarchies based on a data set of medieval illuminations. The system itself gives access to a large image data set by providing different entry points to understand complex relations between manuscripts, images and the ways that these images have been studied by generations of art historical scholarship. We argue that such a system can be generalized to a number of different cultural heritage collections where metadata gaps prevent

holistic discoverability. It supports the annotation process by combining machine learning methods with interactive visualizations. The resulting annotations and the label hierarchy are in a preliminary stage at present, but can be iteratively refined and used for machine learning tasks where contemporary hierarchies are not appropriate. The results presented here are just a starting point to build bridges between the artificial siloes created in historical research and also to provide more complex, multi-faceted access to the illuminations in later work. Furthermore, we presented cases of how a domain expert would work with such a system, descriptions of what tasks were most appealing to them, a qualitative evaluation of the state of the research, and an assessment of the current limitations and potential for future work in this domain.

The purpose of computing is insight, not numbers.
Richard Wesley Hamming

7 Discussion

Working at the intersection of visual analytics and digital humanities is a challenging endeavor. Due to differences in research practices and vocabulary, it is not easy to gain valuable outcomes for both communities from joined research projects [Jän16]. The specificities of humanities research practice and humanities research material can also lead to several challenges in the design process of a visualization system that assists in answering domain-specific research questions. In order to share our experience and to give some perspectives on how to deal with similar problems, we first reflect on our interdisciplinary projects with a focus on data problems, the design process, and how to achieve valuable outcomes for both communities. After this, we discuss open challenges when working with cultural heritage data regarding labeling and machine learning and how these challenges affect current endeavors.

7.1 Reflection on Interdisciplinary Projects

Participatory visualization design [JKKS20] is based on, but also extends, task-based development [Mun09], since most design considerations and adaptations are discussed in depth among all project members [Wri18]. A side effect of this type of collaboration is the design of visualizations as a speculative process running through multiple iterations by creating several "visualization sandcastles" [HFM18]. The term was coined

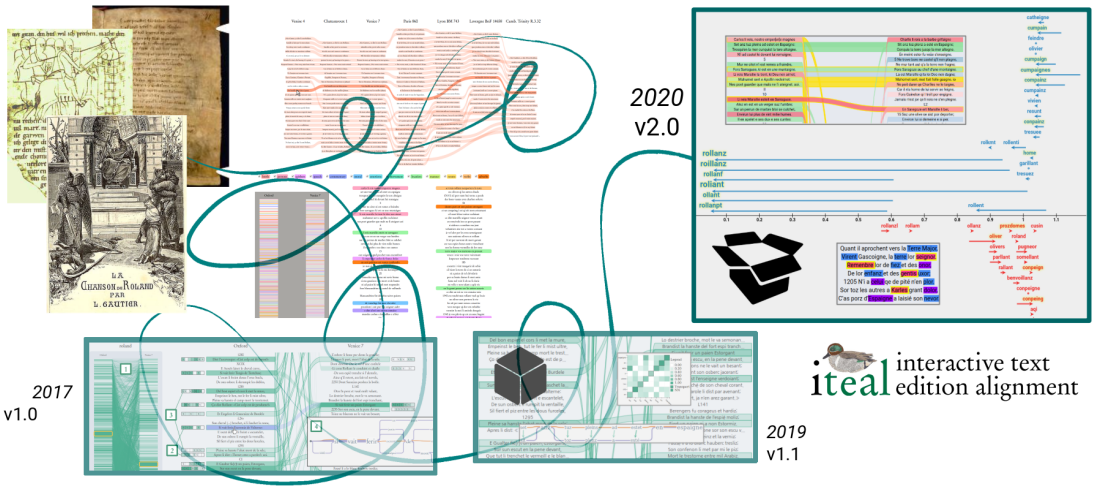


Figure 7.1: The process of creating the interactive text edition alignment system as a speculative process inspired by Hinrichs et al's. [HFM18] "sandcastle" metaphor.

in contrast to the typical approach of thinking of visualization as tools with a means to a certain end. Instead, visualization design can be seen as part of the research and thinking process by iterating through several sandcastles, which can even lead to new research questions for both communities. An example of a speculative process can be seen in Figure 7.1 showing the iterations of the interactive text edition alignment project leading to the version presented in Chapter 4. This type of design process also leads to vibrant reflections on the required adjustments and gives entirely new visual perspectives on the data at hand, as the visualization can also help as a mediator between the disciplines. Furthermore, in our projects, it helped us come up with ways to include domain knowledge as feedback to the visualization system, and thus to engage in the labeling process of cultural heritage data.

7.1.1 Creation of an Interactive Semi-automatic Text Edition Alignment

The first case study is the project that focuses on the alignment of medieval vernacular literature in Chapter 4. To be more precise, the alignment of different versions of the Song of Roland and other works belonging to the genre of French epic poetry [MWJ21]. For this project, many design iterations have already been carried out in a participatory design process between a digital humanities scholar and a visualization scholar [JW17a, JW17b] creating a number of "sandcastle" visualizations. Therefore, a common vocabulary was already established when a second visualization scholar joined the project to include word embeddings to automate the alignment process. However,

the new potential methods and how they worked needed to be discussed with all team members during multiple meetings. After we included these automatic methods, new ways were required to interpret and interact with alignment and also views to communicate changes [MWJ19].

Data Problems

From the computational side, there were several problems with the data source. Starting with text artifacts caused by the OCR process, but also including changes in words because of regional and scribal dialects. The texts of interest were originally passed on orally and later written down in different dialects of medieval French. All of these aspects complicate the alignment process as the same concepts are displayed by different words. Applying word and sentence embedding methods to the data is challenging, as there is no pre-trained language model for medieval French and its dialects, which is a common problem for low-resource and under-resourced languages. This led to training a model from scratch on a small corpus, which can be refined iteratively through user interactions. Furthermore, the whole alignment process of poetry is highly interpretive, which becomes a problem when evaluating the method, since no ground truth is available for an alignment tuple. The only way to evaluate the method was to present the domain expert with two alignments to rate, one before domain knowledge was induced and one after, without them knowing which alignment was which.

Design Process

To include methods to interpret and interact with the alignment, we engaged in a participatory design process by meeting once a week to discuss possible visualization designs. The first challenge was to explain to the domain expert why a particular alignment occurred. For this, we first introduced a heat map showing the similarities of word vectors and explaining the computation of the Word Movers Distance [KSKW15]. With these new means of understanding why an alignment occurred, there also came the need to accept or reject alignments, but because of the highly interpretive nature of the poetry at hand, we decided to use a Likert scale to label the line-level alignments and hence to induce domain knowledge into the alignment process. The general idea was that according to the rating, the underlying word vector distribution should change. For this, we adapted the Rocchio algorithm [Roc71]. The problem with the Likert scale approach was that half-line alignments, i.e., alignments where one line is split into two lines in another version, would get a low score. The reason for this is the usage of different meters in the poems. Due to this, we applied a binning approach to treat half-line alignments differently. This new interaction method of scoring the alignments created the need to see changes in the alignment after an interaction, that is,

which alignments are added and which are removed. Furthermore, more direct and easily understandable methods were needed to change word vectors and to see the changes in the neighborhood of the word vector. For this, we include methods to change the distance between word vectors based on drag and drop, which can be seen as labeling the relation of the words with a numerical value. Then, we also added methods to see the changes in the neighborhood of word vectors over multiple iterations. In order to better communicate which parts of the poem and which words were strongly affected through the interactions, we created word-level heat maps showing either how strongly a word vector or its neighborhood has changed.

Takeaways

The usage of the system showed that at the beginning the labeling of word relations was rarely used, but in becoming more familiar with the system, the domain expert used this labeling method primarily in the end. We conclude that easy interaction methods for labeling with direct feedback, such as moving a word from one position to another, are more appealing than more complex ones that are not easy to grasp, such as applying a scoring method to multiple words and sentence components.

7.1.2 Labeling and Visualizing Entities in Medieval Manuscripts

The second case study is about detecting and visualizing entities in medieval manuscripts (Section 6.2). The idea was to focus on the similarities of images in manuscripts in a similar way to what we did with the textual alignments. For this, we applied computer vision methods and visualization as a starting point for distant viewing on a data set of around 700 medieval illuminations of the French Marco Polo textual tradition. The project constellation was the same as in the first case study: two visualization scholars with experience in several digital humanities projects and one digital humanities scholar/medievalist interested in medieval manuscripts.

Data Problems

In the beginning, we applied object detection with a Faster R-CNN [RHGS15] trained on ImageNet [DDS+09] for feature extraction and Open Images [KRA+20] for detection. The problem is that these data sets and their underlying hierarchies do not match the entities depicted in medieval illuminations. This relates to the contemporary vocabulary used in the hierarchy, such as 'airplane' or 'car', and the depiction of entities, like the cross-depiction problem. Furthermore, the data set was not annotated with bounding boxes, we did not have a list of object classes of interest, and the data set was not large enough to train a new network. But the network trained on natural images still

provides a first impression and a convenient starting point for creating new classes and extracting some initial training data for the classes that are appropriate for the domain and the period.

Design Process

In order to analyze the results, we built a visual interface to explore the classes. For this, it was important to allow browsing the different classes and to compare all depictions of a specific class in a visualization. Furthermore, the domain expert could select contemporary classes to be removed from the hierarchy. On an image level, it is important to allow all detection results to be shown as bounding boxes with their confidence score and to allow filtering based on class and confidence to prevent visual clutter. For a specific bounding box, it is also possible to display the most similar bounding boxes, e.g., to see the most similar faces to a specific Human Face in the data set. In addition, the possibilities to draw new bounding boxes, create new classes, and relabel existing bounding boxes were needed.

Takeaways

The project led us to think about other larger visual corpora, including the paintings in Chapter 5 and Paris Bibles in Chapter 6. It also showed that there is a need for label hierarchies with period-specific terminology and methodologies to unify the vocabulary of different data sets. This becomes even more complicated when you have data that was digitized in different institutions. Furthermore, manual labeling of these data sets takes a lot of time, so there is a need to integrate visual interactive labeling for multi-label problems or weak supervision [IFYA18] to reduce the amount of manual work.

7.1.3 Designing a Virtual Museum

Following the previous projects, two visualization scholars initially had the idea of applying the concept of quantitative text analysis to image data. Thus, we combined computer vision methods with interactive visualizations for a large art collection. During the collaboration with a museum exhibition designer, which we presented in Chapter 5, we saw an opportunity to improve the museum visitors experience by including interactive visualizations in a virtual museum tour. Leading to several discussions on how to present the artworks using visualizations.

Data Problems

The data problems for this project were similar to the problems with the medieval manuscripts. We used the WikiArt data set and applied the same Faster R-CNN as for the medieval manuscripts for object detection. For this project, the problem was not the underlying hierarchy but different art styles, e.g., Cubism leading to missing out on some entities or misclassifying some. There was also again no ground truth to evaluate how well the detection worked on the WikiArt data set.

Design Process

In the beginning, the two visualization scholars created a wide range of two-dimensional visualizations to highlight different facets of the WikiArt data set. Each visualization generated a new quantitative perspective on art collections, including methods to show how the representation of an object has either changed over time, or from artist to artist. When the exhibition designer joined the process, the initial research question changed. The new question was: if a three-dimensional environment can be used to give access to this large art collection through different visualizations and how can we encourage serendipitous discoveries? As a result of this, a joined virtual museum concept was created that allowed the exploration capabilities of the visualizations to be made accessible to the general public.

Takeaways

In the process of working on this project, it became clear that many virtual museums try to recreate a digital representation of the physical space and that current solutions cannot replace a real visit to the museum. But interactive visualizations can still be useful to augment a real visit and therefore to give new perspectives to the material. These visualizations can lead to serendipitous discoveries, give context to a specific artwork, or could be used by curators to present more facets of the material. The collaboration showed us that the concept of Distant Viewing [AT19] can be of interest to museum visitors by showing connections and perspectives that were not possible by seeing a single artwork.

7.1.4 Creation of a Hierarchical Classification for Medieval Illuminations

The last case study in Chapter 6 is about combining two different multi-modal data sets of very similar material. The goal was to create a common label vocabulary of two data sets of medieval illuminations that belong to different institutions and are annotated with different types of metadata. As well as, to create a label hierarchy that is appropriate based on domain- and time-specific properties for the classification and detec-

tion of entities inside the illuminations. The constellation of the first two projects was extended by a medievalist interested in studying depictions in medieval manuscripts. Bringing expertise in codicology, scribal practices, and medieval manuscripts and illuminations.

Data Problems

The first data problem we faced was the automatic extraction of illuminations from the manuscript pages. Each digital image illustrates one or two pages of a specific manuscript and can contain one or several illuminations depicting various scenes and objects. This process is hard to standardize because of the different quality of the background material, as well as some illuminations or ornaments can take up almost the whole page.

Another problem lies in the vocabulary used for the annotations. Both data sets were created at different times and with different priorities. However, the vocabulary of both data sets was initially based on the "Thésaurus Garnier" [Gar84], but both institutions deviated from this controlled vocabulary. This divide between data sets is based on the history of the different institutions and is a major problem when scholars want to see relationships in the larger picture of data from cultural heritage, as inhomogeneities in the vocabulary complicate the application of computational methods. Furthermore, annotators have paid attention to different details for similar images from the same unit of text, while some images have no or only a few annotations; others have more than 50 different annotations, which complicates the application of machine learning methods. The inconsistencies within the data sets are also based on different decisions of several annotators and an annotation process that ranges over several years. This is something that happens often in the GLAM sector, as institutions preserve different assets of cultural heritage and therefore have access to different materials. Furthermore, manual labeling of the data sets takes a lot of time, so there is a need to integrate either visual interactive labeling for multi-label problems or weak supervision to reduce the amount of manual work. However, using an active learning approach to label different classes is complicated to apply here for multiple reasons. Cultural data often has a long-tail distribution and relations between the labels. This imbalance would be problematic for a machine learning model, even with methods like over- or under-sampling, or even when only focusing on a subtree in the hierarchy.

Design Process

We started by creating exploratory visualizations for the different facets of the data sets. Including presenting the images through timelines, image clouds, set visualizations of annotation combinations, comparing the annotation of different manuscripts, and a faceted search based on annotations and metadata. However, these methods were not

as helpful as we imagined at this stage of the project to answer specific research questions. The reason for this was the previously mentioned inconsistencies in the annotation and metadata because of different institutions and people working on the data. This led to focusing on resolving inhomogenities, and engaging in the creation of a label hierarchy and the labeling of the data.

So, we started to focus on grouping the images first by manuscripts, as well as different similarity measurements. The result was a graph visualization that helped select subsets of manuscripts according to different criteria. From there, a point cloud shows the images projected into a two-dimensional space based on text or image features. This helped to select a subset of images to enrich their labels simultaneously and to build a hierarchy of the vocabulary. We decided to focus on labeling multiple similar images at the same time to help detect similar depicted themes and objects that are not used consistently over both data sets, thus also helping to bring the data sets together. The recommendation of terms also helps, as the vocabulary of the metadata is rather large and not all terms are known by the domain experts. This is probably a similar problem the original annotators of the data had, as the labeling is not always consistent in the data source.

Takeaways

This research project led to several possible ideas for future research directions in terms of human-in-the-loop scenarios for image classification and object detection, as well as the comparison of different graphs and hierarchies, putting labels in context. Nevertheless, a paper about the project was initially rejected at a high-quality visualization venue, with the recommendation to submit the "Greatly written and didactically prepared paper" to a domain-specific venue. A paper about this topic may have a higher impact in a domain-specific venue, but we still had new ideas based on this initial rejection about possible research directions that could be of interest to the visualization community, that can be evaluated, and could help the domain experts in the final goal of studying the illuminations. The major problem lies in a missing evaluation with a baseline for labeling, which is hard to do, as most systems do not tackle the problem of multi-label classification at the scale of several thousand possible classes. Even state-of-the-art approaches such as VIAL were not applied prior to this type of data. The problem of missing evaluation methods often arises in the GLAM sector or will arise more often when advanced machine learning methods are applied to cultural heritage data, therefore, standard solutions are needed. At the current stage, we do not have a classifier trained on the data to suggest candidates for labeling. So, in contrast to active learning and VIAL, we embed the data, so the whole process is in an unsupervised fashion and purely uses the similarity of images and their metadata to explore.

7.1.5 Valuable Outcome for both Communities

Finding valuable outcomes for both the humanities community and the visualization community is not always an easy task [Jän16] as the needs and wants of the two communities can be slightly different [BSC21]. Good communication and a participatory design process can help to find a common research question or several questions that could be of interest to both communities. Transparency about the challenges of publishing work in the respective fields can also help prevent conflicts of interest in academic goals and find a common place to publish the results [SRF+19]. Finding a common vocabulary also helps bridge the gaps between different research areas [EAGJ+16]. However, from a visualization perspective, finding ways to evaluate the results is often challenging, as quantitative evaluation does not always play a role in humanities research, and existing standards do not always apply to the challenges and complexity of cultural data. Furthermore, when there is no ground truth for a cultural heritage collection, it becomes almost impossible to evaluate the applied methods, which is, in the case of a human-in-the-loop approach, already difficult enough when there is ground truth available [BBC+20].

Furthermore, some fields in the humanities are still reserved about using computational methods for their research work. An example is the domain of art history [LBT+18, Dru13]. Although this domain has a history of using visualization dating back to Aby Warburg's Mnemosyne Atlas in the 19th century, such methods are rarely applied [WD16]. A way to bridge the gap is to augment traditional workflows with additional insights, rather than replacing them. An example is ARIES [CRD+17] where the operations art historians apply to physical light boxes to curate a collection are imitated in a visualization system.

7.2 Challenges When Working with Cultural Heritage Data

In the following, we discuss open challenges when working with cultural data with regard to labeling and machine learning. Solving these challenges would enable domain experts to conduct large-scale studies on cultural heritage materials and would even allow for a more in-depth visual analysis. In particular, we discuss the incompleteness of cultural data, the lack of multi-label solutions, data imbalance, conflicting vocabulary in data sets, the cross-depiction problem, the cultural gap in computer vision methods, intangible heritage, and multi-modality.

7.2.1 Incompleteness of Cultural Heritage Data

The first challenge is the data itself. Often, only a limited size of data is available or is of interest to the domain-specific research question, making the application of machine learning methods often not feasible. Other problems related to data quality are incompleteness of the data, such as missing metadata [KKFJ19], damaged material [KBJ22], or imprecision, for example, an OCR approach for text, or artifacts in a three-dimensional model. Another problem occurs when there is data available but the data is not labeled, i.e., without assigned ground truth. In these cases, either manual labeling is needed, which can be supported in some cases by visual analytics methods, or only unsupervised methods can be applied, which again are limited to models trained on other data sources or the data at hand. When manual labeling of data sources is needed to perform an analysis of the data, active learning [WSZ+20] can be applied to reduce the amount of manual labeling. VIAL [BZSA18] as a concept to combine active learning with visualization systems to explore and select data points for labeling can also give great results for problems with a small number of classes [ABB18, SJS+21, RAZ+18], but it has not yet been applied to multi-label problems with a larger number of classes, such as entities or scenes depicted in visual material of cultural heritage.

7.2.2 Multi-Label Classification of Cultural Heritage Data

Collections of cultural heritage are imbalanced in nature, since some objects, entities, or themes are included more often than others. This leads to many cultural heritage data sets suffering from long-tail distributions of their labels, as well as many other real-world applications. This is problematic when training a classifier on these data sets, as the skewed distribution leads to a poorly trained classifier that overfits the high-frequent classes. For example, an image in the Paris Bible data set has on average 6 labels assigned, but some images have no labels, while others can have more than 30, similarly an annotation appears on average 38 times, but some only appear once, while others appear over 1000 times. For natural image data sets, it would be possible to generate new images, but this is not possible for historical assets. Methods like oversampling low-frequent categories or undersampling high-frequent categories can help in this process, next to other data augmentation methods like style transfer, by generating or removing images to create a more balanced training set.

Especially when working with visual material from the GLAM sector, multi-label methods are needed in order to assess the variety of depictions on a page or image of interest. A problem for multi-label classification is that the distribution of label combinations can be even more skewed than the label distribution, resulting in problems for

oversampling methods such as MLSMOTE [CRdJH15]. The reason for this is that in a multi-label setting, the distribution of the label combinations also needs to be balanced, and this can be hard, as some combinations could not be part of the training data. This is also a problem when creating a balanced split of training, test, and validation data. One option would be to remove rare classes, but this is not really a viable option because these classes are still of interest. These cases can be even more interesting than the other classes for methods to generate more labeled data like visual interactive labeling, or for specific research questions. Methods such as iterative stratification [STV11], second-order iterative stratification [SK17], or EvoSplit [FR21] can help to guarantee a specific quality of the data split in multi-label settings, but cannot tackle the problem in completeness.

7.2.3 Conflicting Vocabulary in Computer Vision

A problem that arises with collections of cultural heritage curated at different institutions is that not always a specific standard is followed. This can lead to cases where enough data is available and labels and metadata are present, but there is a lack of controlled vocabulary between different data sources [KKFJ20b]. For example, the vocabulary can then drift apart from another, resulting in siloed databases, making it hard to combine the data from these institutions for large-scale analysis. In these cases, visualization can help in the process of combining the different vocabularies and resolving inhomogeneities. A process that otherwise is often done by hand. It is even possible to rewrite the meaning and significance of collections of digital cultural heritage using crowd-sourced approaches and other forms of participation to contextualize and annotate the collection [GMD15]. Efforts such as the International Image Interoperability Framework [Con11] try to solve the lack of controlled vocabulary by providing a standard to describe images and present their metadata.

The problem of conflicting vocabulary also exists between the classes of common image classification and object detection data sets for natural images, such as ImageNet and OpenImages, and the domain-specific vocabulary needed for cultural heritage data. One way to still use neural networks trained on natural image data sets is to create a mapping between the two hierarchies. Thus, mapping the specific domain vocabulary to the more general vocabulary of an existing data set. An example of the PASCAL VOC data set and the Paris Bible data set from Chapter 6 can be seen in Figure 7.2. In this figure, both hierarchies can be seen side-by-side, while creating a colored mapping. Based on word vector similarity, words can also be recommended for a specific mapping. This approach can even be used for data sets that are annotated in different languages, as long as aligned word embeddings for the labels exist, such as

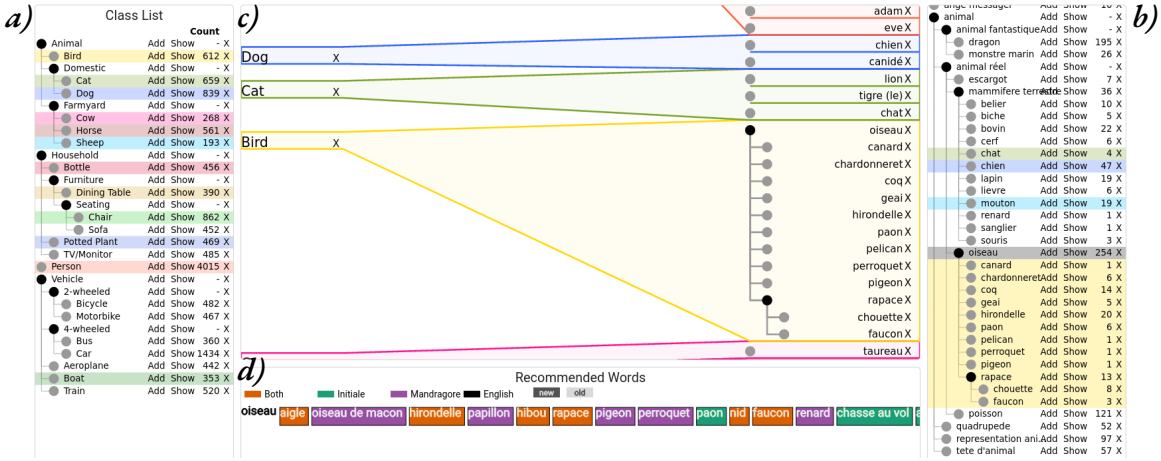


Figure 7.2: A prototype to create a mapping between the vocabulary of different label hierarchies. (a) shows a contemporary hierarchy, while (b) shows the hierarchy created in Chapter 6. (c) shows a potential mapping between both hierarchies and (d) similar words in both hierarchies for a selected word.

MUSE [CLR+17]. Together with weakly supervised object detection [IFYA18] and style transfer [ZPIE17] this approach can be applied to images to generate bounding boxes of object classes that can also be found in natural image data sets. This proof of concept can also be extended to larger data sets such as OpenImages [KRA+20].

7.2.4 Cross-Depiction & Cultural Gap in Computer Vision

As previously discussed, there are different types of uncertainty, such as imprecision and incompleteness. Both are important to address for cultural data and machine learning. One form of imprecision that occurs when applying computer vision methods to cultural data is the cross-depiction problem [HCWC15]. This problem addresses the need to recognize visual objects by computer vision methods that are agnostic to the depictive forms they take, e.g., photographed, painted, or drawn. Although there are existing computer vision data sets that address the cross-depiction problem by including photos and paintings from different art styles, such as the PhotoArt50 [WCH14] and PeopleArt [WCH16] data sets, they only provide ground truth for people and a limited number of objects, and thus do not address the needs of computer vision methods for cultural heritage data. Cross-depiction can also be a problem for annotators, for example, in medieval illuminations, often zoomorphic objects or human-animal hybrids are depicted, which can make it hard to decide which labels should be assigned to a specific object or entity. Examples are given in Figure 7.3.

Figure 7.3: Examples of human-animal hybrids and zoomorphic objects in Paris Bibles.

Another form of imprecision and incompleteness is given in computer vision by the cultural gap, that is, the "lack of coincidence between the information that one can extract from the visual data and the interpretations that the same data have for cultural groups across time" [vN22]. This extends the concept of a semantic gap [SWS+00] between two representations of an object with the cultural aspect. This is particularly acute for iconic images, where the meaning and interpretation of an image go beyond the depiction and is given by the social and temporal context of the observer, which makes automatic extraction with machine learning models difficult, as this cannot be extracted from the content of the image. In the same way that text reuse methods can be used to help study intertextuality in text, computer vision methods could help study intericonicity in iconic images. Currently, there is a lack of suitable algorithms to address this gap, which would be important for studying cultural data. Most computer vision benchmarks focus on natural images from high-income countries and were collected using an English vocabulary [DVMWVdM19], therefore inducing bias in the models, which are then used as feature extractors for more complex downstream tasks such as object detection and image segmentation. This is problematic when dealing with cultural data beyond the problem of not matching vocabulary; even if the vocabulary is the same, some objects could not be detected in the case of an object detection approach because the cultural asset does not match the learned depiction.

7.2.5 Intangible Heritage

Another open challenge is the analysis of intangible heritage. Most of the time, when cultural heritage data is visualized, the focus is on tangible assets [WFS+18]. Only a small number of works focus on intangible assets, such as performing arts, expressions,

customs, or rites. This is even more acute when looking at machine learning methods. In order to apply machine learning and visual analytics methods to intangible heritage, such as a specific dance or oral history, a digital representation needs to be created, such as a text, image, or video, which could then be enriched by human-annotated labels. Examples of intangible heritage where a tangible asset was created are texts that were passed on orally in vernacular languages until someone wrote them down, such as the Song of Roland [JW17b], which we discussed in Chapter 4. Other examples are the comparison of dance movement through recordings [EVHB20, AKD+21], or biographies for prosopographical research [WSK+17, MJ18, KKFJ20a]. However, these methods cannot include all aspects of intangible heritage. Part of the ongoing discussion is to find appropriate forms, together with efficient methods, to document intangible heritage and to communicate knowledge inextricably linked to people, including recording, representing, and reviving of the living nature [HKP+22]. This is primarily important, as intangible cultural heritage is what communities today recognize as part of their cultural heritage and exists only in the present [UNEU03].

7.2.6 Multi-Modal Heritage

Another important aspect of cultural heritage is multi-modality. A cultural asset can be represented by several digital representations with different modalities. For this, methods are needed to address or even combine all these facets of cultural data. The simplest approach is to use one classifier for each modality and then to compute an agreement or disagreement score [KHP18]. Some other works have already used cultural heritage data with different modalities for classification tasks. For example, the combination of sound and text features can help in sentiment analysis [FW22] and in finding similar songs. There were also methods proposed for the cross-modal retrieval of photos either solely by textured three-dimensional models [GMM+17], or by combining three-dimensional models with sketches drawn by domain experts [LKL+19]. Images of cultural objects are often presented together with their textual metadata and descriptions. This information can be used to classify different categories or to find similar objects [AS12]. Furthermore, multi-modal classifiers can be used to predict missing metadata [RMD+22]. One example would be to predict the author or workshop that created a medieval illuminated manuscript, by combining the illuminations together with textual, temporal, and geographic metadata. For the creation of multi-modal machine learning data sets from unstructured data sources of contemporary cultural data with heritage values and attributes, Bai et al. [BNLPR22] proposed a framework.

A mind cannot be independent of culture.

Lev Semyonovich Vygotsky

8 Summary

This dissertation presented four projects at the intersection of visual analytics and digital humanities, focusing on exploratory analysis and labeling of cultural data. Three of them were conducted in close collaboration with (digital) humanities scholars engaging in a participatory design process.

First, we presented two visualization systems to explore different facets of rap lyrics and rap artists with a strong focus on text reuse. The systems allow a user interested in rap music to detect allusions between songs, to find artists similar to an artist of interest based on their lyrics, and to explore the network of musicians. The process of creating the systems also gave us new ideas to focus on the multi-modality of songs and cross-lingual similarities.

Then we engaged in a more complex case of text reuse, the collation of medieval vernacular text editions. The collaborating domain expert engaged in a human-in-the-loop process that used word embeddings and interactive visualizations to perform textual alignments on under-resourced languages. Visualizations enabled the expert reader to feed domain knowledge into the system at multiple levels with the aim of improving both the product and the process of text alignment. This showed how visualization can augment complex modes of reading in the humanities. In particular, we focused on the alignment of different versions of the Song of Roland and other sources belonging to the genre of French epic.

We then moved away from textual data to visual data by presenting a Virtual Museum that combines interactive visualizations and computer vision to explore a collection of artworks. In collaboration with a museum exhibition designer, we created a three-dimensional space where artworks are presented with several visualizations. Visualizations contextualize the artwork of interest by allowing a visitor to focus on specific objects detected by neural networks, observing the oeuvre of an artist, and finding similar images and artists to an image or artist of interest through serendipitous discoveries. The evaluation of this space gave us insight into the increasing desire of humanities scholars to quantitatively analyze art collections.

With the lessons learned from the previous projects, we then engaged in the labeling and analysis of medieval illuminations of Paris Bibles. This resulted in a visual analytics framework that combines machine learning methods and visualizations that were previously used for textual data with computer vision methods. The results of this collaboration were a label hierarchy for medieval illuminations that is still enriched and extended, and several research directions to further focus on combining metadata from different legacy sources in the GLAM sector.

In addition, we shared our experience working on interdisciplinary projects at the intersection of visual analytics and digital humanities. We gave some perspectives on how to deal with similar domain problems by reflecting on our projects with a focus on data problems, the design process, and how to achieve valuable outcomes for both communities. We then discussed open challenges when dealing with cultural heritage data, focusing on labeling and machine learning. Thus, addressing issues of incompleteness of data, data imbalance, conflicting vocabulary, intangible heritage, multi-modality, cross-depiction, and the cultural gap in computer vision. Furthermore, we pointed out that there is a lack of projects in the GLAM sector focusing on multi-label classification and human-in-the-loop processes like visual interactive labeling in order to bridge the gap between machine learning, digital humanities, and visualization research. This opens up research potential with many interesting questions in regard to machine learning, human-computer interaction, and visualization design. Solving the discussed challenges would enable domain experts to conduct large-scale studies on cultural heritage materials and would even allow for a more in-depth visual analysis.

In this thesis, we wanted to give a perspective on how visualization and machine learning can be combined to produce a better analysis of cultural heritage data. Our projects were and will continue to be carried out as a participatory design process, from which all members gain valuable inspiration that can be carried out in the respective research areas. With our documented process, we hope to inspire other visualization researchers to engage in participatory design, which pays off in the form of numerous ideas for future research directions.

References

[AB08] Spotify AB. Spotify, 2008. https://www.spotify.com/ (Accessed 2023-02-22).

[AB18] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[ABB18] Shivam Agarwal, Jürgen Bernard, and Fabian Beck. Computer-supported interactive assignment of keywords for literature collections. In *Proceedings of the Machine Learning from User Interaction for Visualization and Analytics Workshop at IEEE VIS*, 2018.

[ABSMJ23] Yasmin Hawar Abo Bakir Shuan, Christofer Meinecke, and Stefan Jänicke. Image classification of paris bible data. In *to appear in DH 2023: "Collaboration as Opportunity"*, 2023.

[ACL20] Mohammad Saqar Alharbi, Tom Cheesman, and Robert S Laramee. Transvis: Integrated distant and close reading of othello translations. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[AEA+16] Bharathi Asokarajan, Ronak Etemadpour, June Abbas, Sam Huskey, and Chris Weaver. Visualization of latin textual variants using a pixel-based text analysis tool. In *Proceedings of the International EuroVis Workshop on Visual Analytics, EuroVA*, volume 16, 2016.

[AEA+17] Bharathi Asokarajan, Ronak Etemadpour, June Abbas, Samuel J Huskey, and Chris Weaver. Textile: A pixel-based focus+ context tool for analyzing variants across multiple text scales. In *EuroVis (Short Papers)*, pages 49–53, 2017.

[Ahm06] Yahaya Ahmad. The scope and definitions of heritage: from tangible to intangible. *International journal of heritage studies*, 12(3):292–300, 2006.

[AKD+21] Vasiliki Arpatzoglou, Artemis Kardara, Alexandra Diehl, Barbara Flueckiger, Sven Helmer, and Renato Pajarola. Dancemoves: A visual analytics tool for dance movement analysis. *EuroVis 2021*, 2021.

[Ale99] Patrick H Alexander. *The SBL handbook of style: for Ancient Near Eastern, Biblical, and early Christian studies*. Hendrickson Publishers, 1999.

[AMA+16] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. The state-of-the-art of set visualization. In *Computer Graphics Forum*, volume 35, pages 234–260. Wiley Online Library, 2016.

[ARLC+13] Alfie Abdul-Rahman, Julie Lein, Katharine Coles, Eamonn Maguire, Miriah D Meyer, Martin Wynne, Chris R Johnson, Anne E Trefethen, and Min Chen. Rule-based visual mappings – with a case study on poetry visualization. *Computer Graphics Forum*, 32(3pt4):381–390, 2013.

[ARRO+17] Alfie Abdul-Rahman, Glenn Roe, Mark Olsen, Clovis Gladstone, Richard Whaling, N Cronk, Robert Morrissey, and Min Chen. Constructive visual analytics for text similarity detection. In *Computer Graphics Forum*, volume 36, pages 237–248, 2017.

[AS98] James Andrews and Werner Schweibenz. New media for old masters: The kress study collection virtual museum project. *Art Documentation: Journal of the Art Libraries Society of North America*, 17(1):19–27, 1998.

[AS05] Derrick P. Alridge and James B. Stewart. Introduction: Hip hop in history: Past, present, and future. *The Journal of African American History*, 90(3):190–195, 2005.

[AS12] Nikolaos Aletras and Mark Stevenson. Computing similarity between cultural heritage items using multimodal features. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 85–93, 2012.

[AS19] Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, 2019.

[AT19] Taylor Arnold and Lauren Tilton. Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, 34(Supplement_1):i3–i16, 2019.

[BBC+20] Nadia Boukhelifa, Anastasia Bezerianos, Remco Chang, Christopher Collins, Steven Drucker, Alexander Endert, Jessica Hullman, Chris North, and Michael Sedlmair. Challenges in evaluating interactive visual machine learning systems. *IEEE Computer Graphics and Applications*, 40(6):88–96, 2020.

[BC64] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[BCGC18] Lorenzo Baraldi, Marcella Cornia, Costantino Grana, and Rita Cucchiara. Aligning text and document illustrations: towards visually explainable digital humanities. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1097–1102. IEEE, 2018.

[BCS+16] Arianna Betti, Thom Castermans, Bettina Speckmann, Hein Van Den Berg, Kevin Verbeek, et al. Glammapping trove. In *Proc. VALA Biennial Conf. Exhibition*, 2016.

[BD19] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 638–647. IEEE, 2019.

[BDH21] Mark-Jan Bludau, Marian Dörk, and Frank Heidmann. Relational perspectives as situated visualizations of art collections. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii17–ii29, 2021.

[BEC+18] Adam James Bradley, Mennatallah El-Assady, Katharine Coles, Eric Alexander, Min Chen, Christopher Collins, Stefan Jänicke, and David J. Wrisley. Visualization and the digital humanities:. *IEEE Computer Graphics and Applications*, 38(6):26–38, Nov 2018.

[Bed01] Benjamin B Bederson. Photomesa: a zoomable image browser using quantum treemaps and bubblemaps. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 71–80, 2001.

[BEM+19] Katy Börner, Oyvind Eide, Tamara Mchedlidze, Malte Rehbein, and Gerik Scheuermann. Network visualization in the humanities (dagstuhl seminar 18482). In *Dagstuhl Reports*, volume 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[BGC10] Daniele Borghesani, Costantino Grana, and Rita Cucchiara. Surfing on artistic documents with visually assisted tagging. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1343–1352, 2010.

[BGJM17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[BH03] Stephan Baumann and Oliver Hummel. Using cultural metadata for artist recommendations. In *Proceedings Third International Conference on WEB Delivering of Music*, pages 138–141. IEEE, 2003.

[BHZ+17] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics*, 24(1):298–308, 2017.

[Bie21] Bernadette Biedermann. Virtual museums as an extended museum experience: Challenges and impacts for museology, digital humanities, museums and visitors–in times of (coronavirus) crisis. *Digital Humanities Quarterly*, 15(3), 2021.

[BJP+19] Martin Baumann, Markus John, Hermann Pflüger, Cornelia Herberichs, Gabriel Viehhauser, Wolfgang Knopki, and Thomas Ertl. An Interactive Visualization for the Analysis of Annotated Text Variance in the Legendary Der Heiligen Leben, Redaktion. In *LEVIA 2019: Leipzig Symposium on Visualization in Applications*, 2019.

[BKNS00] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[BKSK12] Michael Behrisch, Miloš Krstajic, Tobias Schreck, and Daniel A Keim. The news auditor: Visual exploration of clusters of stories. In *Proc. EuroVA International Workshop on Visual Analytics*, pages 61–65. Eurographics, 2012.

[BM13] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013.

[BMHC16] Adam James Bradley, Hrim Mehta, Mark Hancock, and Christopher Collins. Visualization, digital humanities, and the problem of instrumentalism. In *1st Workshop on Visualization for the Digital Humanities, IEEE VIS 2016, Baltimore, Maryland, USA*, 2016.

[BNLPR22] Nan Bai, Pirouz Nourian, Renqian Luo, and Ana Pereira Roders. Herigraphs: a dataset creation framework for multi-modal machine learning on graphs of heritage values and attributes with social media. *ISPRS International Journal of Geo-Information*, 11(9):469, 2022.

[Bod17] Katherine Bode. The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history. *Modern Language Quarterly*, 78(1):77–106, 2017.

[BOH11] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[BPF+18] Mafkereseb Kassahun Bekele, Roberto Pierdicca, Emanuele Frontoni, Eva Savina Malinverni, and James Gain. A survey of augmented, virtual, and mixed reality for cultural heritage. *Journal on Computing and Cultural Heritage (JOCCH)*, 11(2):1–36, 2018.

[BSC21] Adam James Bradley, Victor Sawal, and Christopher Collins. Approaching humanities questions using slow visual search interfaces. In *4th Workshop on Visualization for the Digital Humanities, IEEE VIS 2019, Vancouver, Canada*, 2021.

[BTC10] Paolo Brivio, Marco Tarini, and Paolo Cignoni. Browsing large image datasets through voronoi diagrams. *IEEE transactions on visualization and computer graphics*, 16(6):1261–1270, 2010.

[Bur02] John Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287, 2002.

[BZSA18] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. Vial: a unified process for visual interactive labeling. *The Visual Computer*, 34(9):1189–1207, 2018.

[CAB+11] Yang Chen, Jamal Alsakran, Scott Barlowe, Jing Yang, and Ye Zhao. Supporting effective common ground construction in asynchronous collaborative visual analytics. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 101–110. IEEE, 2011.

[Cam16] Hastings Cameron. Diddy's little helpers, 2016. https://www.villagevoice.com/2006/11/14/diddys-little-helpers/ (Accessed 2023-02-22).

[CBC+20] Mohammad Chegini, Jürgen Bernard, Jian Cui, Fatemeh Chegini, Alexei Sourin, Keith Andrews, and Tobias Schreck. Interactive visual labelling versus active learning: an experimental comparison. *Frontiers of Information Technology & Electronic Engineering*, 21(4):524–535, 2020.

[CBY10] Yang Chen, Scott Barlowe, and Jing Yang. Click2annotate: Automated insight externalization with rich semantics. In *2010 IEEE symposium on visual analytics science and technology*, pages 155–162. IEEE, 2010.

[CDA+17] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics, 2017.

[Cha11] Kim D Chanbonpin. Legal writing, the remix: Plagiarism and hip hop ethics. *Mercer L. Rev.*, 63:597, 2011.

[CK04] Pedro Cano and Markus Koppenberger. The emergence of complex network patterns in music artist networks. In *Proceedings of the 5th international symposium on music information retrieval (ISMIR)*, pages 466–469. Citeseer, 2004.

[CL18] Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018.

[CLR+17] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[CLRP13] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE transactions on visualization and computer graphics*, 19(12):1992–2001, 2013.

[CMB20] Daniel Alejandro Loaiza Carvajal, María Mercedes Morita, and Gabriel Mario Bilmes. Virtual museums. captured reality and 3d modeling. *Journal of Cultural Heritage*, 45:234–239, 2020.

[CMJ+20] Angelos Chatzimparmpas, Rafael Messias Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, volume 39, pages 713–756. Wiley Online Library, 2020.

[Con11] IIIF Consortium. International image interoperability framework. https://iiif.io/, 2011. (Retrieved 2021-03-31).

[CPY+19] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[CPZ15] Elliot J Crowley, Omkar M Parkhi, and Andrew Zisserman. Face painting: querying art with photos. In *Proceedings of the British Machine Vision Conference (2015)*. BMVA Press, 2015.

[CRD+17] Lhaylla Crissaff, Louisa Wood Ruby, Samantha Deutch, R Luke DuBois, Jean-Daniel Fekete, Juliana Freire, and Claudio Silva. Aries: enabling visual exploration and organization of art image collections. *IEEE computer graphics and applications*, 38(1):91–108, 2017.

[CRdJH15] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015.

[Cro16] Damon Crockett. Direct visualization techniques for the analysis of image data: the slice histogram and the growing entourage plot. *International Journal for Digital Art History*, 2, 2016.

[CRT+22] Armelle Couillet, Hélène Rougier, Dominique Todisco, Josserand Marot, Olivier Gillet, and Isabelle Crevecoeur. New visual analytics tool and spatial statistics to explore archeological data: The case of the paleolithic sequence of la rocheà-pierrot, saint-césaire, france. *Journal of Computer Applications in Archaeology*, 5(1), 2022.

[Cru19] Mark Cruse. Novelty and diversity in illustrations of marco polo's description of the world. *Toward a global Middle Ages*, pages 195–202, 2019.

[CSB+20] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. Explaining digital humanities by aligning images and textual descriptions. *Pattern Recognition Letters*, 129:166–172, 2020.

[CT05] Kristin A Cook and James J Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.

[CYCD18] Sarah Campbell, Zheng-Yan Yu, Sarah Connell, and Cody Dunne. Close and distant reading via named entity network visualization: A case study of women writers online. In *Proceedings of the 3rd Workshop on Visualization for the Digital Humanities. VIS4DH*, 2018.

[CZ14a] Elliot J Crowley and Andrew Zisserman. In search of art. In *European Conference on Computer Vision*, pages 54–70. Springer, 2014.

[CZ14b] Elliot J Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *Proceedings of the British Machine Vision Conference (2014)*. BMVA Press, 2014.

[Dan14] Matt Daniels. The largest vocabulary in hip hop, 2014. https://pudding.cool/projects/vocabulary/ (Accessed 2021-10-27).

[Dar20] Jordan Darville. Report: Three 6 mafia launch $6.45 million lawsuit against $uicideboy$ over samples, 2020. https://www.thefader.com/2020/09/08/report-three-6-mafia-launch-s645-million-lawsuit-against-suicideboys-over-samples (Accessed 2021-10-27).

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[Dir13] Jos Dirksen. *Learning Three.js: the JavaScript 3D library for WebGL*. Packt Publishing Ltd, 2013.

[DM11] Ronald H Dekker and Gregor Middell. Computer-supported collation with collatex: managing textual variance in an environment with varying requirements. *Supporting Digital Humanities*, pages 17–18, 2011.

[DMG+20] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 836–837, 2020.

[DMTS14] Bruno Dumas, Bram Moerman, Sandra Trullemans, and Beat Signer. Artvis: combining advanced visualisation and tangible interaction for the exploration, analysis and browsing of digital artwork collections. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pages 65–72, 2014.

[DPC17] Marian Dörk, Christopher Pietsch, and Gabriel Credico. One view is not enough: High-level visualizations of a large cultural collection. *Information Design Journal*, 23(1):39–47, 2017.

[DPDT18] Chiara Di Pietro and Roberto Rosselli Del Turco. Between innovation and conservation: The narrow path of user interface design for digital scholarly editions. *Bleier, Klug, Neuber, and Schneider*, pages 133–63, 2018.

[DPLM+16] Roberto De Prisco, Nicola Lettieri, Delfina Malandrino, Donato Pirozzi, Gianluca Zaccagnino, and Rocco Zaccagnino. Visualization of music plagiarism: Analysis and evaluation. In *2016 20th International Conference Information Visualisation (IV)*, pages 177–182. IEEE, 2016.

[DPT+12] Vincenzo Deufemia, Luca Paolino, Genoveffa Tortora, Antonella Traverso, Viviana Mascardi, Massimo Ancona, Maurizio Martelli, Nicoletta Bianchi, and Henry De Lumley. Investigative analysis across documents and drawings: visual analytics for archaeologists. In *Proceedings of the international working conference on advanced visual interfaces*, pages 539–546, 2012.

[dreddtdCndlrsSdme12] Institut de recherche et d'histoire des textes du Centre national de la recherche scientifique Section des manuscrits enluminés. Initiale catalogue de manuscrits enluminés. http://initiale.irht.cnrs.fr/, 2012. (Retrieved 2023-02-22).

[Dru13] Johanna Drucker. Is there a "digital" art history? *Visual Resources*, 29(1-2):5–13, 2013.

[DSK21] Yashar Deldjoo, Markus Schedl, and Peter Knees. Content-driven music recommendation: Evolution, state of the art, and challenges. *arXiv preprint arXiv:2107.11803*, 2021.

[DVKSW12] Kasper Dinkla, Marc J Van Kreveld, Bettina Speckmann, and Michel A Westenberg. Kelp diagrams: Point set membership visualization. In *Computer Graphics Forum*, volume 31, pages 875–884. Wiley Online Library, 2012.

[DVMWVdM19] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[DZ19] Abhishek Dutta and Andrew Zisserman. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2276–2279, 2019.

[EAGJ+16] Mennatallah El-Assady, Valentin Gold, Markus John, Thomas Ertl, and Daniel A Keim. Visual text analytics in context of digital humanities. In *1st IEEE VIS Workshop on Visualization for the Digital Humanities as part of the IEEE VIS 2016*, 2016.

[EAKC+19] Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. Semantic concept spaces: Guided topic model refinement using word-embedding projections. *IEEE transactions on visualization and computer graphics*, 26(1):1001–1011, 2019.

[EASS+17] Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel Keim, and Christopher Collins. Progressive learning of topic modeling parameters: A visual analytics framework. *IEEE transactions on visualization and computer graphics*, 24(1):382–391, 2017.

[EB12] Micheline Elias and Anastasia Bezerianos. Annotating bi visualization dashboards: Needs & challenges. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1641–1650, 2012.

[EE16] Timothy AM Ewin and Joanne V Ewin. In defence of the curator: maximising museum impact. *Museum Management and Curatorship*, 31(4):322–330, 2016.

[ERT+17] Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, I Díaz Blanco, and Fabrice Rossi. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, volume 36, pages 458–486. Wiley Online Library, 2017.

[Eth18] Kawin Ethayarajh. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100, 2018.

[EVHB20] Miguel Escobar Varela and Luis Hernández-Barraza. Digital dance scholarship: Biomechanics and culturally situated dance analysis. *Digital Scholarship in the Humanities*, 35(1):160–175, 2020.

[FD17] The Data Face and Matt Daniels. The language of hip hop, 2017. https://pudding.cool/2017/09/hip-hop-words/ (Accessed 2021-10-27).

[FDB18] Cristian Felix, Aritra Dasgupta, and Enrico Bertini. The exploratory labeling assistant: Mixed-initiative label curation with large document collections. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 153–164, 2018.

[FDJ+15] Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*, 2015.

[FKP+20] Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102 – 108, 2020.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[FOJ03] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003.

[FR21] Francisco Florez-Revuelta. Evosplit: An evolutionary approach to split a multilabel data set into disjoint subsets. *Applied Sciences*, 11(6):2823, 2021.

[Fra63] Paul Fraisse. *The psychology of time.* Harper & Row, 1963.

[Fra84] Paul Fraisse. Perception and estimation of time. *Annual review of psychology*, 35(1):1–37, 1984.

[FW22] Tao Fan and Hao Wang. Multimodal sentiment analysis of intangible cultural heritage songs with strengthened audio features-guided attention. *Journal of Information Science*, page 01655515221114454, 2022.

[GAFM11] Doron Goldfarb, Max Arends, Josef Froschauer, and Dieter Merkl. Revisiting 3d information landscapes for the display of art historical web content. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*, pages 1–8, 2011.

[Gar84] François Garnier. Thesaurus iconographique. *Système descriptif des reprèsentation, París: CNRS*, 1984.

[GBC11] Costantino Grana, Daniele Borghesani, and Rita Cucchiara. Automatic segmentation of digitalized historical manuscripts. *Multimedia Tools and Applications*, 55(3):483–506, 2011.

[GBG+18] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

[GCL+15] Zhao Geng, Tom Cheesman, Robert S Laramee, Kevin Flanagan, and Stephan Thiel. Shakervis: Visual analysis of segment variation of german translations of shakespeare's othello. *Information Visualization*, 14(4):273–288, 2015.

[GGLB18] Nicolas Gonthier, Yann Gousseau, Said Ladjal, and Olivier Bonfait. Weakly supervised object detection in artworks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[Gib11] Marek Gibney. Music-Map, 2011. https://www.music-map.de ((Accessed 2021-10-27).

[GKNV93] Emden R Gansner, Eleftherios Koutsofios, Stephen C North, and K-P Vo. A technique for drawing directed graphs. *IEEE Transactions on Software Engineering*, 19(3):214–230, 1993.

[GMD15] Katrin Glinka, Sebastian Meier, and Marian Dörk. Visualising the» unseen «: Towards critical approaches and strategies of inclusion in digital cultural heritage interfaces. *Kultur und Informatik (XIII)*, pages 105–18, 2015.

[GMM+17] Robert Gregor, Christof Mayrbrugger, Pavlos Mavridis, Benjamin Bustos, and Tobias Schreck. Cross-modal content-based retrieval for digitized 2d and 3d cultural heritage artifacts. In *GCH*, pages 119–123, 2017.

[GPD17] Katrin Glinka, Christopher Pietsch, and Marian Doerk. Past visions and reconciling views: Visualizing time, texture and themes in cultural collections. *Digital Humanities Quarterly*, 11(2), 2017.

[GRN19] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Context-aware embeddings for automatic art analysis. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 25–33, 2019.

[GRZL05] David Gleich, Matt Rasmussen, Leonid Zhukov, and Kevin Lang. The world of music: Sdp layout of high dimensional data. *Info Vis*, 2005:100, 2005.

[GSP+19] Ivan Giangreco, Loris Sauter, Mahnaz Amiri Parian, Ralph Gasser, Silvan Heller, Luca Rossetto, and Heiko Schuldt. Virtue: a virtual reality museum

experience. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, pages 119–120, 2019.

[GV17] Michael Grieves and John Vickers. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary perspectives on complex systems: New findings and approaches*, pages 85–113, 2017.

[GV18] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[GvTGD18] Flavio Gortana, Franziska von Tenspolde, Daniela Guhlmann, and Marian Dörk. Off the grid: Visualizing a numismatic collection as dynamic piles and streams. *Open Library of Humanities*, 4(2), 2018.

[HB10] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 203–212, 2010.

[HBN16] Masaki Hayashi, Steven Bachelder, and Masayuki Nakajima. Automatic generation of personal virtual museum. In *2016 International Conference on Cyberworlds (CW)*, pages 219–222. IEEE, 2016.

[HCC05] Chun-Rong Huang, Chu-Song Chen, and Pau-Choo Chung. Tangible photorealistic virtual museum. *IEEE computer graphics and applications*, 25(1):15–17, 2005.

[HCWC15] Peter Hall, Hongping Cai, Qi Wu, and Tadeo Corradi. Cross-depiction problem: Recognition and synthesis of photographs and artwork. *Computational Visual Media*, 1(2):91–103, 2015.

[HDS+19] Amir Hazem, Béatrice Daille, Dominique Stutzmann, Jacob Currie, and Christine Jacquin. Towards automatic variant analysis of ancient devotional texts. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 240–249, 2019.

[HEAB+17] Uta Hinrichs, Mennatallah El-Assady, Adam James Bradley, Stefania Forlini, and Christopher Collins. Risk the drift! stretching disciplinary boundaries through critical collaborations between the humanities and visualization. In *2nd Workshop on Visualization for the Digital Humanities, IEEE VIS 2017, Phoenix, Arizona, USA*, 2017.

[HFM18] Uta Hinrichs, Stefania Forlini, and Bridget Moynihan. In defense of sandcastles: Research thinking through visualization in digital humanities. *Digital Scholarship in the Humanities*, 34:i80–i99, 10 2018.

[HKBE12] Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012.

[HKP+22] Yumeng Hou, Sarah Kenderdine, Davide Picca, Mattia Egloff, and Alessandro Adamou. Digitizing intangible cultural heritage embodied: State of the art. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3):1–20, 2022.

[HM13] Nadav Hochman and Lev Manovich. Zooming into an instagram city: Reading the local through social media. *First Monday*, 18(7), 2013.

[HNBG22] Sorin Hermon, Franco Niccolucci, Nikolas Bakirtzis, and Svetlana Gasanova. Digital twins in cultural heritage. *Social Science Research Network*, 2022.

[HNH+12] Benjamin Höferlin, Rudolf Netzel, Markus Höferlin, Daniel Weiskopf, and Gunther Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 23–32. IEEE, 2012.

[Hof20] Sheila K Hoffman. Online exhibitions during the covid-19 pandemic. *Museum Worlds*, 8(1):210–215, 2020.

[HPK+21] Dongyun Han, Gorakh Parsad, Hwiyeon Kim, Jaekyom Shim, Oh-Sang Kwon, Kyung A Son, Jooyoung Lee, Isaac Cho, and Sungahn Ko. Hisva: A visual analytics system for studying history. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4344–4359, 2021.

[Hri16] Stefka Hristova. Images as data: cultural analytics and aby warburg's mnemosyne. *International Journal for Digital Art History*, 2, 2016.

[IFYA18] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.

[Inc14] Genius Media Group Inc. Genius.com, 2014. https://genius.com/ (Accessed 2023-02-22).

[Jän16] Stefan Jänicke. Valuable research for visualization and digital humanities: A balancing act. In *Workshop on Visualization for the Digital Humanities, IEEE VIS*, 2016.

[Jän18] Stefan Jänicke. Timages: Enhancing Time Graphs with Iconographic Information. In *LEVIA'18: Leipzig Symposium on Visualization in Applications*, 2018.

[Jar19] Ali Jarrahi, Mohammad Hossein und Eshraghi. Digital natives vs digital immigrants. *Journal of Enterprise Information Management*, 32(6), 2019.

[JBR+18] Stefan Jänicke, Judith Blumenstein, Michaela Rücker, Dirk Zeckzer, and Gerik Scheuermann. Tagpies: Comparative visualization of textual data. In *Visigrapp (3: Ivapp)*, pages 40–51, 2018.

[JDJ21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3), 2021.

[JFCS16] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. Visual Text Analysis in Digital Humanities. In *Computer Graphics Forum*. Wiley Online Library, 2016.

[JFS15] Stefan Jänicke, Josef Focht, and Gerik Scheuermann. Interactive visual profiling of musicians. *IEEE transactions on visualization and computer graphics*, 22(1):200–209, 2015.

[JGBS14] Stefan Jänicke, Annette Geßner, Marco Büchler, and Gerik Scheuermann. Visualizations for Text Re-use. In *Information Visualization Theory and Applications (IVAPP), 2014 International Conference on*, pages 59–70. IEEE, 2014.

[JGF+15] Stefan Jänicke, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony, and Gerik Scheuermann. Traviz: A visualization for variant graphs. *Digital Scholarship in the Humanities*, 30(suppl_1):i183–i199, 2015.

[JGS15] Stefan Jänicke, Annette Geßner, and Gerik Scheuermann. A distant reading visualization for variant graphs. *Proceedings of the Digital Humanities*, 2015, 2015.

[JKKS20] Stefan Jänicke, Pawandeep Kaur, Pawel Kuźmicki, and Johanna Schmidt. Participatory Visualization Design as an Approach to Minimize the Gap between Research and Application. In *The Gap between Visualization Research and Visualization Software (VisGap)*. The Eurographics Association, 2020.

[JLC19] Liu Jiang, Shixia Liu, and Changjian Chen. Recent research advances on interactive machine learning. *Journal of Visualization*, 22(2):401–417, 2019.

[JLK+16] Markus John, Steffen Lohmann, Steffen Koch, Michael Wörner, and Thomas Ertl. Visual analytics for narrative text-visualizing characters and their relationships as extracted from novels. In *VISIGRAPP (2: IVAPP)*, pages 29–40, 2016.

[JOV+20] Pauline Junginger, Dennis Ostendorf, Barbara Avila Vissirini, Anastasia Voloshina, Timo Hausmann, Sarah Kreiseler, and Marian Dörk. The close-up cloud. *International Journal for Digital Art History*, 5:6–2, 2020.

[JVHB14] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 9(6):1–12, 06 2014.

[JW17a] Stefan Jänicke and David Joseph Wrisley. Interactive visual alignment of medieval text versions. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 127–138. IEEE, 2017.

[JW17b] Stefan Jänicke and David Joseph Wrisley. Visualizing mouvance: Toward a visual analysis of variant medieval text traditions. *Digital Scholarship in the Humanities*, 32(suppl_2):ii106–ii123, 2017.

[Kab17] Katerina Kabassi. Evaluating websites of museums: State of the art. *Journal of Cultural Heritage*, 24:184–196, 2017.

[KAF+08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[Kas18] Andrei Kashcha. word2vec graph. "https://github.com/anvaka/word2vec-graph", 2018. (Retrieved 2023-02-22).

[KBD15] Florian Kräutli and Stephen Boyd Davis. Revealing cultural collections over time. In *EG UK Computer Graphics & Visual Computing*, pages 19––22, 2015.

[KBJ22] Richard Khulusi, Stephanie Billib, and Stefan Jänicke. Exploring life in concentration camps through a visual analysis of prisoners' diaries. *Information*, 13(2):54, 2022.

[KHP18] Ridwan Andi Kambau, Zainal Arifin Hasibuan, and M Octaviano Pratama. Classification for multiformat object of cultural heritage using deep learning. In *2018 Third International Conference on Informatics and Computing (ICIC)*, pages 1–7. IEEE, 2018.

[KJW+14] Steffen Koch, Markus John, Michael Wörner, Andreas Müller, and Thomas Ertl. Varifocalreader—in-depth visual analysis of large text documents. *IEEE transactions on visualization and computer graphics*, 20(12):1723–1732, 2014.

[KK89] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[KK17] Andrey Kutuzov and Elizaveta Kuzmenko. Building web-interfaces for vector semantic models with the webvectors toolkit. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 99–103, 2017.

[KKFJ19] Richard Khulusi, Jakob Kusnick, Josef Focht, and Stefan Jänicke. An interactive chart of biography. In *2019 IEEE Pacific Visualization Symposium (PacificVis)*, pages 257–266. IEEE, 2019.

[KKFJ20a] Richard Khulusi, Jakob Kusnick, Josef Focht, and Stefan Jänicke. musixplora: Visual analysis of a musicological encyclopedia. In *VISIGRAPP (3: IVAPP)*, pages 76–87, 2020.

[KKFJ20b] Jakob Kusnick, Richard Khulusi, Josef Focht, and Stefan Jänicke. A timeline metaphor for analyzing the relationships between musical instruments and musical pieces. In *VISIGRAPP (3: IVAPP)*, pages 240–251, 2020.

[KKM+20] Richard Khulusi, Jakob Kusnick, Christofer Meinecke, Christina Gillmann, Josef Focht, and Stefan Jänicke. A survey on visualizations for musical data. In *Computer Graphics Forum*. Wiley Online Library, 2020.

[KKP16] Chairi Kiourt, Anestis Koutsoudis, and George Pavlidis. Dynamus: A fully dynamic 3d virtual museum framework. *Journal of Cultural Heritage*, 22:984–991, 2016.

[KKZE19] Mosab Khayat, Morteza Karimzadeh, Jieqiong Zhao, and David S Ebert. Vassl: A visual analytics toolkit for social spambot labeling. *IEEE transactions on visualization and computer graphics*, 26(1):874–883, 2019.

[KLB14] Artjom Kochtchi, T von Landesberger, and Chris Biemann. Networks of names: Visual exploration and semi-automatic tagging of social networks from newspaper articles. In *Computer graphics forum*, volume 33, pages 211–220. Wiley Online Library, 2014.

[KMJ20] Kati Kallio, Eetu Mäkelä, and Maciej Janicki. Historical oral poems and digital humanities: Starting with a finnish corpus. *Folklore Fellows' Network*, 54, 2020.

[Kop02] Baruti N Kopano. Rap music as an extension of the black rhetorical tradition:" keepin'it real". *Western Journal of Black Studies*, 26(4):204, 2002.

[KPSK17] Kostiantyn Kucher, Carita Paradis, Magnus Sahlgren, and Andreas Kerren. Active learning and visual analytics for stance classification with alva. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):1–31, 2017.

[KRA+20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956––1981, 2020.

[KS13] Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1–21, 2013.

[KSD+21] Ryad Kaoua, Xi Shen, Alexandra Durr, Stavros Lazaris, David Picard, and Mathieu Aubry. Image collation: Matching illustrations in manuscripts. In *International Conference on Document Analysis and Recognition*, pages 351–366. Springer, 2021.

[KSKW15] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015.

[KTT09] Joon Hee Kim, Brian Tomasik, and Douglas Turnbull. Using artist similarity to propagate semantic information. In *ISMIR*, volume 9, pages 375–380, 2009.

[KYL+19] Nadezda Katricheva, Alyaxey Yaskevich, Anastasiya Lisitsina, Tamara Zhordaniya, Andrey Kutuzov, and Elizaveta Kuzmenko. Vec2graph: a python library for visualizing word embeddings as graphs. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 190–198. Springer, 2019.

[LA18] Susie Lu and John Akred. History of Rock in 100 Songs, 2018. https://svds.com/rockandroll/#thebeatles (Accessed 2023-02-22).

[Lal01] Élisabeth Lalou. Une base de données sur les manuscrits enluminés des bibliothèques: collaboration entre chercheurs et bibliothécaires. In *Bulletin des bibliothèques de France (BBF)*, volume 4, pages 38–42, 2001.

[LB21] Derek Lim and Austin R Benson. Expertise and dynamics within crowdsourced musical knowledge curation: A case study of the genius platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 373–384, 2021.

[LBT+18] Houda Lamqaddam, Koenraad Brosens, Frederik Truyen, Rudy Jos Beerens, Inez De Prekel, and Katrien Verbert. When the tech kids are running too fast: Data visualisation through the lens of art history research. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[LCHJ16] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691. Association for Computational Linguistics, 2016.

[LGLE18] James Jaehoon Lee, Blaine Greteman, Jason Lee, and David Eichmann. Linked reading: Digital historicism and early modern discourses of race around shakespeare's othello. *Journal ISSN*, 2371:4549, 2018.

[LGS+14] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992, 2014.

[LH21] François Lévesque and Thomas Hurtut. Muzlink: Connected beeswarm timelines for visual analysis of musical adaptations and artist relationships. *Information Visualization*, 20(2-3):170–191, 2021.

[LHP+22] Stefan Lengauer, Peter Houska, Reinhold Preiner, Elisabeth Trinkl, Stephan Karl, Ivan Sipiran, Benjamin Bustos, and Tobias Schreck. Interactive annotation of geometric ornamentation on painted pottery assisted by deep learning. *it-Information Technology*, 64(6):217–231, 2022.

[Lig99] Alan Light. *The Vibe History of Hip Hop*. Three Rivers Press, 1999.

[Lig12] Laura Light. The thirteenth century and the paris bible. In *The New Cambridge History of the Bible: Volume 2: From 600 to 1450*, pages 380–391, 2012.

[Lim07] SoundCloud Limited. Soundcloud, 2007. https://soundcloud.com/ (Accessed 2023-02-22).

[Lim08] WhoSampled.com Limited. Whosampled, 2008. https://www.whosampled.com/ (Accessed 2023-02-22).

[Liu20] Alan Liu. Toward a diversity stack: Digital humanities and diversity as technical problem. *PMLA*, 135(1):130–151, 2020.

[LKK+20] Stefan Lengauer, Alexander Komar, Stephan Karl, Elisabeth Trinkl, Ivan Sipiran, Tobias Schreck, and Reinhold Preiner. Semi-automated annotation of repetitive ornaments on 3d painted pottery surfaces. In *18th Eurographics Workshop on Graphics and Cultural Heritage: EG GCH 2020*. Eurographics Assoc., 2020.

[LKL+19] Stefan Lengauer, Alexander Komar, Arniel Labrada, Stephan Karl, Elisabeth Trinkl, Reinhold Preiner, Benjamin Bustos, and Tobias Schreck. Sketch-aided retrieval of incomplete 3d cultural heritage objects. In *3DOR@ Eurographics*, pages 17–24, 2019.

[LKM04] Beth Logan, Andrew Kositsky, and Pedro Moreno. Semantic analysis of song lyrics. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 2, pages 827–830. IEEE, 2004.

[LKS+18] Jean-Luc Lugrin, Florian Kern, Ruben Schmidt, Constantin Kleinbeck, Daniel Roth, Christian Daxer, Tobias Feigl, Christopher Mutschler, and Marc Erich Latoschik. A location-based vr museum. In *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–2. IEEE, 2018.

[LLL+18] Shenglan Liu, Xiang Liu, Yang Liu, Lin Feng, Hong Qiao, Jian Zhou, and Yang Wang. Perceptual visual interactive learning. *arXiv preprint arXiv:1810.10789*, 2018.

[LO18] Sabine Lang and Björn Ommer. Attesting similarity: Supporting the organization and study of art image collections with computer vision. *Digital Scholarship in the Humanities*, 33(4):845–856, 2018.

[Log07] William S Logan. Closing pandora's box: human rights conundrums in cultural heritage protection. In *Cultural heritage and human rights*, pages 33–52. Springer, 2007.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LSB+04] Fotis Liarokapis, Stella Sylaiou, Anirban Basu, Nicholaos Mourkoussis, Martin White, and Paul F Lister. An interactive visualisation interface for virtual museums. In *VAST*, pages 47–56, 2004.

[LTF+04] Céline Loscos, Franco Tecchia, Antonio Frisoli, Marcello Carrozzino, Hila Ritter Widenfeld, David Swapp, and Massimo Bergamasco. The museum of pure form: touching real statues in an immersive virtual museum. In *VAST*, pages 271–279, 2004.

[LZS16] Huibin Li, Jiawan Zhang, and Jizhou Sun. A visual analytics approach for deterioration risk analysis of ancient frescoes. *Journal of Visualization*, 19:529–542, 2016.

[LZSW17] Huibin Li, Jiawan Zhang, Jizhou Sun, and Jindong Wang. A visual analytics approach for flood risk analysis and decision-making in cultural heritage. *Journal of Visual Languages & Computing*, 41:89–99, 2017.

[MA20] Tom Monnier and Mathieu Aubry. docextractor: An off-the-shelf historical document element extraction. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 91–96. IEEE, 2020.

[Mal54] André Malraux. *Le musée imaginaire de la sculpture mondiale*, volume 3. Gallimard, 1954.

[Man16] Lev Manovich. The science of culture? social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*, 1(1):11060, 2016.

[MBM14] Raheleh Makki, Stephen Brooks, and Evangelos E Milios. Context-specific sentiment lexicon expansion via minimal user interaction. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 178–186. IEEE, 2014.

[MBV17] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*, 2017.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[McK] Jessica McKinney. Hip-Hop Becomes Most Popular Genre In Music For First Time In U.S. History – VIBE.com. https://www.vibe.com/music/music-news/hip-hop-popular-genre-nielsen-music-526795/ (Accessed 2023-02-22).

[MCS17] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191, 2017.

[MDG16] Fintan McGee, Marten During, and Mohammad Ghoniem. Towards visual analytics of multilayer graphs for digital cultural heritage. *Towards Visual Analytics of Multilayer Graphs for Digital Cultural Heritage*, 2016.

[Mei22] Christofer Meinecke. Labeling of cultural heritage collections on the intersection of visual analytics and digital humanities. In *7th Workshop on Visualization for the Digital Humanities, IEEE VIS 2022, Oklahoma City, USA*, 2022.

[Mey82] Paul Meyer. Étude sur les manuscrits du roman d'alexandre. *Romania*, 11(42/43):213–332, 1882.

[MGWJss] Christofer Meinecke, Estelle Guéville, David J. Wrisley, and Stefan Jänicke. A visual analytics framework for composing a hierarchical classification for medieval illuminations. In *Digital Scholarship in the Humanities*, in press.

[MHA17] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[MHJ22a] Christofer Meinecke, Ahmad Dawar Hakimi, and Stefan Jänicke. Explorative visual analysis of rap music. *Information*, 13(1):10, 2022.

[MHJ22b] Christofer Meinecke, Chris Hall, and Stefan Jänicke. Towards enhancing virtual museums by contextualizing art through interactive visualizations. *Journal on Computing and Cultural Heritage (JOCCH)*, 2022.

[MHM18] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[MHR12] Eetu Mäkelä, Eero Hyvönen, and Tuukka Ruotsalo. How to deal with massively heterogeneous cultural heritage data–lessons learned in culturesampo. *Semantic Web*, 3(1):85–109, 2012.

[MHSG18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[Mil95] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[MJ18] Christofer Meinecke and Stefan Jänicke. Visual analysis of engineers' biographies and engineering branches. In *LEVIA 2018: Leipzig Symposium on Visualization in Applications*, 2018.

[MJ21] Christofer Meinecke and Stefan Jänicke. Detecting text reuse and similarities between artists in rap music through visualization. In *LEVIA 2020: Leipzig Symposium on Visualization in Applications*. OSF Preprints, 2021.

[MMF19] Joyce Ma, Kwan-Liu Ma, and Jennifer Frazier. Decoding a complex visualization in a science museum–an empirical study. *IEEE transactions on visualization and computer graphics*, 26(1):472–481, 2019.

[MMOS+20] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

[Mor13] Franco Moretti. *Distant Reading*. Verso Books, 2013.

[MRB18] Francisco J Melero, Jorge Revelles, and Maria Luisa Bellido. Atalaya3d: Making universities' cultural heritage accessible through 3d technologies. In *GCH*, pages 31–35, 2018.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[MSEW22] Christofer Meinecke, Jeremias Schebera, Jakob Eschrich, and Daniel Wiegreffe. Visualizing similarities between american rap-artists based on text reuse. In *LEVIA 2022: Leipzig Symposium on Visualization in Applications*, 2022.

[MT14] Narges Mahyar and Melanie Tory. Supporting communication and coordination in collaborative sensemaking. *IEEE transactions on visualization and computer graphics*, 20(12):1633–1642, 2014.

[Mun09] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009.

[Mun14] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[MWJ19] Christofer Meinecke, David J. Wrisley, and Stefan Jänicke. Automated Alignment of Medieval Text Versions based on Word Embeddings. In *LEVIA 2019: Leipzig Symposium on Visualization in Applications*, 2019.

[MWJ21] Christofer Meinecke, David J Wrisley, and Stefan Janicke. Explaining semi-supervised text alignment through visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[MWJ22] Christofer Meinecke, David J. Wrisley, and Stefan Jänicke. From modern to medieval: Detecting and visualizing entities in manuscripts of marco polo's devisement du monde. In *DH 2022: "Responding to Asian Diversity"*, 2022.

[MWL+22] Eva Mayr, Florian Windhager, Johannes Liem, Samuel Beck, Steffen Koch, Jakob Kusnick, and Stefan Jänicke. The multiple faces of cultural heritage: Towards an integrated visualization platform for tangible and intangible cultural assets. In *2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, pages 13–18. IEEE, 2022.

[MYZ13] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[MZ18] Christoph Musik and Matthias Zeppelzauer. Computer vision and the digital humanities: Adapting image processing algorithms and ground truth through active learning. *VIEW Journal of European Television History and Culture*, 7(14):59–72, 2018.

[ndF03] Bibliothèque nationale de France. Mandragore, base des manuscrits enluminés de la bnf. http://mandragore.bnf.fr, 2003. (Retrieved 2023-02-22).

[OCF+15] Jorge H. Piazentin Ono, Débora Christina Corrêa, Martha Dais Ferreira, Rodrigo Fernandes Mello, and Luis Gustavo Nonato. Similarity graph: visual exploration of song collections. In *SIBGRAPI*. IEEE, Institute of Electrical and Electronics Engineers United States, 2015.

[Opg21] Loes Opgenhaffen. Visualizing archaeologists: A reflexive history of visualization practice in archaeology. *Open Archaeology*, 7(1):353–377, 2021.

[OSEAS15] Sergio Oramas, Mohamed Sordo, Luis Espinosa-Anke, and Xavier Serra. A semantic-based approach for artist similarity. In *Müller M, Wiering F, editors. Proceedings of the 16th International Society for Music Information Retrieval (ISMIR) Conference; 2015 Oct 26-Oct 30; Malaga, Spain.[Sl]: International Society for Music Information Retrieval; 2015. p. 100-6.* International Society for Music Information Retrieval (ISMIR), 2015.

[Ots79] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[PBT14] Denis Parra, Peter Brusilovsky, and Christoph Trattner. See what you want to see: visual user-driven approach for hybrid recommendation. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 235–240, 2014.

[PdJ78] Louis Petit de Julleville. *La Chanson de Roland : Traduction nouvelle rythmée et assonancée*. Alphonse Lemerre, Classiques français, 1878.

[PE16] Hermann Pflüger and Thomas Ertl. Sifting through visual arts collections. *Computers & Graphics*, 57:127–138, 2016.

[Pec99] Aaron Peckham. Urban dictionary, 1999. https://www.urbandictionary.com/ ((Accessed 2023-02-22).

[PKL+17] Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmqvist. Conceptvector: text visual analytics via interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics*, 24(1):361–370, 2017.

[QMSM17] Mohammed RH Qwaider, Anne-Lyse Minard, Manuela Speranza, and Bernardo Magnini. Find problems before they find you with annotatorpro's monitoring functionalities. In *CLiC-it*, 2017.

[RASJ14] Rafael P Ribeiro, Murilo AP Almeida, and Carlos N Silla Jr. The ethnic lyrics fetcher tool. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):27, 2014.

[RAZ+18] Christian Ritter, Christian Altenhofen, Matthias Zeppelzauer, Arjan Kuijper, Tobias Schreck, and Jürgen Bernard. Personalized visual-interactive music classification. In *EuroVA@ EuroVis*, pages 31–35, 2018.

[RGP+12] Patrick Riehmann, Henning Gruendl, Martin Potthast, Martin Trenkmann, Benno Stein, and Bernd Froehlich. Wordgraph: Keyword-in-context visualization for netspeak's wildcard search. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1411–1423, 2012.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Ric88] James Rice. Serendipity and holism: The beauty of opacs. *Library Journal*, 113(3):138–41, 1988.

[RMD+22] Luis Rei, Dunja Mladenic, Mareike Dorozynski, Franz Rottensteiner, Thomas Schleider, Raphaël Troncy, Jorge Sebastián Lozano, and Mar Gaitán Salvatella. Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*, pages 1–23, 2022.

[Roc71] Joseph Rocchio. Relevance feedback in information retrieval. *The Smart retrieval system-experiments in automatic document processing*, pages 313–323, 1971.

[RPSF15] Patrick Riehmann, Martin Potthast, Benno Stein, and Bernd Froehlich. Visual assessment of alleged plagiarism cases. In *Computer Graphics Forum*, volume 34, pages 61–70, 2015.

[ŘS10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[Sam19] Mark Sample. The black box and speculative care. *Debates in the Digital Humanities*, pages 445–448, 2019.

[SAN+17] Jiankai Sun, Deepak Ajwani, Patrick K Nicholson, Alessandra Sala, and Srinivasan Parthasarathy. Breaking cycles in noisy hierarchies. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 151–160, 2017.

[Sch04] Werner Schweibenz. Virtual museums. *The Development of Virtual Museums, ICOM News Magazine*, 3:3, 2004.

[Sch15] Kevin Schramm. Wer hat den größten?, 2015. https://story.br.de/rapwortschatz/ (Accessed 2021-10-27).

[Sch17] Susan Schreibman. Versioning machine. "http://v-machine.org/", 2017. (Retrieved 2023-02-22).

[SCO11] Joseph Schlecht, Bernd Carqué, and Björn Ommer. Detecting gestures in medieval images. In *2011 18th IEEE International Conference on Image Processing*, pages 1285–1288. IEEE, 2011.

[SFKP09] Sylaiou Styliani, Liarokapis Fotis, Kotsakis Kostas, and Patias Petros. Virtual museums, a survey and some issues for consideration. *Journal of cultural Heritage*, 10(4):520–528, 2009.

[SFM+20] Gjorgji Strezoski, Lucas Fijen, Jonathan Mitnik, Dániel László, Pieter de Marez Oyens, Yoni Schirris, and Marcel Worring. Tindart: A personal visual arts recommender. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4524–4526, 2020.

[SGBW18] Gjorgji Strezoski, Inske Groenen, Jurriaan Besenbruch, and Marcel Worring. Artsight: an artistic data exploration engine. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1240–1241, 2018.

[SH12] Markus Schedl and David Hauger. Mining microblogs to infer music artist similarity and cultural listening patterns. In *Proceedings of the 21st International Conference on World Wide Web*, pages 877–886, 2012.

[Shn96] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pages 336–343. IEEE, 1996.

[SIVA17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[SJS+21] Rita Sevastjanova, Wolfgang Jentner, Fabian Sperrle, Rebecca Kehlbeck, Jürgen Bernard, and Mennatallah El-Assady. Questioncomb: A gamification approach for the visual explanation of linguistic phenomena through interactive labeling. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–38, 2021.

[SK17] Piotr Szymański and Tomasz Kajdanowicz. A network perspective on stratification of multi-label data. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 22–35. PMLR, 2017.

[SKW05a] Markus Schedl, Peter Knees, and Gerhard Widmer. Discovering and visualizing prototypical artists by web-based co-occurrence analysis. In *ISMIR*, pages 21–28, 2005.

[SKW05b] Markus Schedl, Peter Knees, and Gerhard Widmer. A web-based approach to assessing artist similarity using co-occurrences. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05)*. Citeseer, 2005.

[SLB+09] Ray Siemens, Cara Leitch, Analisa Blake, Karin Armstrong, and John Willinsky. "it may change my understanding of the field": Understanding reading tools for scholars and professional readers. *Digital Humanities Quarterly*, 3(4), 2009.

[SLK+19] Luke S Snyder, Yi-Shan Lin, Morteza Karimzadeh, Dan Goldwasser, and David S Ebert. Interactive learning for identifying relevant tweets to support real-time situational awareness. *IEEE transactions on visualization and computer graphics*, 26(1):558–568, 2019.

[SLT17] Yunjia Sun, Edward Lank, and Michael Terry. Label-and-learn: Visualizing the likelihood of machine learning classifier's success during data labeling. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 523–534, 2017.

[SOR04] Corina Sas, Gregory O'Hare, and Ronan Reilly. A performance analysis of movement patterns. In *International Conference on Computational Science*, pages 954–961. Springer, 2004.

[Spo18] Spotify. Spotify Artist Explorer, 2018. https://artist-explorer.glitch.me/ (Accessed 2023-02-22).

[SRF+19] Victor Schetinger, Kathrin Raminger, Velitchko Filipov, Nathalie Soursos, Susana Zapke, and Silvia Miksch. Bridging the gap between visual analytics and digital humanities: Beyond the data-users-tasks design triangle. In *4th Workshop on Visualization for the Digital Humanities, IEEE VIS 2019, Vancouver, British Columbia, Canada*, 2019.

[SSB+19] Gjorgji Strezoski, Arumoy Shome, Riccardo Bianchi, Shruti Rao, and Marcel Worring. Ace: Art, color and emotion. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1053–1055, 2019.

[SSKEA19] Fabian Sperrle, Rita Sevastjanova, Rebecca Kehlbeck, and Mennatallah El-Assady. Viana: Visual interactive annotation of argumentation. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 11–22. IEEE, 2019.

[SSZW08] Tobias Schreck, Michael Schüßler, Frank Zeilfelder, and Katja Worm. Butterfly plots for visual analysis of large point cloud data. In *16th International Conference in Central Europe on Computer Graphics, WSCG*, pages "33–40, 2008.

[STN+16] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, 2016.

[STT81] Kozo Sugiyama, Shojiro Tagawa, and Mitsuhiko Toda. Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 11(2):109–125, 1981.

[STV11] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.

[SW18] Gjorgji Strezoski and Marcel Worring. Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–21, 2018.

[SWS+00] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

[SYN+14] Shoto Sasaki, Kazuyoshi Yoshii, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima. Lyricsradar: A lyrics retrieval system based on latent topics of lyrics. In *Ismir*, pages 585–590, 2014.

[TGBvdB14] Martin Tröndle, Steven Greenwood, Konrad Bitterli, and Karen van den Berg. The effects of curatorial arrangements. *Museum Management and Curatorship*, 29(2):140–173, 2014.

[THC12] Alice Thudt, Uta Hinrichs, and Sheelagh Carpendale. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1461–1470, 2012.

[Tim05] Sebastiano Timpanaro. *The genesis of Lachmann's method*. University of Chicago Press, 2005.

[TL19] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[TMdS19] Shuai Tang, Mahta Mousavi, and Virginia R de Sa. An empirical study on post-processing methods for word embeddings. *arXiv preprint arXiv:1905.10971*, 2019.

[Tow20] NLP Town. Bert base multilingual uncased sentiment, 2020. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment (Accessed 2023-02-22).

[Twi93] Michael Twidale. Redressing the balance: the advantages of informal evaluation techniques for intelligent learning environments. *Journal of Artificial Intelligence in Education*, 4:155–155, 1993.

[Und17] Ted Underwood. A genealogy of distant reading. *Digital Humanities Quarterly*, 11(2), 2017.

[UNEO01] Scientific United Nations Educational and Cultural Organization. Universal declaration on cultural diversity. In *adopted by the 31st Session of the General Conference of UNESCO, Paris, 2 November*, 2001.

[UNEU03] Scientific United Nations Educational and Cultural Organization (UNESCO). Convention for the safeguarding of the intangible cultural heritage, 2003.

[ÚSN+20] Ignacio Úbeda, Jose M Saavedra, Stéphane Nicolas, Caroline Petitjean, and Laurent Heutte. Improving pattern spotting in historical documents using feature pyramid networks. *Pattern Recognition Letters*, 131:398–404, 2020.

[Vav17] Frederic Vavrille. LivePlasma, 2017. http://www.liveplasma.com/ (Accessed 2023-02-22).

[VCPK09] Romain Vuillemot, Tanya Clement, Catherine Plaisant, and Amit Kumar. What's being said near "martha"? exploring name entities in literary text collections. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 107–114. IEEE, 2009.

[VKCA20] Maria Vayanou, Akrivi Katifori, Angeliki Chrysanthi, and Angeliki Antoniou. Cultural heritage and social experiences in the times of covid 19. In *AVI² CH@ AVI*, 2020.

[vN22] Nanne van Noord. A survey of computational methods for iconic image analysis. *Digital Scholarship in the Humanities*, 2022.

[vW59] Walther von Wartburg. *Französisches etymologisches Wörterbuch*. F. Klopp, 1959.

[VZAA20] Joris Van Zundert, Smiljana Antonijević, and Tara Andrews. 6. 'black boxes' and true colour — a rhetoric of scholarly code. *Digital Technology and the Practices of Humanities Research*, pages 123–162, 02 2020.

[WCH14] Qi Wu, Hongping Cai, and Peter Hall. Learning graphs to model visual objects across different depictive styles. In *European Conference on Computer Vision*, pages 313–328. Springer, 2014.

[WCH16] Nicholas Westlake, Hongping Cai, and Peter Hall. Detecting people in artwork with cnns. In *European Conference on Computer Vision*, pages 825–841. Springer, 2016.

[WCW06] Krzysztof Walczak, Wojciech Cellary, and Martin White. Virtual museum exbibitions. *Computer*, 39(3):93–95, 2006.

[WD16] Martin Warnke and Lisa Dieckmann. Prometheus meets meta-image: implementations of aby warburg's methodical approach in the digital era. *Visual Studies*, 31(2):109–120, 2016.

[WDC+17] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T Brown, and Alex Endert. Podium: Ranking data using mixed-initiative visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):288–297, 2017.

[WDS+20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics.

[WFH+01] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H Witten. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292, 2001.

[WFS+18] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE transactions on visualization and computer graphics*, 25(6):2311–2330, 2018.

[WHHA11] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. Commentspace: structured support for collaborative visual analysis. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 3131–3140, 2011.

[Whi15] Mitchell Whitelaw. Generous interfaces for digital cultural collections. *Digital Humanities Quarterly*, 9(1), 2015.

[WJ13] Dana Wheeles and Kristin Jensen. Juxta commons. In *In Proceedings of the Digital Humanities 2013*, 2013.

[WMM+20] Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. Urban dictionary embeddings for slang nlp applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4764–4773, 2020.

[Wri18] David Joseph Wrisley. Pre-visualization. In *3rd Workshop on Visualization for the Digital Humanities, IEEE VIS 2018, Berlin, Germany*, 2018.

[WRZ+15] Chaoli Wang, John P Reese, Huan Zhang, Jun Tao, Yi Gu, Jun Ma, and Robert J Nemiroff. Similarity-based visualization of large image collections. *Information Visualization*, 14(3):183–203, 2015.

[WS20] Melvin Wevers and Thomas Smits. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1):194–207, 2020.

[WSK+17] Florian Windhager, Matthias Schlögl, Maximilian Kaiser, Ágoston Zénó Bernád, Saminu Salisu, and Eva Mayr. Beyond one-dimensional portraits: A synoptic approach to the visual analysis of biography data. In *BD*, pages 67–75, 2017.

[WSZ+20] Jian Wu, Victor S Sheng, Jing Zhang, Hua Li, Tetiana Dadakova, Christine Leon Swisher, Zhiming Cui, and Pengpeng Zhao. Multi-label active learning algorithms for image classification: Overview and future promise. *ACM Computing Surveys (CSUR)*, 53(2):1–35, 2020.

[WVL+15] Andy T Woods, Carlos Velasco, Carmel A Levitan, Xiaoang Wan, and Charles Spence. Conducting perception research over the internet: a tutorial review. *PeerJ*, 3:e1058, 2015.

[XEJJ14] Weijia Xu, Maria Esteva, Suyog D Jain, and Varun Jain. Interactive visualization for curatorial analysis of large digital collection. *Information Visualization*, 13(2):159–183, 2014.

[XYX+19] Shouxing Xiang, Xi Ye, Jiazhi Xia, Jing Wu, Yang Chen, and Shixia Liu. Interactive correction of mislabeled training data. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2019.

[YJ20] Tariq Yousef and Stefan Jänicke. A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[YMCO10] Pradeep Yarlagadda, Antonio Monroy, Bernd Carque, and Björn Ommer. Recognition and analysis of objects in medieval images. In *Asian Conference on Computer Vision*, pages 296–305. Springer, 2010.

[YMDK11] So Yamaoka, Lev Manovich, Jeremy Douglass, and Falko Kuester. Cultural analytics in large-scale visualization environments. *Computer*, 44(12):39–48, 2011.

[You19] Tariq Yousef. Ugarit: Translation Alignment Visualization. In *LEVIA 2019: Leipzig Symposium on Visualization in Applications*, 2019.

[YQFF+20] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[YTRC19] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–16, 2019.

[ZGB+16] Jian Zhao, Michael Glueck, Simon Breslav, Fanny Chevalier, and Azam Khan. Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations. *IEEE transactions on visualization and computer graphics*, 23(1):261–270, 2016.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[Zum92] Paul Zumthor. *Toward a medieval poetics*. U of Minnesota Press, 1992.

List of Figures