

White Paper Report

Report ID: 98438

Application Number: HJ5005610

Project Director: Dan Edelstein (danedels@stanford.edu)

Institution: Stanford University

Reporting Period: 1/1/2010-3/31/2011

Report Due: 6/30/2011

Date Submitted: 8/30/2011

NEH WHITE PAPER

White Paper

GRANTHFJ00018976

Digging Into the Enlightenment: Mapping the Republic of Letters

Dan Edelstein and Paula Findlen

Stanford University

August 29, 2011

Project Narrative

When the Mapping the Republic of Letters project began in 2008, our goals and research questions were fairly straightforward. With advice and technical support from computer scientists and digital artists, we hoped to gain insight into the scholarly networks of the early modern era. Humanities scholars unanimously acknowledge that scholars in the early-modern period had extensive correspondence networks and these networks formed an important part of what we call the Republic of Letters. However, the vast amounts of data also made it clear to us that the normal interpretive toolsets of a humanities scholar could not sufficiently grapple with the big questions: What did these networks look like? How geographically extensive were they in reality? Did these networks connect or overlap? Was there any evolution in the configuration of scholarly networks over time, from the beginnings of the Republic of Letters in the Renaissance to its flourishing in the Enlightenment?

Approaching these questions required confronting the challenges of our data. Gathering data about correspondences is a process that will probably never be entirely completed. Additionally, for every letter captured, we found that we had access to varying and inconsistent metadata. Thankfully, others had already begun to undertake these difficult tasks. We found an invaluable partner in Robert McNamee, who directs the Electronic Enlightenment Project (EEP) at Oxford, and who subsequently became our English collaborator for the Digging into Data grant. EEP had digitized more than a dozen correspondences from the eighteenth century (using modern editions) and structured the metadata for roughly 50,000 letters in an exemplary fashion. In addition to capturing biographical and chronological metadata, they had also recorded (when available) location information for both sender and recipient. With EEP as a partner, geo-coding and mapping this kind of data became a realistic endeavor.

Initially, Stanford computer-science graduate students produced our demo visualization (produced in Adobe Flex), which allowed us to map the EEP metadata. Although this initial system was not nearly perfect, it helped us establish a list of functionality requirements, which have been essential moving forward. One of the first discoveries was actually not what we visualized, but what we could not visualize. There are about 19,000 letters in Voltaire's correspondence, but we are only able to map about 13% of them. The rest simply don't appear because our data is so incomplete: we do not have geographical or chronological metadata for the vast majority of the letters in these correspondences. From a visual perspective, this meant that we need to indicate more clearly what was not being seen and we needed to develop a visual language for representing data that is uncertain. Visualizing degrees of uncertainty has since become a major research question, springing from our initial visualizations.

In many respects, these initial visualizations only skimmed the surface of what our data could reveal. One of the most attractive features of our data is its multidimensionality. We not only have a significant amount of metadata for each letter, in addition –thanks to the painstaking work of EEP – we have a rich collection of metadata for each correspondent (such as gender, nationality, religion, occupation, and social status). We wanted to incorporate this data into our visualizations. However, dealing with multidimensional data can be very difficult. A new collaboration between the Stanford and Oklahoma teams was essential to broaching these new issues.

The recently established field of visual analytics aims to transform craft into science by casting a wider eye on the entire process of engaging data visually. The field is domain-agnostic, interested in applications to virtually any sufficiently rich data set of interest to scholars, scientists, students, and even the general public (such as social networks). In fact, many visualization techniques that are useful in one domain are also useful in others. Applications of visualization techniques to business logistics and intelligence analysis can often translate to humanities scholarship because they all are driven by a desire to understand complex structures and behaviors of individuals and groups across multiple data dimensions.

The goal of visualization design is to find that combination of graphical representations and interactive capabilities that allows a data explorer/analyst to pursue an open-ended inquiry about the system behind their data. The challenges of visualization design are manifold. To be relevant, the explorer/analyst must be able to translate their desired human-form questions into machine-form queries. To be usable, expressing queries must be fast and easy. To be useful, it must be possible to express complex queries about space, time, categories, names, quantities, and combinations of these. To be important, the visual results of queries and the overall process of visual inquiry must be interpretable.

In these challenges lies the central technical research goal of our collaboration: to initiate a highly interactive and computationally augmented visual externalization of the open-ended and fundamentally interpretive reasoning processes at the core of humanities research. Externalization of interpretation is the key idea here. Visualization is certainly important if the results of queries can be understood and interpreted, and those interpretations shared with others by traditional, including digital, means. However, this is only a portion of visualization's potential. We propose that its fuller potential is in helping explorer or analysts to realize their interpretations of data by folding them back into visualizations themselves.

The Mapping the Republic of Letters Project has provided mutual research benefit for all collaborators. The humanities scholars have gained new tools for exploring data, and the computer scientists interact with enthusiastic users (and their absorbing data sets) in order to evaluate new visualization approaches. Together we are designing a variety of interactive visualization tools for exploring and analyzing the EEP metadata. We are also developing a new architecture for

designing visualizations with rich annotation capabilities based on what we've learned about how people want to interact with the EEP metadata visually. (It should be noted here that different agencies have funded us at different start points and for different durations throughout this project. The computer science side of the project is currently at eleven months out of two years.)

We realized that to continue this exchange and extend it beyond the walls of our own lab, we would need to design interactive visualizations that can be shared via the Web and that are easily understandable, particularly for our more technophobic colleagues. To do this we needed visual tools accessible to the casual user, not only for the analyst. While *Improvise* provides us with rich analytic tools in the form of a desktop application, we could not translate it to a Web environment for easy distribution and we could not expect colleagues unfamiliar with our data and data visualization techniques to grasp the complexity of the *Improvise* interface.

To make visualizations that can be shared on the Web in the most open format possible we have looked to new research from the Stanford Visualization Group in the Department of Computer Science. As an alternative to Adobe Flash, or the more sophisticated Flex package, we have been using *Protovis*, *D3* and, for map-making, *Polymaps*. One advantage of these three tools over the very powerful Adobe tools is that with them we can create animated or interactive visualizations that work in the browser without requiring any special proprietary plugins. To require our colleagues to download and install a special application or be faced with a blank web page is just too great of an obstacle to our goal of broad access. Another advantage of non-commercial tools is that, unlike Adobe Flash, they can be viewed on the mobile devices (e.g., smart phones and tablet computers) that are becoming ubiquitous on campus among both students and faculty.

To make visualizations that are immediately understandable to colleagues unfamiliar with visual representations of data and analytic tools, we turned to the expertise of Density Design Research Lab (*Density*). *Density* is a group of researchers from the Communication Department at the Polytechnical Institute in Milan. Their research focus is the graphic representation of complex phenomena. We initially worked with them to develop a tool to explore Athanasius Kircher's complex correspondence network. *Density* proposed a Sankey flow diagram to show the proportional distribution of letters, allowing us to compare a number of different factors, including the sender's location, religion, social status, expertise, gender, and community of practice as well as the language of the letter.

Launched in spring 2011, the Athanasius Project has built open-source software that allows for visual browsing of very large, heterogeneous data sets of seventeenth- and eighteenth-century correspondence, people, travel, events, places, and publications. This visual browser will be made available to all intellectual historians and literary scholars who are exploring the emergence of ideas in the early modern period. The Athanasius Project uses our shared code base for tools on Github, which allows for efficient, collaborative development of software. Each of the developments described above has proceeded as part of both the Athanasius

Project and the Mapping the Republic of Letters project. The Athanasius Project has helped us to clarify and resolve some of the problems inherent to comparing historical data from a variety of projects

As we look forward to our next phases, we have two primary objectives. First, we are convinced that elegant and intuitive design, along with practical functionality, are the keys to guaranteeing that our visualizations will reach a broad audience and have a strong impact on the traditional humanistic fields from which most of us come. To this end, we will continue to partner with Density Design on a number of other visualizations that we are planning. Secondly, while we already have more than enough data to fuel our individual research projects, we want to make our visualizations available to as many other projects as possible. Unfortunately, this is not as simple as making our toolset open access. The data specifications are extremely precise, and it can be extremely difficult to format new data to work with our visualizations. We are tackling this larger problem by partnering with another Oxford project, “Cultures of Knowledge,” and a Dutch project, “Circulation of Knowledge,” to create a larger visualization and data processing platform for early-modern scholarly metadata. Dubbed SCHOLA (for “Scholarly Online Archive”), this platform will consist primarily of a metadata catalogue, which will be run by the Oxford team, a text-mining operation, which our Dutch collaborators are setting up, and Stanford’s visualization tools. Future projects wishing to use our tools will first have to contribute their metadata to the Oxford catalogue; at that stage, data specifications will be checked, ambiguities and redundancies will be resolved, and the metadata will be parsed in a standardized fashion. We can then design our visualizations so that they conform to the metadata structure of the Oxford catalogue, thereby ensuring cross-compatibility. This arrangement, moreover, should incite projects to pool their metadata, which in itself will be a great accomplishment.

Explanation of Tools Developed

Our demo-visualization, produced in AdobeFlex, allowed us to represent correspondence volume through the width of lines connecting cities allows users to immediately identify the major highways of communication. The visualization could be refined using a sliding timescale that lets users focus on a particular period. Given that most of the humanists working on this project are historians, it was essential that our ensure that our visualizations preserved a temporal dimension as of inquiry. This need led us to explore how correspondence volume shifts over time, so the tool can graph this relation and combine it with the timescale slider. Another useful trait of this demo-visualization is the ability to view directionality. We know at a glance, for example, that while we have approximately 16,000 of the letters Voltaire sent, we have only around 3,000 letters that he received. Finally, an additional feature of this visualization is the ability to compare networks that may or may not have occupied the same times and places. For example, John Locke’s correspondence is markedly Anglo-Dutch (and tellingly colonial), whereas Voltaire’s correspondence is far more continental (with fewer overseas contacts).

For our practice of using visualizations as an inquiry or search tool needed a we have adopted the term *ampliation* (from the Latin *ampliare*, meaning “adding to that which is already known”) to evoke the notion of interpretation-driven extension of data through visual interaction. Such rich, natural annotation in situ will allow users to correct data, add commentary, and incorporate markers for instances or groups of locations, dates, names, and other data types. Annotation of the associations between dimensions will enable recording of hypotheses and interpretations as parallel data, making it possible for individual and collaborating scholars to revisit, share, explore, and analyze interpretive histories in the very visualization tools used to do these things with “primary” data.

Our more recent visualization tools produce tables, maps, and histograms. One table (“Letter Recipients”) shows the number of letters each letter recipient received. A map (“Cities”) then shows straight paths from letter source cities to destination cities. In instances in where letter path information is incomplete, the available source or destination information is indicated by the city point alone (unconnected to a path). The color of the path indicates the number of letters that traveled between the connected cities. Checkboxes above the map let the user restrict the visible paths to include only the letters sent to a particular individual or group of individuals, written by senders of a particular nationality or set of nationalities, or sent from a particular source country. Another table (“Letters Sent by Nationality”) shows the number of letters written by people of a particular nationality. A checkbox at the top allows the user to restrict the letters shown to only those received by the recipient selected in the “Letter Recipients” table.

We’ve also built in a table (“Letters Sent by Source Country”) that shows the number of letters sent from particular countries. As with the previous table, a checkbox at the top allows the user to restrict the letters shown to only those received by the recipient selected in the “Letter Recipients” table.

We use a two-dimensional histogram (“Sender Nationality or Letter Source Country by Year”) to show the changes in source country and sender nationality of letters over time. A bivariate color scheme shows the number of letters of selected source countries (green) relative to the number of letters of selected nationalities (purple). A checkbox in the upper right allows the user to control whether the histogram displays source country information or nationality information. Ovals indicate years in which source country information is missing (when nationality patterns are being displayed) or in which nationality information is missing (when source country patterns are being displayed). The users can pan-and-zoom interactively to view the entire timeline all at once, getting a high-level perspective, or to view a smaller area to focus on a subset of years and source countries or nationalities.

The table “Letters” is always available to display the letter metadata. Checkboxes at the top allow the user to restrict visible letters to those selected in the other three tables or the histogram. Clicking a row in the table opens a browser window that shows the results of searching the Electronic Enlightenment for that

letter or similar letters (such as letters written in the same language or the same year).

The Athanasius Project is the visualization toolset we have been working on most recently with Density. This new toolset is open source and freely available for download. Athanasius combines the visual filtering method with the selection, comparison and diachronic features of the original demo prototype. The interactive features of this new visualization provide scholars much greater flexibility in exploring the data. One innovation is a set of tools we call “Basic Stats.” With the Basic Stats you can see the general shape of the data based on dates, location information, and biographical details. The Basic Stats view shows a tremendous amount of data at different scales, depicting not only the data we have, but also where our data are incomplete. The calendar view, for example, combines a 300-year snapshot of the data with a day-to-day view of the volume of correspondence over a period from 1500-1800. By showing the letters by day and month, seasonal patterns are revealed from a distance, but we can select any number of days to see a list of the individual letters as well.

We have also added to the Anthanius Project toolset the ability to choose the point of entry for analysis of a data set; you can begin with all the data viewable or you can start with a blank canvas and add data. Wherever you choose to begin, you can make use of a visual filtering technique that allows you to see, at a glance, the breakdown of the data before making a selection. Then you can directly select, with a mouseclick, the subset of data to be visualized. For example, if we are looking at correspondence we can immediately see the percentage of letters that are Voltaire’s correspondence. If we select that set, we can see how many he sent and how many he received. From that point you can choose to see only received letters, etc.

One of the limitations of evaluating the character of a correspondence network with a geographically-based is that the axial or radial pattern is distorted by the movements over time of the primary author. With our new visualizations, we can move between a geographic view and a interpretive view that puts Voltaire at the center, with the lines of correspondence fanning out to points that are either individuals, individuals grouped by nationality, nations, cities, or letters grouped by distance from Voltaire. This allows us to explore many more dimensions of a correspondence collection, while also addressing, in part the problem of incomplete data. Though we may not know the source location of 1,000 letters in Voltaire’s correspondence, we can see how many of those letters were written by Englishmen.

In addition to standard two-dimensional representations of our data, we have begun to experiment with three-dimensional visualizations. To do this, we have shifted our frame of reference from maps to globes. Although we are still only in the initial stages of three dimensional visualizations, it has already been impressive to watch as the topography of the Republic of Letters emerges from the globe.

The toolsets of Athanasius rely on our new, web-friendly data storage system that uses Mongo DB. Mongo DB is a schema-less, document-oriented database. Instead of using pre-structured tables for data, it stores data in the form of documents (using JSON format). In addition it does not force us to adopt a unique schema for all the documents we store. Mongo enables us to access all of our data and move beyond the complications of incomplete and sometimes incongruent data sets.

Since the Athanasius Project draws from all Mapping the Republic of Letters datasets, we have compiled a full bibliography of citations for each of the individual projects. After consulting with several experts, we determined that the needs of our project were not met by any preexisting citation formats. Our full bibliography captures spreadsheets that we created and relates them in a parent-child hierarchy to the sources from which they were assembled. This format acknowledges the work of individual data compilers as well as the authors of published collections. Our bibliography is built into the visualization toolset of Athansius and is also available on our updated websites to any scholar interested in the project.

Successes and Lessons

Although we have alluded in the project narrative to some of our successes and lessons learned, we would like to explain the value of these revelations and accomplishments in greater detail. Our initial visualization, which allowed us to establish some basic functionality guidelines, also taught us that visualization could also serve as a kind of search form. Where standard search forms lead you “from nothing to something,” that is, from a blank series of fields to a set of results, based on an inputted query, cartographic visualizations proceed in the opposite direction, namely “from everything to something,” or from a representation of all the data to a selection. This selection can be made through filtering, but there are also advantages to just showing everything, and letting the visualization direct the user’s eye.

To provide an example, we noticed that we were constantly drawn to more peripheral letters, those sent to distant, exotic places (e.g., Voltaire’s letter to Kazakhstan), or to less commonly frequented countries (e.g., in Eastern or Northern Europe). These letters may not be representative of the broader correspondence, but they can still be of individual interest. The visualization, in this regard, can serve a heuristic purpose, leading the user toward less known corners of the dataset, or even prompting new research questions (e.g., how *did* one send a letter to Kazakhstan in the 18th c. century?). The visualization literally serves as a search engine, since by highlighting a correspondence pathway, you can click back to the full text of the letters in question in the EEP database. This is a feature that our EEP collaborators were most excited about and which they are planning on incorporating in their site. The potential for this kind of search functionality becomes even more exciting when imagining a visualization that draws on a broad range of datasets. In this instance, the visualization becomes not only a finder, but also a means of searching across distinct corpora.

While inherent and contingent problems with our dataset have prevented us from definitively visualizing the big picture of the Republic of Letters, we nonetheless found that the ability to map individual correspondences has been extremely productive. Even a basic, synchronic snapshot view yields a great amount of information. For example, it immediately draws the eye not only to hotspots of correspondence, but also to “coldspots,” or areas with little or no correspondence. Identifying such coldspots can have important scholarly ramifications. Dan Edelstein commented that his first reaction to seeing Voltaire’s correspondence mapped was to wonder why there were not more letters to and from England. Voltaire’s time in England is usually taken as his formative experience in becoming a *philosophe*, and more generally, the Anglo-French vector is traditionally seen as central conduit of the Enlightenment. However, there are hardly any letters to or from England – less than 1% of Voltaire’s correspondence, as it turns out. This simple observation is obvious when you look at a map, but it lay hidden from view in the hundred-plus volumes of his correspondence. In itself, this absence requires interpretation: the lack of correspondence alone does not disprove the common views about Voltaire. But with this new information in hand, we could go back and read his comments about England as well as his few English letters, and discovered that England (and the English) played a much more marginal role in Voltaire’s conception of history and philosophy than is usually thought.

Visualizing early modern correspondence has also helped us to question the nature and function of a “global network” in this period. Caroline Winterer has argued that even a figure whose correspondence has a global appearance, such as Benjamin Franklin, is in fact more cosmopolitan than global. Franklin writes predominantly to (and from) only three countries: the American colonies, and subsequently, States; Britain; and France. The flurry of letters across the Atlantic give the correspondence a global feel, but, Winterer notes, it actually takes just as long for a letter to go from London to Edinburgh as it does for London to Philadelphia. The scarcity of truly global correspondence networks raises the question of whether there might be necessary preconditions for developing one. Paula Findlen has suggested that such a correspondence network may need the support of a global trade network. Simon Schaffer has shown, for instance, how the data points for Newton’s *Principia* were taken from ports and other places along major British trade routes. Bentham, for his part, was writing in a new age of colonialism; whereas for Kircher, his network rested on the pre-existing network of Jesuit missionaries. We mention these hypotheses less as findings of our project, and more as examples of how visualizations have triggered new and important research questions.

Working with data visualizations for this project has not also generated new interdisciplinary research questions, it has also changed the way that all parties involved do research. For one thing, formulating humanities research questions in terms of machine-readable queries required us to think systematically, across our areas of expertise and our individual research interests, and come up with broad, generalizable concepts. Humanistic inquiry—by contrast— is freeform, fluid, and

NEH WHITE PAPER

exploratory; not easily translatable into a computationally reproducible set of actions. Much of our work throughout this project has been bridging this cultural divide: learning more about each other's underlying assumptions in order to communicate the needs of humanistic research within the constraints of visual analytics.

On the other hand, having data-driven images at the center of this project has brought faculty and student scholars working in different fields, with somewhat different disciplinary assumptions, together into dialogue far more easily and immediately than circulating a paper. When our Stanford-Berkeley research team meets and looks at the same image: the exchange of letters; the movements of travelers on the grand tour; or a comparison between 17th and 18th century libraries, that image becomes the focal point and catalyst for discussion. Winterer pointed out how, as an intellectual historian working on early America, she rarely has the opportunity to talk at this level with fellow historians working on Europe, but now we can bring visiting scholars into our lab for a few minutes and spark a lively discussion that cuts across individual research interests. This Digging Into Data project has brought us out of our narrowly defined academic fields and led us to engage in a larger, fundamentally cross-disciplinary conversation about history and the nature of historical data.