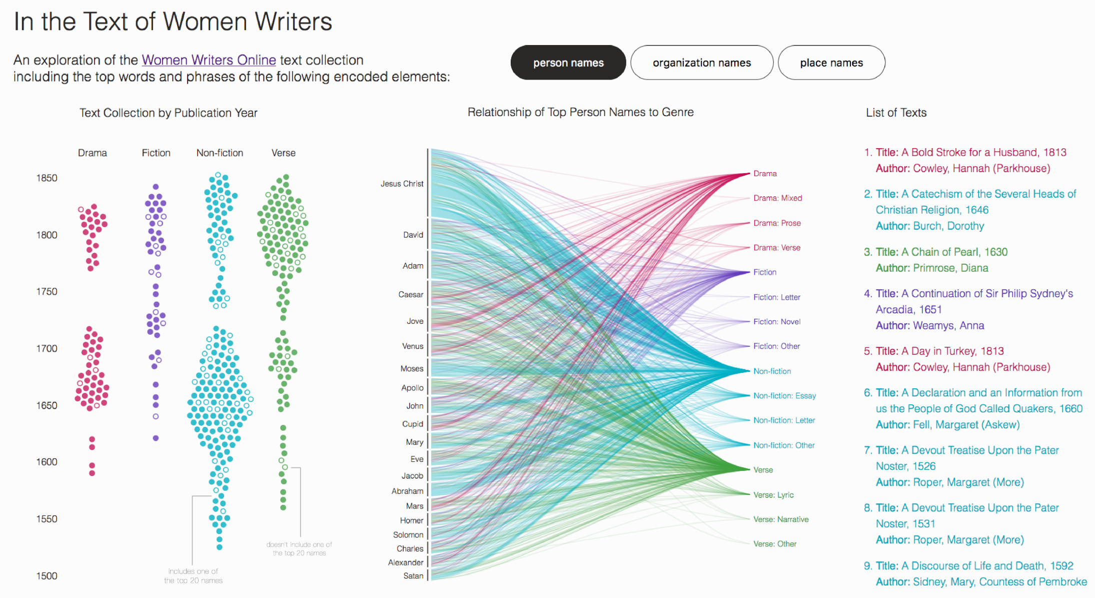
Close and Distant Reading via Named Entity Network Visualization: A Case Study of Women Writers Online

Sarah Campbell, Zheng-yan Yu, Sarah Connell, and Cody Dunne



In the Text of Women Writers

An exploration of the Women Writers Online text collection including the top words and phrases of the following encoded elements:

1. **Title:** A Bold Stroke for a Husband, 1813
   Author: Cowley, Hannah (Parkhouse)
2. **Title:** A Catechism of the Several Heads of Christian Religion, 1646
   Author: Burch, Dorothy
3. **Title:** A Chain of Pearl, 1630
   Author: Primrose, Diana
4. **Title:** A Continuation of Sir Philip Sydney's Arcadia, 1651
   Author: Weamys, Anna
5. **Title:** A Day in Turkey, 1813
   Author: Cowley, Hannah (Parkhouse)
6. **Title:** A Declaration and an Information from us the People of God Called Quakers, 1660
   Author: Fell, Margaret (Askew)
7. **Title:** A Devout Treatise Upon the Pater Noster, 1526
   Author: Roper, Margaret (More)
8. **Title:** A Devout Treatise Upon the Pater Noster, 1531
   Author: Roper, Margaret (More)
9. **Title:** A Discourse of Life and Death, 1592
   Author: Sidney, Mary, Countess of Pembroke

Fig. 1. Three linked visualizations showing a named entity network queried from the Women Writers Online textbase. Left: a beeswarm visualization shows the genre of each publication by year. Middle: a bipartite node-link visualization of the top 20 named entities connected to the genre & subgenre of texts they reside in. Right: a list of texts that include at least one of the top 20 named entities, ordered alphabetically and linked to the full text. Marks are colored categorically by genre: drama is pink ●, fiction is purple ●, non-fiction is blue ●, and verse is green ●. Empty circles show texts that do not include the top 20 named entities, e.g. ○.

Abstract—Close reading and distant reading are widely used in digital humanities and can benefit from information visualizations. Digital humanities scholars have curated numerous TEI-encoded textual collections which provide the data necessary for blending both close and distant reading – however we do not have tools to support general users in conducting these blended analyses. In this paper we focus on one such collection: Women Writers Online (WWO). We contribute the design and implementation of a multiple coordinated view network visualization to facilitate close and distant reading in WWO and a transparent view into our iterative design process to help guide future designers and humanists in applying our approach to other textual collections.

Index Terms—Bipartite network visualization, Women Writers Online, Text visualization

1 Introduction

Close and distant reading are two important tools in the digital humanities toolbox which can both benefit from information visualization [6]. **Close reading** is the traditional method for literary criticism. Several visualizations have been developed to support close reading, but existing approaches can benefit from adding supplementary **named entity** information, especially acting persons and their relationships [6]. **Distant reading**, introduced by Moretti [11], alternatively focuses on an abstract view of global features of one or more texts. Network/graph visualizations can be particularly useful for examining relationships between these features and texts for corpus analysis [6]. We designed an interactive visualization to support a blend of close and distant reading – both explorations at scale and text-level investigation.

This paper focuses on the application of our visualization approach in service to the Women Writers Project (WWP). The WWP is a long-term digital humanities research project at Northeastern University, devoted to early modern women's writing and electronic text encoding. The goal of the project is to bring texts from pre-Victorian women writers out of the archive and make them more accessible to a wide audience of teachers, students, scholars, and the general user. We focus on the WWP's major textual collection, Women Writers Online (WWO). WWO is a full-text collection of early women's writing in English. It currently includes full transcriptions, encoded following the standards of the Text Encoding Initiative (TEI), of 407 texts published between 1526 and 1850. In addition to the collection's broad chronological framing, the texts in WWO also represent a very diverse set of genres, ranging from prophecies, religious meditations, petitions, and recipe books to novels and dramas in both prose and verse. The WWP has basic visualization prototypes published through the project's WWO Lab [17], and was interested in developing more dynamic visualizations to take advantage of the detailed information available in the TEI markup for each text.

Women Writers Online is one of numerous TEI-encoded textual collections and thus can be used to model a common challenge faced

- *Sarah Campbell, Zheng-yan Yu, Sarah Connell, and Cody Dunne are with Northeastern University. Emails: [campbell.sar | yu.zheng]@husky.neu.edu, [sa.connell | c.dunne]@northeastern.edu.*

Proc. 3rd Workshop on Visualization for the Digital Humanities (VIS4DH), co-located with IEEE VIS 2018.

by many digital humanities projects: how to represent the very detailed information captured in the encoding without overwhelming or confusing readers. While prior research reiterates the use of projections for understanding heterogeneous networks (multiple types of nodes) with node-link visualizations, we find that they are not intuitive for a general audience to understand and that they lose information by collapsing variables [18]. Moreover, interactive network visualizations for distant reading can easily suffer from poor layout, visual clutter, and complexity. This presents an opportunity for applying alternate network representations and layouts. Finally, in order to support interactive data exploration and serendipitous discovery it is necessary to incorporate interactivity into the visualization design.

Cognizant of existing visualization limitations and to support the goals of digital humanities researchers, in this paper we contribute:

- the design and implementation of a multiple coordinated view network visualization to facilitate close and distant reading in a digital humanities textbase and

- a transparent view into our iterative design process to help guide designers and humanists combating similar problems.

Our visualization is available integrated with Women Writers Online (WWO) [1] and open source[2]. We also provide a supplemental video which demonstrates the visualization in action[3]. The rest of this paper discusses related work, data used, our design goals, visualization design and implementation, and preliminary evaluation results.

2 Background and Related Work

Here we discuss related work from multiple perspectives relating to the digital humanities, network analysis, visualization, and design.

We already mentioned the survey by Jänicke et al. [6] which summarizes visualization approaches for close and distant reading in the digital humanities, as well as Moretti's introduction of distant reading [11]. An example of distant reading is Posavec's work [13] visualizing the text of Jack Kerouac's *On the Road*. A common way to analyze a text is by visualizing named entities. While names of places are often represented geospatially (with uncertainty issues) [6], names of people are often visualized as a network. For example, Marcus Bingenheimer et al. [2] analyzed the relationships between Buddhist monks from biographies as a node-link visualization and Klein [7] illustrated the social network of Thomas Jefferson using an arc diagram.

Existing arc diagram approaches were particularly inspiring for us, especially one of frequently used user interface commands by Matejka [8]. It presents a network in list form with additional encoded elements. *Voices that Care* by Studio Terp [16] also shows an arc diagram but combines it with a parallel coordinates plot to show a **bipartite network** – a heterogeneous network with two node types. Our underlying data from Women Writers Online is a heterogeneous network consisting of many node types, but we can limit our representations to two types at a time. Borgatti & Everett [4] analyzed bipartite network data in various ways, including detecting clusters and measuring centrality. Melamed [10] and Bongiorno et al. [3] provide algorithms for community detection in bipartite networks. A common way to analyze a bipartite network is creating a **projection** of it into a network with a single node type so as to reduce the complexity of the data and facilitate the application of traditional analysis techniques. Banerjee et al. [1] present results on the topological properties of projected networks and Zhou et al. [18] propose a weighting method of edges for bipartite network projection based on resource-allocation. However, projections cause a loss of data [18] and the results are not intuitive to a general audience, so we endeavor to design a non-projected bipartite visualization. One classic way of visualizing these heterogeneous networks related to parallel coordinates is Jigsaw [15].

Other visualizations served more as design guidance. *textexture* by Nodus Labs [12] and *Traces* by Fathom Information Design [5] are two diverse approaches for visualizing a body of text, but both incorporate interactions to connect the user directly to locations in the original text. This was a goal in our work as well. Another goal was to represent a dense and complex archive in a simple, clear way. *Linked Jazz* by Pattuelli [14] visualized links between each jazz musician in the associated archive, but the resulting structure is more cluttered than we prefer. *Citeology* by Matejka et al. [9] served as a valuable reference for how to structure a text-heavy network based on such an archive.

3 Data

The data we queried from the textbase[4] included metadata for each text along with the contents of select elements. The available metadata for each of the texts (401 at the time) includes the title, author, publication location, publication year, and main genre. For element selection, we relied on the list of elements and definitions on the Women Writers Online website. Below is a list of elements we selected to visualize:

- `<orgName>`: proper names of organizations

- `<placeName>`: proper names of places

- `<persName>`: proper names of human beings

We transformed and cleaned the queried data using R and Excel. The Women Writers Project does not regularize original spellings or expand common abbreviations. We performed further named entity de-duplication manually.

4 Design

Our visualization has three linked views shown in Fig. 1: a beeswarm overview of text genres by year, a bipartite network visualization of edges between named entities and genres edges, and a list of texts. The overview (left) represents metadata for all 401 texts then in the collection. The bipartite network visualization (middle) shows the relationships between the top 20 named entities of the selected type and the genres of texts in which they appear. Last, the list (right) shows alphabetically the texts that have at least one of the top 20 named entities with links to the full texts on the Women Writers Online website. In this section, we present our design goals, design iterations, and our resulting visual and interaction design for visualizing this textbase.

4.1 Design Goals

Our ultimate design goal was to support a blend of close and distant reading of the texts in the Women Writers Online (WWO). To this end we created several sub-goals:

To support close reading we want to ***DG1:*** display for a single text the named entities extracted from the TEI metadata and their relationships to genres and ***DG2:*** provide the full text of the source material. To support distant reading we want to ***DG3:*** display for a corpus the named entities extracted from the TEI metadata and the relationships between named entities, texts, and genres. To make WWO more available to a wide audience we aim to ***DG4:*** prevent visual clutter as much as possible and design easily comprehensible visual representations. Finally, we also want to encourage data exploration and serendipitous discovery by providing ***DG5:*** multiple coordinated views of different aspects of the data and ***DG6:*** interactive, dynamic visualizations.

4.2 Iterative Design: Analog Sketches

Our analog sketches in Fig. 2 show several styles of network visualization that we explored in order to display different variables in the data. The sketch in the top left is an arc diagram of a network that shows how values of one named entity type (organization names) connect with each other if they appear in the same text. The node-link

[1] `https://wwp.northeastern.edu/wwo/lab/names-viz/`

[2] `https://github.com/VisDunneRight/WWOVis`

[3] `https://www.youtube.com/watch?v=uGfDw5yK71w`

[4] We used the WWP's routines for selectively extracting plain text from XML files with XQuery and XSLT (github.com/NEU-DSG/wwp-public-code-share/tree/master/fulltext). These routines lightly regularize the encoded text by, for example, preferring corrections to errors and expansions to abbreviations, modernizing *i/j* and *u/v* shifts, removing forme work such as page numbers and catchwords, and reconciling words that break over lines or pages.

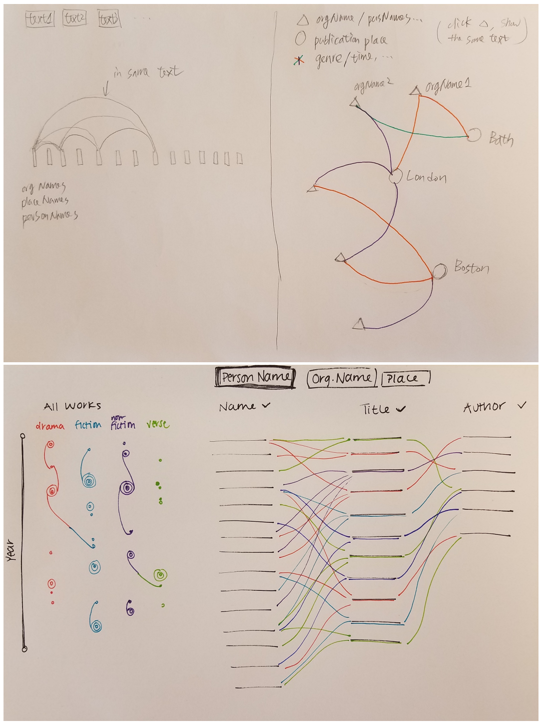


Fig. 2. Two initial sketches show potential visualization designs.

visualization in the top right focuses on displaying the relationships between publication location and organization names, with publications laid out using publication location.

Based on explorations of this dataset using standard tools, we decided it was necessary to include more than one view of the data to better convey the interesting complexity and connections. The bottom sketch experiments with combined visualizations: a view of the metadata and a network visualization highlighting a named entity. The metadata view shows every text in the collection across time and genre. One circle represents one text – for texts of the same genre and publication year, circles are placed concentrically to prevent overplotting. Edges represented by arcs connect circles of texts that are by the same author. Color encodes the genre of the text, which is also encoded on the arcs for network edges. The bipartite network visualization connects named entity values to the texts they are found in. A third variable is added to aggregate texts by author. The values of these variables are presented in a list form like parallel coordinates or Jigsaw [15], with each variable having its own column. We progressed with this sketch for our first digital implementation, but had to reconsider the treatment of these variables since the sketch does not take into account the sheer number of values they have.

4.3 Iterative Design: Digital Implementation

We refined the sketch designs digitally and iteratively. In the overview visualization, we deleted the arcs connecting the circles for each text and positioned each circle using force properties. In the network visualization, we focused on the most frequent named entities in the collection, sorted by frequency, and connected these names to genre metadata. Lastly, we added a third view: a text list, which immediately gives readers access to text titles, authors, and the full text online. Iteration details for specific views are discussed in more detail below.

4.3.1 Overview Visualization

A primary goal for the overview visualization was to enable users to explore and interact with each text. Therefore, it was vital to make each circle accessible to the user. Fig. 3 illustrates three iterative designs for this view of the collection. We visualized the texts by their main genre and the year of publication, where each circle represents one of the 401 texts in the collection. The left-hand image shows the first digital implementation of the visualization. The decreased opacity of the circles reveals a high level of overplotting, limiting the user's ability to interact with each circle. We first attempted to resolve this with concentric circles, found in the middle image and in the bottom-left sketch in Fig. 2. While this design begins to resolve the overplotting problem, the visualization is still cluttered and does not encourage user interaction at the text level as individual circles are difficult to select. In our third and final iteration, we implemented a beeswarm design using d3.js forces to prevent any circle overlap.

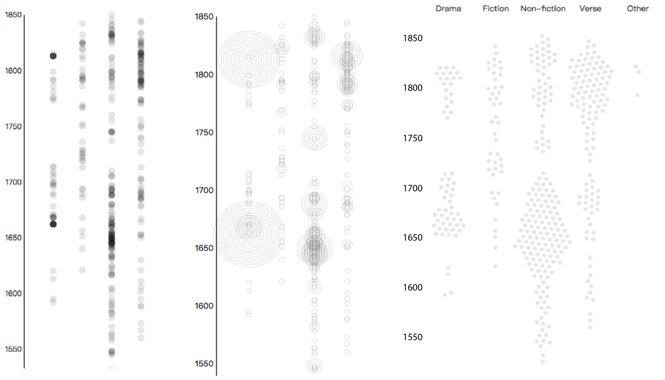


Fig. 3. Design iterations of the overview visualization, showing each text by main genre and publication year.

4.3.2 Node-Link Visualization

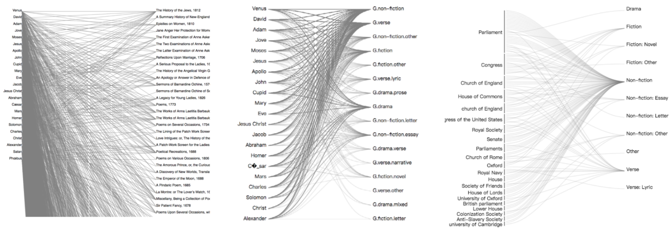


Fig. 4. Design iterations of the named entity network visualization.

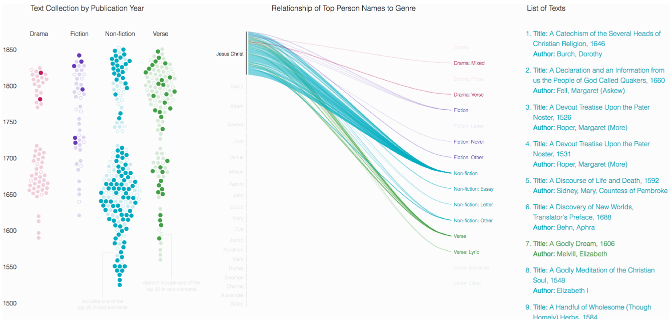
The node-link network visualization underwent several design iterations before reaching the final stage, summarized in Fig. 4. First (left), we filtered the in-text person named entities to those that most frequently occurred in the collection and connected them to the titles of the texts they occurred in. This design proved infeasible due to the number of texts that could be potentially listed. In our second iteration (middle), we instead connected the named entities to the genre metadata. This change enabled us to have a common variable across the overview and node-link visualizations to facilitate connections between metadata and named entities across the two views. Due to the limitation of screen height, we chose the top 20 entities to display in the visualization. Focusing only on the most frequent named entities provides a sense of which entities are dominant in the language of the collection and any variation across genres. We distributed the named entities and genres uniformly along vertical axes and curved the edges. However, this iteration still suffers from dense overlapping of edges.

The third iteration is shown in the right-hand image of Fig. 4. While this image displays <orgName>, instead of <persName> like the previous two iterations, it still accurately reflects the design changes implemented. In this iteration, we allocated proportions of the height of each axis to the named entities based on how many texts they occur in and then distributed those edges across that designated height. For instance, the image indicates that 'Parliament' was the most frequently occurring organization name and the multiple edges connecting 'Parliament' and 'Non-fiction' are visible. We added thin bars to visually clarify the given height for each value. For a given name, the edges are ordered by the publication date of the text. We experimented with ordering the edges for each name by genre. However, `<persName>` and `<placeName>` are far more dense in their number of edges for the top 20 named entities, so the groupings resembled more of a Sankey diagram and the individual edges were lost. Therefore, we sacrificed minimizing edge crossings in order to maintain the individuality of each connection.

4.4 Final Design and Interactions

The final version of the visualization combines the three views as shown in Fig. 1. Categorically coloring all marks by genre became the cohesive piece that visually ties the three views together – drama is pink ●, fiction is purple ●, non-fiction is blue ●, and verse is green ●. In the overview visualization, the horizontal separation of the circles is used alongside color to represent the main genre of the text. The color legend for genre is implicitly provided by this visualization. The circles are either solid – e.g. ● – or hollow – e.g., ○. The solid circles indicate that the text includes at least one of the top 20 named entity values that appear in the bipartite network. This encoding is explained with two annotated examples at the bottom of the visualization.

The bipartite visualization maintains the genre color encoding for the edges and genre names. Several texts include a more granular categorization of genre; thus we show the four potential categories for each primary genre while maintaining the four-color scale. The list of texts includes the title and author for each text which is also color-coded for immediate connection to the text's main genre. Unlike the overview visualization, not all texts are represented. If the text includes any named entities visualized in the bipartite network, it will show up in the list; the list corresponds to the solid circles in the metadata view.



1. **Title:** A Catechism of the Several Heads of Christian Religion, 1646
   Author: Burch, Dorothy
2. **Title:** A Declaration and an Information from us the People of God Called Quakers, 1660
   Author: Fell, Margaret (Askew)
3. **Title:** A Devout Treatise Upon the Pater Noster, 1526
   Author: Roper, Margaret (More)
4. **Title:** A Devout Treatise Upon the Pater Noster, 1531
   Author: Roper, Margaret (More)
5. **Title:** A Discourse of Life and Death, 1592
   Author: Sidney, Mary, Countess of Pembroke
6. **Title:** A Discovery of New Worlds, Translator's Preface, 1688
   Author: Behn, Aphra
7. **Title:** A Godly Dream, 1606
   Author: Melvill, Elizabeth
8. **Title:** A Godly Meditation of the Christian Soul, 1548
   Author: Elizabeth I
9. **Title:** A Handful of Wholesome (Though Homely) Herbs, 1584

Fig. 5. Hovering over a named entity highlights related genres and texts.

When the user hovers over one of the solid circles, that text's corresponding information is highlighted across the three views by fading irrelevant marks. The hovered circle will be highlighted, the named entity values that exist in the text will be highlighted in the bipartite network along with their edges, and the text's title and author will be highlighted in the text list and scrolled to the top of the list. The same interactions occur when the user hovers over one of the texts in the list. These interactions highlight one text at a time. Hover interactions built into the bipartite network hold the potential to draw insight about the text collection as a whole. As shown by Fig. 5, if a user hovers over one of the named entity values, the corresponding edges will be highlighted, the corresponding solid circles will be highlighted, and the text list will be filtered to show only the corresponding texts. A user exploring this collection also has the opportunity to access the texts in full. If they click on a text in the text list, a new window will open to the Women Writers Online (WWO) website, showing the digitized text in its entirety. WWO is a subscription-based collection but free trials and research accounts are available upon request.

The buttons above the visualization are used to select one of the three types of named entities to display: `<persName>`, `<orgName>`, and `<placeName>`. E.g., by clicking the place names button, the visualizations will update to reflect the most frequent place names and their connections in the text collection. All 401 texts remain present in this visualization but their encodings update accordingly.

5 Evaluation

We conducted a small, informal evaluation of our exploratory visualization. The evaluation consists of feedback from two groups: (1) two digital humanities scholars (one is a co-author on this paper) who are immersed in the data on a daily basis, but not trained in information design; and (2) four MFA graduate students at Northeastern University who are trained in visualization and information design but had no prior knowledge of the Women Writers Online textbase.

The digital humanities scholars responded quite positively to the visualization design and the responsiveness of the functionality. They found the overall organization effective for displaying a large amount of information and the interactions intuitive. For them, the linking to full texts in the Women Writers Online was particularly important – it enabled readers to navigate between the visualization and the texts themselves to encourage both explorations at scale and text-level investigation. They noted potential future work to expand the visualization to show other named entities we did not include and other aspects of the collection's metadata. A future improvement they recommended involves stacking interactions, so one could click a particular element value then interact with that subset of texts.

The feedback received from the graduate students were more related to the design and functionality. Participants gave positive feedback on the color choice and the organization of information, which one participant described as "clean, clear, and easy to navigate." Feedback on improvements primarily revolved around clarity of information and existing functionality. For instance, several participants found the legend for the hollow and filled not immediately clear and occurring too low on the page. One participant mentioned how he wished he could search for a particular author with a search bar. Feedback received from this informal evaluation helps direct our attention to the next improvements to be made.

6 Conclusion

We present a visualization that enables students and researchers alike to explore the named entities in a collection of texts. Creating an exploratory visualization for a digital text archive targeted at general users is a challenging task. Digital archives are usually large, complex, and not easily accessible for a general audience. In addition, network visualizations, especially when created for exploratory purposes, can easily become complex and cluttered. The visualizations we propose helps to resolve these two problems: providing a view into the complex nature of digital archives and a design that limits the density of network visualizations. Our solution enables users to move from the birds-eye view of the entire corpus to individual texts, encouraging both close and distant reading. We present our design iterations in an effort to make transparent our design process. We hope this proves beneficial for designers and humanists that combat similar problems within the digital humanities.

Our implementation is available online integrated directly with the Women Writers Online (WWO) textbase. WWO is one of many midsize TEI encoded corpora that are used in digital humanities analysis. For this research, the representativeness of the corpus is often at stake and has a serious impact on the kinds of claims that can be made. Our visualization thus pairs well with other kinds of analysis because it provides an effective way of seeing and understanding key analytical components of the corpus as a whole (genre, publication date, the names of entities, and the actual texts). E.g., this visualization shows that WWO's fiction is strongly slanted toward the 18th and 19th centuries and that mid-17th-century nonfiction and 19th-century verse are particularly well represented. This work could generalize well to other analytical categories represented in TEI metadata as well as other elements to provide a blueprinting mechanism for a corpus.

Acknowledgments

The authors wish to thank Ashley Clark for WWP integration assistance and Northeastern University for internal funding.

References

[1] S. Banerjee, M. Jenamani, and D. K. Pratihar. Properties of a projected network of a bipartite network. In *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0143–0147, 2017. doi: 10.1109/ICCSP.2017.8286734

[2] M. Bingenheimer, J.-J. Hung, and S. Wiles. Social network visualization from TEI data. *Literary and Linguistic Computing*, 26(3):271–278, 2011. doi: 10.1093/llc/fqr020

[3] C. Bongiorno, A. London, S. Miccichè, and R. N. Mantegna. Core of communities in bipartite networks. *Phys. Rev. E*, 96:022321, 2017. doi: 10.1103/PhysRevE.96.022321

[4] S. P. Borgatti and M. G. Everett. Network analysis of 2-mode data. *Social Networks*, 19(3):243 – 269, 1997. doi: 10.1016/S0378-8733(96)00301-2

[5] Fathom Information Design. Traces. https://fathom.info/traces/, 2009.

[6] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. Visual text analysis in digital humanities. *Computer Graphics Forum*, 36(6):226–250, 2017. doi: 10.1111/cgf.12873

[7] L. F. Klein. Social network analysis and visualization in 'The Papers of Thomas Jefferson'. *Proc. Digital Humanities*, 4(9):12, 2012.

[8] J. Matejka. Command usage arc diagrams. https://www.autodeskresearch.com/projects/command-usage-arc, 2010.

[9] J. Matejka, T. Grossman, and G. Fitzmaurice. Citeology: Visualizing paper genealogy. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pp. 181–190, 2012. doi: 10.1145/2212776.2212796

[10] D. Melamed. Community structures in bipartite networks: A dual-projection approach. *PLOS ONE*, 9(5):1–5, 2014. doi: 10.1371/journal.pone.0097823

[11] F. Moretti. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

[12] Nodus Labs. textexture: visualize any text as a network. http://textexture.com/, 2012.

[13] S. Posavec. Literary organism: A visualization of part one of On the Road by Jack Kerouac. http://www.stefanieposavec.com/writing-without-words/, 2007.

[14] Semantic Lab at Pratt. Linked Jazz: Revealing the relationships of the jazz community. https://linkedjazz.org/, 2013.

[15] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008. doi: 10.1057/palgrave.ivs.9500180

[16] Studio Terp. Voices that care. http://www.studioterp.nl/voices-that-care/, 2017.

[17] Women Writers Project. Women writers online lab. https://wwp.northeastern.edu/wwo/lab/.

[18] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang. Bipartite network projection and personal recommendation. *Phys. Rev. E*, 76:046115, 2007. doi: 10.1103/PhysRevE.76.046115