Lecture Notes for Week 2, Class 1

Readings

- Rosenberg, Daniel. 2013. "Data before the Fact." *Raw Data Is an Oxymoron*, 15–40.
- Ramsay, Stephen. 2014. "The Hermeneutics of Screwing Around; or What You Do with a Million Books." In *Pastplay: Teaching and Learning History with Technology*, edited by Kevin B. Kee, 111–20. Ann Arbor: University of Michigan Press.

Agenda

1. Humanities Data (after Miriam Posner's "Humanities Data: A Necessary Contradiction")

Today I'd like to talk about **the ways in which humanists think about data**, and how that's distinct from the ways in which scientists and social scientists think about it.

I'll start with an **anecdote**. As part of my work here at Princeton, I have to consult with professors, staff and students and plan out the workflow for a DH project that they envisage. The passion and the research question are often very clear from the start. "Great," I respond. "Let's see your data." "Data?" they say. **"Oh, I don't have any data."**

This is not because we're stupid or naïve; it's that **humanists have a very different way of engaging with evidence** than most scientists or even social scientists. And we have different ways of knowing things than people in other fields. **We can know something to be true without being able to point to a dataset, as it's traditionally understood.**

> For example, we can know, that early silent film relied on the conventions of melodrama to create legible narratives, not because we have a spreadsheet somewhere, but because we've immersed ourselves so deeply in our source material that we're attuned to its nuances.
>
> Or, another example, we can know that the language of Chaucer's time was different from the language of Shakespeare's time, not because we have a dataset of word frequencies, but because we've read so much of it that we can feel the difference in our bones.

That's why humanists sometimes think you can make a visualization without data; because they want to illustrate ideas and movement, not necessarily data points as we've been discussing them here.

So, when we talk about data in the humanities, we're talking about a very different kind of thing than when we talk about data in the sciences. And that's not to say that one is better than the other, or that one is more rigorous than the other. It's just to say that they're different. In fact, very few traditional humanists would call their source material "data."

You may have heard of this explosive piece in the LA Review of Books in October 2012: **Literature Is not Data: Against Digital Humanities by Stephen Marche**. Although the wording is quite exaggerated, it does

a good job in expressing the sentiments of many humanists towards the idea of data in relation to their actual work.

When you call something data, you imply that ...

- it exists in discrete, fungible units
- that it is computationally tractable
- that its meaningful qualities can be enumerated in a finite list
- that someone else performing the same operations on the same data will come up with the same results.

This is not how humanists think of the material they work with.

This is not a perfect analogy, but imagine that someone called your family photograph album a dataset. It's not inaccurate per se, but it suggests that this person just fundamentally doesn't understand why you value this artifact. And it's the same with humanists. With a source, like a film or a work of literature, you're not extracting features in order to analyze them; you're trying to dive into it, like a pool, and understand it from within.

Or: imagine if someone referred to your personal diary as a database. Technically, this isn't wrong, but it implies a significant misunderstanding of the diary's sentimental and personal value to you.

Let's take my silent film example again. It would be possible to enumerate all of the filmic conventions that recall the conventions of melodrama. Is there a villain? Is there a heroine? Are good and evil depicted in stark, black-and-white terms? You could even build a dataset like this and use it to show how film changed over time.

But, seriously, who cares? **There's just such a drastic difference between the richness of the actual film and the data we're able to capture about it.**

A dataset like this is so much less interesting than the trained judgment of someone who's seen many of these films and can turn a nuanced observation of these changes into a real argument.

However. Things are changing, in ways both obvious and not. **All of our stuff is on our computers now — all of it, from books to movies to archival documents.** This is why, more than anything else, I think digital humanities is here to stay. If you can analyze something computationally, I think it's going to be really hard to tell people that they shouldn't.

This state of affairs has created some real problems for humanists, and, I would say, some real opportunities for libraries. If you speak to any historian who works in an archive, I guarantee you that they have hundreds, maybe even thousands of photos shot in an archive that look like this:

This is it! This is how historians are organizing hundreds of archival photographs!

It's not just historians who have a problem. Literature scholars, film scholars, **everyone's dealing with lots of journal articles, video clips, and other sources, and are really struggling to organize them so that they can produce scholarship.**

So humanists — even those who aren't digital humanists — desperately need some help managing their stuff, and libraries are in a great position to help them.

In many ways, digital humanists will have similar data-management needs to scientists and social scientists — they'll have spreadsheets, images, and video, and will probably at least know what metadata is. In addition, various funding agencies now requires a data-management plans -- this is something that humanists are going to have to start thinking about.

Just to give you a sense of the kinds of things humanists might do with structured data, I'll give you an example: Old Bailey Online. This is a **database of 197,000 criminal trials** held at London's central criminal court between 1674 and 1913. The database is structured in such a way that you can ask questions like, "How did the length of trials change over time?". One result of this work is that we now know that the length of trials has been increasing over time, and that this increase is not due to the increasing complexity of the cases. This is a really interesting result, and it's the kind of thing that you can only get from structured data.

Or take the example of the Google Books Ngram Viewer. This is a tool that allows you to search for words and phrases in a corpus of books that Google has digitized. You can see how the frequency of a word or phrase has changed over time. For example, you can see that the term "car" has overtaken the term "automobile" in the English language around 1910.

It requires some real soul-searching about what we think data actually is and its relationship to reality itself; where is it completely inadequate, and what about the world can be broken into pieces and turned into structured data? I think that's why digital humanities is so challenging and fun, because you're always holding in your head this tension between the power of computation and the inadequacy of data to truly represent reality.

2. N-grams and the Google Books Ngram Viewer

- Paper
- Student examples
    - Where did you get inspiration from?
    - What did you find?
    - What did you learn?
    - What did you find surprising?
    - What did you find confusing?
- University
- (My Shakespeare vs. Chaucer example)

5. General discussion on the readings

6. Could we ask ChatGPT to create these terminology narratives?

   e.g. "ChatGPT, can you explain to me the evolution of the term 'car' versus 'automobile' in the English language? What are the most significant events that explain the shifts in usage of these terms? When does the term 'car' start to be used more frequently than 'automobile'?"

7. Data Biorgaphy assignment explanation

- Real world example of Krause