Mining the Knowledge Base: Exploring Methodologies for Analysing the Field of Electronic Literature

Paper for Digital Methods Winter School, Amsterdam 22-25 January 2013

By Jill Walker Rettberg and Scott Rettberg
Electronic Literature Research Group
University of Bergen, Norway

This is a work-in-progress report from an exploration of the intersection between the fairly conventional digital humanities method of creating a database - specifically, the ELMCIP Electronic Literature Knowledge Base (http://elmcip.net/knowledgebase) and the digital methods strategy of directly analysing online, digital content. We are testing out different methods of analysing data about conference series harvested from the Knowledge Base, using social network analysis to visualise the connections between people, events and works, tag analysis to examine changes over time in the type of creative work and scholarly presentations and using IssueCrawler to analyse URLs associated with the conference series and associated works.

Electronic literature is digitally native literature that, to quote the definition of the Electronic Literature Organization, "takes advantage of the capabilities and contexts provided by the stand-alone or networked computer". Most, though not all, electronic literature is only published online and thus digital methods would seem an obvious technique for studying and analysing the genre. Electronic literature works are multimodal, highly variable in form and only rarely present linear texts, so datamining the works themselves would be difficult. Instead, we wish to take what Franco Moretti has called a "distant reading" approach, where one looks for larger patterns in the field rather than in individual works (Moretti 2005, 2011). However, the lack of established metadata for electronic literature makes this difficult as there is no central repository or bibliography of electronic literature.

The ELMCIP Knowledge Base of Electronic Literature is a human-edited, Drupal database consisting of cross-referenced entries on creative works of and critical writing about electronic literature as well as entries on authors, events, exhibitions, publishers, teaching resources and archives (Scott Rettberg 2012). All nodes are cross-referenced so you can see at a glance which works were presented at an event, and follow links to see which articles have been written about those works or which other events they were presented at. Most entries simply provide metadata about a work or event, but increasingly we are also gathering source code of works and videos of talks and performances. The Knowledge Base provides us with a growing wealth of data that we are beginning to analyse and look at in terms of visualisations, social network analysis and other methods.

As a database, the ELMCIP Knowledge Base is a typical digital humanities project. Rather than applying digital methods to text that is already online, we gather and organise metadata, and in some cases, archive source files and original documentation. It may seem a paradox that electronic literature, as a digitally native genre, should not already be a rich ground for digital methods. As a deliberately structured database of metadata, the ELMCIP

Knowledge can be seen as digitized, not natively digital, despite almost all its entries being about natively digital content. A digital methods approach would not carefully craft a structure, but would analyse existing structures in web data to find patterns not immediately visible. As Richard Rogers suggests, “May the Web deliver structured data after all? In this way of thinking, Web services – search engines, collaboratively authored wikis and social networking platforms – become the data filterers, cleaning and ordering the data for end usage as well as perhaps for research.” (Rogers 2013, forthcoming). We have begun using digital methods to analyse the field of electronic literature (J. W. Rettberg 2012) and in the work described in the current paper, we are looking to develop more general methods that will work with the kind of data we have access in and through (in the case of URLs connected to entries) the Knowledge Base.

But electronic literature, although digitally native, does not have any obvious structure. There are some blogs, such as Leonardo Flores’ *I E-Poetry*, which presents a hundred word reading of a new digital poem every day, or Jason Nelson et. al.’s *Netpoetic*, or the now retired blog *GrandTextAuto*. There are websites of individual authors - creative works and papers are often published on author websites. There are some online journals (creative work is published in journals like Beehive, Drunken Boat, Poems That Go, New River, Hyperrhiz and critical writing in the field in journals like Electronic Book Review and Dichtung Digital) and there are a few collections (e.g. The Electronic Literature Collection Volumes 1 and 2, The ELMCIP Anthology of European Electronic Literature). There are databases of electronic literature other than ELMCIP. The two largest are The Electronic Literature Directory, which emphasises high quality descriptions but has a very limited number of records, and NT2, which has an excellent overview of French language works.

But there is no central clearing house system of DOIs or ISBNs of works of electronic literature. Libraries generally catalogue electronic literature unless it is published on CD-ROM with an ISBN, as the hypertext publisher Eastgate did in the 1990s and early 2000s. Because electronic literature is so experimental, without very strong genre conventions, there is no obvious way of automatically detecting from a website or a file that a given object is a work of electronic literature, in the way that you can detect that something is an image or a video or a text or even a sonnet or a movie script.

Perhaps we could use Twitter or Facebook as a data filterer in the digital methods sense. There is a lot of discussion in a very active Facebook group on electronic literature, and some discussion on the Twitter uses the hashtag #elit. Another very important venue for electronic literature is conferences and festivals, where new works are read, performed, and exhibited and new research is presented.

Perhaps then the ELMCIP Knowledge Base is—or will become—precisely the collaboratively written web service that will allow us to filter the digitally native data that exists in abundance. The Knowledge Base carries with it a bias in that each entry is added and edited by humans. Although we have an open contribution policy, the Knowledge Base does not and can perhaps never offer a true, complete snapshot of the field of electronic literature. We will always be limited by the focus and knowledge of the people who contribute to the resource.

Are there any ways in which digital methods could provide such a thing? In this project, we want to explore ways of using the admittedly still incomplete data in the Knowledge Base as a starting point for understanding the larger patterns of the field, both by analysing the data we have in the Knowledge Base itself, and by using the Knowledge Base as a starting point that can lead us to existing online data. While there is a great deal of semantically structured information in the database proper, it for instance also links to thousands of URLS of resources on the Web, and to the majority of dissertations that have been published in the area in the past two decades. The data mining possibilities of these resources are promising.

Because conferences are a major point of contact and communication in the field, as one aspect of our research we are documenting them and researching how the thematic focus of creative works and critical writing presented at these events has changed over time. In the Knowledge Base, we enter information about which (if any) event each creative or critical work is presented at, so we can see what was on the program for each conference, and can access keywords assigned to the works (in some cases by the authors themselves, in other cases these keywords are assigned by contributors to the Knowledge Base) and information about authors, references made in the writing presented, and so forth.

But with 110 events, ranging from individual readings, to performance nights in 2003 to Brazilian digital art shows with only a few presentations really relevant to electronic literature, and of course also conferences and workshops dedicated to electronic literature, the data was just too messy and heterogenous. So we tried selecting a more limited data set, just looking at conferences that are part of established, multi-year series, where electronic literature was the main focus or consistently covered over a series of years.

Our list now included a total of 44 events, including

- the Digital Arts and Culture conferences, which ran every year from 1998-2001 and then every other year until 2009
- The E-Poetry conferences, which have run every other year since 2001, and
- The ELO conferences, 2002, 2007, 2008, 2010, 2012
- The ELMCIP project events
- The MLA events

This is an admittedly Anglo-American selection, and we will need to look more closely at other nationalities and language groups later.

Unfortunately, we discovered that not only did the Knowledge Base lack entries for a lot of the program at some of these events, but also several of the conference websites had also disappeared from the web altogether. Fortunately, we have been able to access backups from the organizers and are in the process of entering the information in the Knowledge Base. This means that the data we have is quite flawed at this point, but we have still begun testing out various ways of analysing it in order to test the methodologies. Later we will have to re-analyse a more complete dataset.

Tag analysis

At the time of analysis, we had 138 records of conference presentations at the Electronic Literature Organization (ELO) conference series. Creating a tag cloud from the tags

assigned to the papers generates a sort of picture of the topics of the field, as shown in figure 1.

Figure 1: A tag cloud showing all tags associated with 138 research papers presented at the ELO conference series from 2002-2012.

Not all the papers had tags assigned to them, and not all papers presented at the conference series are recorded, so this is a very preliminary snapshot of the field, and yet it does quite clearly show the kinds of topics that the conference series deals with.

More interesting uses of this methodology might be comparing tags over time, for instance seeing how the emphasis of research in a field likely shifts over the course of a decade. We might also compare the tags used at different conference series. Presumably the E-Poetry conferences have less “fiction” and “narrative” than the ELO conferences, but do they have different emphasis on “history” or “user-generated”, for instance?

Several of the most recent conferences have used the EasyChair system for submissions of abstracts and papers, and in this system, authors are asked to provide keywords for their submissions by default. Being able to access this author-provided data directly will add value to the tags for future conferences.

Another way of using tag analysis to see the field from a distance is to look at tags that appear together. Figure 2 shows tags (or, to be precise, words that are part of tags) that are assigned to works that are *also* tagged “hypertext”. The word “hypertext”” is reduced so as not to completely dominate the tag cloud.

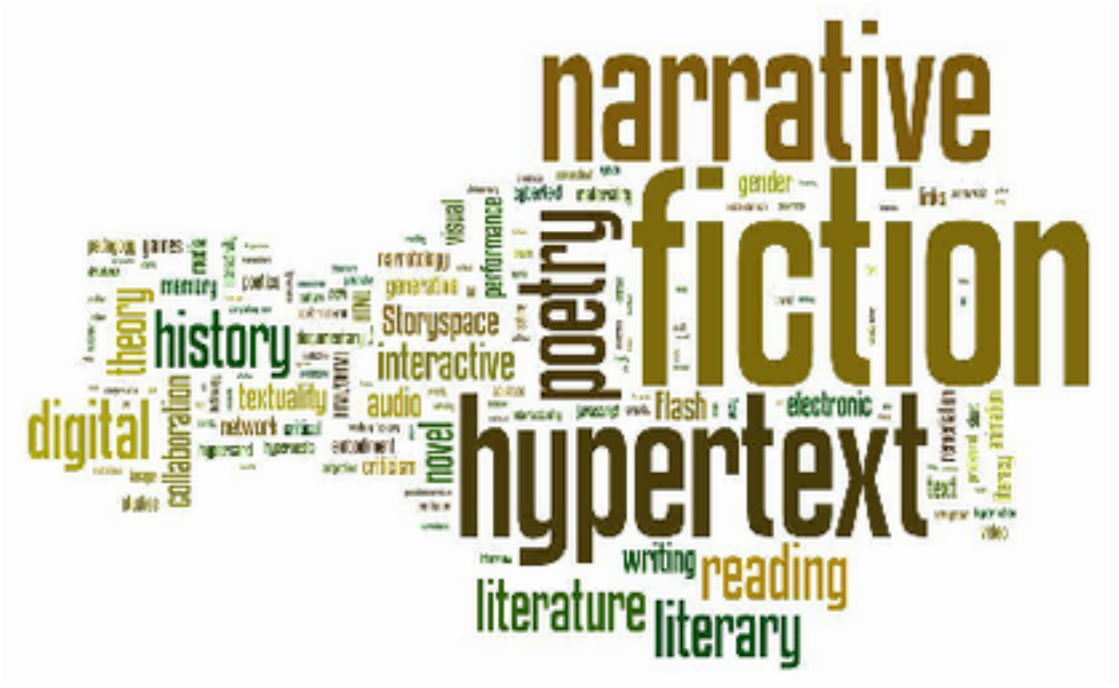

Figure 2: Tags that appear with hypertext, from creative works in the ELMCIP Knowledge Base.

Tag analysis like this could lead to a methodology for extracting or considering a map of genres in electronic literature, which could for instance then be compared to a more qualitative way of discussing genres.

Of course, tag clouds are a fairly primitive form of textual analysis, as has been pointed out by, among others, data journalist Jacob Harris of the New York Times, who decries the over-simplistic journalism that tag clouds make possible when poorly used:

> For starters, word clouds support only the crudest sorts of textual analysis, much like figuring out a protein by getting a count only of its amino acids. This can be wildly misleading; I created a word cloud of Tea Party feelings about Obama, and the two largest words were implausibly “like” and “policy,” mainly because the importuned word “don’t” was automatically excluded. (Harris 2011)

And yet, as a way of gaining an overview of possible trends in a large set of data, tag clouds can be a useful and simple approach.

Social Network Analysis

We are also beginning to visualise the social networks implicit in the data we have. We hope to be able to find ways to answer a broad range of research questions, for example:

- What are the community structures in the field? Do actors form close-knit clusters or are groupings more random and transient? Are these structures stable over time?

- Is there a connection between productivity and particular types of position in the social network of the field (e.g. being part of a closely knit group or being highly connected)? What characterises the community participation of actors whose works

are highly referenced?

- What common characteristics do actors who frequently interact and thus belong to a group share? E.g. nationality, residence, age, language, gender, genre they work in?

- Has the speed with which ideas and theoretical paradigms are developed and disseminated internationally increased with the adoption of network technologies? Can literary genres in new media be understood as “memes” which circulate and are developed virally on a transcultural basis?

- Can necessarily reductionist, quantitative, semantically structured approaches to mapping literary and cultural practices enable richer, more expansive analyses of individual artifacts represented within an unfolding historical context?

Presumably we will not be able to find answers to all these questions, and the reasons why we are not able to answer them will be key in developing a theoretical methodology for the field. We will also presumably find answers to questions we were not able to envision before having computationally processed and visualized the dataset.

At this stage, we are just beginning network analysis of the data in the Knowledge Base. We began by generating a graph simply by exporting a spreadsheet where all authors and all conferences in the Knowledge Base are represented as nodes, and another spreadsheet which describes an “edge” (a link or connection) between an author and a conference at which he or her has presented a paper or a creative work. We imported these spreadsheets into Gephi, an open source social network analysis tool.

Unsurprisingly, this created a graph showing a lot of unconnected nodes around the edges (all the authors in the Knowledge Base who had not presented at any of the selected conferences) and a big connected component in the middle. There was also a separate grouping (shown a little to the right of the top of figure 3) which turned out to be the Brazilian FILE conferences, which, at least as the various conferences are currently documented in the Knowledge Base, appear to have little crossover with other electronic literature conferences, despite some arguably related content matter.

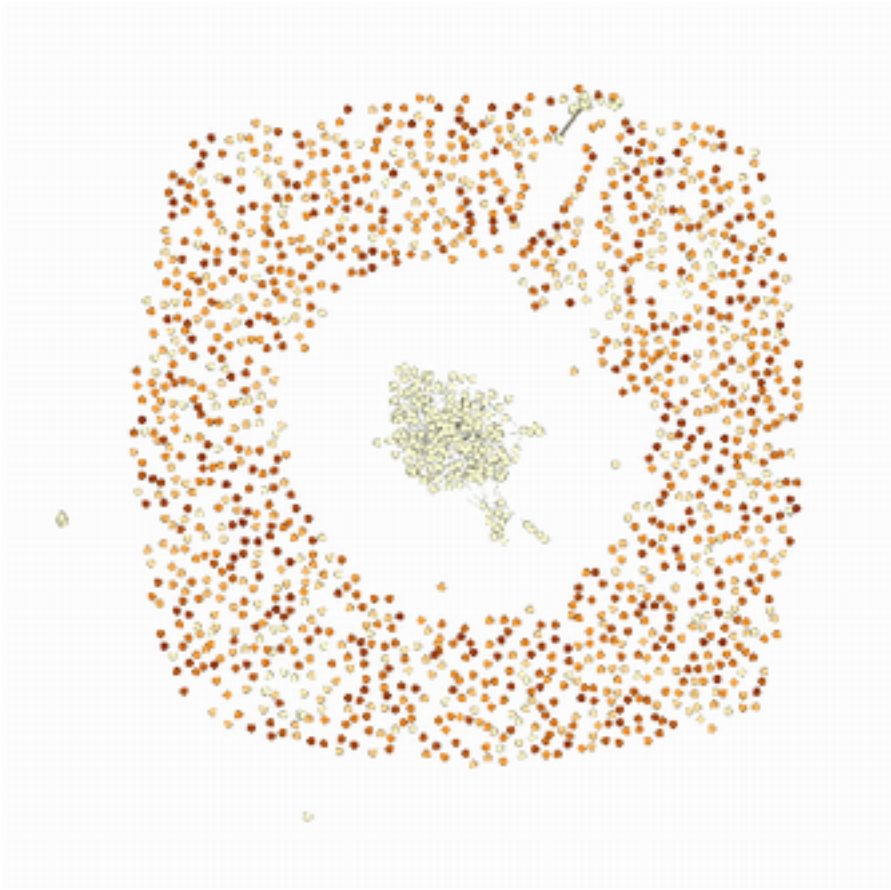

Figure 3. The graph of all authors and conferences in the ELMCIP Knowledge Base.

Running a modularity algorithm, it's easy to see communities forming around individual conferences (see Figure 4) and we can also see that some people have far more presentations than others, but this is stil not the most useful diagram. It is hardly surprising though that a modularity algorithm would find a community around each specific conference, given that the edges are all between the conferences and the people who have presented at them.

Figure 4: Having hidden the nodes representing the authors who are not recorded as having presented at the selected conferences, we ran a modularity algorithm to find community structures in the remaining authors and conference nodes. Conferences are labelled, author nodes are unmarked.

Next, we ran an eigenvector centrality algorithm. Eigenvector centrality measures how close a node is to central nodes, and is run over and over again much as Google's PageRank gives more weight to websites connected to highly-connected websites. This allowed us to see a ranked list of authors by eigenvector centrality. John Cayley, an influential scholar, showed up at the top of the list, but again, this may be an example of the bias of the data. We encourage scholars and authors in the field to contribute their own and others' works to the Knowledge Base and Cayley is not only one of the most active scholars and authors in the field, he is also an active contributor to the Knowledge Base.

Seeing this ranking of authors - who are living people - also raises ethical questions about such research. We are only using publicly available data about published authors and their works and papers. Many of these authors may be self-published, but as artists, authors and scholars they explicitly make their work public and wish to be read. However, while working on a project like this we must continually remain aware of the ethical challenges inherent in making visible structures and connections that may not be readily apparent. As Charles Kadushin puts it in his discussion of the ethics of social network analysis, "the data were

already there for ‘all to see’ but in fact, without first collecting data from various sources and putting them in a data base and only then analyzing and graphing them, the data would have remained invisible” (2005) The very act of entering information into a database carries ideological and ethical considerations. When designing a database, you must make choices about what fields to include and which to leave out, and this necessary simplification inevitably results in the obscuring of certain nuances and details. Additionally there may be a bias in the way the data was gathered in the first place. Kadushin concludes that although science and society may benefit from network analysis, the individuals who make up the network rarely do. In this case, the individuals are for are published scholars and authors who desire their work to be more broadly known, so we expect the benefit to outweigh any costs, but will constantly consider ethical implications of new analyses.

Branching away from the focus on conferences, we have also tried a network graph showing all the creative works in the Knowledge Base with edges connecting them to their authors. Figure 5 shows an overview, and a PDF, the text of which is more legible if viewed at 1200%, can be downloaded from http://jilltxt.net/images/works_and_authors.pdf.

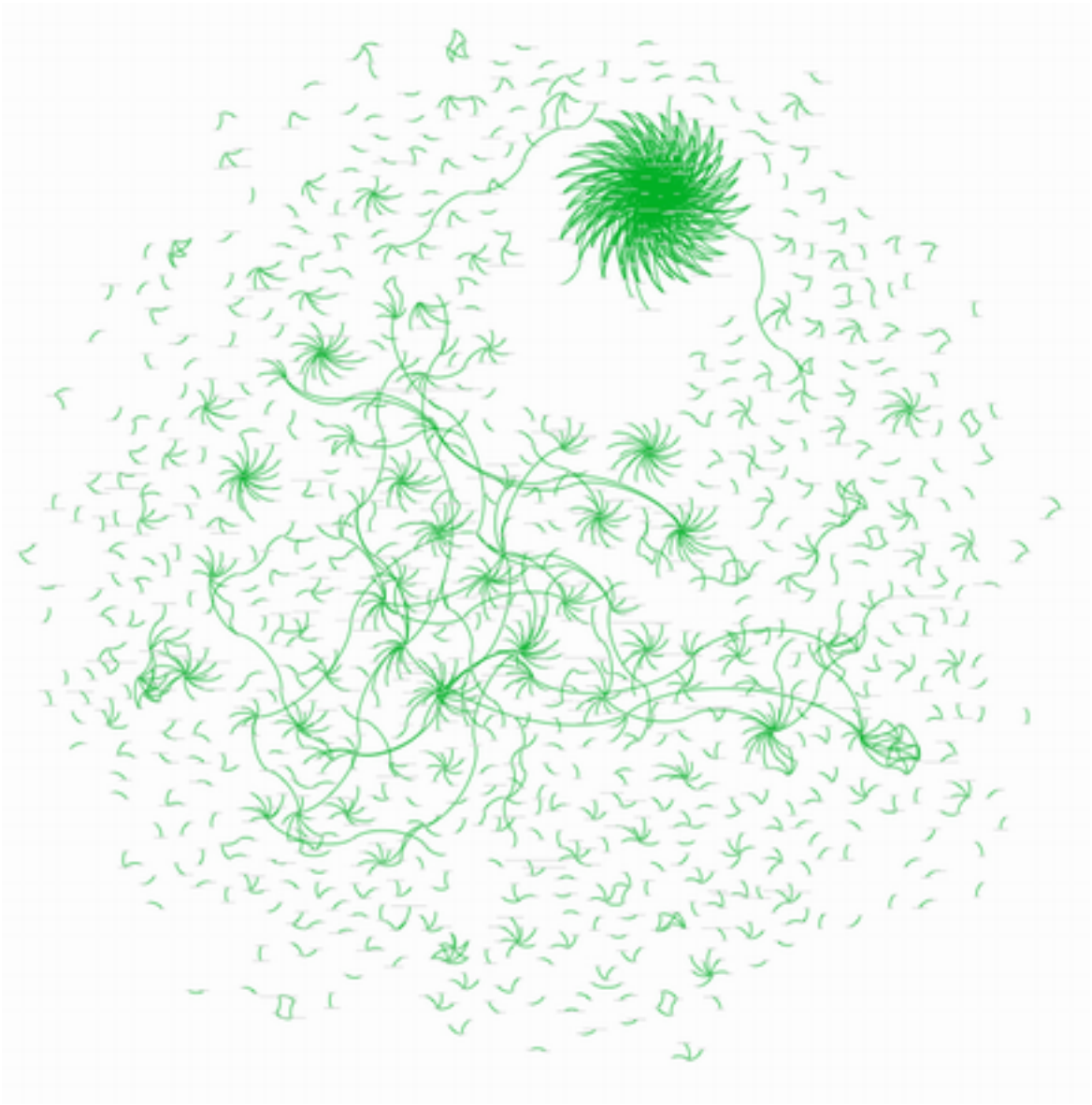

Figure 5: An overview of a visualisation of all creative works and authors in the ELMCIP Knowledge Base, with lines connecting the authors with their works.

Here swirls are created when an individual authors has written several works, and clusters and criss-crossing lines appear when more than one author has collaborated on a work. The large cluster shows the works of the extremely productive South Korean collective Young Hae Chang Heavy Industries, a group that has both authored many, many works and was the focus of a student project very thoroughly documenting the works in the Knowledge Base.

The most interesting parts of the visualisation are the ones showing connections between authors who have collaborated on works, such as the cluster of German-language authors shown in Figure 6:

Figure 6: A closer view of a cluster of works. The central authors here are René Bauer, Beat Suter and Johannes Auer.

This visualisation may be most useful as an alternative introduction to the field, or a sort of road map for scholars and students of electronic literature. Panning across the graph provides an interesting alternative view of how electronic literature is written. It also gives those who know the field a good idea of areas where data is missing and that need better documentation in the Knowledge Base. If we had the data, it might be interesting to compare the collaboration structure in the field of electronic literature to collaborative networks and clusters in other fields, such as visual art, electronic art or cinema.

In analysing a social network, or an information network, the choice of what should be a node and what should be an edge is key. In the next iteration of graphing the social networks of electronic literature, we will try other data constellations. Instead of seeing both conferences and authors as nodes, we'll graph the authors as nodes and draw edges between them when they have presented at the same conference.

This is similar to the approach taken by Dan Wang in an analysis of sociological articles taught in a large number of university courses (Wang 2012). Whenever two articles were taught in the same week of a course, Wang drew an edge between them, and thus generated a network diagram showing clusters of articles, and also articles that were "bridges" or "brokers" between the clusters. According to network theory, these brokers would be where information moves between clusters. Wang writes: "in culling a set of canonical references from this network representation, we privilege not only those references that are most emblematic of a given tradition, but also the bridging references that give these different territories of economic sociology some measure of coherence and mutual relevance."

Next Step: Using Digital Methods Tools

Our next method will be to try using digital methods tools such as those developed by the Digital Methods Initiative in Amsterdam to analyse the websites that our Knowledge Base references. We have collected many URLs in the Knowledge Base and it would be interesting to use tools such as the Google Scraper (Lippmannian Machine) or the Issuecrawler to explore related sites, such as the websites of authors who had presented at a conference, or conference websites themselves.

At this point we have unfortunately not yet begun this work, and we are researching possibilities here.

Exploration as Method

Our method is by necessity exploratory. We want to use software tools to find patterns in the network structures and data about the field that we gather in the ELMCIP Knowledge Base, and when patterns are found, we will form hypotheses and interpretations and perform further experiments to test them.

As Matthew Kirschenbaum wrote about the nora project (which grew into MONK) at the Maryland Institute for Technology in the Humanities (MITH), "A significant component of our work is therefore basic research in the most literal sense: what kinds of questions do we seek to answer in literary studies and how can data mining help, or— more interestingly—

what new kinds of questions can data mining provoke?" (Kirschenbaum 2011)

This work-in-progress paper is a report from our early explorations in analysing the material collected in the ELMCIP Knowledge Base, and we will present more developed analyses in the next months and years.

Bibliography

Harris, Jacob. 2011. Word Clouds Considered Harmful. *New York Times*, 13 Oct. http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/

Kadushin, Charles. 2005. Who Benefits from Network Analysis: Ethics of Social Network Research. *Social Networks* 27 (2):139-153.

Kirschenbaum, Matthew. 2011. Poetry, Patterns, and Provocation: The nora Project. In *Reading Graphs, Maps, and Trees: Responses to Franco Moretti.* Anderson, SC: Parlor Press.

Moretti, Franco. 2005. *Graphs, Maps, Tree: Abstract Models for Literary History.* London: Verso.

Moretti, Franco. 2011. *Network Theory, Plot Analysis.* Pamphlet. Stanford Literary Lab. http://litlab.stanford.edu/?page_id=255

Rettberg, Jill Walker. 2012. "Electronic Literature Seen from a Distance: The Beginnings of a Field" in *Dichtung Digital* 41. http://www.dichtung-digital.org/2012/41/walker-rettberg/walker-rettberg.htm

Rettberg, Scott. 2012. "ELMCIP Research Project Goals, Results, and Impact." http://elmcip.net/critical-writing/elmcip-research-project-goals-results-and-impact-presentation-remediating-social

Rettberg, Scott. 2011. "The ELMCIP Knowledge Base and the Formation of an International Field of Literary Scholarship and Practice." *OLE Officina di Letteratura Elettronica: Lavori del Convegno.* Naples, Italy: Atelier Multimediale Edizioni. http://elmcip.net/sites/default/files/attachments/criticalwriting/rettberg_knowledgebase.pdf

Rogers, Richard. Forthcoming / 2013. *Digital Methods*. Cambridge, MA: MIT Press. Excerpt from book as published in *Big Data* compendium for the Digital Methods Summer School 2011.

Wang, Dan. 2012. "Is there a Canon in Economic Sociology?" in *ASA Economic Sociology Newsletter* 11(2), May 2012.