Overview Visualizations for Large Digitized Correspondence Collections: A Design Study

Laura Swietlicki, Pierre Cubaud

Overview visualizations for large digitized correspondence collections : a design study

Laura Swietlicki and Pierre Cubaud

Centre d'étude et de recherche en informatique et communication (CEDRIC),
Conservatoire national des arts et métiers (CNAM), Paris, France
cubaud@cnam.fr

Abstract. Overview visualization is a useful alternative to search engines in digital libraries. We describe such a tool in the context of a large correspondence collection : the Godin-Moret archive (20,000 letters). After a review of previous works, we describe our interface design and how it has been derived with the help of an online co-creation workshop. The interface is organized with a specific representation called the correspondence matrix, coordinated with more standard visualizations like tag clouds, maps and bar graphs.

Keywords: Digital libraries, Cultural Heritage, Data visualization, UX design.

1 Introduction

Data visualization is recognized nowadays as an essential tool for digital libraries. One of the most important feature of visualization is the capability to provide for the users an overview of the data under study [1]. Designing an overview interface is however a difficult task in the context of digital libraries, because each category of digitized documents (books, photos, music, software, etc.) implies specific browsing and searching practices. We study in this paper the case of correspondence collections.

Correspondence archives are a very important source for a wide spectrum of academic researches in history and sociology. Many correspondence archives have already been digitalized, generally through a joint scholar effort (see for instance a list of over 50 recent projects in [2]). However, as we shall see further, only a few of these projects have developed specific visualization tools. The *FamiliLettres* project gave us an opportunity for such a study. FamiliLettres is the ongoing publishing project for the private correspondence of Jean-Baptiste Godin (1817-1888) and Marie Moret (1840-1908), Godin's collaborator and companion. Godin was a French industrialist, creator of social experimentation inspired by Fourierism that was the Familistère, located at Guise in the north of France. The collection includes more than 20,000 copies of letters[1] archived at CNAM and the Familistère, accounting for about 3,000 correspondents. The

[1] As was customary at their time, Godin and Moret kept copies of their letters. These copies were produced using a special mechanical press, then bound together in registers and archived. The collection also includes a thousand original letters received by Godin and Moret.

digitized collection is hosted by EMAN, a national research platform for digital humanities [3]. It will be available for public use in 2023.

2 Related work

One can find in [4] a general review of works dated from 2004 to 2018 on visualization for cultural heritage (CH) collections, but only one of the 166 selected references addresses correspondences specifically (i.e. [5]). The taxonomies of usage and tasks derived by the authors, ranging from casual browsing to specialized scholar research, are however fully relevant to our subject of study. The authors also identify some important challenges for digital libraries design : we need alternatives to search tools that support accidental discovery (serendipity) and help the user's understanding of the collections. Designing "generous" overview interfaces can help to fulfill these challenges [6].

Correspondence letters do have specificities as textual documents. They are most of the time rather short, hand written, signed and dated documents, written for specific addressee(s). Formal metadata grammars have been defined: see for instance the work of the TEI Correspondence SIG [7]. The richness of these metadata impacts upon the visualization tools that must be provided for the digitalized collections. The "social network" identified by the letters needs to be represented as such. Locations of actors ask for maps. Timelines can be used for analyzing the flow of writings.

Table 1 summarizes, in chronological order of publication, the only papers we found that specifically address interactive visualization for correspondence collections. All of them are based on specific data sets, of very different sizes. The more recent works have benefited from the global progress of computer graphics. One can also note the great variety of the visualization tools, without consensus for any overview organization.

Table 1. Works on visualization for correspondence collections

Project, date and ref.	**Collection description**	**Visualizations**
Compus, 2000 [8]	100 TEI-encoded manuscript letters, XVIth century	text mining oriented, space filing, color for letter structure, bar charts
Emily Dickinson, 2006 [9]	300 XML-encoded letters , XIXth century	text mining oriented, color for letter topics, scatterplots
Mapping the republic of letters, 2008-17 [5]	Six sub projects on Franklin, Kircher, Locke, etc. ±3000 letters each ?	specific interactive views for each project : networks, bar graphs, maps (online)
Emigrant letters, 2014 [10]	4000 TEI-encoded letters, 1740-2010	multiple (uncoordinated) views : scatterplots, maps, networks, timelines
Diggers diaries, 2016 [11]	337 authors, 688 diaries and letters, 81763 pages, WW1 period	page grid, facetted view, color for letter topics, timeline, bar graphs (online)
Tudor Networks, 2019 [12]	37101 unique names, 132747 letters, XVI-XVIIth century	interactive tool for network exploration, maps (online)
Raoul Hausmann Sammlung, 2019 [13]	1400 documents (4500 images) including letters	multiple coordinated view : thumbnails within timelines, network (online)

3 Collaborative design

In order to define our own visualization overview for FamiliLettres, we choose to rely at first on co-creation sessions with its stakeholders. We have invited 8 persons with different level of implication in the project (4 historians and 4 curators). All are proficient users of digital libraries and 5 have expert knowledge of Godin's life and work. Unfortunately, meetings were forbidden at this period of time (spring 2021) due to the COVID-19 crisis in France. Two sessions were organized online with collaborative tools, organized the same day. Both had a scheduled duration of 90 min.

For the first session, the panel was asked to analyze a selection of 8 websites with overview interfaces that are representative of CH visualization's current state of art. Work was divided in 4 groups, for 2 sites each, in the order given in Table 2. The analysis was guided by a short usability questionnaire. We asked the groups to designate up to 3 noteworthy visualizations for each site, and to evaluate their relevance for FamiliLettres. The session ended with a debrief of each team and a general discussion. Participants found useful components in 7 of the 8 sites analyzed, with variations summarized in Table 2. Every group found at least one reusable visualization tool. Coordinated views within the visualizations were the most noted feature during the debrief.

The purpose of this first phase was also to fuel a creative session for our overview definition. Because of the online situation, we had to rely on whiteboards (Miro). Drawing activities of the panel were simplified to the extreme, focusing on collage of screenshots of selected elements from the sites studied in the first phase (fig. 1). Two equal groups were formed, tutored by each of us. They were instructed to test as many collage combinations as necessary in order to reach a satisfactory overview definition. The session lasted 1 hour and was followed by a collective debrief. The two groups finally produced quite different overviews, but all were organized as coordinated views using tag clouds, maps and the timeline made with thumbnails of [16]. Network visualizations were used only in one group. It was agreed in the debrief that such representation would be of minor use for the FamiliLettres archive because of the great difference of size between active and passive correspondence.

Table 2. Sites studied during the workshop. URLs are in the reference section.

Projects for each group	Noted component(s)	Relevance
Tudor Networks [12]	Direct access to the correspondents network Geographic section	maybe yes
Raoul Hausmann Sammlung [13]	Timeline with access to document detail Keywords search section Correspondance section (network)	yes definitely yes definitely yes
Reading Traces [14]	Continuous zoom from collection viz. to pages	maybe
Speculative W@nderverse [15]	Multiple coordinated views Expert and casual navigation Many stackable filters	yes definitely yes yes
Past Visions [16]	Timeline and keywords coord. with thumbnails Commentated timeline	definitely yes yes
Atlante Calvino [17]	none (but "very interesting visual impact")	no
Deutsche Digitale Bib. [18]	Timeline coordinated with network Places and sectors section	yes yes
Diggers diaries [11]	Explore by time and Explore by diary sections	yes

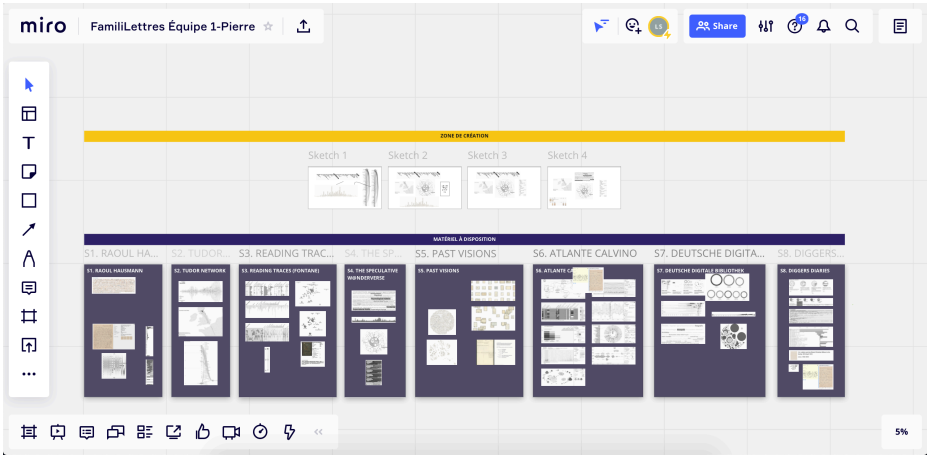

Fig. 1. Online collaborative sketching session.

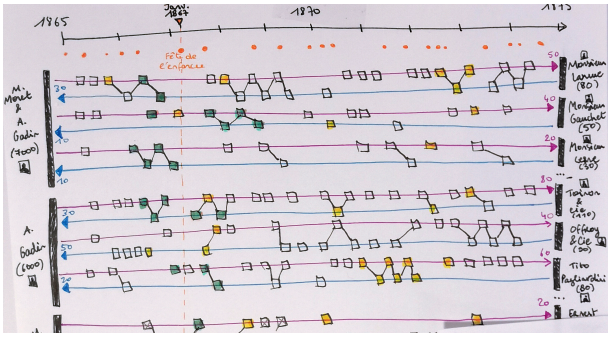

Fig. 2. Early sketch of the correspondence overview.

4 The correspondence matrix

The co-creation workshop was followed by other design iterations. We reproduce fig. 2 one of the first sketches that were produced for the overview representation. We chose to arrange the letters chronologically on an horizontal axis with their respective sender and addressee on the vertical axis. The letters are represented by a small selectable square, possibly colored to convey extra metadata. However, when applied to the Godin-Moret archive, it appeared clearly that this representation couldn't scale to the real amount of letters or would require noticeable user's scrolling. As E. Tufte eloquently demonstrated, dense representations can only be obtained through massive decrease of the "ink/data ratio" in the drawings [19]. So we investigated some more abstract representations like matrix plots.

Fig. 3 show the resulting matrix for 8836 letters sent by Godin from 1840 to 1890 (the amount of letters metadata already prepared at time of this experiment, roughly half of the collection). Time is on the horizontal axis. There are 2496 addressees on

vertical axis. The big amount of points requires sub-pixeling, therefore a standard scatterplot (fig. 3 - left) can not express, because of the overlapping of dots, situations when intense writing occurred for some adressee in a short period of time. A quick statistical investigation showed that this happens quite often in the Godin archive. The underlying model for this might be some kind of power-law distribution, since a few addressees monopolize most of Godin's attention during rather large periods of time. A jittered plot is much more efficient in this respect (fig. 3 - right), although this is considered as bad practice for distribution representation [20]. The jittering effect can be produced according to various drawing schemes. We have used here a recursive randomized algorithm where the dot is randomly pushed around its original position until a free white space is available. The "beeswarm" scheme is also very interesting, especially when zooming into a line of the matrix (fig. 4). Beeswarm uses simulated particule physics for dots positions and is implemented in many popular visualization packages, such as D3.js and Rawgraphs. It has also been used in the *Atlante Calvino* project [17].

The correspondence matrix cannot be used alone. Fig. 4 shows the general organization of the visualization interface[2]. In this wireframe view, all shapes and colors are arbitrarily defined. Currently selected items by the user are colored in purple. Following our workshop conclusions, we have completed the matrix with : 1) a list of keywords, arranged as a tag-cloud 2) a map of the addressees locations 3) a bar graph summarizing the yearly number of letters 4) idem for each adressee 5) thumbnails views of the letters. These five elements are coordinated : the impact of a user's selection within one of them is computed for the others. Selections can be repeated by the user to narrow her/his search, this can be reset. The matrix can be zoomed by the user, down to the level of each letter (fig. 5).

This design has been validated by the FamiliLettres stakeholders, but two issues remain to be clarified until the end of the project. How can we represent the letters having unknown date or addressee? For now, they are displayed in a specific column on the left of the matrix. Should we integrate further the visualizations of passive and active correspondence? At present, they are only considered as separated views, triggered by switches.

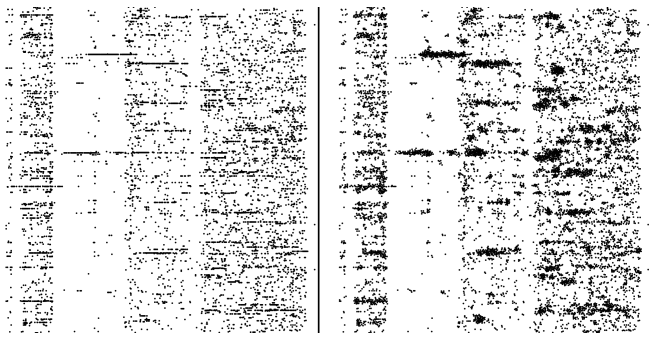

Fig. 3. Correspondence matrix. Left : standard dot plot. Right : jittered plot, revealing activity.

[2] A video mockup is also available at https://youtu.be/_130fXdQ35k

Fig. 4. The overview interface (wireframe). Left : tag-cloud for the collection's keywords and map of letters destinations. Center : correspondence matrix. Right : letters thumbnails.

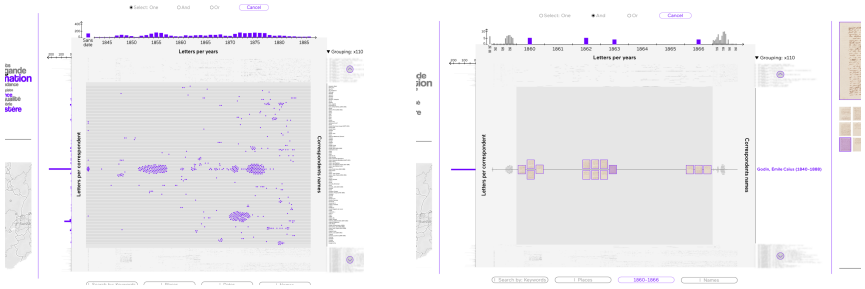

Fig. 5. Zoom in the correspondence matrix. Left : the first level of zooming also reveals a group of correspondents names. Right : zooming further for direct access to the letters thumbnails.

5 Conclusion

We have presented in this paper a novel overview visualization for large correspondence collections. Although specifically designed for FamiliLettres, the interface can certainly be used for the visualization of other collections with possibly bigger dimensions. Before that, the wireframe design that we presented must be implemented as a user interface. Since EMAN relies on Omeka, we plan to prepare a specific Omeka plugin that would be re-usable for the other large correspondence projects hosted by EMAN. Data from other open collections (as in [12]) could also be used to test, via some controlled experiment, our hypothesis concerning the patterns that are revealed by the matrix view, and its effectiveness for the users understanding of the collections.

This work has been partly funded by the GIS CollEx-Persée.

References

1. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In Proc. 1996 IEEE Symposium on Visual Languages, 336-343 (1996).
2. Walter, R.: L'édition numérique de correspondances : guide méthodologique. Consortium Cahier. 2018. https://cahier.hypotheses.org/guide-correspondance
3. EMAN, https://eman-archives.org/
4. Windhager, F., Federico, P., Schreder, G., Glinka, K., Dork, M., Miksch, S., Mayr, E.: Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Trans. Visual. Comput. Graphics. 25, 2311–2330 (2019).
5. Mapping the republic of letters, https://republicofletters.stanford.edu/
6. Whitelaw, M.: Generous Interfaces for Digital Cultural Collections. Digital Humanities Quarterly (DHQ), vol. 9, no. 1 (2015).
7. Stadler, P., Illetschko, M., Seifert, S.: Towards a Model for Encoding Correspondence in the TEI: Developing and Implementing <correspDesc>. Journal of the Text Encoding Initiative. Issue 9, Sept. 2016 - Dec. 2017.
8. Fekete, J.-D., Dufournaud, N.: Compus: visualization and analysis of structured documents for understanding social life in the 16th century. In Proc. of the fifth ACM conference on Digital libraries - DL'00, 47–55 (2000).
9. Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M.G., Smith, M.N., Clement, T., Lord, G.: Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In Proc. of the 6th ACM/IEEE-CS joint conference on Digital libraries - JCDL'06, 141-150 (2006)
10. Moreton, E., O'Leary, N., O'Sullivan, P.: Visualising the Emigrant Letter. Revue européenne des migrations internationales. 30, 49–69 (2014).
11. Nualart Vilaplana, J., Pérez-Montoro, M.: Diggersdiaries: Using text analysis to support exploration and reading in a large document collection. In Proc. Eurographics Conf. on Visualization (EuroVis), Posters Track (2017) http://diggersdiaries.org
12. Ahnert, R. Metadata, Surveillance and the Tudor State. History Workshop Journal, 87, 27 - 51. (2019) http://tudornetworks.net/
13. Bludau, M.J., Dörk, M., Heidmann, F. Relational Perspectives as Situated Visualizations of Art Collections. In Proc. ADHO Conf. on Digital Humanities (DH'2019). https://uclab.fh-potsdam.de/hausmann/
14. Bludau, M.-J., Brüggemann, V., Busch, A. and Dörk, M. Reading Traces: Scalable Exploration in Elastic Visualizations of Cultural Heritage Data. Computer Graphics Forum, 39: 77-87. (2020). https://uclab.fh-potsdam.de/ff/
15. Hinrichs, U., Forlini, S., Moynihan, B. Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections. IEEE Transactions on Visualization and Computer Graphics. 22. 1-1. (2015). https://vimeo.com/236169311
16. Glinka, K., Pietsch, C., & Dörk, M. Past Visions and Reconciling Views: Visualizing Time, Texture and Themes in Cultural Collections. Digit. Humanit. Quat., 11. (2017). https://uclab.fh-potsdam.de/fw4/en/
17. Atlante Calvino: literature and visualization, 2020. http://atlantecalvino.unige.ch/?lang=en
18. Dörk, M., Pietsch, C., Credico, G.: One view is not enough: High-level visualizations of a large cultural collection. Information Design Journal. 23. 39-47. (2017) https://uclab.fh-potsdam.de/ddb/
19. Tufte, E.: The visual display of quantitative information. Graphics Press, 2001.
20. Wilkinson, L.: Dot plots. The American Statistician. 53-3, 276-281. Aug. 1999.