Speaking notes, True North Science Bootcamp 2018
Montreal, Quebec
24 May 2018
Holly Hendrigan

Digital humanities and STEM librarianship, or why I stopped rolling my eyes at word clouds

[slide 1: overview] In the next 20 minutes, I will provide you with the background to the project, a STEM librarian onboarding technique, and my methodological failures and successes. I'll tell you what emerged, touch on some theory about its implications, my reservations, as well as some other possible uses for the application.

[slide 2] This methodology came about in response to a real-world problem I encountered: how do I get up to speed in my new portfolio area? Our library underwent a significant organizational change and in January 2017, I switched liaison areas from Education and World Literature to Applied Sciences. So how does a librarian with an undergraduate degree in English Literature and liaison experience in humanities and social sciences disciplines map out what her university's engineers and computer scientists are doing? For simplicity's sake, I'm going to focus on one department, Mechatronics Systems Engineering, but this technique would work with any STEM department.

[slide 4: literature review] But first, as an evidence based practitioner, what does the literature say about becoming a STEM librarian? Maness and Tobin-Cataldo mention traditional methods such as networking with STEM librarian colleagues, reading websites, selective literature searches, building relationships. I did all that: joined the Vancouver network of science and engineering librarians, met with colleagues who formerly liaised with the STEM departments and the faculty representatives.

I poured over the applied sciences collections profiles from our vendor, and course lists.

[slide 5] and read departmental web pages. As you might be aware, faculty web pages are particularly unhelpful, because they are written to attract students, not educate librarians.

"**WHAT IS MECHATRONIC SYSTEMS ENGINEERING?**
Mechatronics is a dynamic, multidisciplinary subject combining three engineering fields: mechanical, electrical and software engineering. This highly integrated approach creates smart, inventive and evermore efficient solutions for a wide range of high-tech engineering problems."

When I looked at the websites of individual faculty, they weren't much more helpful. For example:

Research Interests of Professor X:

- Mechatronics,
- Biomechatronics,
- Biomedical technologies,
- Wearable technologies
- Biorobotics

The "recent publications" section of individual faculty members was pretty hit and miss, with some revealing none at all, to some being fairly recent (2017). Quite a few citations from 2011, 2013, 2014.

[slide 6] The Mechatronics web page indicate on the web page that it is a multidisciplinary school, with six subspecializations:

[slide 7] But there was very little in any of the web pages that helped me with ordering decisions.

[slide 8] Since the department website was so little help, I went to the individual researcher level, and set up Google Scholar search alerts. I got an email everytime a new citation by member of my departments appeared.

[slide 9] I cut and pasted the Google Doc metadata into Excel spreadsheets, which was a lot of work. But the spreadsheets were a different type of problem from the department website pages: this was too much information.

[slide 10] But then I remembered a presentation I went to at ACRL in 2017, "Text mining and data visualizations" (Gao and Wallace) and went back to that.

[slide 11] I looked at their methodology, which involved obtaining a corpus of their entire university's citations from Scopus, then running it through topic modeling software, Mallet. In this context, a "topic" consists of a cluster of words that frequently occur together." (Mallet website http://mallet.cs.umass.edu/topics.php).

[slide 12] Gao and Wallace then entered the topics from the Mallet output into Tableau.

[slide 13] So, with the help of my in-house technical support (in other words, my husband), I tried this on a smaller scale, with my departments. I agreed with Gao and Wallace to focus on titles, for "most people expect a title to indicate the subject of the intellectual content of the item" (Hagler, 30). I downloaded Mallet, got my husband to

write a script that pared each citation to its title only, created separate files for each title, and input the folder into Mallet's "topic modeling jar."

[slide 14] This view shows 10 topics with 5 terms.

[slide 15] Essentially, none of the permutations of topics really worked for me in Mallet. Not the 20-topic, 10 word; not the 10-topic, 10 word or 10-topic, 4 word list. Without a background in engineering, I could not conceptualize or name many of these topics (eg, "through flow catalyst based analysis"): what is that topic about? Research conducted by subject experts can decode and interpret Mallet's word clusters. However, in my case, these topics made less sense to me than the organizational categories that Mechatronics had on their own department website.

[slide 16] But, as I was reading about digital humanities applications in the Historian's Macroscope,

[slide 17] I also learned about Voyant Tools. It's been described as a Gateway Drug of text mining and data visualizations. For a number of technical reasons, it's not appropriate for a huge Wikileak-like dataset, but it seemed right for my relatively small dataset.

[slide 18] And by then, I had abandoned the Google Doc idea, because I thought Web of Science would be better. It allows exporting CSV files rather than cutting and pasting, and better quality control. And, luckily for me, Web of Science's "Address" field worked quite well--it captures citations by department.

[slide 19] Restricting the date from 2015 to 2018, I retrieved 234 records from Mechatronics; 269 records from Engineering. I exported the citations of each department, put them in a spreadsheet, and copied the title column into a separate document. I then had my corpus for the two departments, which Graham calls a "bag of words," for analysis. I input this corpus into Voyant Tools, and voila!

[slide 20] [screenshot of Mechatronics and Engineering cirrus view]

The Cirrus view is the "positions the words such that the terms that occur the most frequently are positioned centrally and are sized the largest." I have some personal issues with word clouds, which I will discuss later, but I do find it striking how little overlap there is between these two engineering departments. "Fuel Cells" and "membranes" dominate Mechatronics; "Optical coherence tomography" (an imaging technique) dominates Engineering. I really didn't get that from the website.

Thing is, Voyant is much more than a word cloud generator. Right under the cirrus view is the corpus summary statement, which provides a more quantifiable analysis.

In Mechatronics, Fuel cells came up in 43 word frequencies over a corpus of 234 records. This represents 18% of the research output, which is certainly significant, but

still does not constitute a majority of citations. To get a complete picture, librarians must also run searches on less frequently used terms, research groups, and individual researchers.

Let's also look at the Colocates graph, which "represents keywords and terms that occur in close proximity as a force directed network graph." The collocates graph provides more nuance than the cirrus, showing the network of higher frequency terms that occur in proximity. This graphic echos Mallet's topic models, but it is more interactive, allowing users to search or click on a term to view its colocates

Finally, Voyant Tools provides a keyword in context tool, which acts as a concordance. Here's a look at the term "thermal"

This is useful, again, for seeing terms in context and learning key phrases of your research areas. Note to self: thermal conductivity, thermal contact resistance, thermal loads: relevant to the mechatronics engineers.

[slide 24] Since the departments are so multidisciplinary, though, it's vitally important to recognize this by running searches on research groups or individual faculty members.

[slide 25] On the web site, the language of her research interests is quite broad: patient specific technologies; accident reconstruction and injury analysis.

[slide 26] Here is a word cloud of a faculty member in the Bio-mechatronics group. When you compile a corpus of her 21 recent article titles, however, you notice that the terms "spinal cord" are the most frequent. It's beyond the scope of this presentation to compare the language of the web page to that of the research output, but I will say that the more concrete and specific terminology in the research output corpus is very handy to understand the specifics of her research, and this specific language is invaluable when doing title by title book selection.

To broaden the lens from this very specific use case of Voyant Tools, however, what are some of the theoretical underpinnings of this experiment?

[slide 27] First of all, it is a Big Data approach to understanding my complex departments. As Graham says, "big data is simply more data that you could conceivably read yourself in a reasonable amount of time – or, even more inclusively – information that requires computational intervention to make new sense of it." In my case, I have neither the time nor the inclination to read 500 recently published articles of the Engineering Science and Mechatronics Systems departments. I'll add that making "new sense" of the research output is quite generous; the language is so technical that I argue that a non-expert could use a shortcut or two to learn the vocabulary.

[slide 28] And when we are dealing with big data, we have to change the research metaphor. We have to move away from the process of finding clues in the details of an object, as though we are looking through microscope. Instead, we need to think of using

a macroscope that compiles and rearranges the data from the object level to reveal patterns in a broader view.

[slide 29] Another way to understand this approach is via the literary critic Franco Moretti, who is interested in world literature. Scholars of world literature cannot be close readers of individual texts **or** genres **or** periods **or** nationalities. There are simply too many published books on a global and historical scale to operate on the level of an individual item. Experts in world literature have to think bigger, and examine themes or motifs across historical, national, and linguistic boundaries. He called this "Distant Reading," as opposed to what's commonly known as "close reading" and I think it's a interesting way to approach trying to understand the modern multidisciplinary complexity of a academic department.

[slide 30] And LIS professionals have never been close readers, anyway: we have always been experts in metadata; we are very proficient in skimming and scanning! But in the case of using Voyant Tools, we are flipping the process. Normally, WE skim and scan documents and assign terms, OR we skim and scan results lists based on the terms WE input. With Voyant Tools, we relinquish control of the output, and it provides US with the natural language frequencies. As Graham says, it "lets the sources speak to you"

[slide 31] It's quite possible that librarians might have some "issues" with natural language. My former professor in library school, Ronald Hagler, wrote that librarians have been shunning natural language access points since the 19th century, in favour of controlled vocabularies. And when I look at the Mechatronics wordbag, horrors! There's both 3-D and "three dimensional," which--for good reason--makes our blood pressure go through the roof. Never mind synonyms--what to do about those? At this point, though, I'll say take a deep breath, and don't let the perfect be the enemy of the good. Build in time to normalize and clean your data as well as possible.

[slide 32] Again, back to Ronald Hagler, he argues that scientists' scholarly language is precise enough to mimic controlled vocabulary. Engineers aren't fond of puns or wordplay (at least in their article titles); again, according to Ronald Hagler, "titles of articles in academic journals tend to avoid connotative references and to include many substantive words, a practice encouraged by the long-standing custom among the A & I services of using title words to index the contents of such articles" (Hagler 34).

[slide 33] Engineers will title an article "Numerical collapse analysis of Tsuyagawa Bridge damaged by Tohuko Tsunami," not "Bridge over troubled water."

[slide 34] To wrap up, Voyant Tools has made me reconsider the utility of natural language visualizations for my STEM liaison area. I think it has a lot of potential for other uses, too: analyze the column of journal titles the faculty frequently publish in, or curriculum documents, or graduate theses. Since we like controlled vocabularies so much, we could map the high frequency natural language terms to the lexicon of our

collections profiles. I'm finding my uploaded "word bags" VERY HELPFUL in my collections work.

[slide 35] Voyant is a cheap and easy, yet powerful tool that knowing a thing or two about word frequencies renders me qualified to slip on an iron ring. Nothing could be farther from the truth. But I do think, in our business, words are important, and Voyant Tools makes them much easier to learn.

Works Cited

Gao, W., & Wallace, L. (2017). Data mining, visualizing, and analyzing faculty thematic relationships for research support and collection analysis. In *ACRL 2017 Conference Proceedings* (pp. 171–178). Baltimore: ALA.

Graham, S., Milligan, I., & Weingart, S. (2015). *Exploring big historical data: the historian's macroscope*. London: World Scientific Publishing.

Hagler, R, & Simmons, P. (1991). *The bibliographic record and information technology.* Chicago: ALA.

Maness, J. M. (2016). Engineering and applied science librarianship. In K. Sobel (Ed.), *Mastering subject specialties: practical advice from the field* (pp. 31–38). ABC-CLIO.

Moretti, F. (2013*). Distant reading*. London : Verso.

Tobin Cataldo, T., Tennant, M. R., Sherwill-Navarro, P., & Jesano, R. (2006). Subject specialization in a liaison librarian program. *Journal of the medical library association* 94(4), 446–448.