Visualization and Literary Analysis

Katherine Faull (Bucknell University)

openpublishing.psu.edu/textanalysis/visualization

Within the field of literary analysis, visualization is nothing new. Whenever we read, we visualize fictional worlds; we imagine, for example, what the characters might look like, put ourselves in their place, build up scenes in our heads from the descriptive passages given us by the author. One of the major distributors of "[digital] books on tape" uses precisely this function of literature to invite potential customers to download audio files and "make the story come alive." As readers (and, increasingly, listeners), we are asked to visualize the character's point-of-view, see through her or his eyes; and we could perhaps argue that this is one of the most powerful aspects of literature—its ability to ask the reader to suspend her disbelief and make the imagined world real.

Within the field of DH, visualization can mean several things. We can, for example, find digital means to visualize a fiction world, much in the way described above. We can make virtual worlds that correspond to the descriptive passages in a work. We can spatialize a character's journey (where, for example, does Clarissa Dalloway walk through Bloomsbury on the day of her party? What path does Goethe take on his "Italienische Reise"?) We can spatialize the plot of a novel. We can ask, what role do landscape, geography, and place play in the action of a novel or drama? And indeed, the earliest ventures into thinking about visualization and literature are not at all digital, but rather focussed on both the graphical rendering of plot and character and also the extraction of metadata from collections of documents.[1]

We might well ask, along with Franco Moretti, what such visualizations of place, plot, or character add to the interpretive act. Questions such as, "what specific form of knowledge does visualization reveal?" or, "what do literary maps *do*?" are crucial to ask as we enter into the work of visualization. And we must also ask ourselves, how we avoid running the risk of reducing the infinite world of the literary imagination to the quantifiable and positivistic representation of an inferred certainty in order to render it visually. For example, if we map Clarissa Dalloway's movements through London, what effect does this have on our interpretation of the literary text? What is the heuristic value? Does the visualization produce new knowledge or is it a barrier to further interpretation?

These are criticisms that have certainly been levelled against the ideas of, for example, Franco Moretti, whose revolutionary work in the last 15 years has challenged literary scholars to rethink the way in which they assume the stability of the "canon" when faced with analyses emanating from making visual the "big data" of publishing history.[2] Fending off accusations of positivism and formalism, DH scholars who have successfully used visualization in their analyses of literature argue for the new hermeneutic of digital visualization, the way in which their visualizations produce both new knowledge and also invite ambiguity, the traditional province of humanistic critical thinking. Johanna Drucker, in her most recent works, has made a strong, deep, and powerful case for scholars in DH to recognize that a) the humanities has always produced visualizations of knowledge systems, and b) that when we visualize digitally that we should follow principles, long established in design circles, that rest upon concepts of depth, color, and line.[3] Manuel Lima, the Portuguese-born designer and thinker, invites us to revisit Edward Tufte's principles of design in the digital age through his elegant and scholarly works, in which he, like Drucker, argues for the experience of visualization as itself a producer of

knowledge: the phenomenological becomes epistemological.[4] Whether we are using pen and ink or a computer to produce visualizations, these principles are the same.

What to Visualize?

Bearing this in mind, we might then well ask, what is it that we want to visualize within literary analysis? What should we not attempt to visualize? What is our research question that can be best addressed through visualization? And, how does visualization add to or challenge a traditional literary analysis? In this short introduction to visualization and literary analysis, I draw on illustrative examples from my own work with a large corpus of manuscript materials that come from the Moravian Church (also known in Germany as the Herrnhuter). These materials are in multiple languages; however, I deal primarily with German and English language texts.

Literary Language

The analysis of language is a traditional act of literary analysis. And the digital visualization of lexical patterns within literary works can be an integral part of the "multiple lenses" of reading that Tanya Clement has discussed.[5] We might ask what the most frequent words are in a corpus, how do they differ between works in an author's oeuvre, what kind of lexicon do different characters use in a work, how does translation change the lexical patterns of the source text, is there a significant difference in the lexica used by male and female authors, how does word use reveal patterns of class, race, and gender privilege?

There are user-friendly web-based platforms, such as Voyant,[6] that will allow the uploading of a corpus and speedy first-pass visualizations that are based on word frequencies,

such as would be found in a traditional concordance, KWIC (keyword in context) searches, frequency patterns across a work and a corpus by means of radio lines, and a quasi-network visualization ("Links") that reveals the connectedness of terms within the corpus. Stop words are automatically excluded from the visualizations but can be modified and ignored. The enormous advantage of the Voyant platform is that it allows for the textual analysis of non-Latin, Unicode UTF-8 alphabets.

In my own work , I have enhanced the initial insights afforded by Voyant through the use of the more statistically-based tools, such as "Antconc"[7] that will visualize (albeit less beautifully) KWIC searches, clusters, and "keyness" measures. Running my corpus of women's memoirs through other platforms has allowed for the visualization of topics (Mallet), style (Lexos), sentiment (Jigsaw) and, in poetry, sound (Poemage).[8]

<insert Fig. 3.1 near here>

Comment [Office1]: correct

Again, each of these web based platforms, allows for the "algorithmic" reading of literary language, and serves as an entry point into the field of computational and corpus-based linguistic analysis of literary texts. However, without an understanding of how language works, what the pragmatic functions of speech are, whether fictional or non-fictional, the visualization of literary language can seduce the researcher into looking only at the phenomenon of the image and not ask, what this might really mean. We must always keep in mind, in what way do these methods enhance, challenge, make more complex, the original research question(s). For my own research, for example, running contrastive analyses of memoirs written by men and women has

allowed me to test whether gender is a differentiating factor in the composition of a life story. Similarly, I have been able to run a similar analysis for memoirs by Africans and non-Africans.

Spatial Analysis

The "spatial turn" over the last ten years within the humanities has produced exciting, challenging and provocative questions about the literary canon and the role of place and spatial knowledge within the literary imagination. Again, the question of space is not a new one. Scholars have long looked at the question of the "where" in literature. We classify groups of writers by geographical locations: the British Romantics, the Jena Romantics, the Harlem Renaissance, for example, belying a perhaps implicit consciousness that a binding factor in literary production is the physical environment in which we find ourselves.

We have moved far beyond the realm of nineteenth century literary criticism, however, which is rife with essentialist readings of writers and place. Within DH, spatial analysis focuses on questions not of identity derived from place, but rather on what new knowledge about a literary text or texts can be derived from a sophisticated analysis of a work in three or four dimensions. The kinds of research questions that might be asked that require spatial analysis might be; how does the movement of characters in a text reveal, challenge, or underscore the action of the work? Does the spatialization of the action reveal significant groupings of character or action? Can we reveal spatial segregations of plot lines, characters, that are otherwise difficult to elicit from the text? Can a spatial analysis invite the reader further into the literary imagination? To quote Moretti again, "Do maps add anything to our knowledge of literature?" (Moretti 2005, 35)

What is Spatial Data in Literature?

There is a variety of ways in which we can extract spatial data from a literary text and then visualize it. Again, the methods we choose depends very much on what question we are asking. The most obvious place to start is to identify literary locations as actual geographical locations and then map them using a coordinate system. In projects such as the *Map of Early Modern London*[9] (see later in this volume) an historical map provides a visual spatial interface to, among other things, a searchable gazetteer of people and places that appear in early modern texts. Or we can take biographical data and spatialize it on a map, to explore the possible shared locations of writers, places about which they wrote, and thus produce a literary landscape/cityscape.[10] Metadata associated with collections of literary works can also be mapped (adding the fourth dimension of time) such as in *The Book Biography Machine* from Harvard's MetaLab.[11]

These projects sometimes consist of drawing on very large collections of texts, pulling the geographical information out of them and mapping them onto a base layer that is either a digital map or digitized manuscript map. In a well conceptualized visual interface, the spatialization of large amounts of data can reveal patterns at the macro, meso and micro level and allow the user to "drill down" into the source data. Visualization produces new knowledge not just as a static representation of data points but rather as an interactive and shapeable interface where the reader/user can make interpretive choices about metrics of analysis.

In my own work with a team of historians and literary studies faculty from the University of Gothenburg, Sweden, we have taken this approach as a first step in visualizing the metadata from 60,000 Moravian memoirs.[12] Starting with a subset of memoirs that formed the basis of an earlier work[13], dates and places of birth and death were extracted from much larger digital files in the Unity Archives in Herrnhut, Germany and Bethlehem, Pennsylvania in the US and then

Deleted: 2

Deleted: the

visualized on a basemap. Users can choose to visualize places or movements of either men or women, or both over a specific time period. Through clicking on the places, users can gain access to the specific record. Users can also download JSON file of the results, should they wish to make their own analyses. Future steps in the project will allow access to the original memoir manuscript and provide contextual information.

Deleted: "drill down"

<Insert Fig. 3.2 near here>

Comment [Office3]: yes

Deleted: 1

However, on another level, the spatial visualization of literature is not only about extracting geolocations from a text, a collections of texts, or a large database of associated metadata. The complex and suggestive work of the geographer Anne Knowles (and her students) that focuses on experiences of the Holocaust, models questions of spatial experience for the humanities scholar that once again take us back to the provenance of the literary imagination. Questions she and her students ask focus on the experience of space and the emotional inflection of that experience. For example, drawing on the field of critical cartography, Knowles (and other scholars in her field, such as Margaret Pearce and Todd Presner) deconstructs the spatial boundedness of the Cartesian co-ordinate system and visualizes the subjective point of view. Projects such as "Geographies of the Holocaust" rely on what Knowles has termed "inductive visualization", a method that she has found to be more suitable "to the acutely perceived but imprecise, often highly relational spatial content in the kinds of sources humanists rely on."[14] Knowles' prizewinning work invites literary scholars to rethink the representation of spatial experience in a rich, suggestive manner.

Another geographer whose work has been highly influential in the realm of literary GIS (Geographic Information Systems) is Ian Gregory, whose methodology for the mapping of emotions provides a more quantitative approach than Knowles' inductive methods.[15] In Gregory's work we see the adaptation of classic cartographic methods (such as chloropleth density analysis) applied not to social scientific data, such as population density or income level, but rather to data extracted using the methods of corpus linguistics in order to measure the intensity of emotion associated with, for example, place. Working with historical accounts of the Susquehanna River in the mid-18th century, students Steffany Meredyk and Bethany Dunn extracted quotations and locations in order to visualize the sense of fear that pervaded the area. Using scale as a means to represent intensity, their resultant "inductive visualization" was begun in Arc GIS and then finally rendered in Adobe Illustrator.

<insert Fig. 3.3 near here>[16]

Comment [Office4]: correct

The spatial visualization of literature (whichever method is used) can sometimes be aligned with the closest of close readings. Whether creating <placeName> tags in TEI, geo-locating on a coordinate system places that occur in texts, or devising a sophisticated and suggestive signification system for the experience of place within a text or set of texts, spatial visualization is, at its best, labor intensive. However, as Knowles' Holocaust project shows, its results can produce stunning new knowledge about the human experience of space.[17]

Sentiment

Related to the above, the visualization of sentiment or emotion within a literary text is a complex form of analysis, not least because of the way in which emotions are aroused, experienced and represented. We could argue that within the Western canon, since at least the writings of Aristotle, the function of literature in all its forms has been to emote. Whether in order to effect catharsis, to represent historical causality, to morally instruct, to express love, to make people laugh, the reading of literary texts has an emotional effect upon the reader. So, when analysing and visualizing sentiment, some clear decisions must be made. Whose sentiment is being analysed? the author's? the narrator's? the characters'? the reader's? And, if we identify the subject (or experiencer) of emotion or sentiment, we also need to separate this from the emotion experienced by that subject, and also the object that arouses the emotion in the subject.

Some sentiment analysis algorithms simply identify terms in a text that are then placed on a scale from -1 to +1, where -1 is negative or angry and +1 is positive or happy. For example, within some entity recognition software, a text is defined as either angry or happy based on the relative frequency of negative or positive sentiment words and then visualized using color intensity (usually red for angry and blue for happy). Such analyses need to be accompanied with KWIC searches in order to identify the presence of negative particles, such as "not", in order not to return false results.

Deleted: (Jigsaw)

<insert Fig. 3.4 near here>

Comment [Office5]: correct

However, this type of sentiment analysis and visualization does not account for either the subject or the object of emotion. Other projects have attempted to map emotional responses (see

Knowles and Gregory cited above) and represent (geo)graphically the link between emotional states of the observing subject and the latter's spatial location.

As the literature on the history of emotions tells us, not all historical periods and cultures interpret and represent emotion the same way. Thus adding a temporal variant to sentiment analysis either within the narrated time or narrative time of a single text, or across literary periods could reveal new cross-cultural insights about how humans experience and represent emotions.

Timelines

In their work on the history of representing time, Daniel Rosenberg and Anthony Grafton describe geography and chronology as the "double eyes" of history.[18] The visualization of change over time, revolutionized in the 18th century by Joseph Priestley with the introduction of scale as a measure of time and significance, has become so commonplace that we often do not step back to consider the assumptions behind its construction. First, as Rosenberg and Grafton point out, is the selection of data to visualize: "From the beginning, the biggest challenge of the time chart was not to include more data, but to clarify a historical picture—to offer a form that was intuitive and mnemonic, and that functioned well as a tool of reference."[19] Thus the visualization of events is also an interpretation; adding points on a timeline can immediately present the viewer with a possible sequentiality and inferred causality of events. Further, as both Manuel Lima and Anthony Grafton have argued, by visualizing time as a line that might stretch from left to right, we are replicating the motion the eye follows in reading a line of text in English or German. We assume that the beginning is on the left and that the end is towards the right. However, if we read Arabic, would time then be represented in the opposite direction?

Bearing these caveats in mind, constructing a timeline can be a very instructive and illuminating component of a literary analysis. Sometimes, in addition, timelines around a narrative can include contextual events. Priestley's famous 'A Specimen of a Chart of Biography' (1765) places political figures on a timeline with writers, inviting an analysis of the interdependence of political periods and literary production.

In addition, including time as an axis within other analyses allows for rich comparative and dynamic visualizations. As noted above, the *Book Biography Machine* from Harvard's MetaLab, uses time as a z-axis in its representation of the spatial distribution of printed editions of works accessed from the WorldCat API. As the developers say, the platform "is an interface that permits humanities scholars to map the diffusion of written works across geographic space and time in order to ask new questions about the history of literature…. Time is represented on the vertical axis, while space is represented across the horizontal plane."[20]

Deleted: , currently in development,

Networks

In the field of historical research, the drawing of (social) networks of "actors" in a historical and geospatial field can reveal new insights into spheres of influence (trade networks; correspondence networks; political influence and flows of money). Drawing networks can enable us to measure, for example, the strength of ties between "nodes" (which can be people, places, organizations) and the determination of communities of nodes (through the addition of attributes to those nodes) as well as density and size of networks, For the historian, these are all useful and revealing visualizations of historical data. Software, such as the freeware program "Gephi",[21] has made the graphical representation of all such metrics possible to those who wish to learn how to construct datasets and matrices.

However, what does the construction of a network diagram mean in literary analysis? Certainly, within the field of literary history datasets can be constructed of publishing trends, citation indices, literary influence, correspondence, and the genealogies of fields of knowledge. All these analyses rest upon the extra-literary metadata associated with the field of literary studies. However, what might the "graph" (as Moretti termed it) of the plot reveal and how would we construct it?

One approach literary scholars have taken to the visualization of social networks within a novel, for example, is to construct a database of characters (nodes) within a work. Then connections or contacts between those characters can be constructed in an "edges" table that can also be "weighted". In other words, if we wanted to construct a network diagram of Frederick Douglass' 1845 *Narrative* we might first identify all the characters in the work, then apply attributes to them (enslaved African American, African American, white American). We could then construct an "edges" table that is based on contact (we could further define the "nodes" as helpful or hostile). Then we could visualize these relationships as a network (perhaps also time-enabled) to see what it added to our understanding of Douglass' path to freedom. Which communities of people were helpful? Which were hostile? How did this change, or not? Some of my students in an introductory DH class performed a similar exercise with an excerpt from 18th century manuscript diary of the Moravian mission at Shamokin, PA. By creating tables of nodes (people) and then creating connections/edges between these nodes, students were able to make sophisticated analyses of the relationships between the European missionaries, the area's Native Americans, and the white settlers. By adding a time axis, we were also able to see the different modes of connectivity between the first missionaries and the last ones.[22]

Deleted:

<insert Fig. 3.5 near here>

Comment [Office6]: correct

In my own work on Moravian archival materials from the 18th century, the construction of a network visualization based on data extracted from a collection of thirty memoirs by women has revealed interesting results. Whereas many historians might think that a religious group would keep its members (especially its women) segregated from the outside world, a network analysis performed with Gephi shows that women in the movement had significant contact with non-Moravians, many of them highly influential. Such findings have important implications for rethinking the agency of women in, for example, Colonial America and 18th century Britain and Germany.

<insert Fig. 3.6 near here>

Comment [Office7]: correct

Within textual analysis, network visualization can also produce representations of topic clusters. David McClure's "TextPlot"[23] uses textual analysis methods outline above in conjunction with graph theory to produce what he terms "mental maps' of texts and their topics. The methodologically sophisticated work of Micki Kaufman on Henry Kissinger's text and voice memos, uses network analysis to produce time-enabled graphical representations of the topics in Kissinger's memos. Exemplary in her explanation of methodology, Kaufman's project can serve as a model for the combination of multiple forms of visualization within the analysis of text.[24]

Deleted:

Network visualizations can count among some of the most intriguing and aesthetically pleasing in DH-inflected literary analysis. But we must tread carefully. At the forefront of our

"reading" of these visualizations we must ask, in what way are they answering our research questions.

Conclusion

This short chapter has argued that visualization is about both representation and interpretation. Whether digital or not, the way in which we make visual our qualitative and quantitative arguments demands that we articulate clear research questions at the very outset of our investigations. The methods that we choose must fit those research questions, and we must be aware throughout our data analysis that our methods will also shape our results. We are, most definitely, caught in a digital hermeneutical circle.

However, even with that caveat in mind, visualization and literary analysis can make for a robust duo in both dissemination and analysis, allowing others to see our arguments in new ways, and allowing us to develop those arguments in new ways. In order to "lay bare" the risk of loading the dice in our enquiries, clear methodological discussions should be an integral part of any project write up. As part of a knowledge community, DH literary critics share their methodologies for others to replicate and develop. Despite the fact that there are many "black box" platforms available, ones that allow us to upload a corpus, or a .csv, or .txt file, and produce a visualization almost instantaneously, if we do not understand the analytical principles behind the visualization, or we can't explain them to our readers, then we cannot write up our methodologies and share them with the community. Although the "click/aha!" effect is seductive, it can unfortunately lead to less than robust analyses and readings.

Deleted:

Visualization can enrich and challenge conventional literary analysis. Whether by means of more quantitative methods, such as text analysis based on corpus linguistics, literary mapping emanating from either a GIS, geo-tagging, or inductive visualization, network analysis based on mathematical graphing, or a combination of these methods, digital methods of making us see our texts and their fictional worlds are powerful. With them, we clarify our argument, and invite the reader into the process of interpretation. As one student put it, concluding an assignment on the network visualization of recent political rhetoric, "at the end of the day, visualizations mean nothing without context. So giving the viewer the keys to [draw] their own conclusions, opens up a new realm of discoveries, even beyond what the original creator could have imagined." (Robert Cowen, Bucknell University, '17)

Deleted: Bobby

[1] The work of Franco Moretti is fundamental to the reconceptualization of literary criticism in a digital age. See, inter alia, his *Graphs, Maps, Trees* (London: Verso, 2005) and *Distant Reading* (London: Verso, 2013) .

[2] The fact that Moretti's work has been labelled "heretical" has brought with it both positive and negative responses, even from the highbrow non-academic press. See, for example, Joshua Rothman "An Attempt to Discover the Laws of Literature" *New Yorker* March 20, 2014, http://www.newyorker.com/books/page-turner/an-attempt-to-discover-the-laws-of-literature

Deleted:

[3] See, Johanna Drucker, *Graphesis: Visual Forms of Knowledge Production* (Cambridge, Mass.: Harvard UP, 2014).

[4] See Manuel Lima, *Visual Complexity: Mapping Patterns of Information* (Princeton: Princeton Architectural Press, 2011) and *The Book of Trees: Visualizing Branches of Knowledge* (Princeton: Princeton Architectural Press, 2014).

[5] Tanya Clement, "Text Analysis, Data Mining, and Visualizations in Literary Scholarship" *Literary Studies in the Digital Age: An Evolving Anthology,* MLA Commons, 2013. *(https://dlsanthology.commons.mla.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/)*

[6] Web-based platform http://voyant-tools.org/ by Stéfan Sinclair and Geoffrey Rockwell

[7] Antconc is a freeware corpus analysis tool for concordancing and text analysis. http://www.laurenceanthony.net/software/antconc/

[8] Mallet http://mallet.cs.umass.edu/index.php; Lexos http://lexos.wheatoncollege.edu/upload; Jigsaw http://www.cc.gatech.edu/gvu/ii/jigsaw/; Poemage http://www.sci.utah.edu/~nmccurdy/Poemage/

Deleted:

[9] *The Map of Early Modern London* (https://mapoflondon.uvic.ca/) is based at the University of Victoria, BC and directed by Janelle Jenstad.

[10] See, for example, *Authorial London*, https://authorial.stanford.edu/

[11] http://metalab.harvard.edu/2016/03/the-book-biography-machine-at-the-medieval-academy-of-america/

[12] A first step in this project can be found at http://moravianlives.org/.
[13] Katherine Faull, *Moravian Women's Memoirs: their Related Lives 1750-1820* (Syracuse: Syracuse University Press, 1997).
[14] Anne Kelly Knowles, Levi Westerveld & Laura Strom, "Inductive Visualization: A Humanistic Alternative to GIS", *GeoHumanities*, 1:2, (2015): 233-265.
[15] Ian Gregory and A. Hardie. "Visual GISTing: Bringing together corpus linguistics and geographical information systems." *Literary and Linguistic Computing* 26:3 (2011): 297–314 and Cooper, D. and Gregory, I. N. (2011), Mapping the English Lake District: a literary GIS. Transactions of the Institute of British Geographers, 36: 89–108.
[16] http://bit.ly/2upPMRu
[17] Knowles, A. K., T. Cole, and A. Giordano, eds. *Geographies of the Holocaust*. Bloomington: Indiana University Press. 2014.
[18] Daniel Rosenberg and Anthony Grafton, *Cartographies of Time: A History of the Timeline* (Princeton: Princeton Architectural Press, 2012) 17.
[19] Rosenberg and Grafton, 246.
[20] http://metalab.harvard.edu/projects/
[21] https://gephi.org/
[22] These assignments and their results are discussed in "Doing DH in the Classroom: Transforming the Humanities Curriculum through Digital Engagement" (with Diane Jakacki) *Doing Digital Humanities: Practice, Training and Research*. Richard J. Lane, Raymond Siemens, and Constance Crompton, eds. Abington, UK: Routledge, 2016.
[23] http://dclure.org/essays/mental-maps-of-texts/
[24] See Kaufman's site http://blog.quantifyingkissinger.com/

Deleted: https://katiefaull.files.wordpress.com/2013/11/meredyk-paxton-map.jpg

Deleted:

Deleted: Forthcoming