Supporting Exploration of Women's Print History Project Data via Interactively Constructing Networks of Interest

Parnian Taghipour
parnian_taghipour@sfu.ca
Simon Fraser University
BC, Canada

Maryam Rezaie
Simon Fraser University
BC, Canada
maryam_rezaie@sfu.ca

Michelle Levy
Simon Fraser University
BC, Canada
mnl@sfu.ca

Thomas Shermer
Simon Fraser University
BC, Canada
shermer@sfu.ca

Sheelagh Carpendale
Simon Fraser University
BC, Canada
sheelagh@sfu.ca

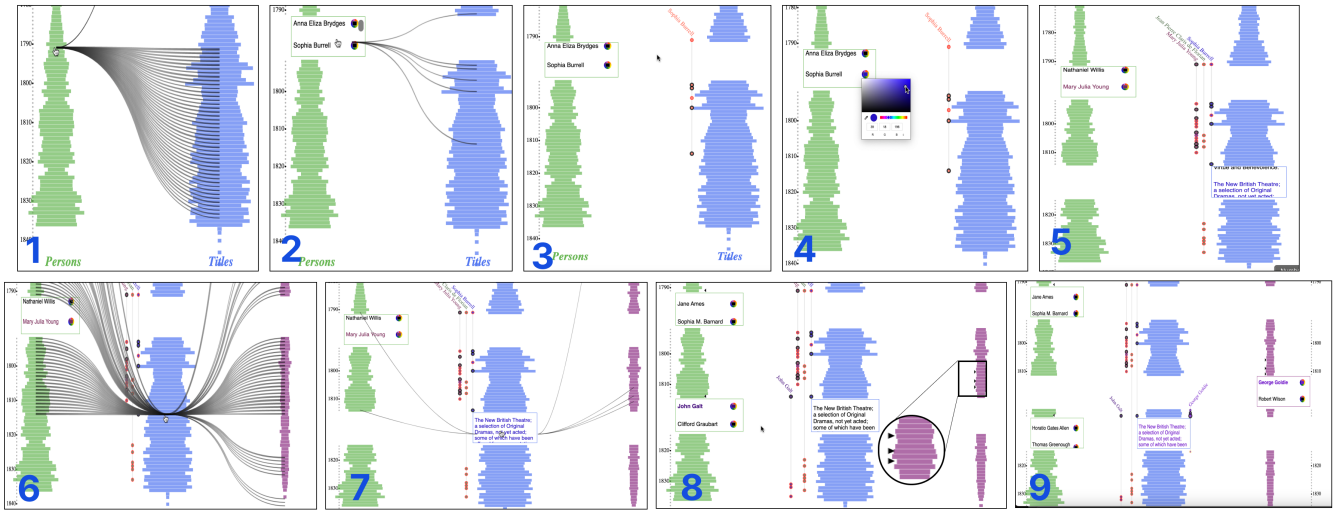

Figure 1: Exploring network complexities with WPHPVis: 1) Hovering over a timeline rectangle displays connections for all contributors for that year; 2) Clicking on a rectangle opens a scrollable list containing persons' names, and hovering over a name reveals their contributions; 3) Clicking on a name displays the timeline of that person's contributions; 4) Differentiating contributors' timelines is achieved using a color palette; 5) Multiple timelines can be opened, each assigned a distinct color; 6) Relationships between titles in one year and their contributors can be explored; 7) Opening titles in a specific year results in a scrollable list that auto-scrolls to on-screen titles, assigning them the same color as their respective timelines; 8) Small triangles aid in finding contributors' years; and 9) The font for individuals and firms related to the clicked title is bold.

ABSTRACT

We designed, developed, and studied a visualization, WPHPVis, to support exploration of the Women's Print History Project (WPHP) data. WPHP are manually-collecting a bibliography that spans the years 1700 to 1836 recording information about books in which women have been involved through a number of roles including as authors, editors, translators, publishers, printers and booksellers. By working directly with WPHP experts to focus on their understanding, and their research practices and needs, we co-designed WPHPVis using interactive construction of network links to support exploration of their data. Through our qualitative study with both experts and non-experts, we learned about how the tool supported the WPHP experts' research practices as well as about how to improve overall interactive experience. We conclude by discussing the importance of representing missing data, the advantages of striking a balance between visualization structure and explorability, and the opportunities enabled by co-design with domain experts.

AVI 2024, June 03–07, 2024, Arenzano, Genoa, Italy

ACM ISBN 979-8-4007-1764-2/24/06
https://doi.org/10.1145/3656650.3656697

CCS CONCEPTS

• **Human-centered computing** → *Information visualization*; • **Applied computing** → *Arts and humanities.*

KEYWORDS

Digital Humanities, visualization

ACM Reference Format:

Parnian Taghipour, Maryam Rezaie, Michelle Levy, Thomas Shermer, and Sheelagh Carpendale. 2024. Supporting Exploration of Women's Print History Project Data via Interactively Constructing Networks of Interest. In *International Conference on Advanced Visual Interfaces 2024 (AVI 2024), June 03–07, 2024, Arenzano, Genoa, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3656650.3656697

1 INTRODUCTION

In today's data-driven world, the collection and analysis of historical data has become increasingly important in gaining valuable insights and in understanding various aspects of our past. Historical dataoffers a wealth of knowledge about our cultural heritage, societal developments, and the lives of individuals who have shaped our history. However, extracting meaningful information solely from looking through historical records can be challenging. Therefore, visualization methods can be helpful to bring hidden patterns to the surface. In this paper, we work towards this by exploring some of the challenges associated with visualizing historical data, focusing specifically on the Woman's Print History Project (WPHP) dataset [27]. The WPHP dataset is a growing hand-assembled dataset that aims to provide comprehensive information on women in publications between 1700 and 1836. It is already a valuable resource for understanding the contributions of women to the print industry during this period. Women contributed to the titles in the WPHP data through different roles, for example, a woman might have contributed as an author, illustrator, editor, publisher, etc. The dataset has three main sections: people [26], titles [41], and firms [13] each of which is being manually collected. In this collaboration between English and Visualization researchers, we are working towards making an explorable tool for WPHP data, to bring hidden information and patterns to the surface.

Researchers from WPHP are gathering all the possible information and are exploring its reliability and trustworthiness, verifying accuracy whenever possible using different sources. However, a difficult aspect of the data collection process for researchers is that the information is often not available in digital format, forcing manual data collection. Because it was historically collected by different people it may not be totally coherent, and there are inevitable gaps in the dataset, some of which reflect gaps in the historical knowledge. Consistent formatting might be difficult because the original information is so varied. Besides, this dataset includes information from marginalized groups that have been overlooked in the past. Thus much useful information may have been overlooked or not considered important enough to record. Therefore, the completeness of the data may vary significantly depending on the sources from which it was collected. Also, we need to consider that authors and contributors at that time may not always share the same amount of information about themselves. For example, some authors may have used pseudonyms or female authors may have published under male names, their husband's name, or even simply as a *'by a lady'*. Additionally, there may be missing information about dates, with only a possible range or even no year known. In short, since this data is modeling complicated human/object relationships over time and space, there are interesting data complexities, and there are inevitable gaps in the dataset. Although these shades of missing data make it hard to work with this data and visualize it, they are a rich part of data and the manner in which data is missing helps to shed light upon that era. By visualizing the data, the researchers aim to identify and analyze these variations and incompleteness to gain a more comprehensive understanding of the historical context, and the challenges women in print faced during the 18th and 19th centuries. Furthermore, the variations and incompleteness in the dataset can provide valuable insights into the social, cultural, and economic factors that influenced women's contributions to print during this time. Also, knowing about these resources and the relationships between data and these resources can help data experts to collect better-quality data. The combination of these challenges and finding a place for visualization in dealing with these challenges was the motivation behind this project.

The challenges associated with visualizing historical data are at the confluence of several active research directions in the field of visualization. This project involves working with domain experts, leveraging personally collected data, dealing with various types of missing data values, and handling a growing dataset with over 20,000 entries. Although 20,000 entries may be considered small in the context of big data [29], it still presents challenges in terms of screen space usage and interactions. Consequently, designing an effective visualization requires careful consideration of these limitations in working towards providing meaningful insights.

Our research is a collaboration between data experts and visualization experts. We first worked toward understanding the domain needs and tasks. Next we focused on design and implementation, which we then studied to learn more about how it answers the needs of data experts and its usefulness in general. We explain our design goals and justified our design choices based on the design goals. Third, we studied our design with two different groups, data experts and non-experts, to get feedback about the effectiveness of this tool in the expert's research as well as the general user experience. Finally, we reflected on lessons we learned regarding visualizing missing data, balance between structure and explorability in visualization, and co-design sessions with data experts.

2 RELATED WORK

As information visualization applications extend across various fields and prove their usefulness, the intersection between humanities and visualization has witnessed increased and varied research [3]. For example, Hinrichs et al. [20] visualized a specific literary collection. In their research process, they framed the design space of their literary data within four aspects: audience, use case (result or process), approach (qualitative and quantitative) and visual qualities. Although many tools and research showcase results in humanities, particularly for general users, the utilization of visualization in the research process is less common [16]. Researchers have explored why collaboration between humanities and visualization is not more effective [24], and have used human-centered design processes to engage humanities scholars and better understand their needs [37]. Moreover, Lamqaddam et al.'s paper in 2021 [25] explored humanities researchers' views on visualization and discussed how to enhance collaboration. They emphasized the

importance of reducing the semantic distance between visual representation and the data. To achieve this, they suggested considering data transparency and using terminology familiar to experts. We recognize that showing data experts their actual data can assist them in using a visualization tool. Some studies argue that presenting data in its raw form, without cleaning, provides the advantage of assessing data quality and gaining insights into missing and incomplete data [30, 44]. Keeping missing data and errors in the visualization can significantly impact the experience of domain experts and their sense-making process [33].

Another relevant topic in the digital humanities is the improvement of investigative analysis and exploration [20]. Various concepts have evolved over time in terms of explorability. For example, in Bohemian Bookshelf, interlinked visualizations are used to give various perspectives on a collection, aiming to encourage serendipitous discoveries by providing multiple access points, offering flexible exploration pathways, using abstract representations, and enabling a playful approach to information exploration. One of their visualizations is a time visualization that has two timelines that are simply showing the relation between two aspects of data [36]. There are also works that show publications on timelines [5] in different ways [23]. Dörk et al. [12] proposed four designs for their humanities collection, and noted that incorporating time as an aspect of visualizations (having both network view and timeline view separately in one tool) seemed to make it easier to filter and explore the data. Whitelaw [42] suggested the need for more generous interfaces for humanities research. However, Windhager [43], in 2017, reviewed 70 papers related to digital humanities and observed that exploration concepts are often designed for casual users. In this project, we focus on exploration possibilities for the domain experts, while still considering casual usability.

Furthermore, most tools designed to assist humanities researchers focus on research tasks that deal with large documents [1, 10, 39]. In contrast, we are working with a network dataset. Some tools show networks within the text, such as Bingenheimer et al. [4], who extracted a network of historical Chinese Buddhist monks from texts and created a visualization of their social network. Our dataset is multi-model data and visualizing a multi-model graph is recognized as challenging due to the complexity of the data and the limitations of the screen [11]. One example that uses an aggregating method for networks is MMGraph [15]. In terms of publication data, there are papers that show the relation between scientific papers and collaboration; such as eDBLP [6] that uses a network tool, and Chen et. al [8] that uses map, network and radar visualization. Other tools are more focused on visualizing networks, such as Gephi [2], Palladio [22], and Pajek [31]. However, networks are more commonly uni-dimensional, while ours is multi-dimensional (meaning vertices in the network are from different categories). Multi-dimensional network visualization can be complex and clutter tends to increase as the size of the dataset grows.

3 UNDERSTANDING THE DOMAIN

First, we delved into the data experts' research process and the nature of their data. To understand this information, we frequently met with data experts, discussing and exploring their WHWP dataset.

3.1 Data Set

Our focus in this project centers on the Women's Print History Project (WPHP) dataset. The WPHP is an ongoing collection of titles printed between 1700 and 1836— a pivotal period in both women's history and print history. It consists of data gathered from over 100 sources, resulting in a dataset that comprises more than 25,000 titles, their associations with persons and firms, and other relevant metadata. Within the WPHP dataset, we concentrated on three critical tables: Titles (25,073 entities), Persons (12,178 entities), Firms (6,934 entities). Every Title has a minimum of 2 connections (to a Person and a Firm) and usually has many more connections including all the people in different roles. This forms a dense network graph, where drawing all edges would entirely fill the display. According to WPHP's methodology, the core dataset revolves around Titles, while also capturing contributions from Firms and Persons. For people contributors the diverse roles include authors, publishers, booksellers, printers, editors, translators, engravers, introductions, illustrators, compilers, and composers. Firm contributorsare identified as publishers, printers, and booksellers. Titles are linked to both people and firms separately, resulting in two types of edge: title-people and title-firms. All connections between People and Firms occur through the intermediary of Titles.

Data Challenges The dataset is challenging in terms of being inconsistent, incomplete, and including a dense network structure:
C1) Challenge: Inconsistent and Missing Data: Within the WPHP data there is a significant amount of missing and/or partially incomplete data. Our discussions with data experts revealed both interest and challenges associated with these gaps. The incomplete data can be interesting because it is indicative of the historical social setting of the time. For example, an author maybe simply entered as "written by a woman" or as "Mrs. Henry Smith". Also, consideration around missing data can be exemplified by the following quote: *"Yeah, empty is better than inaccurate, especially because the database. People look at databases and they're [if it is in the data, they think] like that is fact, it's like somebody has said that this is true."*. The WPHP commitment to preserve historical context and the intriguing aspects of data inconsistencies precludes traditional data wrangling and cleaning practices. This challenge requires addressing programming and data representational challenges that arise from data discrepancies.
C2) Challenge: Hairball problem: While not large in terms of big data, this dataset is large in terms of hand-collected data.Its size in entities and edges definitely presents hairball/spaghetti challenges in terms of screen space and interactions.

3.2 Understanding the Domain and Tasks

In our initial engagement with domain experts, we identified two main research streams within the WPHP project:
Gathering and Verifying Data: The primary focus of WPHP is to collect and compile data from various sources to showcase the significant contributions of women to the publishing industry. Three teams are responsible for gathering information related to people, firms, and titles. There is a verification process for each record in which, if possible, another person cross-checks the information using different resources to ensure accuracy. If all the information is successfully verified using two sources, the record is marked as

verified. If complete information cannot be found, the record is marked as attempted verified. Ensuring data quality is a priority for the WPHP team. The dataset initially focused on records from 1750 and onwards, so data from those years is more complete. In cases where information on a title's first page is insufficient, researchers use other reprinted versions or additional contributions by the same person to complete the details. To ensure accuracy, connections between different variables are used to verify the information.

Writing Spotlights and Making Podcasts: In addition to data gathering and verification, the WPHP project actively communicates its findings through podcasts and spotlights. These resources provide additional context and insights beyond the dataset itself, shedding light on specific historical patterns and contributors.

4 DESIGN GOALS AND PROCESS

Our project aims to address the challenges of visualizing historical data while supporting the specific research needs of the WPHP team. Our design goals center on creating an explorable visualization tool that complements the objectives of data experts.

- **DG1** To use an interactive constructive visualization approach to support discovering content in this complex network,
- **DG2** To work towards creating an exploration approach that supports delving into aspects of fascination,
- **DG3** To leverage timelines to support coherence across people, publications, and firms,
- **DG4** To provide access to missing, partial, and anomalous data,

Designing effective tools for data experts requires a human-centered approach [19]. Our design methodology is rooted in User-Centered Design (UCD) [17, 28], which emphasizes the active involvement of domain experts to better understand user and task requirements. To address the unique needs of data experts, we adopted an iterative design approach and conducted qualitative studies to refine our tool based on their feedback. Our design process began with brainstorming, sketching, and an in-depth review of existing work in the domain. We leveraged the 10+10 design sketching approach [18] to create initial sketches of our tool. Furthermore, we asked domain experts to participate in the sketching session, in order to understand more about their ideas and reach the mutual language. These early sketches laid the foundation for our subsequent discussions with data experts.

4.1 Early Design Choices

During the initial sketching, we identified time as one key aspects of the dataset and decided to focus on the temporal data, which offered rich and diverse information, that is closely linked to historical context, a primary concern for data experts (DG3). The prototype includes three distinct timelines corresponding to the main sections in the WPHP dataset: persons, titles, and firms (DG3). While the decision to keep the timelines central to the visualization remained constant, the fact that the actual data only contained years as dates – causing many entries to fall on the same date – required making multiple entries evident.

We opted for a vertical timeline because vertical scrolling is considered more favorable [40]. Data entries that did not have a date, and thus could not be placed on the timeline were preserved on top of the timelines – respecting their importance (DG4).

Showing the existing dense network in the dataset requires tackling the hairball problem (see Data Challenges) While many approaches to the dense edge problem have been tried such as: degree-of-interest layouts, which support filtering based on a combination of declared interest in the data and graph topology [14]; fisheye approaches that magnify selected spatial regions [7]; use of edge bundling [21], or link curvature [34]; and many different types of filtering approaches [32], dense graphs remain an issue [32]. We took an alternate approach, leveraging ideas from constructive visualization to provide interactions so that network relationships can be revealed as needed (DG1).

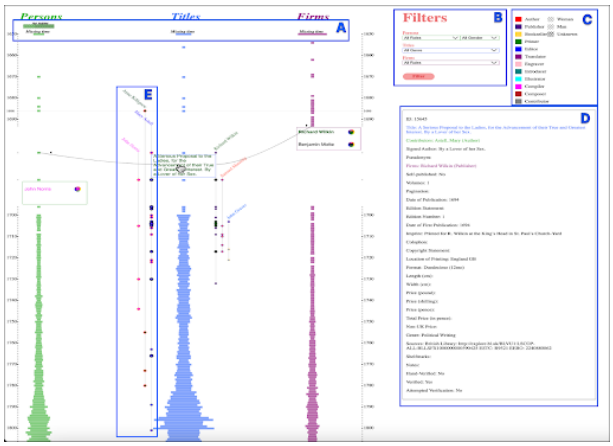

Figure 2: An overview of WPHPVis. A) The subset of the data that has missing dates are placed on top of the screen to make interaction still possible, B) Filters enable exploring points of interest, C) Color Legend, D) Information box, E) Contributor timelines appear in parallel, with size of circles indicating the contributions in that year.

5 WPHP VISUALIZATION

The WPHP Visualization uses structure to organize the entities and interactive construction to reveal edges of choice. Its main structure is a triple timeline.Hover-and-click interactions support revealing networks of interest. Each of the three timelines is dedicated to one of the primary entities in the dataset – from left to right – Persons, Titles, and Firms (Fig 2).

The Triple Timeline: The triple timeline (People, Titles, Firms), our primary focus, is displayed on the left side of the screen. Data is positioned on these timelines according to specific criteria: Persons are placed by their first contribution year, Titles by their publication date, and Firms by their establishment year. When the date is missing for an entry, it is still included but placed at the top of the timeline. Titles are placed for each reprint to accommodate the interests of data experts (DG2). The rectangle that represents each year is sized according to the number of items in that year, offering an overview of historical occurrences. This same abstraction is used to show the missing data (DG4). Different types of missing data, such as titles with missing time information or unknown contributors, were categorized and represented using distinct visual patterns

on top of the Persons timeline. This respects the significance of these data gaps, and ensures that missing data is not lost.

Control and Information Panel: Clicking on a person, title, or firm displays full data details in the information box beside the triple timeline. This feature supports details on demand and is included for data verification and to invite further exploration (DG2).

Constructive Network Interactions: By hovering over a date rectangle on any of the persons, titles, or firms timelines, connections between the adjacent timelines will be displayed. The relationships between timeline entities are revealed temporarily by hover, while a click will make their indications persistent. As in the WPHP dataset, Person edges connect to related Titles, Title edges connect to both Persons and Firms, and Firm edges connect to Titles. These edges are drawn with curved lines to visibly point to the related years.

To get the desired information on the screen step by step (DG1), users can hover on any of the timeline rectangles. This will reveal the relations between that timeline rectangle and the associated timeline by showing the lines connecting to the related dates on the adjacent timeline (Fig 1(1)). To get more details, and to proceed in the process of constructively building up a network of interest, users can also chose to reveal a scrollable list of that year's items within the selected rectangle. Opening up a given date on either person or firm timelines shows the scrollable list of the names of individuals and firms (Fig 1(2)), along with a color picker tool (Fig 1(4)). With this scrollable list open, one can look for more specific connections. Hovering over a name in the list displays its specific connections to years in which that person or firm contributed to titles. Clicking on a name provides further details, including a contributions timeline (Fig 1(3)) and fills the information box on the right (DG2).

The contribution timeline is a vertical timeline with circles for relevant years. If a contribution is associated with more than one publication in a given year, a black border around the contribution's circle and the size of the circle indicate this. Clicking on circles with a black border reveals a spiral of related titles. The border of the circles indicating one contribution show the role of firm/person. For the titles timeline, hovering over a title reveals connections to different years on the person/firm timelines (Fig 1(6)). Simultaneously, the publication name in the title timeline is highlighted using the same color as the corresponding contribution timeline (Fig 1(7)). Clicking on a title shows an information box with additional details and keeps the timeline's year indicator (Fig 1(8)) visible.

Updating the Timeline's Time Scale when Opening and Closing a Year. To open a rectangle on any of the three timelines, the rectangles below it move down, ensuring that all data remains on the screen. Additionally, the entire timeline scale adjusts, maintaining consistency. Closing a timeline restores the previous state, providing a coherent times.

6 LEARNING ABOUT WHPHVIS IN USE

To understand more about WHPHVis in use, we conducted a qualitative study. Since a visualization that uses interactive construction to support one's ability to slowly reveal network specifics set within their historical timelines would be novel to both domain experts and non-experts, we designed a qualitative study to allow us to observe the process of discovery. Using a qualitative approach allowed us to observe the details about the reality of discovery and exploration within constructive network revealed as an interactive approach to a exploratory network visualization. The primary objective was to gain insights about the nuances of user discovery of tool's usability, collect user feedback on the design and interaction features, and understand how it facilitates data exploration for both domain experts and non-experts.

Participants: included domain experts (P1, P2, P4, P9, P11, P12 (5 female)) and non-domain experts (P3, P5, P6, P7, P8, P10 (4 female)). The domain experts included 4 graduate students and 2 undergraduate students from our university's Department of English. They exhibited a wide range of experience with the dataset, spanning from 6 months to 5 years. The non-domain experts were invited through email and word of mouth from diverse backgrounds, including visualization, archaeology, cognitive science, psychology, English, and environmental science. The educational backgrounds of this group varied and included 1 post-doctoral researcher, 1 PhD student, and 4 undergraduate students.

Experimental Setting: the study took place in a Lab environment using MacBook Pro laptops with 16-inch display. The WHPHVis was accessed through a Chrome browser, emulating a realistic research environment. A consent form was signed, and participants filled out pre-experiment questionnaires. Participants were introduced to the experimental procedure and WPHPVis. The main study sessions, which included the completion of tasks, personal interactive exploration, semi-structured interviews, and sketching activities, were recorded in both audio and video formats.

6.1 Procedure

Activity 1: Tasks First activity comprised seven carefully designed tasks (derived from domain expert research questions) that were intended to inform participants about different interactions and aspects of the tool. These tasks were designed to help the participants discover functionality and let us see how well they could understand and utilize the tool's visual variables and interactions. Some of the tasks involved repetition to gauge whether participants could apply similar interactions after encountering them once (See supplementary materials for more details).
Activity 2: Free Exploration This part allowed participants to use the tool freely based on their own research interests, facilitating exploration of the dataset in ways they commonly encounter during their work with data.
Activity 3: Semi-Structured Interviews After completing the tasks, each participant took part in a fifteen-minute semi-structured interview to provide insight into their experiences using the tool (See supplementary materials for more details).
Activity 4: Questioners Finally, participants filled questionnaires about their background and their opinion on the tool.

6.2 Data Analysis

The experimenter observed and gathered notes throughout the sessions. The session and interviews were transcribed. Following guidelines about inductive data analysis [9], these notes and transcriptions were open-coded. The first emerging topics were: network reveals (how while the network is not initially shown they can expand networks of interest); visual features triggering exploration (examples: large data rectangles, time base surprises, data

anomalies); leveraging previous knowledge; and people mentioning other uses of this visualizations. Then through rounds of discussion with two other researchers the categories of findings were refined which are presented in the following section.

7 FINDINGS

Based on the post-study questionnaire, all participants found WPHPVis useful in identifying time-related patterns and easy to use for these tasks. 5 experts explicitly stated that it was helpful in discovering relationships between persons and firms. 4 non-data experts also said it was valuable in identifying relationships between persons, firms, and titles. 11/12 participants indicated that they were able to obtain the desired information from the dataset. Moreover, all data experts expressed a strong agreement that they would use WPHPVis to explore their dataset further and mentioned their intention to recommend the tool to other researchers on their team. The following explains on our findings from our data analysis.

7.1 Using Constructive Network Reveal

Demonstrating the network behind the data was a main focus of this project. While we decided against showing the complete network with all the edges in an overview, which would have resulted in a fully filled screen, we provided interactions to allow a user to constructively build their network of interest. We observed that most participants were not initially involved with the network aspect of the data. However, this changed as they began to grasp the interaction capabilities. Many participants (7/12) expressed their opinions about the network behind this visualization. For example, P12 said: *"Cool [...] I love the network ability of it. [...] if I click on her now and then I get to see all her different stuff and, whoa, she's involved in a couple of different ways"*. While saying this, she indicated with a hand gesture showing the step-by-step process of exploring the network. Also, P1 stated *"It's sort of stacked. When data is sort of stacked in this form, it tells me that, me as a user has to go through and click through things."* Participants also thought that it is beneficial that the network is structured and does not show all of the edges all the time. For example, P10 mentioned *"I think it's organized Well. At first, before using it, I didn't really understand. Like before you start asking those questions. I didn't. I don't even know how to go about finding that stuff, but I think the more you use it, the more it makes sense"*, or P5 mentioned *"And although there are already many variables, I think those are quite easy to distinguish. And the interface is not very messy, it's not overwhelming."*.

7.2 Exploration Triggers

We also observed different factors that prompted exploration.

The triple timeline often played a crucial role for individuals who were less familiar with data, by providing helpful starting points for their explorations. For instance, when indicating the large date rectangles, P8 said *"I am very curious about these kind of prolific years."*. The extreme dates, another visual clue, also grabbed attention; for example, P11 became interested in the latest publications saying, *"These ones go all the way to 1840, which is interesting."*

The filters were used by data experts to explore and narrow down the dataset based on their own research questions. Non-data experts used them to look for people in the publication industry that they knew, for example, Shakespeare and Jane Austen.

7.3 Supporting Experts' Research Practices

Working with data anomalies. The WPHPVis included display of the data as is including messiness and missingness. Data experts spotted relatively more of these data anomalies. P2 said, *"It's funny I keep opening titles and seeing, oh, this needs to be fixed."* and P11 noticed that the person was a woman when she opened a record from missing data with an unknown gender timeline. Data experts sometimes used the features that surfaced missing data to focus the timeline specifically on those entries. P1 said, *"When I clicked on missing persons, it showed me that many of the titles, for example, from the later 18th century; and that's really useful to know because then that could be a place where we might start to continue my recovery work rather than starting all the way from the 17th century."* For example, there is a data discrepancy in firm start dates being later than their first recorded publication. Data experts mentioned that the way lines were used made it easy to spot such issues. P4 stated, *"we noticed that there are clearly problems with the firm data. The way that you show firms and titles over time makes that insight really easy to see and find because I can just scroll along firms and everywhere the line is kind of curving up except missing title."*

Using WPHPVis raises uncertainty and helps verification. Non-data experts accepted visuals and relations as straightforward, while data experts approached them skeptically, questioning the underlying information. For example, when looking at the overview showing a concentration of publications in later years, non-data experts may interpret it as a flourishing era for print. However, data experts consider other possibilities, such as the likelihood of focusing more on publications from that specific period. P11 expressed this uncertainty, mentioning the need to examine whether the data is representative of the actual publishing industry during that time or if it was influenced by the data imported.

In addition, data experts have uncertainty about the way data is inputted into the database. During task completion, data experts frequently used the information box to verify title information to address their uncertainty. Based on the WPHP approach of saving the titles as they are, they can check if the fields are filled properly. This provided them with more confidence in their results. WPHPVis also facilitated additional verification. For example, when firm start dates are later than their first recorded publication, P12, interacted with WPHPVis and realized that the firm name in the imprint did not match the firm assigned to the title, indicating the possibility of multiple records for the same firm with different addresses.

WPHPVis and idea generation. Moreover, the fresh perspectives derived from doing the study sometimes led to research ideas. For instance, one of the data experts(P12) observed lines connecting the title to other columns, prompting them to ponder which firms collaborated and at what stage they were involved. This observation was intriguing: *"You get 2 lines. That tells me there are the people who printed this.[...] They started being active in 1745 or whatever but there were also a firm involved that became active in 1725. so that's almost you can know how established the firm is. Who published it? That's cool. I really like that. That's really needed."* Afterward, their focus shifted to the title column, and they remarked: *"I'm really*

interested in how you can see the different firms. This is fascinating to me because this tells me that you've got firms working together who've been active at very different times, right? like You've got a firm who started in 1765 working with a firm that started in 1785. Like you've got someone really new to the business working with someone who hasn't been in the business all that long. That's really cool to know. And you've got kind of people at all sorts of stages. We have a lot of different firm groups that would work together, who knew each other and they would publish together over a really long time."

The experts also started to have conjectures that firms usually have various roles during their life in comparison to persons. They stated: *"He was a printer and a bookseller, is that right? The yellow and the green... fascinating So that's interesting because it means that our firms kind of show up more easily as having multiple roles. Because they more commonly had multiple rules. That's cool."*

Starting from known data problem. Data experts used their knowledge of data to find answers more quickly by optimizing their starting points. They usually formulated their own specific questions. They initiated their exploration with a focus on detail, using filters and specific criteria to narrow down the dataset. For instance, when looking for a person with 10 publications, P1, a data expert, began in the year 1750 because they knew it was a prolific period for the print industry. Similarly, when looking for an author active over 7 years, they might focus on people they were familiar with and knew to have many contributions. P11, used the main title, instead of the information box, to find the firm contributor because she knew that titles can contain the firm names.

7.4 Arising future design suggestions

Both groups of participants suggested improvements for WPHPVis.

Adjustable arrangement of visual encoding We also observed and discussed data experts' needs and desired features in using the system. For example, P12 discussed how showing the relocation of firms over time can help their research. As in their database, relocation led to having two different records per firm. They suggested merging those firms in the visualization in a way that they are separable and also relatable. Some data experts also mentioned that there are different prints of books in different places during the time, and being able to visually see the location or filter it can be useful, especially since they recently started to work separately on different locations of prints. They pointed out that currently they have separate people to work on different publications based on location and it will be helpful to add a location filter.

7.5 Experts' and non-data experts' future use

Some data experts expressed their enthusiasm for future use, as exemplified by P12 who said, *"This is so neat. I could play around with this for hours."* P2 also mentioned *"I thought it was really fun to use. [...] I think it will be really useful for other researchers working on their projects."*. All experts mentioned other uses for the interactions. Some experts were impressed by what the visualization can do; for example, P9 said *"First of all, we've been talking about visualization for so long, and I had no idea it could be so informative and good."*.

Similarly, a non-data expert mentioned a use case in psychology for categorizing patients in an organized way with a lot of data: *"This is definitely a good tool for a lot of information, I think, can be very useful in psychology. And I could see this kind of tool being used for maybe categorizing and having a very large set of data. And maybe patients with different variables here because I think sometimes it's very hard to find a visualization that it's somewhat, you know, clean and useful when it comes to a lot of data."* A non-data expert with an Archaeology background suggested a use case of showing the bones and their information and also mentioned the possibility of having it as a tool that can be adjusted to different first-level courses *"I'm thinking that it would be really easy to apply this to different datasets [...] in my discipline. Like the name of the bone, and then the features of the bone, and then what is the reference? Where is that defined? or something like that. So it'll be really easy[...] or if you're studying a class in first-year biology, and you want to study for your classrooms like you make one thing here, and then next year you have another class. Like an app, and then you use the same tool, but just for different things."* Two participants mentioned possible use of this tool in research in general, such as by showing the researchers and their research interests. One participant suggested making it phone-friendly so that it will be more accessible.

8 DISCUSSION

The results of our study emphasize several factors raised by using the tool. In this section we are going to discuss this project from five different aspects – the data messiness and missingness, the interplay between structure and explorability, using interaction to reveal a network visualization, bringing sketching and co-design into play when working with data experts, and future directions.

Data Messiness and Missingness: Although common practices in information visualization often start with wrangling and cleaning the data, in contrast, our focus was to try to help the data experts in their research by showing them their actual data respecting the importance of both messiness and missingness. These data discrepancies often conveyed important and subtle information about the women's print history and was an important aspect of the data for WPHP project. As expected, the biggest concern when working with uncleaned data is implementation. If there are unmatched formats in one field of data, you have formulate differently for each of them in the code. It also will limit your visualization ideation process as for some of the files there will be limited room for creativity. In our case we decided to use time as the main aspect of our visualization but we could not go more in detail than the year as we did not have more detail for all records and still there were many books missing a publication year. We placed data with missing time stamps at top of timelines to give the experts the opportunity to interact with them.

As noted in our results the data experts liked the fact that they could see the actual data when working with the dataset as it was. In some cases of exploring the data set, they encountered some part of the data that needed to be changed. We suggest that it is important to consider keeping data messiness as much as possible in visualization, as not only they are the truth of data, but also that they can be helpful in improving data in the future.

Research Objective: **Should we always clean data? When is it important to preserve the messiness of data in visualization?**

The Interplay between Structured and Explorable Visualizations: In our research, the need to organize data for data experts

while preserving missing data led us to consider the design of the underlying network structure. As we delved into related work, we noticed that various famous designs, such as the use of parallel coordinates, have been employed to organize networks [35]. We decided to combine the concepts of parallel coordinates and timeline design to visualize the network data. Our study indicated that this combination effectively assists end-users in discovering patterns and contributes to maintaining a clean and organized design. Although results from our study demonstrate that end-users found this system useful, there is a strong need to try different visualization that can address the problem of showing the network behind the historical datasets in a way that it is organized.
Research Objective: **How to have a structured screen as well as an explorable network?**

Interactive Construction of a Network Visualization: Instead of starting from an overview of the complex network in the WHPH dataset and developing interactive ways of clarifying through various action like bending, bundling [21], filtering, etc. we chose to start with an organization of the entities and provide tools through which one can interactively construct the network. Through our study we did notice that our participants were surprised by no overt display of the network initially. Practicing interactive construction seemed to come naturally and was used with enthusiasm. The edge reveal on hover seemed to be essential for people to become aware of the possibility of displaying the network edges and they did have to learn how clicking would open up the network.
Research Objective: **How to extend the concept of a constructive approach to network visualization?**

Leveraging Sketching and Co-Design: In our collaboration we had on-going discussions, made extensive use of the WPHP website to learn as much as we could about the WPHP project, but most fruitful of all we held sketching sessions. When possible these sessions included WPHP experts, and visualization experts. Starting with small hand accessible sample datasets, we would create and discuss multiple possible sketched visualizations. Working from a small sample datasets, we encouraged the whole group including the data experts to express their visualization needs – including hopes and fears – more informally. Sketching sessions have been very informative a several points during that design process and have been useful sparking and maintaining expert interest when we had no concrete designs. In these sessions, both our research team and the data experts sketched their visualization ideas, allowing for a more interactive and understandable exchange of concepts.

We used an iterative approach, and continuously seeking feedback from the data experts throughout the design and implementation process. By involving them at every stage and incorporating their insights, it is possible to guide the final visualization tool towards meeting their needs. In this process, interleaving implementation and design simplified the process of explaining the design to experts and provided more opportunities for the experts to influence the visualization as it progressed.
Research Objective: **How can we expand on possibilities of co-design without over taxing domain experts?**

8.1 Limitations and Future Work:

WPHPVis is currently limited by the number of contribution timelines that can be opened on simultaneously. Presently one can open 10 contribution timelines from both Persons and Firms, however, at this point screen space becomes an issue. While gratifying that we can already see that the domain experts would use more capacity for complexity, this is a current limitation we are looking to address. We are exploring screen space expansion possibilities.

Our design prioritized direct interaction to uncover data of interest and unexpected information. However, discussions with the domain experts are suggesting integrating search functionality into this visualization. As one participant (P1) expressed, "So this is interesting because it gives me the visual component, but in terms of finding information quickly for me, I would maybe struggle with this a little bit." However, search engines also can have issues with data messiness and missingness. Future work could investigate finding a solution that strikes more balance between search capabilities and constructive interaction, discovery and exploration features.

Continuing to work with WPHP data experts, we are deploying the tool on the Women's Print History Project (WPHP) website. This deployment will enable us to conduct a long term study [38], which will provide valuable insights about the use of visualization integrated into a manual data collection project.

9 CONCLUSIONS

In this paper, we have presented the design, development and study of a visualization tool, WPHPVis, which we co-designed with some members of the Women's Print History Project (WPHP) research process. Through our sketching plus co-design process, we arrived at a visualization that takes a fresh look at edge congestion problem. Instead of working from a visually presented full set of edges and employing various, filtering, bundling, bending and highlighting methods, we prove the tools for the viewer to constructively build that specific network of immediate interest.Both data experts and non-experts quickly grasped the process and could make active use of it. We found that our sketching co-design process offered ways to better communicate within the team about the data, the data related research questions and challenges. This in turn led to new ideas focused directly on the WPHP data. Through our qualitative study, we saw that people (experts and non-experts alike) quickly appreciated the exploration possibilities within the tool and non-expert suggested creating tools like this for their domains. We also contribute four new research objectives: 1) Should we always clean data? When is it important to preserve the messiness of data in visualization? 2) How to have a structured screen as well as explorable network? 3) How to extend the concept of a constructive approach to network visualization? 4) How to expand on possibilities of co-design without over taxing domain experts?

ACKNOWLEDGMENTS

We thank our colleagues and reviewers for their thoughtful comments and our participants for their time and invaluable input. This research was funded in part by NSERC Discovery Grant Interactive Visualization RGPIN-2019-07192, and Canada Research Chair in Data Visualization CRC-2019-00368.

REFERENCES

[1] [n. d.]. Voyant Tools. https://voyant-tools.org/.

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*.

[3] Alejandro Benito-Santos and Roberto Therón Sánchez. 2020. A Data-Driven Introduction to Authors, Readings, and Techniques in Visualization for the Digital Humanities. *IEEE Computer Graphics and Applications* 40, 3 (2020), 45–57. https://doi.org/10.1109/MCG.2020.2973945

[4] Marcus Bingenheimer, Jen-Jou Hung, and Simon Wiles. 2011. Social network visualization from TEI data. *Lit. Linguistic Comput.* 26 (2011), 271–278.

[5] Matthew Brehmer, Bongshin Lee, Benjamin Bach, Nathalie Henry Riche, and Tamara Munzner. 2017. Timelines Revisited: A Design Space and Considerations for Expressive Storytelling. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (2017), 2151–2164. https://doi.org/10.1109/TVCG.2016.2614803

[6] Michael Burch, Abdullah Saeed, Alina Vorobiova, Armin Memar Zahedani, Linus Hafkemeyer, and Marco Palazzo. 2020. EDBLP: Visualizing Scientific Publications. In *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction* (Eindhoven, Netherlands) *(VINCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages. https://doi.org/10.1145/3430036.3430052

[7] Sheelagh Carpendale and Catherine Montagnese. 2001. A framework for unifying presentation space. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. 61–70.

[8] Rex Chen and Chiming Chen. 2015. Visualizing the world's scientific publications. *Journal of the Association for Information Science and Technology* 67 (09 2015). https://doi.org/10.1002/asi.23591

[9] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[10] Anthony Don, Elena Zheleva, Machon Gregory, Sureyya Tarkan, Loretta Auvil, Tanya Clement, Ben Shneiderman, and Catherine Plaisant. 2007. Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (Lisbon, Portugal) *(CIKM '07)*. Association for Computing Machinery, New York, NY, USA, 213–222. https://doi.org/10.1145/1321440.1321473

[11] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. 2012. GraphTrail: Analyzing Large Multivariate, Heterogeneous Networks while Supporting Exploration History. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2207676.2208293

[12] Marian Dörk, Christopher Pietsch, and Gabriel Credico. 2017. One view is not enough: High-level visualizations of a large cultural collection. *Information Design Journal* 23 (07 2017), 39–47. https://doi.org/10.1075/idj.23.1.06dor

[13] Isabella (Belle) Eist. 2023. Hidden in the Imprints: Introducing Ann Vernor, Bookseller and Publisher, Active 1793-1807. https://womensprinthistoryproject.com/blog/post/117. [last access: Apr-15-2023].

[14] George W Furnas. 1986. Generalized fisheye views. *Acm Sigchi Bulletin* 17, 4 (1986), 16–23.

[15] Sohaib Ghani, Bum Chul Kwon, Seungyoon Lee, Ji Soo Yi, and Niklas Elmqvist. 2013. Visual Analytics for Multimodal Social Network Analysis: A Design Study with Social Scientists. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2032–2041. https://doi.org/10.1109/TVCG.2013.223

[16] Fred Gibbs and Trevor Owens. 2012. Building better digital humanities tools. *DH Quarterly* 6, 2 (2012), 1–14.

[17] John D. Gould and Clayton Lewis. 1983. Designing for Usability—Key Principles and What Designers Think. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '83)*. Association for Computing Machinery, New York, NY, USA, 50–53. https://doi.org/10.1145/800045.801579

[18] Saul Greenberg, Sheelagh Carpendale, Nicolai Marquardt, and Bill Buxton. 2011. *Sketching user experiences: The workbook*. Elsevier.

[19] Kyle Wm. Hall, Adam J. Bradley, Uta Hinrichs, Samuel Huron, Jo Wood, Christopher Collins, and Sheelagh Carpendale. 2020. Design by Immersion: A Transdisciplinary Approach to Problem-Driven Visualizations. *IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE InfoVis 2019)* 26, 1 (2020), 109–118. https://doi.org/10.1109/TVCG.2019.2934790

[20] Uta Hinrichs, Stefania Forlini, and Bridget Moynihan. 2016. Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 429–438. https://doi.org/10.1109/TVCG.2015.2467452

[21] Danny Holten. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics* 12, 5 (2006), 741–748.

[22] Humanities + Design Lab. 2014. *Tutorials and FAQs*. https://hdlab.stanford.edu/palladio/help/ Accessed: 2023.

[23] Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D. Stolper, Michael Sedlmair, Jian Chen, Torsten Möller, and John Stasko. 2017. Vispubdata.org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics* 23, 9 (2017), 2199–2206. https://doi.org/10.1109/TVCG.2016.2615308

[24] H. Lamqaddam. [n. d.]. When the tech kids are running too fast: Data visualisation through the lens of art history research. *3rd Workshop on Visualization for the Digital Humanities at IEEEVIS* ([n. d.]).

[25] Houda Lamqaddam, Andrew Vande Moere, Vero Vanden Abeele, Koenraad Brosens, and Katrien Verbert. 2021. Introducing Layers of Meaning (LoM): A Framework to Reduce Semantic Distance of Visualization In Humanistic Research. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1084–1094. https://doi.org/10.1109/TVCG.2020.3030426

[26] Michelle Levy. 2022. Ann Williams: Postmistress, Poetess, Sericulturist. https://womensprinthistoryproject.com/blog/post/106. [last access: Apr-15-2023].

[27] Michelle Levy and Kandice Sharren. 2023. The Women's Print History Project. https://womensprinthistoryproject.com/. [last access: Apr-15-2023].

[28] Ji-Ye Mao, Karel Vredenburg, Paul W. Smith, and Tom Carey. 2005. The State of User-Centered Design Practice. *Commun. ACM* 48, 3 (mar 2005), 105–109. https://doi.org/10.1145/1047671.1047677

[29] Vivien Marx. 2013. The big challenges of big data. *Nature* 498, 7453 (2013), 255–260.

[30] Nina Mccurdy, Julie Gerdes, and Miriah Meyer. 2019. A Framework for Externalizing Implicit Error Using Visualization. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 925–935. https://doi.org/10.1109/TVCG.2018.2864913

[31] Andrej Mrvar and Vladimir Batagelj. 2016. Analysis and visualization of large networks with program package Pajek. *Complex Adapt Syst Model* 4, 1 (2016), 6. https://doi.org/10.1186/s40294-016-0017-8

[32] Carolina Nobre, Miriah Meyer, Marc Streit, and Alexander Lex. 2019. The state of the art in visualizing multivariate networks. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 807–832.

[33] Georgia Panagiotidou, Ralf Vandam, Jeroen Poblome, and Andrew Vande Moere. 2022. Implicit Error, Uncertainty and Confidence in Visualization: An Archaeological Case Study. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2022), 4389–4402. https://doi.org/10.1109/TVCG.2021.3088339

[34] Nathalie Henry Riche, Tim Dwyer, Bongshin Lee, and Sheelagh Carpendale. 2012. Exploring the design space of interactive link curvature in network diagrams. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 506–513.

[35] Ross Shannon, Thomas Holland, and Aaron Quigley. 2008. Multivariate graph drawing using parallel coordinate visualisations. *University College Dublin, School of Computer Science and Informatics, Tech. Rep* 6 (2008), 2008.

[36] Alice Thudt, Uta Hinrichs, and Sheelagh Carpendale. 2012. The bohemian bookshelf: supporting serendipitous book discoveries through information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1461–1470.

[37] Paola Valdivia, Paolo Buono, Catherine Plaisant, Nicole Dufournaud, and Jean-Daniel Fekete. 2021. Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 1 (2021), 1–13. https://doi.org/10.1109/TVCG.2019.2933196

[38] Eliane R. A. Valiati, Carla M. D. S. Freitas, and Marcelo S. Pimenta. 2008. Using Multi-Dimensional in-Depth Long-Term Case Studies for Information Visualization Evaluation. In *Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel EvaLuation Methods for Information Visualization* (Florence, Italy) *(BELIV '08)*. Association for Computing Machinery, New York, NY, USA, Article 9, 7 pages. https://doi.org/10.1145/1377966.1377978

[39] Fernanda B. Viegas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1137–1144. https://doi.org/10.1109/TVCG.2009.171

[40] Pawan Vora. 2009. *Web application design patterns*. Morgan Kaufmann.

[41] Angela Wachowich. 2022. (Unidentified) Woman not Inferior to Man: 'Sophia,' Proto-Feminism, and the Anonymous Female Writer. https://womensprinthistoryproject.com/blog/post/97. [last access: Apr-15-2023].

[42] Mitchell Whitelaw. 2015. Generous Interfaces for Digital Cultural Collections. *Digit. Humanit. Q.* 9, 1 (2015). http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

[43] Florian Windhager, Paolo Federico, Günther Schreder, Katrin Glinka, Marian Dörk, Silvia Miksch, and Eva Mayr. 2019. Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. *IEEE Transactions on Visualization and Computer Graphics* 25, 6 (2019), 2311–2330. https://doi.org/10.1109/TVCG.2018.2830759

[44] Ruojin Zhang, Vimukthi Jayawardene, Marta Indulska, Shazia Wasim Sadiq, and Xiaofang Zhou. 2014. A Data Driven Approach for Discovering Data Quality Requirements. In *International Conference on Interaction Sciences*.

Received 24 January 2024; revised 3 April 2024; accepted 5 March 2024