Refelctions on Data Cleaning

- Reflection:
  - Schöch, Christof. “Big? Smart? Clean? Messy? Data in the Humanities.” *Journal of Digital Humanities*, vol. 2, no. 3, 2013.
  - Rawson, Katie, and Muñoz Trevor. “Against Cleaning.” *Debates in the Digital Humanities*, University of Minnesota Press, 2019, pp. 279–92.
  - Broman, Karl W., and Kara H. Woo. “Data Organization in Spreadsheets.” *The American Statistician*, vol. 72, no. 1, 2018, pp. 2–10.

James Sowerby

1:28 PM The readings for Thursday all revolved the interesting tension between "big" and "smart" data, most clearly elucidated by Schöch's article. On the one hand, big and messy data reveals patterns and details that could be lost in the kind of "data cleaning" that Rawson talks about. The example of potatoes au gratin was particularly relevant to this. Instead of combining all of the different menu listings for this one dish, they were able to theorize about the project's "scaleability" using a framework from Anna Tsing. All three articles seemed to suggest something similar—smart data, which is commonly thought of as the only alternative to big and messy data, is not helpful either. It encapsulates and disguises the questions of ethical preservation and presentation that we talked about in other readings. It appears then that neither approach is really helpful. Big data requires strong media literacy and computing power, while smart data can present facile depictions. That's why I thought that Schöch's suggestion of "smart, big data" or "big smart data" was well-deserved. I'm still not exactly sure what that would look like (perhaps a useful topic in class discussion?) but think it has potential.

Pippa LaMacchia

4:06 PM The articles for Thursday were fascinating to read in tandem because I feel like each touched on the same subject with different depth and application. To begin with, the questions in the Schoch article become a baseline what we are studying and discussing — what does data in the humanities even mean? Who has the authority to make decisions about how data is digitized and represented? I was particularly struck by the question of whether or not we can even use the word “data” for research in the humanities. The notion of finding a balance between “big” and “smart” data trickled into the Rawson article because there is a similar distinction between “clean” and “messy” data. Rawson concludes that perhaps messiness is simply another aspect of the humanities and therefore a balance between the two is the most accurate and logical representation of digitized data (in conjunction with Schoch’s conclusion). Finally, the Broman article grounded our readings in a concrete understanding of what these two sides of data gathering even look like. Again, I came away with the understanding that these discussions must happen in order to make the transition to digitized study one of ethical and moral value through its accuracy and clarity.

Alison Fortenberry

7:57 PM I really enjoyed the Rawson and Munoz piece. The idea of cleaning data does seem to go against classic conceptions of what research in the humanities should look like, but classic approaches to the humanities naturally need to evolve with the advent of new technologies and the introduction of data into humanistic study. While I understand Schoch’s argument about big, smart data, and see the importance of having clear datasets, Rawson and Munoz's argument that the process of cleaning implies that there is a right form of data and a wrong form of data, which I would agree is not true in the humanities. As they further argue, even though an aspect of data may seem irrelevant to a specific project, it doesn't mean that that form of data is inherently irrelevant. I really like their idea of embracing the mess of humanistic data studies rather than trying to evolve the humanities into something they aren't. I wonder if there is a middle ground to these approaches. Maybe maintaining raw, non-smart/cleaned datasets and making versions of these datasets that are smart for research clarity. Something like that could balance the mess of the humanities with the need for organization. Is Rawson and Munoz's indexing system an answer to this debate, or is it still too unclean/unsmart? At first glance, I really appreciated its ability to help document non-scalable data and balance the mess of the humanities with the need for order with large datasets, but I'm not sure how practical of a solution this would be for authors like Schoch who want very clear datasets.

Pia Bhatia

2:19 PM (for the 20th) The readings outlined many challenges in digitizing humanities data for me -- it appears as though each object would need to be treated with an exacting precision (and there seems to be a wide scope for misuse and misrepresentation here), and also additional training that ensures that the public can fully access these hypothetical resources. I'm not sure if the result these authors propose would really rectify the issues they are laying out, both ethically and practically. I also think these articles showed me how opaque the digitization processes really are to me, and I think I would have to understand what can and cannot be done with a clean dataset prior to evaluating whether the data should have been cleaned

Clay Glovier

4:47 PM I enjoyed reading Schoch, Rawson, and Broman's articles. Schoch in particular raised some intriguing questions, asking whether data replaces books, paintings, movies, and other forms of art. I think that data is a representation of these things, rather than their replacement. When one reads a book and then adds it to a dataset, they are not just absorbing the work itself, but allowing it to be represented as a datapoint in a separate study. The book is not being lost or replaced, rather its essence is being represented to help further a project. The reader of the book and the data can appreciate the text's content and the more informed model that its presence as a datapoint facilitates. In Rawson's article, the author discusses how scholars are skeptical of digital humanities in part because of the process of data cleaning, which they fear may reduce the complexity of their work. I would argue that data cleaning actually can broaden the scope of the information collected, as by sorting different variations of a phrase into one value, models ensure that atypical spellings are included in the result. If the data was left "messy" these different spellings could remain obscure and not be included in the research being conducted. Broman's text was relevant for me as I have used Excel often for work. The software definitely has flaws, in that it is easy to make errors and hard to notice and correct them. Rawson's tips for how to keep data sets clean and accurate will be helpful as I use Excel once more this summer!

Colin Brown

9:15 PM The readings this week took a zoomed-out perspective on the characteristics and roles of data in the humanities. Schoch, Rawson, and Munoz all make intriguing insights and arguments, but the biggest thing that stood out to me was that these articles were taking this perspective in the first place. From an engineering perspective, I've rarely heard an engineer or scientist talk philosophically about their relationship with data and the use of it as a relationship between them and their object of study. As such, it brings a new perspective when humanists encounter data and bring their methods of analysis to the matter.

I think a neat advantage of using data in the humanities versus STEM is that the practitioners of this data are now more inclined to analyze metaknowledge, stories, and social impacts. In other words, humanists can perform humanistic work on their data and create a whole new analysis out of that. As someone who majors in stuff that centers around gathering and studying data, these papers made me think about that work in a new light.

On a more specific note to the readings, I believe that Rawson and Munoz make valid claims about what can be lost through data cleaning, but I wish they had more of a focus on distinguishing between when data cleaning helps and when it hurts. For instance, data cleaning is essential for, say, curating a training data set of medieval texts for a machine learning model, but would be devastating for a study that tries to understand culturally-specific nuances in indigenous writings. In other words, I think this paper is better seen for its ability to point out that both data cleaning and raw analysis are useful, they are just used to yield different findings.

Raphaela Gold

10:01 PM The theme of this week's readings that most stuck with me was that of the intersection between language and digital humanities. All three of the articles emphasized the language that scientists, humanists, and the general public typically use to discuss data. Schoch, for example, focuses on phrases like "big data" and "smart data", both of which have entered the vernacular with very little context or understanding. I was struck by the lengths to which Schoch went to define these terms which, while not exactly self-explanatory, ultimately seemed to mean exactly what they sounded like. While I hadn't known about the "3 V's" with regards to big data, they all make a lot of sense as factors to consider. I was also curious about why these specific authors chose to focus on language, and thought that perhaps it indicated that humanists were their target audience and these authors were using language as a lens through which humanists could appreciate data. This brings me to another point: I've noticed that many of the articles we've read so far have assumed that humanists are fundamentally against data. This really surprised me, because ironically, the authors don't back up their claims with data. For example, in "Against Cleaning", the authors write, "when humanities scholars recoil at data-driven research." Do humanities scholars recoil at data-driven research? There is definitely skepticism toward digitization in the field, but I'm not sure that's the same as skepticism toward data. I was also left with a lot of questions after the data cleaning article about the relationship not just between data and humanities, but between data and the humans interfacing with it. I mentioned this in an annotation, but I thought it was interesting how while we might consider digitization a step society is taking to eliminate the human in favor of the digital, they can actually work in collaboration with one another. A human is presented with data, and based on the data, the real human being must make a decision. But how does one use data and make these decisions? The third article, "Data Organization in Spreadsheets," definitely answered that a bit for me. I noticed that it was the first article that provided concrete rules for data analysis and use. How did these rules originate? Is there any disagreement over these rules within the field, or are these just the basics which everyone subscribes to? (edited)

Melissa Woo

12:24 AM I find myself resonating with certain aspects of each of the articles while also holding some reservations. Schöch's perspective on the value of embracing the messiness of data in the humanities makes a lot of sense. I believe it's important to preserve the richness and complexity of human culture and history, which aligns with his argument for the potential insights that can be derived from unstructured data. I agree that traditional cleaning and standardization processes may not always be suitable for all types of humanities data, and that new methods and tools are needed to effectively analyze such data. Similarly, Rawson and Muñoz's caution against overzealous cleaning of data also strikes a chord with me. Their emphasis on the potential loss of valuable cultural and historical context through subjective cleaning processes is a valid concern. I agree with their proposal for approaches that preserve the messiness of the data to allow for more inclusive and nuanced interpretations. But, while there are indeed risks associated with over-cleaning and the potential loss of valuable information, I believe that there are instances where a certain level of data cleaning may be necessary to ensure the accuracy and reliability of analysis, especially when dealing with large and complex datasets. It seems that, in practice, a balanced approach is needed, one that acknowledges the value of messy data while also recognizing the importance of thoughtful and transparent data cleaning processes when necessary. This would promote our ability to derive meaningful insights from humanities data while preserving its inherent richness and diversity.

Anya Kalogerakos

1:14 PM In reading, “Against Cleaning,” the first thing that struck me was how normalized the process of data cleaning is in other fields besides humanities. Although it is called data cleaning, I don’t often think of the process of correcting and narrowing fields of data in more data/science-driven fields as a process of erasing important data. Perhaps this is because most scientific research with data is trying to prove a very narrow point, as opposed to, say, digitizing an aspect of culture. As “Against Cleaning” mentions, typically, these procedures are executed in tandem with some sort of warning about errors and outliers. I think it would be interesting to see this type of warning used in humanities data, although I fear it would be more complex, as there is not a simple formula to explain what might have gone wrong during the data collection process. I felt the conclusion of this article had the most valuable points, as it referenced how cleaning often prioritizes structure over data integrity and that this sort of structure, while allowing for scalability, suppresses diversity. While diversity has always seemed to be the enemy of a data set (such as the dreaded outlier in an otherwise well-distributed dataset), I am excited to see the field of humanities challenging that notion by finding ways in which these outliers can matter and be used for reflection on the structure of our data and the communities they represent. The distinction between big and clean data also may be able to help resolve the issues of outliers, as outliers can be valuable in big data and both are necessary for optimal humanities research, as Schoch suggests.

Andrew Huo

2:47 PM This week's readings and the data biography project have opened my eyes to the importance of data formatting. Coming from a strictly humanities background, it was interesting to understand what Trevor Owens said, that data is “a multifaceted object which can be mobilized as evidence in support of an argument.” That data itself is only a small part of the whole picture. Who collects it, why, and what format from the spreadsheet specifics of Broman and Woo’s article to the distinction between CSV and JSON, and GitHub as a public domain that constantly evolves and changes how data is observed and accessed. Another point that fascinated me in Schöch’s article is that the “apparent empiricism of data-driven research in humanities seems at odds with principles of humanistic inquiry.” The idea of ‘context-dependent interpretation’ is imperative in the study of humanities and thus, this conflict of simplifying information to serve humanities is an interesting conflict. All in all, it makes realise the importance of data biographies especially in the humanities.

Talia Goldman

1:02 PM While reading for this week, I was particularly interested in how both Schöch's Big? Smart? Clean? Messy? Data in the Humanities and in Rawson and Muñoz's Against Cleaning address the apparent tension between data and meaning in context, a crucial consideration in all humanities fields that Rawson and Muñoz would describe as "nonscalable." I appreciated–especially in Against Cleaning–the head-on tackling of this issue. I was intrigued by Rawson and Muñoz's suggestion of an index as an approach to embracing "messiness" in the digital humanities, and wonder how this interacts with Schöch's idea of taking ideal qualities from both "big" and "smart" data, which considers messiness/flaw/context but seems to advocate more strongly for cleaning or organization. Together with these readings, the "Data Organization in Spreadsheets" reading was a welcome look into concrete steps DHers can take to minimize errors in their data, which may help mitigate issues posed in the other two readings, like that idea that big data can mask flaws or the problem of inconsistencies in inputting data into databases as in the What's on the Menu? project. Additionally, some of the ideas in this week's reading reminded me of the podcast featuring Dot Porter, especially the idea of scalability and nonscalability, since the physical properties of a book are (though zooming in and other possibilities help) nonscalable. The Rawson and Muñoz reading further impacted my understanding of what digitization struggles to capture in discussing the experience of looking at a menu, leaving me with the question: how might DH be able to account for how people interacted with objects at the times of their making?

Ethan Haque

2:33 PM Out of the readings this week, I liked Against Cleaning and Data Organization in Spreadsheets and felt conflicted about the Big? Smart? Clean? Messy? Data in the Humanities article. I was also a little skeptical of the Data Organization in Spreadsheets article because dogma in technical fields is never meant to be strictly adhered to but is presented that way. I think the Against Cleaning article puts several ideas I've encountered before in a eloquent way and frames the idea of cleaning data well. The author started off by saying something about how data cleaning is just a stand in for a huge amount of work that goes into data intensive research. That something like 80% of the work on these projects is just getting the data into the right format and the 20% is the analysis. Right after they said that, I felt like I strongly agreed with the viewpoint of the author and enjoyed the rest of the article. The data processing steps are so time intensive and require so much work but there is often so little to show for them and people only care about that 20% of the work that is the analysis. I've dealt with tons of awful data in the past and I think the author of the Data Organization in Spreadsheets makes several great points that so many researchers could benefit from. I think if there is one thing to take away from that paper it's that consistency is the most important aspect of preparing data for analysts. I don't think the rest of the more fine-grained rules they present are true all the time, but the need to be consistent is something that is always true.

Layla Williams

7:52 PM Data is a complicated concept! It extends beyond the debate of what exactly we consider as data and into the methods in which we interact with the data. For example, this example of "cleaning" data is something I find helpful to investigate; however, I do not quite know where I might stand on the debate of whether or not we call it "cleaning" the data. I see where the negative connotations of having "messy" data that needs cleaning can come from. I imagine how it might also discourage researchers from recording additional data when they do not know exactly what to do with it for fear of having to "clean" it in the future. Despite this, I think the action of "cleaning" data can be especially fruitful (from at least what I understand based on the articles). You can narrow down the questions that you are interested in based on how you end up refining data in future iterations of the research. It also familiarizes you with your dataset by allowing you to conduct a preliminary analysis. Maybe the opposition to “cleaning” data simply comes from the language we use to discuss it. For example, what if we called it “dressing” the data? As we refine our ideas, we are dressing up the data in a way for it to be presented. And this implies that we are contributing to what we already have rather than trying to chisel away. This is a wild suggestion!

Helen Gao

10:23 PM The article “Big? Smart? Clean? Messy? Data in the Humanities” defined some important terms, and I found the description of ‘smart data’ to be especially interesting, since it was a new term for me. Additionally, some of the points (such as the amount of written materials that have yet to be digitized, the challenges of dealing with different formats for data, and the difficulties of OCR) reminded me of the discussion that we had with Will Clements last week. “Against Cleaning” was an interesting article as well, and I enjoyed reading about the authors’ experiences with data cleaning using a tool that we are going to use in class, though I was surprised by OpenRefine’s apparent limitations. I also appreciated how the article considered the different kinds of messiness that arose during the process, which highlighted just how much choice is actually left to people working with data (as well as how arbitrary data cleaning can be sometimes). The statistics on the rates of spreadsheet errors that “Data Organization in Spreadsheets” mentioned were very shocking! As I was reading the article, I thought that the suggestions seemed relatively common-sense, but keeping the statistics in mind, I wondered if they are necessary because people performing data entry may not necessarily be trained in the area and/or lack experience working with data, so they might not understand why rules such as consistent variable names are so important. This article reminded me of “Datasheets for Datasets”, in the sense that both are comprehensive lists of considerations when working with data (though this one was less open-ended and more prescriptive), and I have a similar concern that it will be difficult to convince people to actually adhere to these suggestions.

Yaashree Himatsingka

3:30 AM I appreciated how both Schöch’s article and Rawson and Muñoz’s piece encourage us to accept and explore the natural messiness of data in the humanities. It seems to me that embracing that messiness (I’m intrigued by @Layla Williams’s suggestion of “dressing” rather than “cleaning” data) can actually allow us to uncover richer, more accurate insights into human culture and experiences. Schöch specifically considers the debate around big and small data and suggests that the focus shouldn’t just be on how much data we have, but also on how we organize and understand it. And because data in the humanities is supposed to reflect a range of conditions or cultures, it’s bound to be a bit messy. Rawson and Muñoz also push back against the impulse to fit data into a neat mold. Their argument that the process of “cleaning” data, which involves removing outliers or inconsistencies and might actually strip away important details, was especially compelling — because analytically convenient or “clean” data can be (evidently often is) inaccurate and misleading when it comes to representing the full spectrum of human experience. (edited)