Laying the Foundation

White, John W., Gilbert, Heather

Published by Purdue University Press

4 | *Curating Menus*: Digesting Data for Critical Humanistic Inquiry

Katherine Rawson

INTRODUCTION

Beginning in 2011, people across the United States came to the *What's On the Menu?* website and typed in snippets of text—names of dishes and prices on menus that ranged from the 1850s to the 2000s. They were working from images of menus digitized and held by the New York Public Library (NYPL). Out of curiosity, interest, or school assignment, these people were building a data set of over one million points of information about American dining.

This data set, which continues to grow, is a treasure trove for researchers, particularly those interested in twentieth-century America and its food culture. Anyone can easily download the data set from NYPL's website; however, the data is not easy to use: though the set is structured, the information in it is messy. Because untrained volunteers typed the menu item data in a free-text field, it contains an array of orthographic variations. The menu data, much of which was created by an earlier team of volunteer transcribers working from handwritten catalog cards, is also highly irregular.

Propelled by the promise of the data despite its messy state and by the investments of the many people who created it, *Curating Menus,* the project that is the focus of this chapter, aims to make the data more usable for researchers. It does so by beginning with a framework of critical inquiry about the data.

Curating Menus is an ongoing research and data curation project that relies on the New York Public Library's *What's On the Menu?* data. Its goal is to produce and foster scholarship about food and foodways in the twentieth-century United States by cleaning, indexing, and presenting the *What's*

On the Menu? data for analysis by scholars (including ourselves). The project consists of two key parts: humanities research and data curation. However, these parts are not completely distinctive. They are recursive, shaping and informing each other, and most often, they are integrated. *Curating Menus* is as much about fashioning structures of knowledge as it is about the many technical pieces that we will use in the process.

Curating Menus uses a recursive and integrated structure of knowledge that acknowledges the many people involved in creating the data set. It aims to maintain the information that these different people produced—from the process of collecting the menus in the early twentieth century to the structure of the downloadable files from the NYPL Labs. *Curating Menus*' approach to data curation, then, is deeply informed by humanities methods and theories. In particular, feminist practices and feminist theories of the archive shape our project.

This chapter will explore three interrelated projects, all based in libraries. From 1899 to 1923, volunteer librarian Frank E. Buttolph collected thousand of menus for the New York Public Library. These menus were eventually digitized and became the corpus for *What's On the Menu?,* a crowdsourced transcription project developed by NYPL Labs. This chapter will describe the stakeholders for these projects and reveal the individual contributions to generating and curating the projects' data. This case study in data curation as cultural construction begins with two claims: there are traces of many contributors in our data sets, and a critical engagement requires us to see them. Ultimately, this chapter argues that scholars and librarians can and should structure digital projects in a way that reveals explicit engagement with these traces.

DIGITAL PRODUCTION, FEMINISM, AND CRITICAL HUMANISTIC INQUIRY

Despite its goal of cleaning and using a data set, the first product of *Curating Menus* was an archive-based research essay. The essay examined the life and work of Frank E. Buttolph. Because she collected and curated most of the menus, understanding her positionality and the culture she worked in is important to using data in ways that are rigorous. Beginning with cultural context—and believing that it is central to how we can use data to answer humanities questions—shaped how the *Curating Menus* team approached curating the data as well.

The work of feminist scholars not only framed our understanding of the history of Frank E. Buttolph, but it also provided ways of approaching digital data curation. In "Whence Feminism? Assessing Feminist Interventions in Digital Literary Archives," Jacqueline Wernimont explores how the development and format of two well-known literary digital projects, the Orlando Project and the Women Writers Project, constitutes a "feminist archive" beyond collecting women's writing. She considers the ways that the digital archive facilitates feminist structures. By providing documentation that makes editorial decisions and power visible, these projects push against a single authority in the archive and allow for the imagining of alternative interventions. Further, by presenting the technosocial scene in which these projects developed, Wernimont illuminates the feminist work that collaboration can do, transforming and distributing authority in the archive.[1]

In "Feminist HCI: Taking Stock and Outlining an Agenda for Design," Shaowen Bardzell presents similar structural understandings of how feminist frameworks can shape design in human–computer interactions. Three of the elements that she focuses on—pluralism, participation, and self-disclosure—align with those Wernimont identifies.[2] These principles influenced the approach of *Curating Menus*. Instead of "correcting" data or developing an authoritative data set, the project aims to maintain the contribution of multiple participants and to make those contributions clear—not simply as an acknowledgment of their work, but as a pluralistic and transparent approach to knowledge-making.

HANDS

As the product of 115 years of work and not one but two (maybe three) crowdsourcing projects, the *What's On the Menu?* and *Curating Menus* data is the cumulative work of many people.

Trevor Muñoz and I began *Curating Menus* in 2014. As we began to formulate questions that we could answer using the *What's On the Menu?* data, we wanted to answer the question "What does this data represent?" Armed with years of humanities training, we turned not to the cells in our spreadsheet, but to the people who made this data. Defined both as the origin and the record of origin, provenance is central to using humanities data in ways that are rigorous—to see the ways that it is situated historically, shaped by the people and societies that formed it.

When discussing our project's provenance, I sometimes say that Muñoz was looking for a food scholar to work on the data he'd been curating and that I was lucky to be that person. But our origin story is slightly more complicated. We are not actually filling in gaps for each other: we are both humanities scholars and librarians, with backgrounds in food culture. Despite our different educational credentials, we have worked on a range of digital humanities projects, hold less traditional library positions, and are fairly knowledgeable of and invested in food. I say this because our positionality—who we are professionally and culturally (and even what seem like trivial biographical notes: we were born three months apart)—impacts our research and the ways we clean and sort data for future use. Just as the lives of the other people who are part of this long story of food information shape what we are working with and how it can best be used, so do we.

Muñoz and I also understand that the way we choose to categorize and normalize data for search and analysis will shape what we and other scholars ask and see. Where will we decide to make distinctions? Are Chicken Marsala and Coq au Vin and Chicken with Wine Sauce a collection of related dishes? Or maybe thornier because of what seems—on both sides—so apparent: is a half of a chicken, a quarter of a chicken, and a chicken the same thing?[3] And what are the implications of us deciding so?

As the scholarship of food makes quite clear, our dishes and our meals are intimately tied to how we define ourselves and each other. *Curating Menus* will draw on the knowledge and perspectives of the people working in the many fields our data has implications for: food studies, history, cultural studies, environmental studies, and anthropology.

Before this project, Muñoz had already been working with the data, using it to train colleagues and graduate students in the humanities and in library and information sciences to curate data. After an initial data curation seminar, Muñoz and MLIS student Lydia Zvyagintseva developed a precursor project to *Curating Menus*, in which they began exploring ways to clean the data and categorize it for future researchers.[4] The project was framed as a prototype *for* content-interested researchers; our current work shifts the focus—we are simultaneously researchers using the content and developers of improved data resources.

Curating Menus also collaborates with a set of public librarians from the digital humanities-focused NYPL Labs, who developed and worked on the *What's On the Menu?* project. Over a dozen people at NYPL Labs and

other departments produced the infrastructure for this large-scale crowdsourcing transcription project of the library's menus. Since the project's launch, thousands of volunteers have transcribed and reviewed over 17,500 digitized menus.

A decade before *What's On the Menu?*, twenty-first-century librarians digitized the menus, and another set of volunteers transformed the paper records of the menus into a database. This earlier project understood the immense usefulness of being able to explore the menus by a variety of categories. By transcribing the collection's records from print catalog cards into a database, researchers could search by restaurant, location, and other metadata previously buried in the records.[5]

Both of these digital projects at the New York Public Library, as well as *Curating Menus*, relied on decades of work by librarians who acted as stewards of the collection. These librarians worked with scholars as they sifted through the thousands of sorted-by-date boxes of menus. They accessioned Buttolph's personal papers in the 1980s, including correspondences that trace the development of the collection and include information about the meals they represent.

Each of these digital projects was born from the work of Buttolph and the many individuals who donated the menus, in what was (if one forgives the anachronism) an early twentieth-century crowdsourced project. Buttolph was a teacher and translator from a small town in Pennsylvania who had a deep engagement with how to make and preserve history, particularly social history in the United States. Although she collected a range of materials in the twenty years she volunteered at the New York Public Library, her longest and most significant project was her collection of menus, which she believed, was for "future students of history." To obtain the materials, she corresponded with hundreds of people, placed ads in trade magazines, and worked with newspaper and journal editors to publish stories about the collection that encouraged readers to contribute their menus to grow it even further. She then cataloged and prepared the menus for preservation and access.[6]

These letters, articles, and catalogs are artifacts of the people who made the menus. They are the historical record of the restaurant managers, the cooks, the printers, the people who we are trying to get to, across a hundred years and a passel of formats, with our million points of data. The history of the collection matters because it reflects the ways that the

data was shaped and what it can tell us. For a large data set like this, it is important to understand how it was created and parsed over time. In this case, diving into the provenance provides detailed texture and insights into knowledge organization.

FINGERPRINTS

What traces are left on the data? How do we maintain meaningful traces while making messy data easier to use? It is no surprise that the data based on eight decades of individuals typing and retyping information is full of variation. In fact, the accuracy of the NYPL data is perhaps more impressive. The NYPL's downloadable data set includes information from three places: NYPL's metadata, the menu collection database, and the *What's On the Menu?* transcriptions.

The two key moments that introduced inconsistency in the data points were the earlier volunteer-made menu metadata database and the crowdsourced menu transcription project.

In the menu file of the *What's On the Menu?* data set, for example, researchers might encounter "Waldorf Astoria," "Waldorf-Astoria," "WALDORF ASTORIA," "waldorf astoria," "Waldorf Astoria Hotel," "Hotel Waldorf Astoria," "The Waldorf Astoria," "Waldorf," or simply, "Astoria."[7] Having standardized data that conforms to a controlled vocabulary would allow researchers eventually to run analyses about who used the Waldorf Astoria for their events, what the restaurant served, whether that changed over time or between groups, and how it compared to other similar establishments or to its sister establishment in Philadelphia. The material could also be combined with manuscript materials from the hotel, such as ledgers and recipes.

Collating the data by normalizing to a single name can be a problem. Not all similarly named places signify the same place. Though they stood on the three hundred block of Park Avenue in New York City, the Waldorf, the Astoria, the Waldorf-Astoria, and the Waldorf Astoria are different historical (though interconnected) establishments. Our goal then was to smooth out orthographic inconsistencies while maintaining meaningful variations in the data. This is at the heart of making good humanities data sets that can be machine queried: how do we keep the texture while smoothing out the inconsistencies?

We take two approaches. First, we maintain the original data point, and simply add more information to the data set. Second, for the new, normalized data, we decide what variation was significant. When are transcribers maintaining information that is meaningful, and when are the differences just manifesting differences in transcription methods—keeping capitalization or not, for example?

Curating Menus' solution to normalizing relies on a technical method and a research method. The data set has identifiable features that, almost certainly, do not signify difference. For example, in this set, variation in capitalization is almost never meaningful. These can be removed en masse, computationally. Second, we identify entities we would need to research. Given a list of similar place names, we study historical records—often beginning with the images of the menus themselves—to see if places or organizations are the same.

A similar issue happens with the food items. How do we deal with thirteen ways to describe a half chicken? Again, we can identify the things we are almost certain do not signify difference: "chicken (half)," "half chicken," "half of a chicken," "1/2 chicken," "Half chicken," and "HALF CHICKEN" are probably similar enough to smooth out their differences.[8] However, our data structure also keeps a record of the orthographic differences, in case they are of value to Buttolph's "future historians," who may be invested in representations of fractions or the economic status of word order or preposition use. We are also aware of how different the actual half chickens might have been. We or other scholars may be able to make judgments about the chicken's preparation based on other aspects of the menu, further historical research, or perhaps even an analysis of the other items on the menu.

While tools like Google's *Refine*, now *OpenRefine*, offer solutions for smoothing out these kinds of variation through pattern-based clustering, they can have scale limitations and don't provide a simple way to keep the original orthography and have a clean collection.[9] To find the matching selections of dishes across the data computationally, we built a small piece of software, which relies on *Elasticsearch*, and wrote a query that finds what we call "fingerprints."[10] These are words in a dish, without care to order, capitalization, punctuation, or some prepositions and articles. The name signifies a unique characteristic that identifies a dish (like a human fingerprint). While in the project's software code, these fingerprints allow

us to create more uniform data, they are also reminiscent of the smudges that let us know this data was crafted and shaped by people who had a stake in it being useful, people who believed in its worth.

DUSTING FOR FINGERPRINTS

While Frank E. Buttolph made sure that there were no fingerprints on the menus she collected, often returning submissions that had traces of food or dirt on them, we can still see all sorts of hands in her work. In handwritten and typed letters, in articles from the early twentieth century, and even in which menus are in the collection, we see the people who fashioned it. Our goal is to find ways to add these traces to the data set, while increasing the usefulness of the information in the transcriptions as well.

Curating Menus aims to reveal strata of meaning. Each layer in the data set shapes the experiences of another and provides the kind of rich resource that humanities scholars seek in their research. In addition to adding information, the many people who worked on this data set across the twentieth and twenty-first centuries also structured their data in ways that are significant, not only because they influence the validity of the evidence, but also because they suggest different kinds of questions. Being aware of those implicit structures of knowledge allows scholars to see the landscape of information and knowledge differently. Two of those organizational structures—Buttolph's catalog cards and the "What's On the Menu?" interface—demonstrate different kinds of readings of their objects.

When we started the *Curating Menus* project, the plan was to briefly discuss the contours of the data on our website, a precursor to digging into the data itself. Nonetheless, as Muñoz and I discuss in "When a Woman Collects," we found ourselves digging much deeper into the initial development of the collection, in part because we wanted answers to *why* the collection looked like it did. Given what we learned about the development of this research collection, we have a much clearer idea of the kinds of cultural questions Buttolph would have been interested in.

For example, understanding Buttolph's catalog cards is critical to understanding the overall project. Knowledge is structured in many ways, but metadata is integral to how people research in the digital humanities. Metadata makes it possible to make claims about the data or to perform comparative or other pattern-seeking analytical processes, be they

computational or not. A long intellectual and practical history with metadata is part of why digital humanities make sense in libraries, why librarians are DH scholars, and why DH scholars collaborate with librarians.

The *What's On the Menu?* data comes in four connected CSVs, structured around the menus, menu items (a transcribed dish), menu pages, and "dishes." Each of these has data from multiple sources, including the transcription data, metadata about the transcription and the menu created by the computer application, and bibliographic metadata from the cataloging and database of the menu collection.

In the file for the menus, there are columns for "place," "event," "occasion," "venue," and "notes." The separate category for sponsor and location reflects an important element of the original print collection on which the data set is based, and its origins can be found in Buttolph's catalog collection.

The Frank E. Buttolph menu collection includes eighteen boxes of menus and boxes of catalog cards that match each menu. Buttolph categorized and organized the cards by type of group that was organizing the meal or the occasion for the meal. Then each category (Masonic orders, for example) was organized by place (states, New York City). On each card is the sponsoring organization (the cards are further ordered alphabetically by this piece of information), the date she accessioned the menu, and the date and location of the meal (i.e., June 1, 1918; Bellevue Hotel). If Buttolph had more than one menu from the sponsor, those menus were also listed on the same card, with locations and dates.

In Buttolph's organization, it is more significant that both meals are from the Masons than that the meals occurred next to each other in New York City. The date of a meal is important enough to record, but not an element of organization at all. Although one does not need an explicit understanding of Buttolph's categorization in order to use the *What's On the Menu?* data set, knowing about her organization system may suggest more useful questions for research.

Her schema is simply *recorded* by the catalog cards, but her collecting practices are embedded in the very structure of the collection. This means two things: First, it exposes that there *are* questions that are appropriately answered by the collection at scale, and it gives a sense of what some of those questions could be. Second, it necessitates paying attention to subsetting

the data in ways that are not encumbered by (or conversely could focus on) her interests in social structures and particularly celebration, the nation-state, and civic organizations.

Buttolph's schema is embedded in the data, a featured demonstrated by my own experience with it. Before looking at Buttolph's catalog cards (which are held at the New York Public Library), I began organizing the menu data myself. It was apparent that there were two basic types of menus: (1) menus for ordering and (2) set menus for events. These different constructions of menus—a space for choice and availability versus a description of what would or did take place—reflect different food practices. Food events would often have been confined to particular invited guests who would be eating the same meal at the same time. Conversely, ordering menus are often from public establishments, where people eating together may have different meals and people in different parties would eat at different times. The information the menus include is also dissimilar (prices or not, for example) and signifies differently (event menus reflect decisions about structuring taste and theme, for example).

However, there were numerous menus that fit into a middle space: menus from steamships and railroads, for example. These menus had characteristics of each descriptive type. They were often without prices, and they were sometimes singular in what they offered. The experience of people eating and making food in these places was key to why they didn't seem to fit into my categories. The people on trains and steamships were not invited, like at an event; however, they also did not have access to an array of options, as one does in a cityscape of restaurants. We framed five basic types: restaurant, association/group, person, transit, and hotel. While these categories did not cover all the menus, they seemed to reflect the menus.[11] Buttolph's categories recorded in her catalog cards mapped on to these categories, and her metadata system also encoded the significance of event and daily menus, through both categories of organization and recording location and sponsor. Moreover, she considered the sponsor to be the more significant part of the menu, an organizational structure that suggests a set of questions quite different from those about restaurant development.[12]

Just as Buttolph's collecting and categorization practices shape our data set, so do the decisions of the NYPL librarians and developers as they created the framework and tools for the *What's On the Menu?* project. The group

decided that users would transcribe dishes and prices—the names of food and how much they cost. This information could be cross-referenced with metadata included in the digitization process to learn something about food history in the United States. The information that the NYPL staff decided users would record might seem self-evident for a menu transcription project; however, it reflects decisions to not include other types of information, which may also be important to researchers. There is no way of recording non-dish–related textual content—the taglines of restaurants, phone numbers and addresses, food categories, information about staff and management, any origin stories, pithy phrases, or citations of Bible verses. This kind of text can reveal a great deal about the kind of establishment the food was served at. The group decided not to include this information because it was much less uniform and because they were aiming to collect a volume of information with as little burden on the users who would transcribe the information as possible.[13]

In the *What's On the Menu?* data set, visual information, or design, is also omitted. In fact, many of the twentieth-century discussions of the Buttolph menu collection are about design. Buttolph herself was interested in the menus' pictures and materials: watercolors of airplanes, sketches of literary figures, silk pages, ribbons to bind, a range of handwriting styles and handmade fonts.

The data set omits information about the framing and layout of the menu where the dishes occur: are they listed as desserts, as appetizers, as roasts, as entrees? How do different menus divide their contents? Not having a space for this data in the set is part of the nature of shaping a project: resources are finite; to attend to one part, we jettison another. It also means that the data does not accommodate some kinds of work. However, this kind of information can still be tied to the data. The *What's On the Menu?* data set does this in two ways: it includes a link to the digitized menu page, providing relatively easy access to the image (which could be analyzed by humans or perhaps computer vision), and it includes information about the position of each dish on the page, making it possible to aggregate dishes based on where they are placed on a menu.

CONCLUSION

The decisions data creators and curators make shape what scholars can say and unmasks how digital humanities is formed by human frameworks as much as technological possibility and limitation. *Curating Menus* contends

with and makes accessible the structures of knowledge that we have found within the data set and that we are making. This part of the process of humanities data curation has several features.

First, *Curating Menus* adds information rather than correcting or overwriting it. In this way, it disperses authority and maintains plural notions of knowledge. Second, it aggregates materials that may be able to be added to the data set later, or may be in forms that cannot be added to the data set. This includes things like biographical information about Buttolph, which may ultimately be another feature or classification in the data set, but is currently a narrative. In addition to the images of the menus themselves, *Curating Menus* aims to digitize the letters from the Buttolph collection—mostly written to her, including contextual information about the establishments, menus, and sometimes even the meals they accompanied. Our goal is to link these letters to the menus, just as the dish data is linked to the images of the menus. We also want to include sample images of Buttolph's cards as well as annotated photographs of the catalog card collection in its boxes.

We are aiming to create a different kind of documentation for digital humanities projects. This documentation draws on the characteristics of both technical documentation and archival practices. Like the programming languages and tools we use, it includes documentation that tells about how to use the data and how it was prepared; however, we are also documenting in ways that reflect what the librarians, including Buttolph, have done: including biographical and historical information and analysis of the many people who made this data through essays and bibliographies.

The construction of the project acknowledges and connects knowledge structures. A simple version of this is the data dictionary we wrote in order to clearly identify the materials in the NYPL CSVs, which gives information about each of the categories of the data and where that information comes from. A more complex version of this is indexing that includes and allows for multiple information structures, with information about the provenance of those structures. This allows us to include things from Buttolph's categorization as well as NYPL's, to add our own, and to leave space for future scholars who may want to connect a wealth of other information including dictionaries of organizations, food sources, or environmental data.

NOTES

1 Jacqueline Wernimont, "Whence Feminism? Assessing Feminist Interventions in Digital Literary Archives," *Digital Humanities Quarterly* 7, No. 1 (2013).

2 Shaowen Bardzell, "Feminist HCI: Taking Stock and Outlining an Agenda for Design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010):1301–310.

3 For the record, our current answer is "no."

4 Lydia Zvyagintseva, "Organizing Historical Menus: A Data Curation Experiment," *MITH* (blog), June 31, 2013, http://mith.umd.edu/taxonomizing-historical-menus-a-data-curation-project.

5 Michael Lascarides and Ben Vershbow, "What's On the Menu?: Crowdsourcing at the New York Public Library," *Crowdsourcing our Cultural Heritage*, ed. Mia Ridge (Surrey, UK: Ashgate, 2014).

6 Trevor Muñoz and Katie Rawson, "When a Woman Collects Menus: Sifting Stories and Histories of Frank E. Buttolph's Research Collection," *Curating Menus* (April 2014), http://www.curatingmenus.org/articles/when-a-woman-collects-menus.

7 According to NPYL Labs's Ben Vershbow, 157 variations were encountered.

8 In the NYPL data, each spelling or form constitutes a dish, which leads to overlaps.

9 Trevor Muñoz, "Borrow a Cup of Sugar? Or Your Data Analysis Tools?—More Work with NYPL's Open Data, Part Three," *Trevor Mu*ñoz (blog), January 2014, http://trevormunoz.com/notebook/2014/01/10/borrowing-data-science-tools-more-work-with-nypl-open-data-part-three.html.

10 The normalization is being done with a small piece of *JavaScript* software we developed. This chapter does not cover the technical aspects of *Curating Menus*.

11 For names of sponsors that were ambiguous, we looked on the menus and Googled the name. This produced some surprises: What appear to be men's names are often department stores; "house" is more likely a hotel than a restaurant. It also presented a few conundrums, including this one: in what category is a casino?

12 Two of the most significant scholarly contributions using the Buttolph menu collection, before *What's On the Menu?*, focus on restaurant culture: Andrew P. Haley, *Turning the Tables: The Aristocratic Restaurant and the Rise of the American Middle Class, 1880–1920* (Chapel Hill: University of North Carolina Press, 2011), and a suite of essays by historian Paul Freedman.

13 Lascarides and Vershbow; Trevor Owens, "Digital Cultural Heritage and the Crowd," *Curator: The Museum Journal* 56, No. 1 (2013): 121–30.