Katherine Rawson
*Presentation for "Digital Scholarship in Action: Research"*
*2016 MLA Convention | Austin, Texas*

Today, I am going to talk about building humanities data sets for research.

Humanities data sets are often made by cobbling or accretion, sometimes both. Building these data sets is often an activity that involves many people.

They are
- institutionally built,
- made by collections of scholars,
- made with the help of the public through crowdsourcing.

Examples:
HathiTrust Worksets - digitizing libraries, HathiTrust, scholars
Curating Menus - librarians, librarians, crowds of transcribers, scholars

The peopled nature of this work can be a virtue or a complication (often both).

So if data sets are made by many people: what are the stakes for humanities researchers?

As scholars, we want to:
- maintain research value
- act ethically around the work of others
- attend to the ways that the perspective of those involved shaped the data set

I am going to focus on the later two (in ways that I think support the first).

Attend to the Work of Others

We want to act in ways that are ethically and intellectually responsible.

Things like collaborators' bills of rights can help:
NEH: http://mcpress.media-commons.org/offthetracks/part-one-models-for-collaboration-career-paths-acquiring-institutional-support-and-transformation-in-the-field/a-collaboration/collaborators%E2%80%99-bill-of-rights/
UCLA: http://www.cdh.ucla.edu/news-events/a-student-collaborators-bill-of-rights/

But sometimes the work is already done. What about when collaborators are not living or are not named?  In this case, we might consider narratives of acknowledgement. For example, for Curating Menus, which relies on the work of anonymous people who contributed to NYPL's What's on the Menu project and on librarians and donors who are now deceased, we decided to write the story of these people and to create a data dictionary that includes a place of agents. For each kind of data, we not only describe what it is, but who made it.

Finally, if we are using the work of others, we should — as much as possible — share our data. This is not unlike the notion of the public trust that Lisa Rhody discussed yesterday — only the bar is lower.  It's much easier to share data than maintain tools and projects.

Attend to the Perspective of Others

How do we attend to the ways that the perspectives of the people who made them shape our data sets?

Miriam Posner raised this at Keystone DH as she discussed the nature of diversity and contingency of critical humanities discourse in comparison to the data sources we sometimes have to work with — what do we do about the fact that the census (and numerous agents involved in the census) say that gender is a binary? I think this is an important and promising question for digital humanities scholars. Today, I am going to offer thoughts on one approach.

In answering these kinds of questions for Curating Menus, Trevor Munoz and I did a few things. First, we researched and wrote about where our data came from.  Then we began experimenting with data structures.  We have been working with indexing, because this allows us to maintain the frameworks of the other people who made the data (and to explicitly acknowledge their frameworks), while adding our own.  Instead of modifying the data that came before us, we are layering on top of it. The goal is to do this within a linked open data framework, so that many people can transparently add to, connect, and manipulate the data, while being able to see the hands of others in it.