2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3503-2838-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMAR59233.2023.00027

Who Did What When? Discovering Complex Historical Interrelations in Immersive Virtual Reality

Melanie Derksen[*] TU Dortmund University
Julia Becker Bielefeld University
Mohammad Fazleh Elahi Bielefeld University
Angelika Maier Bielefeld University
Marius Maile Bielefeld University
Ingo Pätzold Bielefeld University
Jonas Penningroth Bielefeld University
Bettina Reglin Bielefeld University
Markus Rothgänger Bielefeld University
Philipp Cimiano Bielefeld University
Erich Schubert TU Dortmund University
Silke Schwandt Bielefeld University
Torsten Kuhlen[†] RWTH Aachen University
Mario Botsch [‡] TU Dortmund University
Tim Weissker[§] RWTH Aachen University

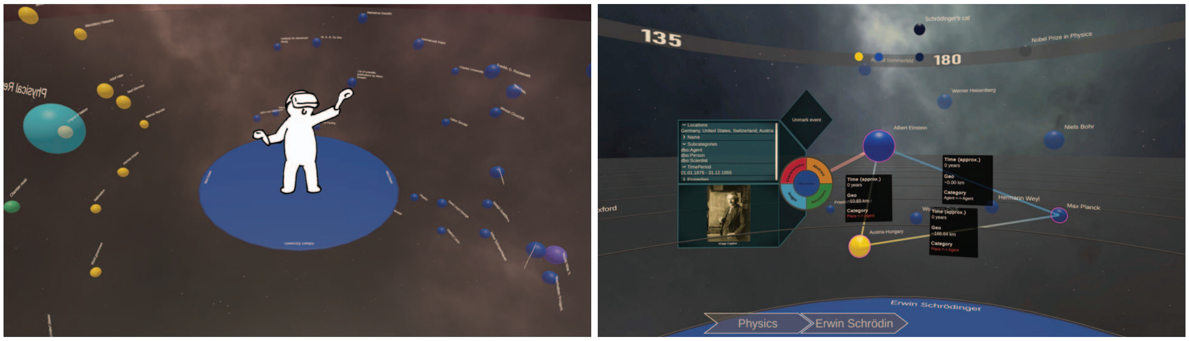

Figure 1: We present a tool for exploring historical data in VR using a head-mounted display. **Left:** The user is placed on a platform representing the main historical fragment (e.g., a person or an event) of their research (here *Albert Einstein*). Other related fragments are represented as colored labeled spheres and are arranged around the user based on their historical interrelationships. Users can select a sphere to make it the new main fragment, thereby allowing for an immersive exploration of the dataset. **Right:** For every sphere, a *detail window* with more information on the corresponding fragment can be opened. An exploration history in the lower field of view shows the order of visited main fragments (here *Physics* and *Erwin Schrödinger*). If a sphere is marked, a corresponding mark appears on a compass view in the upper field of view (e.g., a blue mark for Albert Einstein). The *relation details*, here shown for *Albert Einstein*, *Austria-Hungary*, and *Max Planck*, reveal additional details about their historical connections.

Abstract

Traditional digital tools for exploring historical data mostly rely on conventional 2D visualizations, which often cannot reveal all relevant interrelationships between historical fragments (e.g., persons or events). In this paper, we present a novel interactive exploration tool for historical data in VR, which represents fragments as spheres in a 3D environment and arranges them around the user based on their *temporal*, *geo*, *categorical* and *semantic* similarity. Quantitative and qualitative results from a user study with 29 participants revealed that most participants considered the virtual space and the abstract fragment representation well-suited to explore historical data and to discover complex interrelationships. These results were particularly underlined by high usability scores in terms of attractiveness, stimulation, and novelty, while researching historical facts with our system did not impose unexpectedly high task loads. Additionally, the insights from our post-study interviews provided valuable suggestions for future developments to further expand the possibilities of our system.

Index Terms: Human-centered computing—Visualization—Visualization techniques; Human-centered computing—Visualization—Visualization systems and tools

1 Introduction

Exploring historical relationships requires the analysis and comparison of different sources that are often difficult to interpret without specialized domain knowledge. To assist learners with this process, several digital tools have been proposed that provide curated interactive visualizations like maps, timelines, or graphs. While some tools like *Palladio*[1] are more tailored to the purpose of presenting finished research results, others like *HisVA* [9] focus on an active learning process while interacting with the data, which is claimed to be more beneficial than traditional textbook-based learning. However, historical exploration tools to date are mostly based on classic 2D visualizations which are inherently optimized for demonstrating attribute relationships along only one or two spatial dimensions. Unfortunately, this is often not optimal to discover more complex interrelationships between historical fragments (e.g., persons or events), which still requires considerable manual efforts.

To approach this limitation, we present a novel interactive exploration tool for historical data in VR. The system is not only designed for actors from the humanities and historical sciences but also for laypersons interested in history and people with an affinity for technology. Our approach represents historical fragments as colored

[*] e-mail: melanie.derksen@tu-dortmund.de
[†] e-mail: kuhlen@vr.rwth-aachen.de
[‡] e-mail: mario.botsch@tu-dortmund.de
[§] e-mail: me@tim-weissker.de

[1] http://hdlab.stanford.edu/palladio/

DOI 10.1109/ISMAR59233.2023.00027

labeled spheres, which are arranged around the user to allow for an immersive exploration of the dataset (see Fig. 1). The position of a fragment can be influenced by more than two attributes, which allows to make complex interrelationships more directly visible to the user. In addition to the egocentric viewing position, the user can also switch to an exocentric viewing position to gain a better overall impression of the fragments' structure.

Our system was developed as part of an iterative design process in close collaboration with a team of historians. While an initial expert review published in an earlier stage of development revealed that our visualization of historical data is promising, it was suggested to add more interactive features to increase its versatility and ability to answer more complex historical questions [5]. Following the experts' advice, we therefore expanded the feature set of our system by adding options like detailed information panels on single fragments as well as on their interrelationships (see Fig. 1 right). We then conducted a formal user study to analyze the overall usability and suitability of our system as a historical research tool for laypeople. In summary, our research led to the following contributions:

- a novel way of visualizing and arranging historical fragments in a meaningful way to reveal complex interrelationships,
- a corresponding set of interaction techniques that enables users to interactively adjust the visualization's parameters in order to facilitate the exploration process,
- quantitative results of a user study with 29 historically inexperienced participants demonstrating that our system was suitable for answering historical questions in a particularly attractive, stimulating, and novel way while not inducing unexpectedly high task loads compared to other systems in the literature,
- qualitative results based on interviews showing that both the 3D virtual space and the abstract visualization are well suited to explore historical data and revealing promising areas for future developments.

Our findings suggest that our system has the potential to serve as a viable supplement to conventional historical exploration methods.

2 Related Work

We first discuss the existing landscape of interactive historical data exploration tools for 2D desktop systems (Section 2.1). We then analyze the benefits of 3D and immersive data exploration, as shown by systems in other domains, to motivate our developments (Section 2.2). Finally, we discuss the different reference frames relevant for the design of immersive data exploration systems (Section 2.3).

2.1 Interactive Exploration of Historical Data in 2D

The idea of self-guided learning and knowledge acquisition via digital tools in history has been explored in several prior publications. Visual analytics systems like POLIS [14], VAiRoma [2], and HisVA [9] follow the same approach to help people explore historical datasets by providing individual overviews, mostly consisting of traditional and well-known visualization methods like data tables, timelines or map views, and revealing relationships among events that may not be directly apparent. Each of these approaches focuses on a specific topic, but the presented concepts can also be transferred to different datasets. While POLIS focuses on the sociopolitical landscape of the Ancient Greek world, VAiRoma addresses the Roman history and stands out with its topic view for displaying topic hierarchies, topic content, as well as topic weights. HisVA focuses on an active learning process while interacting with the data, which is claimed to be more beneficial than traditional textbook-based learning. Apart from that, VisKonnect [15] is intended to represent characters involved in common events and comes up with a relationship graph as well as a chat interface that enables typing in a query and generates a short textual answer.

It is to be noted that each of the mentioned tools is a 2D non-immersive desktop application which comes along with a limited space and consequently might not be optimal for visualizations that should offer more complex relationships. That is why we take a closer look at the possible benefits of a third spatial dimension and an immersive virtual environment (VE), which offer a much larger exploration space and can make attribute relationships along more than two spatial dimensions visible.

2.2 Benefits of 3D and Immersive Data Exploration

Lisle et al. [17] propose and investigate the so-called Immersive Space to Think, that helps analysts to better understand large text-based datasets. Here the immersive space is used to organize documents according to individual preferences to support the individual sensemaking process, which is stated to be easier as compared to a traditional desktop or laptop display. McIntire and Liggett [19] analyze several data and information visualization applications to identify which kinds of tasks benefit from 3D representations. They conclude that, among others, 3D visualizations are particularly beneficial for the precise spatial localization of objects, complex imagery analysis, and the manual interaction with data or virtual information. Furthermore, they found out that 3D visualizations can provide performance benefits that seem to reflect cognitive benefits, which provides for an increased understanding of spatial and/or multi-dimensional data. Etemadpour et al. [6] investigate the effect of stereoscopic environments compared to a 2D screen when used for the visual analysis of multi-dimensional data after projection into a 3D visual space. Their user study confirms that distances between individual objects can be perceived better in VR, which leads to an overall improved performance for local analysis tasks that focus on a specific part of the visuals. Allcoat and von Mühlenen [1] directly investigate the effects of interactive learning in VR on performance, emotion, and engagement by comparing it to conventional textbook learning as well as passive learning by watching videos. Their results show that learning in VR leads to the highest engagement, an increase in positive emotions, and a decrease in negative emotions. Kraus et al. [13] investigate the impact of immersion on cluster identification tasks in abstract scatterplot visualizations. They show that the 2D visualization on the screen performs worse compared to the 3D visualizations with respect to accuracy, efficiency, memorability, sense of orientation, and user preference. VR, on the other hand, allows for improved overviews of 3D data due to more natural navigation and better orientation and memorability possibilities. On top of that, Wagner Filho et al. [23] state that exploring 3D scatterplots with an head-mounted display (HMD) leads to a smaller effort in finding information and offers a much larger subjective perception of accuracy and engagement as opposed to desktop applications.

Taken together, all of these results indicate that the visualization and exploration of data in an immersive environment can provide substantial benefits over conventional 2D data representations. Motivated by these insights, we saw merit in the development and analysis of an immersive exploration tool for the field of history and proceeded with more detailed design questions for such a system.

2.3 Frames of Reference for Immersive Data Exploration

Immersive visualization systems may give the user an *exocentric* view onto the data from external viewpoints as well as an *egocentric* view onto the data from within the data themselves. Wagner et al. [22] investigate the effect of these two frames of reference in a more detailed study. While their results indicate that the egocentric frame of reference significantly reduces mental workload, they also observe that the exocentric frame of reference improves user performance in some of the given tasks. They therefore suggest to allow users to switch between both frames of reference based on their intentions, which we adapt in the design of our immersive system for exploring historical data.

3 Immersive Exploration of Historical Data in VR

Our VR system builds upon *DBPedia*[2] and *WikiData*[3] to retrieve important historical fragments as well as connections between them. Using the *Unity* game engine, we embed fragments into a 3D VE to be explored with an HMD and corresponding controllers. While we specialized our developments on the *HTC Vive*[4] and the Vive's dedicated controllers, the underlying concepts are applicable to other HMDs as well.

3.1 System Basics

The system starts with the selection of an initial historical fragment the user is interested in. This fragment then becomes the first *main fragment*, which is represented as a colored circular platform the user is placed on. The DBpedia graph database provides all fragments connected to that main fragment. We take the first 50 of these so as not to overload the visualization. These so-called related fragments are represented as labeled colored spheres and arranged around the platform in a spherical fashion (see Fig. 1 left) in our visualization, which puts the user into an egocentric view from within the data. While the color of each sphere indicates its fragment's category (e.g., agent, event, ...), the sphere's size is related to the number of sources (i.e., `dbo:wikiPageExternalLinks`) available in the corresponding article on DBpedia.

3.2 Arrangement of Related Fragments Around the User

Related fragments are arranged around the user based on two dissimilarity measures, namely the *relative dissimilarity* D_{rel} and the *central dissimilarity* D_{ct}. Both of these high-level dissimilarities are, in turn, modeled as weighted combinations of low-level dissimilarities regarding individual attributes of fragments. D_{rel} is computed for each pair of related fragments and used to position the spheres on a spherical shell around the user so that fragments with a low value for D_{rel} are close to each other. D_{ct} is calculated for each pair of the main fragment with a related fragment, which then determines the individual distance (or radius) of that related fragment to the center.

3.2.1 General Arrangement Algorithm

In the first step of the arrangement algorithm, all spheres are placed on a spherical shell around the user, all having the same distance to the center. To do so, we rely on the *UMAP* algorithm [18], a scalable algorithm for dimension reduction, which is competitive with t-SNE in terms of visualization quality, and arguably preserves more of the global structure with superior run time performance. Additionally, it allows various output metrics, which is suitable for our case, as we want a spherical embedding. With the D_{rel} as the input metric and the attribute values of the fragments as input data, UMAP finds, with the haversine as its output metric, a spherical embedding. As an embedding all over a sphere makes some data hard to see and may lead to dislocations of the neck, we add two additional artificial points in its optimization process that are fixed on the poles and constantly repel all the other data points to bypass constellations in which spheres accumulate directly above or beneath the user. Based on this initial arrangement of spheres the second step of our arrangement algorithm then uses D_{ct} to change the radial distance of each related fragment's sphere according to their dissimilarity value. Thus, spheres are either pushed farther away or placed closer to the center while staying in the interval of a minimal distance of 1.7 m and the maximal distance of 3 m which were considered to be appropriate distances in our design process. In the final layout, longitude and latitude are determined through UMAP based on D_{rel}, while the radial distance to the center is computed based on D_{ct}. This two-step procedure has the benefit, that fragments are less likely placed behind one another, but all are visible from the center.

3.2.2 Computation of D_{rel} and D_{ct}

The high-level dissimilarities D_{rel} and D_{ct} are based on a weighted combination of the fragments' low-level dissimilarities regarding their *temporal* (t), *geo* (g), and either their *semantic* (sem) or *categorical* (c) attributes values. For a visualization that consists of the set of fragments F, we first compute pairwise distances $dist_a$ for all fragments $i, j \in \mathrm{F}$ for each attribute $a \in \{t, g, sem, c\}$. Then we transform them per attribute a to a low-level dissimilarity $d_a \in [0,1]$ where 0 means maximally *similar* and 1 maximally *dissimilar*. This leads to a normalized measure of dissimilarities for each individual attribute. This is needed since our distances for the attributes are measured in different units with different orders of magnitude. Thus, if we would skip the transformation, an attribute with higher order of magnitude would have a greater impact on D_{rel} and D_{ct}, respectively. If a_i is the attribute value for the attribute a of fragment i, then it is considered invalid if no data is available for a_i. More formally, a low-level dissimilarity d_a of two fragments $i, j \in \mathrm{F}$ regarding one attribute $a \in \{t, g, sem, c\}$, with a_i, a_j being their corresponding attribute values, is defined as

$$\mathrm{d}_a(i,j) = \begin{cases} \mathrm{dScaling}_a(i,j) & \text{if both } a_i, a_j \text{ are valid,} \\ 0 & \text{if both } a_i, a_j \text{ are invalid,} \\ 1 & \text{otherwise,} \end{cases}$$

with $\mathrm{dScaling}_a$ being

$$\mathrm{dScaling}_a(i,j) = \begin{cases} \frac{\mathrm{dist}_a(i,j) - distMin}{dist_{max} - dist_{min}} & \text{if } dist_{min} \neq dist_{max}, \\ 0 & \text{if } dist_{min} = dist_{max} = 0, \\ 1 & \text{otherwise.} \end{cases}$$

The value

$$dist_{max} = \max_{m,o \in \mathrm{F}} \mathrm{dist}_a(m,o)$$

represents the maximal distance according to attribute a within the frament set F and

$$dist_{min} = \min_{k,l \in \mathrm{F},\, k \neq l} \mathrm{dist}_a(k,l)$$

is the respective minimal distance according to attribute a. Since the value for $\mathrm{dScaling}_a$ depends on the minimal and maximal distances of the set of fragments F of the current visualization, a pair of two fragments may have different values for d_a if they appear hand in hand for different main fragments and therefore different sets of related fragments with different minimal and maximal distances regarding the attributes. While D_{rel} takes a convex combination of the low-level dissimilarities regarding the attribute values of t, g, c for two fragments i, j into account, D_{ct} makes use of the values for t, g, sem:

$$\mathrm{D}_{rel}(i,j) = w_{rel,t}\mathrm{d}_t(i,j) + w_{rel,g}\mathrm{d}_g(i,j) + w_{rel,c}\mathrm{d}_c(i,j), \quad (1)$$

$$\mathrm{D}_{ct}(i,j) = w_{ct,t}\mathrm{d}_t(i,j) + w_{ct,g}\mathrm{d}_g(i,j) + w_{ct,sem}\mathrm{d}_{sem}(i,j), \quad (2)$$

$$\text{with } w_{rel,t} + w_{rel,g} + w_{rel,c} = 1,\ w_{ct,t} + w_{ct,g} + w_{ct,sem} = 1,$$
$$w_{rel,t}, w_{rel,g}, w_{rel,c}, w_{ct,t}, w_{ct,g}, w_{ct,sem} \geq 0.$$

As a result, if e.g., $w_{rel,g}$ has a high value, fragments will aggregate into clusters if they share locations with a short distance. Additionally, if e.g., $w_{ct,t}$ has a high value, fragments are placed closer to

[2] https://www.dbpedia.org

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

[4] https://www.vive.com/de/product/vive-pro/

the center that took place at the same time. The so-called *weighting triangles* serve as appropriate UI elements that allow to modify the respective weights that influence both dissimilarity estimations independently (see Fig. 2). The user can select a point in one of the triangles with a ray attached to the user's controller. The weights are determined as the barycentric coordinates of that point with respect to the triangle corners. In this way the arrangement of fragments can be modified based on the user's interest and the result provides insight into the relation of surrounding fragments among themselves as well as to the main fragment. In the following we present how we calculate the separate distances for the attributes t, g, sem, c, which are the bases for the corresponding dissimilarities.

Computation of Temporal Distance To compute the temporal distance dist_t between two fragments i, j, the minimal separating interval regarding their time periods t_i and t_j is considered. A fragment's time period t_i is an interval $\left[t_{i,s}, t_{i,e}\right]$ with a start date $t_{i,s}$ and an end date $t_{i,e}$, each of which is converted to a number of days. Two time periods t_i and t_j overlap, if their intersection is not empty: $t_i \bigcap t_j \neq \emptyset$. The temporal distance between two fragments i, j in days dist_t is calculated with

$$\mathrm{dist}_t\left(i,j\right) = \begin{cases} \min\left\{\left|\left|t_{i,s} - t_{j,e}\right|\right|, \left|\left|t_{j,s} - t_{i,e}\right|\right|\right\} & \text{if } t_i \bigcap t_j = \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Computation of Geo Distance To compute the geo distance dist_g between two fragments i, j, the great circle distances dist_{GC} of all their corresponding locations are compared in pairs, and the shortest distance value is taken. In a figurative sense, this provides information about whether two fragments happened or had been in the same place. Since one fragment i may have multiple locations, its set of locations is $\mathrm{G}_i = \left\{g_{i,n}\right\}_{n\in\mathbb{N}}$, where one location $g_{i,n}$ is a pair of longitude and latitude $\left(\phi_{i,n}, \theta_{i,n}\right)$. The geo distance between two fragments i, j in km can thus be calculated by

$$\begin{aligned} \mathrm{dist}_g\left(i,j\right) &= \min_{g_{i,n}\in\mathrm{G}_i} \min_{g_{j,m}\in\mathrm{G}_j} \mathrm{dist}_{GC}\left(g_{i,n}, g_{j,m}\right), \\ \mathrm{dist}_{GC}\left(g_{i,n}, g_{j,m}\right) &= r \arccos\left(\sin\left(\theta_{i,n}\right)\sin\left(\theta_{j,m}\right)\right. \\ &\quad \left. + \cos\left(\theta_{i,n}\right)\cos\left(\theta_{j,m}\right)\cos\left(\phi_{i,n} - \phi_{j,m}\right)\right), \end{aligned}$$

with $r \approx 6371$ km being the Earth's radius. Note that the multiplication with r is only relevant for displaying the shortest distance in km in the relation details (see Fig. 1 right). For the arrangement calculation with D_{rel} and D_{ct}, it can be neglected.

Computation of Semantic Distance As semantic distance, we make use of the *dbo:abstract* entries from DBPedia, and the Sentence Mover's Distance (SMD) similar to the Sentence Mover's Similarity [3], which is a metric for automatically evaluating multi-sentence texts. To calculate the semantic distance dist_{sem} between two fragments i, j, we therefore transform their abstracts into sets of sentence embedding vectors sem_i, sem_j, which serve as the basis for the SMD calculation $\mathrm{SMD}\left(sem_i, sem_j\right)$. The corresponding sentence embedding vectors are taken from the pretrained sBERT model [21] 'all-MiniLM-L12-v2'[5]. As entries for the document vectors used in the SMD calculation, we use uniform weights instead of custom weights based on the number of words per sentence. Furthermore, we choose the cosine distance as the SMD's cost function. Finally, since our concept relies on distances, we omit the last computation step of Clark et al. [3] which would transform the SMD into a similarity. Therefore, dist_{sem} is given by

$$\mathrm{dist}_{sem}\left(i,j\right) = \mathrm{SMD}\left(sem_i, sem_j\right).$$

Computation of Categorical Distance The categorical distance dist_c is 1 if two fragments do not share the same category. Otherwise, it is set to 0. Therefore, if c_i and c_j are the categories of two fragments i, j, then dist_c is defined as:

$$\mathrm{dist}_c\left(i,j\right) = \begin{cases} 1 & \text{if } c_i \neq c_j, \\ 0 & \text{otherwise.} \end{cases}$$

3.3 User Interaction

Our initial expert review presented in earlier work [5] already provided information about which interaction possibilities are desirable for a productive exploration process. That served as a basis for the development of a set of interaction techniques that enable users to browse through our visualizations interactively. Navigation is done exclusively through movements in the real world. However, the application is designed in such a way that everything can be done from one position. In the following we describe what kind of specific interactions and items were developed.

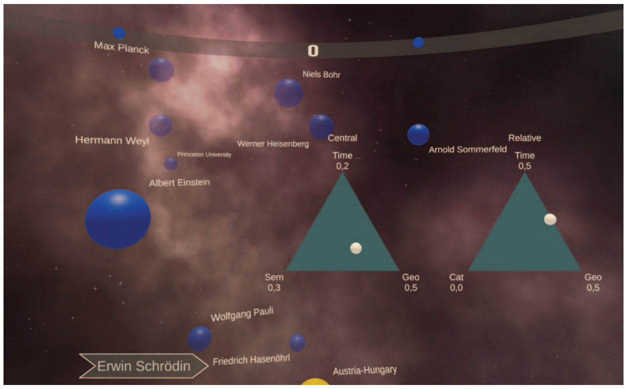

Figure 2: The weighting triangles let the user modify the fragments arrangement. Thereby, the central weights determine according to which attributes the distance of fragments to the center should be calculated, the relative weights determine according to which attributes the related fragments should be arranged among themselves.

Browsing through fragments A related fragment floating around the user can be selected with a ray pointer attached to the user's controller, which makes it the new main fragment and thus rebuilds the visualization. The user's standing platform then represents the newly selected fragment and the surrounding spheres correspond to the respective fragments related to it. Hence, the database can be explored by browsing through the topics and concepts of interest.

Detail window information and translation For each fragment, a movable detail window reveals additional information like its *characteristics* (e.g., name, locations, time period, etc.), an image, and an abstract (see Fig. 1 right).

Relation details The relation details show pairwise distance information regarding the attributes (the gap in time, the shortest distance between their locations and their categorical differences) for the corresponding fragments and thus explain how the spheres' arrangement came about (see Fig. 1 right).

Central weights By adjusting the central weights the user can influence according to which attributes the related fragments should be positioned relative to the center (see Section 3.2.2, Equation 2 and Fig. 2).

[5] `https://www.sbert.net/docs/pretrained_models.html`

Relative weights The relative weights influence the dissimilarity measure used to position the spheres on a spherical shell around the user (see Section 3.2.2, Equation 1) to make interrelationships between related fragments visible (see Fig. 2).

Help menu The help menu is a lookup table that explains all components and the handling.

Compass Spheres can be marked to find them again more quickly. If a sphere is marked, its appearance changes from slightly transparent to opaque, and a mark in the color of its category turns up on the compass, which is a guidance line localized in the upper field of view (see Fig. 1 right and Fig. 2).

Transformation The user can grab and transform the whole constellation of the related fragments (which includes translating, uniform scaling and rotating). Thereby, the user may switch from the initial egocentric into an exocentric perspective.

History A history is localized in the lower area of the field of view (see Fig. 1 right and Fig. 2) and serves as a reminder, showing the order of the main fragments visited so far.

Traverse The user can go back and forth in main fragment selection (which is indicated by the history) at the push of a button.

Screenshots At the push of a button the user can take a screenshot of the current field of view, allowing to retrospectively look back at specific findings.

4 User Study

To learn more about the general suitability of the presented VR application for historical data exploration, to identify its strengths and weaknesses, and to figure out which features are helpful and which might even be disturbing in the exploration process, we conducted a user study that measured both quantitative and qualitative usage data. To make this study comparable across participants, we created seven exemplary tasks that all users had to complete by retrieving specific historical information using our system. These tasks were designed in such a way that they could partially be solved by "brute force" (e.g., by skimming through all historical fragments one by one) or by a more efficient combination of the interactive features provided by our system. By using this approach, we could observe if users were able to build an understanding of our system and apply this knowledge to solve practical tasks efficiently.

4.1 Apparatus

For the study, we used an *HTC Vive Pro* that has a resolution of 1440×1600 pixels per eye and an update rate of 90 Hz. The application is based on the Unity version 2020.3.12f1. The participants had an interaction space of about 2 m × 1 m, which was captured by two wall-mounted base stations. However, physical movements were voluntary since our application offers the opportunity to transform its virtual elements instead of moving oneself.

4.2 Procedure

Participants came to our lab, were informed about the purpose of the study, and agreed to participate voluntarily. After gathering some demographic data, they watched an introductory video of approximately 11 min that explained the purpose of the VR application and the interaction possibilities as described in Section 3. Afterwards, they were given about 5 min of time to familiarize themselves with the system and to test the functionalities. Then, the users had to complete seven tasks (**T1** to **T7**) within the VR application, that were always done in the same order. For each task we logged the number of interaction steps (or feature calls) and the task completion time and how often help was required. The initial main fragment of the study was *Erwin Schrödinger*[6] and the user started in an egocentric frame of reference. The central weights were completely set to semantic, meaning that the distance to the center was based on semantic dissimilarity, while the relative weights were completely set to time, meaning that spheres of fragments clustered around the user that had similar time periods. Furthermore, neither a detail window nor the help menu were opened, no spheres were marked and there was no history of selected main fragments. Based on this initial state (except for **T4**, which directly built upon the state reached in **T3**), the tasks were formulated as follows:

T1: Which places has *Albert Einstein*[7] been to? Solution: The user first needs to find the sphere that is labeled with *Albert Einstein*. To speed up this searching process, it makes sense to take a look into the help menu which offers the color assignments per category so that one gets to know which color is searched for. The user then might maximize the influence of the categorical attribute for the relative weights, which results in a clustering per category. Based on this, the user can now identify the cluster in the corresponding color to find the sphere representing *Albert Einstein*. Then, the detail window of that sphere has to be opened and scanned for the required location information.

T2: Find a picture of *Schrödinger's cat*[8] and make a screenshot together with a picture of *Erwin Schrödinger*. Solution: Similar to **T1**, the correct sphere has to be identified by sorting and then visually scanning the spheres. For this task, the two fragments *Erwin Schrödinger* and *Schrödinger's cat* have to be found and both corresponding detail windows have to be opened. They then have to be placed in a way that both can be seen in the user's field of view before taking a screenshot.

T3: From which category are there the fewest fragments for *Schrödinger's cat*? Solution: *Schrödinger's cat* is now supposed to become the new main fragment. After doing so, the influence of the categorical attribute in the relative weights has to be maximized such that the amount of fragments per cluster can be compared more easily. Then, the user can either look up the categories' colors in the help menu or find the category information in the detail window of a sphere that has the rarest color in the current visualization.

T4: How many fragments of the category *Cultural Artifacts*[9] are related to *Erwin Schrödinger*? Solution: Since this task builds upon the previous one, the system was not reset to the default state as for the other tasks. By doing so, we wanted to emphasize that one could either directly traverse back in main fragment selection by the press of a button or search once again for the sphere that is labeled with *Erwin Schrödinger*. Since this step requires a history of selected main fragments and we did not one task to be too complex, we did not design one big task in which one first creates a history of selected main fragments and then traverses backwards, but separated it into two smaller tasks (namely **T3** and **T4**). Beforehand, we informed the user that we intentionally omit the reset for **T4**. If the categorical attribute in the weights for the relative dissimilarity has been maximized, the next step is to look up in the help menu which color represents the searched category. After locating the correctly colored cluster, the user then has to count the number of spheres to answer the question.

T5: Which person out of the related fragments could *Erwin Schrödinger* never have met, since they did not live at the same time? Solution: To have the persons clustered, the user once

[6] https://dbpedia.org/page/Erwin_Schrödinger

[7] https://dbpedia.org/page/Albert_Einstein

[8] https://dbpedia.org/page/Schrödinger's_cat

[9] https://dbpedia.org/page/Cultural_artifact

again has to set the relative weights to the categorical attribute. To see at a first glance which persons' lifetimes do not overlap with the one of *Erwin Schrödinger*, the weights for the central dissimilarities can be set to the time attribute. This leads to a visualization in which spheres are closest to the center that do have an overlap in their time period with *Erwin Schrödinger*. To give numerical evidence which time periods do not overlap, the user can either open the detail windows to check the time periods for each fragment manually or make use of the relation details, which show up the differences in time in exact numbers. If the number is greater than zero, the periods did not overlap.

T6: Find a fragment that was/has taken place in one of the countries where *Erwin Schrödinger* has been once. Which country is it? Solution: Here, the user has to maximize the weight for the geo attribute in the weights for the central dissimilarities. After that, spheres are positioned close to the center for which the fragments share locations with short distances with the main fragment. Since the category does not matter in this context, any sphere which is closest can be chosen. A look into the detail windows of the chosen sphere and the current main fragment then permits the conclusion which country the fragments have in common.

T7: Find two persons out of the related fragments who were once in the same country and lived at the same time. Solution: The user has to set the relative weights right between the temporal and geo attribute. This leads to a visualization in which spheres are arranged close to each other where the corresponding fragments have likewise a small geo and temporal dissimilarity. Then, after looking up or remembering which color belongs to the searched category, two spheres of the right color need to be found that are close together. Finally, opening the relation details for those two spheres gives evidence about their temporal and spatial relationship. If both values are zero, then an answer is found.

All tasks were completed in this fixed order as we assumed a progressive increase in difficulty due to the increased number of required features participants had to use. The experimenter intervened when help was needed. After each task, participants filled in the Raw TLX Questionnaire [10, 11] and a discomfort scale [7, 20], which consists of the one question "*On a scale of* 0 *to* 10*,* 0 *being how you felt coming in,* 10 *is that you want to stop, where you are now?*", while staying in the VE. Once all the tasks have been completed, participants took the HMD off and filled in a User Experience Questionnaire (UEQ) [16]. They also rated all interaction methods on a scale of 0 – *very disturbing* to 10 – *very helpful*, which was followed by a semi-structured interview to learn more about the perceived strengths, weaknesses, and areas for improvement regarding the application. The whole procedure took approximately 75 min to complete and participants received an expense allowance of 15€.

4.3 Participants

29 participants (8 female, 1 non-binary, and 20 males) between 18 and 58 years of age ($M = 25.79$, $\sigma = 9.41$) participated in the user study. They were recruited on the local university campus and via dedicated mailing lists. Prior experience with HMDs was generally low, with 20 participants reporting to use them less frequently than once a year and only five to use them on a more than monthly basis.

4.4 Results and Discussion

Discomfort The mean discomfort scores reported after each of the tasks (with a possible range from 0 to 10) were all less than 1 with standard deviations of less than 1.5. Most individual scores ($N = 189$) were in the range between 0 and 2, with a clear majority ($N = 156$) reporting no discomfort at all. A few outliers ($N = 14$) were greater than 2, where the largest score of 6 was given by a participant after having difficulties completing the third task. However, their discomfort score returned to 0 for all the following tasks. In the interview three participants reported the fatigue that comes along with the usage of the application as one of its weaknesses and that it therefore could not be used for longer sessions. Nevertheless, these results indicate that our system did not systematically introduce symptoms of discomfort or sickness and that participants were therefore in good overall shape to rate our system in the questionnaires as well as the final interview.

User Experience The UEQ aggregates the individual responses into six overarching scores between -3 and $+3$ to quantify different facets of usability. The results for our system as well as comparisons with the benchmark dataset of the UEQ are shown in Fig. 3. In particular, the most positive scores were obtained for the categories *Attractiveness* ($M = 1.897$, $\sigma = 0.714$), *Stimulation* ($M = 1.862$, $\sigma = 0.943$), and *Novelty* ($M = 2.302$, $\sigma = 0.755$), where our system received results in the highest tier (*Excellent)* that marks the range of the best 10% of systems in the benchmark dataset. The category *Dependability* ($M = 1.664$, $\sigma = 0.849$) was ranked in the following tier (*Good*), indicating that 10% of systems in the benchmark dataset scored better while 75% scored worse. The final two categories *Perspicuity* ($M = 1.405$, $\sigma = 0.812$) and *Efficiency* ($M = 1.302$, $\sigma = 0.964$) received an *Above Average* rating, with 25% of systems in the benchmark dataset scoring better and 50% scoring worse. The high values for attractiveness, stimulation, and novelty show that there was an overall positive impression of the application, that it arouses excitement, and that it stands out due to its innovation. Those findings go hand in hand with the results of our interview. 14 participants described the application as a very intuitive way to quickly compare data and find connections between them, which promotes users understanding. Additionally, seven participants described the application to be interesting, exciting and that it arouses enthusiasm and curiosity through its innovation. The comparably lower values for perspicuity and efficiency are also reflected in the interview. 13 participants mentioned issues with the hardware and its handling since they needed some time to get used to it, which may justify the overall lower values for perspicuity. Two participants also noted the cost factor of the hardware as a weakness. While none of these characteristics are attributed to our software, they indicate general usage barriers of HMDs. We hope that future HMDs will alleviate these issues and therefore lead to more positive ratings of perceived perspicuity and efficiency.

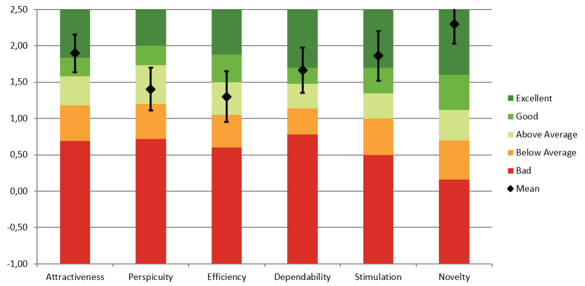

Figure 3: Results of the UEQ in context with the benchmark dataset.

Ease of Learning and Use All participants were able to pick up and use our system after watching the provided 11-minute tutorial video. As participants were completing the tasks of the study, we observed that the number of help requests were still relatively high for **T1** and **T3** while this figure decreased considerably for the last four tasks (see Table 1). This is particularly interesting since the later tasks required participants to discover more complex relationships

Task	T1	T2	T3	T4	T5	T6	T7
Number of Help Requests	15	5	10	1	3	1	2

Table 1: Total accumulated number of help requests per task across all study participants.

by using more features of our system at the same time. These results therefore indicate that participants became more proficient with our system during the course of the study and were then able to complete more challenging tasks without asking for further assistance.

The mean task completion time is maximal for **T3** with 157.8 sec (see Fig. 4). **T4**, on the other hand, had the smallest mean completion time with 89.2 sec and mean number of interaction steps with 9.2. That might be because **T4** involves a similar sequence of steps as **T3** but in reverse order. The mean number of steps is maximal for **T5** with 27.3. One can solve **T5** by going through the fragments in a brute force manner, which would explain the high number of steps. Additionally, when adjusting the weights for the dissimilarities participants tended to test multiple weight combinations within one task and observed its influence on the arrangement. Even though **T6** and **T7** could also be solved in a brute force way, the mean number of steps decreased to 14.1 and 14.4 respectively. While participants in the first half of tasks tended to need fewer interaction steps while needing more time to complete a task, the number of interaction steps increased in the second half while the completion time became less. That can be explained with a learning process and that the participants got used to the system and its handling, which results in an overall shorter task completion time.

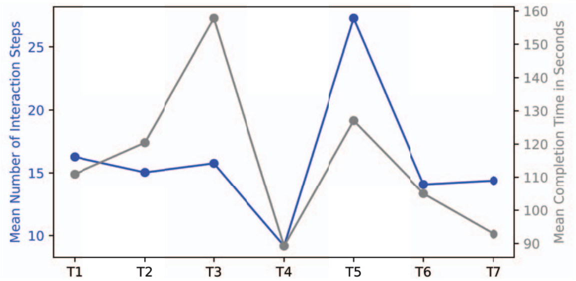

Figure 4: Mean number of interaction steps per task and mean time to complete a task.

Task Load

The distributions of task load scores as measured by the Raw TLX questionnaire are shown in Fig. 5. The mean task loads ranged from $M = 15.9$ (**T7**) to $M = 29.9$ (**T3**) with standard deviations between $\sigma = 13.4$ (**T2**) and $\sigma = 19.1$ (**T3**). As the original publication of the questionnaire does not include absolute benchmark values to judge the severity of scores without a direct comparison condition [11], we used the values provided in the meta analyses of Grier [8] and Hertzum [12] to interpret our results. In particular, we observed that our mean values are considerably smaller than the averages reported in the general datasets of Grier ($M = 45.29$, $\sigma = 14.99$ for the Raw TLX) and Hertzum ($M = 42$, $\sigma = 13$) as well as the smaller but more specific VR dataset of Hertzum ($M = 41$, $\sigma = 15$). The 85$^{\text{th}}$ percentiles of measured task load scores were between 25.7 (**T7**) and 49.9 (**T3**) and therefore still within a comparable range to the mean results in the meta analyses. We therefore conclude that our system does not appear to impose unreasonably high task loads on users and is therefore suitable for exploring historical relationships comfortably. Nevertheless, while our original intention was to design the tasks to be increasingly more difficult as the study progressed, we could not confirm this trend by looking at the task load scores. Contrary to our expectations, we observed that the task loads imposed by **T3** were especially high while the final task **T7** seems to have been completed with relative ease. Considering the previously mentioned findings regarding the decreasing number of help requests and mean task completion time, these findings also support the assumption of a learning process and that the participants got more familiar with the system over time.

We also observed positive correlations between task completion time and the task load, $r = 0.68$, and between task completion time and number of steps, $r = 0.529$, which can be considered large, based on the threshold values of Cohen [4]. Even though it do not imply a causality, it seems natural that a longer task completion time, which might be a consequence of a longer thinking process or trial by error, might lead to a higher perceived task load. The same is true for an increased number of steps performed to complete the task.

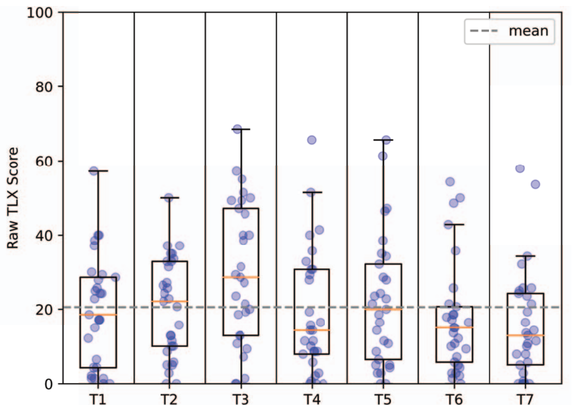

Figure 5: Boxplots illustrating the Raw TLX values per task.

Feature Ratings

The ratings of the features in the system that should support users to solve the tasks are shown in Fig. 6. While the overall mean value is a score of 7.8, the individual mean values per feature range from $M = 5.4$ (compass) to $M = 8.9$ (relation details) with standard deviations between $\sigma = 1.4$ (relation details) and $\sigma = 2.5$ (transformations). Thus, all mean scores are above the neutral score of 5, which means that each feature was on average considered more helpful than disturbing. Most medians range between 7 (traversing, history and screenshots) and 10 (relation details and help menu); only the compass is an outlier with a median of 5. Those findings prove an overall agreement that the relation details are an especially helpful feature. This is also in line with the results of the interview, where it was described as the most interesting and valuable feature, which was asked to be enriched with even more information. The high score of the help menu also makes sense in this respect as it allows to look up feature descriptions without having to leave the VE and the handling, which was described to be difficult at the beginning by some users in the interview. Traversing can also be achieved by searching and switching the fragments again, and the history is more a reminder than an essential feature, and screenshots “only” record what has been experienced. Summarized, these features are rather shortcuts and memory aids, but not essential in the exploration process, which could be the reason for the comparably lower ratings. As stated in the interview, the compass was perceived to be too far up in the field of view, making it difficult to see, and might be more helpful if gets positioned further down. As a consequence, it was barely used and therefore likely resulted in lower scores. While there are clearly more responses on the positive than on the negative rating side, the worst individual score was a 0 for the transformation feature. This was because one participant found the possibility to rotate the constellation around the x-axis (roll) as annoying, although they found the other degrees of freedom of the transformation helpful and made regular use of them during the study. In general, the transformation had the highest standard deviation in the response data, which is likely due to individual preferences and technical affinity. Some participants preferred to move themselves in the real world and did not make use of the transformation feature. Others tended to transform the virtual constellation while remaining in their original position in the real world.

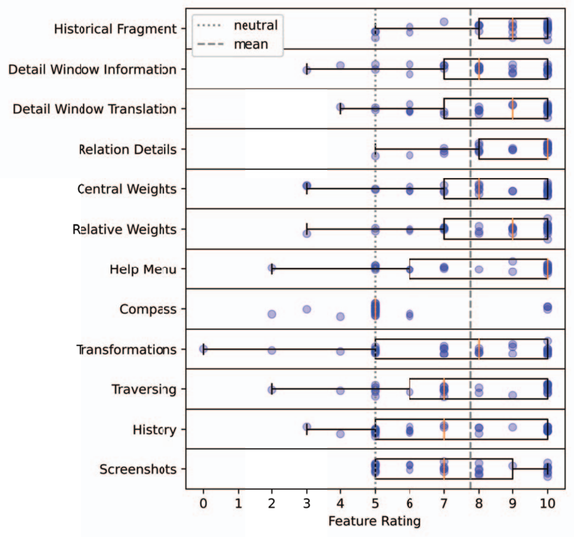

Figure 6: Boxplots illustrating the results of the feature rating, where 0 means that a feature was considered *very disturbing*, 5 that it was *neither helpful nor disturbing*, and 10 that it was *very helpful*.

General Feedback The semi-structured interview gave us further insights on the general applicability, possible cases, the easiness of interpretation of the data representation, and the perceived strengths and weaknesses of interacting with our system.

Regarding possible use cases 17 participants found the software applicable to education purposes and seven were of the opinion that it can also be used in museums. Concerning the suitability of 3D and VR, 25 participants stated that the 3D virtual space is well suited to explore historical data and 13 pointed out that the additional dimension offers a larger space such that more information can be captured at a glance. In addition, 21 considered the abstract visualization in this VR application as suitable for exploring historical data and nine stated that it is more illustrative and comprehensible than e.g., continuous texts in books or web pages. On top of that, ten participants claimed that interpreting the data representation was easy and 13 took one step further and even claimed it to be very easy. Those findings confirm that our data presentation is well understandable and reflect the results found in the UEQ evaluation. We also asked whether some of the weighting triangle attributes (see Fig. 2) should be changed, but 28 clearly said that the attributes are good as they are and that they would not substitute any of them. Nevertheless, three mentioned that the semantic attribute would also be a nice addition to the existing ones for the relative dissimilarity. On the question what additional features would be desirable, seven participants stated that they would like to get more information in the relation details, e.g., information on the amount of overlap of the time periods or on the thematic linkage/relation between two fragments. Three had the idea of integrating a search function to find specific fragments more quickly which might be due to the design of the tasks in which the participants had to find specific fragments at some points. Three more participants mentioned filter functions as a helpful feature to develop to define more specifically which related fragments are of interest and thus should be displayed while others should be faded out. Furthermore, three participants asked for more “visual pleasing content” like different 3D objects and more pictures. Regarding the strengths and weaknesses, five mentioned the interactivity as a good way to experiment. Three especially mentioned that it is good to gain an overview of the data. While two stated that the visualization presents too much data at once, two said it should present even more data at one glance. This discrepancy also supports the suggestion of introducing filter functions to leave it up to the user how much data should be displayed in order to address individual needs.

5 Conclusion and Future Work

We introduced a novel interactive exploration tool for historical data in VR, in which historical fragments are arranged based on their *temporal*, *geo*, and *semantic* or *categorical* proximity. Together with a set of corresponding interaction techniques, our system allows the active user-driven exploration of historical data sets to discover meaningful interrelationships between fragments. The results of our user study show that the presented application arouses excitement and stands out due to its innovation. Additionally, contrary to our expectations, our quantitative results show that more difficult tasks, in which more complex relationships had to be discovered, led to a lower task load than previous tasks that were actually designed to be easier. Thus, those findings suggest a fast learning process, which allowed users to discover complex relationships easily after getting used to the application. Overall, the provided features were perceived to be helpful. Solely the compass was out the line since it was hard to see and thus barely used. Although the handling of the HMD was often classified to be complex initially, a lot of participants stated that the 3D virtual space as well as the abstract visualization were well suited to explore historical data and that the third dimension offers a larger space so that more information can be captured at a glance. Furthermore, the application was considered to promote understanding through comparing and finding connections between data points. Only a few participants mentioned disadvantages like the high costs and fatigue that go hand in hand when using an HMD. Although the previously mentioned advantages seem to outweigh the drawbacks, it makes sense to investigate how the duration of use affects the effectiveness of the system. We hope that future technical developments will reduce the mentioned entry barriers by providing more accessible and comfortable HMDs.

Overall, our interviews provided several valuable recommendations for future work. In particular, more advanced filter functions to control the data presentation as well as more advanced information in the relation details appear promising to gain even deeper insights into the relation between fragments. Once these improvements are incorporated, future studies should also involve a more diverse sample in terms of age and gender. Furthermore, it makes sense to conduct comparison studies evaluating how the proposed VR exploration method and especially the spherical arrangement perform against other approaches, including conventional 2D tools or alternative 3D data analysis techniques, and finally, to explore the advantages of our system over conventional historical teaching tools as well as the prevalent interactive 2D tools. Overall, we believe that novel technologies like VR have the potential to systematically transform historical research and teaching from the passive consumption of texts to interactive exploration experiences that help people to better understand complex interrelationships.

Acknowledgments

This research was supported by the German Federal Ministry of Education and Research (BMBF) through the eTaRDiS project (ID: 01UG2122B). We are grateful to all subjects attending our user study and thank Fabian Herrmann for his efforts.

References

[1] D. Allcoat and A. von Mühlenen. Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology*, 26, 2018. doi: 10.25304/rlt.v26.2140

[2] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky. VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):210–219, 2016. doi: 10.1109/TVCG.2015.2467971

[3] E. Clark, A. Celikyilmaz, and N. A. Smith. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2748–2760, 2019. doi: 10.18653/v1/P19-1264

[4] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd ed.* L. Erlbaum Associates, 1988.

[5] M. Derksen, T. Weissker, T. Kuhlen, and M. Botsch. Towards Discovering Meaningful Historical Relationships in Virtual Reality. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 697–698, 2023. doi: 10.1109/VRW58643.2023.00191

[6] R. Etemadpour, E. Monson, and L. Linsen. The Effect of Stereoscopic Immersive Environments on Projection-Based Multi-dimensional Data Visualization. In *International Conference on Information Visualisation*, pp. 389–397, 2013. doi: 10.1109/IV.2013.51

[7] A. S. Fernandes and S. K. Feiner. Combating VR Sickness Through Subtle Dynamic Field-of-View Modification. In *IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 201–210, 2016. doi: 10.1109/3DUI.2016.7460053

[8] R. A. Grier. How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, pp. 1727–1731, 2015.

[9] D. Han, G. Parsad, H. Kim, J. Shim, O. Kwon, K. A. Son, J. Lee, I. Cho, and S. Ko. HisVA: A Visual Analytics System for Studying History. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4344–4359, 2022. doi: 10.1109/TVCG.2021.3086414

[10] S. G. Hart. NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomic Society annual meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909

[11] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988. doi: 10.1016/S0166-4115(08)62386-9

[12] M. Hertzum. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics*, 64(7):869–878, 2021. doi: 10.1080/00140139.2021.1876927

[13] M. Kraus, N. Weiler, D. Oelke, J. Kehrer, D. A. Keim, and J. Fuchs. The Impact of Immersion on Cluster Identification Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):525–535, 2020. doi: 10.1109/TVCG.2019.2934395

[14] M. Krishnan, J. Ober, M. Pyzyk, and N. Pedrini. POLIS: Designing a Visualization Tool for the Research of Complex Sociopolitical Landscapes. *Parsons Journal for Information Mapping*, 6(2):1–9, 2014.

[15] S. Latif, S. Agarwal, S. Gottschalk, C. Chrosch, F. Feit, J. Jahn, T. Braun, Y. C. Tchenko, E. Demidova, and F. Beck. Visually Connecting Historical Figures Through Event Knowledge Graphs. In *2021 IEEE Visualization Conference (VIS)*, pp. 156–160, 2021. doi: 10.1109/VIS49827.2021.9623313

[16] B. Laugwitz, T. Held, and M. Schrepp. Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger, ed., *HCI and Usability for Education and Work*, pp. 63–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-89350-9_6

[17] L. Lisle, K. Davidson, E. J. Gitre, C. North, and D. A. Bowman. Sensemaking Strategies with Immersive Space to Think. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 529–537, 2021. doi: 10.1109/VR50410.2021.00077

[18] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2018. doi: 10.48550/ARXIV.1802.03426

[19] J. P. McIntire and K. K. Liggett. The (Possible) Utility of Stereoscopic 3D Displays for Information Visualization: The Good, the Bad, and the Ugly. In *IEEE VIS International Workshop on 3DVis (3DVis)*, pp. 1–9, 2014. doi: 10.1109/3DVis.2014.7160093

[20] L. Rebenitsch and C. Owen. Individual Variation in Susceptibility to Cybersickness. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pp. 309–317, 2014. doi: 10.1145/2642918.2647394

[21] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410

[22] J. Wagner, W. Stuerzlinger, and L. Nedel. The Effect of Exploration Mode and Frame of Reference in Immersive Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(9):3252–3264, 2022. doi: 10.1109/TVCG.2021.3060666

[23] J. A. Wagner Filho, M. F. Rey, C. M. D. S. Freitas, and L. Nedel. Immersive Visualization of Abstract Information: An Evaluation on Dimensionally-Reduced Data Scatterplots. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 483–490, 2018. doi: 10.1109/VR.2018.8447558