USING DIGITAL METHODS TO ANSWER HUMANITIES QUESTIONS ABOUT MIGRATION

SAARA KEKKI
PhD, University of Helsinki

KAI FERRAGALLO-HAWKINS
MSc, University of Helsinki

This article examines how digital methods can strengthen replicability, reusability, and collaboration in humanities research while preserving the individuality of human subjects. Using *Community in Motion* — our project on the postwar dispersal of over 120,000 incarcerated Japanese Americans — as a case study, we address the challenges of working with extensive datasets, including War Relocation Authority records and the 1950 US Census. We demonstrate how algorithmic matching, supervised machine learning, and version control systems such as Git enable error detection, workflow control, and scalable analysis. Additionally, we argue that digital methods, when adapted to humanities inquiry, can expand the reach and reliability of scholarship while keeping personhood at its center.

Keywords: digital humanities; historical data; reusability of humanities data; Japanese American incarceration

Introduction

Datafied research can neglect the people it studies, but this should not turn researchers away from the adaptation of digital tools. We argue that incorporating digital tools can be done without sacrificing the individual personhood of those who construct the data while still improving the accuracy, accountability, and quality of the research. In fact, we believe that the ability to contextualize and bring narrative to numbers is precisely what the humanities scholar has to offer to migration research. It is also for this reason that we use the word "data" in a broad sense: not only as the word to describe numeric or quantitative data but also more qualitative material and combinations of several types of data.

Sources related to migration typically lend themselves well to a data-oriented approach: they almost always involve multiple factors (economic, political, cultural, environmental) and thus generate large and varied datasets. Coupling these large amounts of data with humanities questions and practices allows researchers to write the stories of people at scale. While studying macro-level patterns – such as migration flows and networks – we can zoom in on micro-level narratives.

We explore this by addressing how digital methods can best facilitate research reproduction, handle errors in historical data, and create cooperation between researchers with non-digital and digital backgrounds using examples from our project, *Community in Motion*, funded by the Research Council of Finland (2023–2027).

Community in Motion studies the post-World War II migrations of the 120,000 incarcerated (interned) people of Japanese descent in the United States. After Japan attacked Pearl Harbor in Hawai'i in 1941, the United States government deemed any Japanese person on the country's West Coast a potential security threat. Without any evidence of subversion, the government established ten civilian concentration camps, incarcerating anyone with Japanese ancestry. As 2/3 of the Japanese American population had been born in the United States, they were US citizens and thus the incarceration process was unconstitutional. Once in the camps, the inmates were coerced to disperse across the continent or to Japan. (Kekki 2022.)

Data on Japanese American Incarceration

This article focuses on methodological aspects of using historical data. Despite this focus, we will draw examples from the subject matter of the research.

The work currently relies on three datasets, two of them collected by the War Relocation Authority, which administered the camps, and one by the US Census Bureau. The first two are the "entry data" collected at each of the ten permanent concentration camps in the fall of 1942 and the "final roster data" that registered the inmates' departure details. The entry data is more comprehensive in that it asks details such as education, occupation, and religion of individuals, as well as the occupational information of their fathers in the United States and in Japan. These data cover most of the 120,000 incarcerated individuals.

The census, meanwhile, is collected of everyone in the United States every ten years. Census records are closed for 72 years, which means that the 1950 census only became available, and usable, in 2022. Images of the original handwritten records are provided open access, and the private corporation Ancestry gives the ability to view and search through transcriptions of the records for a fee. In order to get access to Ancestry's data for Community in Motion, we turned to the IPUMS USA (originally, Integrated Public Use Microdata Series) at the University of Minnesota, who manages the transcriptions in partnership with Ancestry.

The War Relocation Authority data is straightforward for a researcher to use, as each inmate has an individual number that was also built on a family number. Each family (sometimes taken very broadly to include extended kin or even friends and neighbors) received a five-digit family number (e.g., 12345). Family members were further distinguished by an additional letter of the alphabet, with the head of the household usually designated as "A" (e.g., 12345A), and subsequent letters assigned by order of birth. Using and matching these data thus mostly consists of duplicate and error checks. When it comes to matching War Relocation Authority data to the census data, however, we run into challenges because the census does not use such an identification scheme. Therefore, we created algorithms to ease the matching through the consideration of several factors, such as name, birthyear, and birthplace. The matching is covered in more detail in the second section of the article.

Our handling of these data–as well as the premise for this article–is based on the FAIR principles of data: Data should be Findable, Accessible, Interoperable, and Reusable. In this article, we particularly focus on the reusability of data, which means that data is prepared and documented ("metadata") in a way that allows other researchers to easily use or incorporate it into their own work.

Replicability and Reusability

Replicability – the ability to refollow the steps of other's research – and reusability – allowing the data for a research project to be available and used in new ways – are key elements for creating relevant research. However, the accessibility required for them is often forgotten in the humanities. There is a reason for this: humanities research focuses heavily on individual perspective, both from the researchers and the people they are attempting to understand. If a new work reaches the exact same conclusions without new interpretations or considerations, it would less suggest reflective analysis and more suggest plagiarism.

We argue that striving for replicability and reusability is still deeply relevant to the humanities in all of its forms. To us, the importance of research is that it expands everyone's repertoire of knowledge, and to do so it must engage with the community at large, allowing them to explore the methods and thoughts that built it. There is evidence that this provides benefits back to the research as well, increasing its reach and status (Colavizza et al., 2024).[1] Yet it is not simple for a researcher to write down every thought or record every change during a research process that can stretch months or years. What is a way, then, that greater accessibility and reusability can more easily be achieved?

Within programming, the tool turned to is Git, and it holds similar value for the humanities even when its projects incorporate no programming whatsoever. Git's main benefit is version control, or the ability to track changes to a project's work over time, allowing the researcher to see who changed a file and when, and to reverse to an earlier version of their work. Simpler file for-

1 Additionally, more understandable text, specifically in the social sciences, also appears correlated with the citations that a paper receives, though the evidence of this is more muted in highly established fields (Boghrati et al., 2023; Ante, 2022).

When a research project is built on the premise of replicability from the beginning, it also more naturally lends itself to reusability.

mats, like CSV, TXT, or Markdown function best, as they allow the researcher to see what specifically changed per version, and to have two people's changes to the same file merged together (with some caveats)[2]. However, heavier file formats[3] like Word or Excel still work for the version control benefits of Git.

Git allows sharing the full or a partial version history to others upon the conclusion of the research, and, as each commit – or saved version – allows for notes, it automatically creates an easy log of what was done. It helps solve issues where a mistaken copy-paste, delete, or corrupted file sets back a study's goals, and allows for greater coordination of work for teams. Additionally, Git, while having cloud services like GitHub or Bitbucket, can be set up in personal private servers where confidentiality is key.

We do not suggest that Git is always the right tool. Google Docs and Word have a more limited but still useful form of version control, and a smart backup system could handle many of the same dilemmas that Git solves. It may also not be the right choice: if you are not digitizing your work or your sources, there will not be much benefit. However, it is a tool whose benefits to replicability, reusability, and risk management can make it invaluable.

When a research project is built on the premise of replicability from the beginning, it also more naturally lends itself to reusability. The reusability requirement of FAIR data relies heavily on high-quality metadata. Metadata can be described as "data about data," and its purpose is to inform a viewer what they are looking at. For example, the metadata for census records would include information about the place and time the data was originally collected, where it was stored, and by whom or how it has been edited. Well crafted metadata allows a researcher to assess whether the data they are viewing is suitable for their interests and purposes.

Community in Motion was made significantly easier by the extensive metadata created for the Census and War Relocation Authority datasets and by the documented history that allowed us to understand the assumptions and biases of its creators. The more this information is provided in research, the more replicable, reusable, and useful it can be.

2 Certain changes to CSV rows are difficult for Git to merge, but external packages (like csvdiff) can solve this without much issues.

3 Heavier file formats are designed for the program they run in, not for the general computer. Because Git is not designed with Microsoft tools, it needs to read the file in binary, making it difficult to compare changes between different versions.

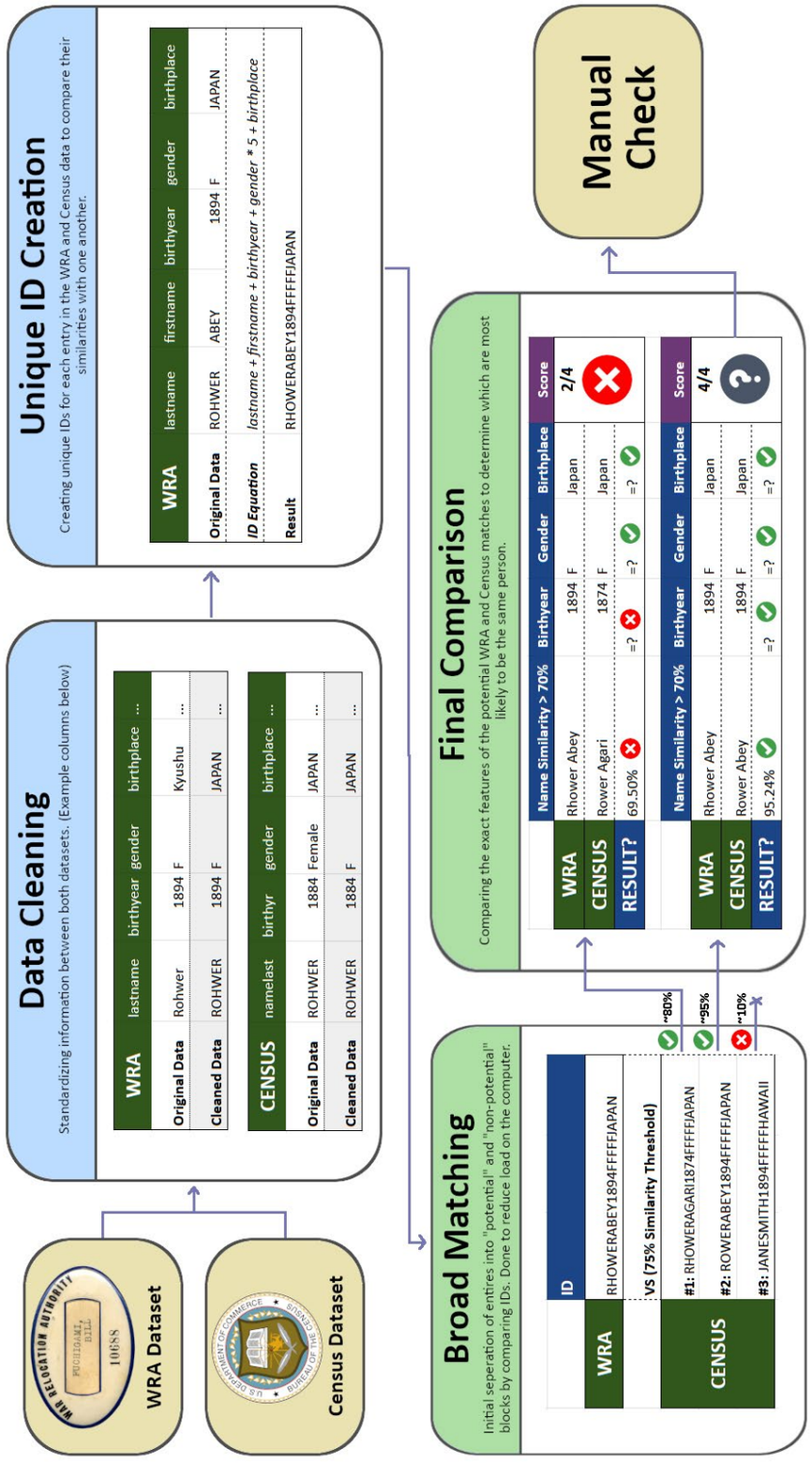

Figure 1. Shows the typical process for the traditional algorithmic approach used for matching individuals in *Community in Motion*.

Digital beliefs that focus so heavily on the results can, when handling people, forget that the numbers are a person and have relevance beyond the immediate information or goal.

Errors in Historical Data

While tracking changes to prevent researcher-created errors and creating metadata can serve the wider research community, they are not going to solve errors that already exist in historical data. These range from errors in the digitization of data (e.g., recording a handwritten table to a digital format) or incorrect spellings by the documenter (e.g., a census taker recording a foreign name wrong) to wear on analyzed documents that prevent a clean reading (e.g., stains on old newspapers or books). These errors weaken our ability to understand the text, and when they exist in great numbers, increase the dangers of contributing to the errors ourselves. For these cases, there are digital methods that can, when designed specifically for a study, help to significantly reduce the workload without taking the researcher's judgement out of the equation.

Within our research project, we have a subset of the 1950 US census records sorted for potential Japanese Americans. We want to match these people with their other entries in concentration camp data to be able to follow their lives. There are more than a million entries[4] in both datasets, making a full manual analysis simply untenable. While we can find a small subset that works with our research, perhaps getting a few hundred matches, there are tens of thousands of matches that exist in the data. Limiting the information we have available would weaken our ability to hone in on unique circumstances (e.g., exploring people with specific levels of education or family dynamics) and to provide this data for others to use.

There are two ways in which we approached person matching in our project. The first is a traditional algorithmic approach, with a goal to have a very exact output of potential matches based on predetermined conditions. After an analysis using the data's primary characteristics, such as name and birthplace, we created groupings of initial potential matches and then ranked those matches based on how many conditions they fulfilled. Because this system is designed by us, those conditions have a level of our own nuance: such that names could be slightly different to pass due to recording errors (e.g., Mary vs. Mari) but that their birthplace had to be in the exact same state (or country, if not in the US); or that the citizenship counts for less points due to less variety in the answers. A diagram of this process for our paper can be seen in Figure 1.

The second method of person matching is through supervised machine learning algorithms. We make decisions on what *could* be features – properties of the data that best explain a match – but leeway on how impactful those features are. Importantly, this is not generative AI, like ChatGPT or Claude, which can often make up plausible sounding but incorrect information when presented with data analysis tasks. This is because those bots are designed foremost to come up with novel ideas, not perform accurate error analysis.

When machine learning systems are trained on specific data or goals, they tend to be much more successful and potentially hold more noteworthy observations than a traditional algorithmic approach. In this case, we wanted to train the model to successfully detect when a person in the concentration camp data is the same person in the census. Our method was gradient boosting, which is a machine learning technique that creates a base prediction and then creates new predictions that correct the errors of the previous one.

To use this, we first needed to perform manual finds so that the model could be trained on what realistic matches are and create the "features" for what could be identifiers between the two entries. For instance, a researcher can add features for a name's exact similarity and phonetic similarity, and through testing, see which is a more successful indicator of a match. We also needed to ensure that our training data is representative of the matches we expect to see, and typically that it has hard negatives, or similar cases that are not matches, to ensure that the model doesn't find false positives. After training, this model was then applied to the data and ranked them probabilistically based on how likely of a match it is.

Both methods require the researcher's concrete knowledge of the corpora to fully contextualize what factors are or could be relevant. Both also require the researcher to look at the final rankings to get truly accurate and useful results. They do this while reducing a functionally impossible or dreary project to something much more achievable.

This research is also only a case study of how these tools can be used for historical error analysis but is not its only use. For instance, if you are handling diaries of migrants to track their journeys across countries, you could create features you think represent relevant differences between the diaries (e.g., the length of prose or emotions expressed), train a machine learning model and then use it to potentially discover new examples in unclassified or misclassified literature. Alternatively, if you have examples of misspelled migrant names in record data and you are trying to understand if a pattern exists, you can set the specific parameters for how you are expecting these misspellings to emerge and algorithmically see if it exists. These approaches can be intimidating without experience, but their potential for error solving makes them an excellent investment.

Perhaps the greatest benefit harkens back to reusability. We aim to provide these results to Densho, a non-profit that manages data on Japanese American incarceration, so that researchers in the future can much more easily follow the lives and stories that they may want to explore but would not be able to without more complex digital systems. This would not have been possible if we did not approach our research with the digital tools at our disposal.

Conclusion

In data science, there is almost always a pre-determined goal in the analysis or production of data. Its approach tries to find the data from the question, getting results where there is none and finding a concrete, actionable answer. This is different from the humanities, which often tries to find the question from the data, approaching it with much less intent to solve anything in particular.

Digital beliefs that focus so heavily on the results can, when handling people, forget that the numbers are a person and have relevance beyond the immediate information or goal. However, there are also takeaways from such a results-focused approach. When we are finding the question from the data, it begs the question of why we are doing this in the first place. Are we trying to help the community we are in, or are we trying to increase knowledge in general? And if the latter is the case, is the approach being used really of value towards accomplishing that goal?

If everyone thinks in the same framework around us, it limits the options one's mind considers even possible for research. It also causes us to miss the depths of a person's thinking educated with different goals and purposes. It is not just that digital tools can help provide greater benefit for humanities researchers – it is that humanities researchers can provide an introspection not always taught for those engrossed in digital fields. It is greater cooperation that can push forward our understanding and research as a whole.

Bibliography

Ante, Lennart. (2022) The Relationship between Readability and Scientific Impact: Evidence from Emerging Technology Discourses. *Journal of Informetrics*, 16 (1). https://doi.org/10.1016/j.joi.2022.101252

Boghrati, Reihane, Berger, Jonah & Packard Grant. (2023) Style, Content, and the Success of Ideas. *Journal of Consumer Psychology*, 33 (4),688–700 . https://doi.org/10.1002/jcpy.1346

Colavizza, Giovanni, Cadwallader, Lauren, LaFlamme, Marcel, Dozot, Grégory, Lecornay, Stéphane, Rappo, Daniel & Hrynaszkiewicz, Iain (2024) An Analysis of the Effects of Sharing Research Data, Code, and Preprints on Citations. *PLOS One*, 19 (10). https://doi.org/10.1371/journal.pone.0311493

Kekki, Saara. (2022) *Japanese Americans at Heart Mountain: Networks, Power, and Everyday Life*. Norman: University of Oklahoma Press. https://library.oapen.org/handle/20.500.12657/57770

4 The incarcerated Japanese Americans constituted over 90 percent of people of Japanese descent in the United States, but the large size of the dataset is due primarily to incomplete enumerations in race and birthplace.