Against Cleaning

KATIE RAWSON AND TREVOR MUÑOZ

Practitioners, critics, and popularizers of new methods of data-driven research treat the concept of "data cleaning" as integral to such work without remarking on the oddly domestic image the term makes—as though a corn straw broom were to be incorporated, Rube Goldberg–like, into the design of the Large Hadron Collider. In reality, data cleaning is a consequential step in the research process that we often make opaque by the way we talk about it. The phrase "data cleaning" is a stand-in for longer and more precise descriptions of what people are doing in the initial phases of data-intensive research. If you work with data or pay attention to discussions among practitioners who do, you have probably heard or read somewhere that 80 percent of that work is "cleaning" (cf. Wickham). Subsequently you likely realize that there is no one single understanding of what data cleaning means. Many times the specifics of data cleaning are not described anywhere at all, but instead reside in the general professional practices, materials, personal histories, and tools of the researchers. That we employ obscuring language like "data cleaning" should be a strong invitation to scrutinize, perhaps reimagine, and almost certainly rename this part of our practice.

The persistence of an element that is "out of focus" in discussions of data-intensive research does not invalidate the findings of such research, nor is it meant to cast researchers using these methods under suspicion. Rather, the collective acceptance of a connotative term, "cleaning," suggests two assumptions: first, that researchers in many domains consider the consequences of whatever is done during this little-discussed (80 percent) part of the process as sufficiently limited or bounded so as not to threaten the value of any findings; and second, relatedly, that there is little to be gained from a more precise description of those elements of the research process that currently fall under the rubric of cleaning.

As researchers working in the domain of data-intensive humanities research, we have found that these assumptions intensify suspicions about knowledge claims that are based on "data." In fields where data-intensive work has a longer history,

researchers have developed paradigms and practices that provide de facto definitions of data cleaning (Newman, Ellisman, and Orcutt). In the humanities, however, these bounds are still unformed. Yet the humanities cannot import paradigms and practices wholesale from other fields, whether from "technoscience" or the nearer "social" sciences, without risking the foreclosure of specific and valuable humanistic modes of producing knowledge. If we are interested in working with data, and we accept that there is something in our work with data that is like what other fields might call "data cleaning," we have no choice but to try to articulate both what it is and what it means in terms of how humanists make knowledge.

Admittedly, this may only be an issue of the present moment, one experienced as a tax on those humanities researchers who wish to adopt new methods by asking them to overexplain their work processes. Once such methods are more widely practiced, the data-intensive humanities researcher may also be able to toss off the shorthand of "data cleaning." For now, however, there is significant value in being arrested by the obfuscation of this phrase. By taking first a descriptive approach (precisely saying what we mean by data cleaning) and, second, speculating on alternative approaches, we intervene in an unresolved conversation about data and reductiveness in the humanities. Cultural critical practices enacted at each stage of data-intensive inquiry reconfigure what has too often been offered as a binary choice in which scholars may choose to work in the tradition of cultural criticism or they may choose to work with data.

Humanities Data and Suspicions of Reduction

When humanities scholars recoil at data-driven research, they are often responding to the reductiveness inherent in this form of scholarship. This reductiveness can feel intellectually impoverished to scholars who have spent their careers working through particular kinds of historical and cultural complexity. The modern humanities have invested mental and moral energy into, and reaped insights from, studying difference. Bethany Nowviskie summarizes this tradition in this volume's Chapter 37, "Capacity through Care": "The finest contribution of the past several decades of humanities research has been to broaden, contextualize, and challenge canonical collections and privileged views. Scholars do this by elevating instances of neglected or alternate lived experience—singular human conditions, often revealed to reflect the mainstream."

From within this worldview, data cleaning becomes maligned because it is understood as a step that inscribes a normative order by wiping away what is different. The term "cleaning" implies that a dataset begins as "messy." "Messy" suggests an underlying order: it supposes things already have a rightful place, but they are not in it—like socks on the bedroom floor rather than in the bureau or the hamper.

Understood this way, suspicions about cleaning are suspicions that researchers are not recognizing or reckoning with the framing orders to which they are

subscribing as they make and manipulate their data. In other data-intensive fields in which researchers have long been explicit about their framing orders, the limits of results are often understood and articulated using specialized discourses. For example, in climate science, researchers confine their claims to the data they can work with and report results with margins of error.[1] While humanities researchers do have discourses for limiting claims (acknowledging the choice of an archive or a particular intellectual tradition), the move into data-intensive research asks humanists to modify such discourses or develop new ones suitable for these projects. The ways in which the humanities engages these challenges may both open up new practices for other fields and allow humanities researchers who have made powerful critiques of the existing systems of data analysis to undertake data-intensive forms of research in ways that do not require them to abandon their commitments to such critiques.

The Value of a Naïve Tool

To contribute to the development of new discourses and the practice of critically attuned data work, we scrutinize cleaning through a reflection on our own work with *Curating Menus. Curating Menus* is a research project that aims to curate and analyze the open data from New York Public Library's *What's on the Menu?*

We set off to answer questions like the following: Can we see the effect of wartime food rationing in what appeared on menus during World War I? Or, can we track the changing boundaries of what constituted a "dish" over time? To do this, we thought we would need to "clean" the "messy" data. What became evident was that cleaning up or correcting values was a misleading—and even unproductive—way to think about how to make the data more useful for our own questions and for other scholars studying food.

Under the rubric of cleaning, we began with a technical solution to what we had imagined was a technical problem. Variation in the strings of text transcribed from menus was obscuring our ability to do things as simple as count how many dishes were in the dataset. Working with Lydia Zvygintseva, we attempted to reduce the number of variations using OpenRefine, an open-source software tool designed for these types of tasks. When the scale of the problem overwhelmed the capabilities of that tool, we discovered that it was possible to run the clustering algorithms popularized by OpenRefine using custom Python scripts (Muñoz). The output of one of these scripts was a list of variant values such as the following:

id

2759 Potatoes, au gratin
7176 Potatoes au Gratin
8373 Potatoes—Au gratin
35728 Potatoes: au gratin

44271 Au Gratin Potatoes
84510 Au Gratin (Potatoes)
94968 Potatoes, au Gratin,
97166 POTATOES:- Au gratin
185040 Au Gratin [potatoes]
313168 Au Gratin Potatoes
315697 (Potatoes) Au Gratin
325940 Au Gratin Potatoes
330420 au-Gratin Potatoes
353435 Potatoes: Au gratin
373639 Potatoes Au Gratin

We were very excited to get lists that looked this way because we could easily imagine looping over them and establishing one normalized value for each set of variants. We had not yet recognized that the data model around which the dataset was organized was not the data model we needed to answer our research questions. At the time, the main challenge seemed to be processing enough values quickly enough to "get on with it."

At that point in the research process, the Python scripts we were using were small, purpose-built command line programs. After some deliberation, we decided to build a simple web application to provide the task-specific user interfaces we would need to tackle the challenge of NYPL's menu data.[2] The piece of software we built does, in some ways, the opposite of what one might expect. A cluster of values like the one for "Potatoes Au Gratin" is presented to the user, and he or she must make a decision about how to turn that cluster of variants into a single value. Our tool sorts the variants by the number of times they appear across the dataset. So the decision may be to simply save the most commonly occurring value as the normalized form: "potatoes au gratin." Or it might be to modify that value based on a set of rules we have defined. Or it might be to supply a new value altogether. The process can end up looking like this:

> What is the authoritative spelling of Buzzards Bay oysters? Let me Google that. . . . Oh, it collapsed an orange juice and an orange sherbet under "orange"; let me flag that. . . . A jelly omelet!?

The tool surfaces near-matches, but it does not automate the work of collapsing them into normalized values. Instead, it focuses the user's attention and labor on exactly that activity. In other words, in our initial version of computer-assisted data curation, we still had to touch each data point. This caused us to doubt that we could use this method with a whole dataset the size of the one from *What's on the Menu?*; however, it was intellectually productive.

In the process of normalizing values, we found ourselves faced with questions about the foods themselves. Choosing the "correct" string was not a self-contained problem, but an issue that required returning to our research questions and investigating the foods themselves. Since the research questions we were asking required us to maintain information about brands and places, we often had to look up items to see which was the correct spelling of the brand or place name.[3] Shellfish and liquor were two particularly interesting and plagued areas for these questions. The process revealed kinds of "messiness" that we had not yet even considered. We realized that the changes we were making were not "corrections" that would "clean up" the original dataset, but rather formed an additional information set with its own data model. What we thought were data points were, in fact, a synthesis or mash-up of different kinds of information within a half-finished or half-articulated data model. This data model was sufficient for the original aim of the project—supporting an application that facilitated crowdsourced transcription—but it is insufficient for scholarly inquiry. To ask research questions, we needed to create our own dataset, which would work in context with the NYPL dataset.

Diversity in Data

The NYPL data is a linked network graph of unique URLs for dishes and menus. The *Curating Menus* dataset is an organized hierarchy of concepts from the domain of food. We made a set of labels that we believe would facilitate humanities researchers' understanding of the scope, diversity, and value of the NYPL's dataset. The two datasets are connected by links from concept labels in our dataset to dishes in the NYPL dataset. Our interaction with the NYPL dataset thus became a process of evaluating which variants in the names of dishes revealed new information, which we should in turn account for in our own data, and which variants were simply accidents of transcription or typesetting. The process freed us to attend to difference and detail, rather than attempting to clean it away. Without cleaning, we could be sensitive to questions about time, place, and subject. This kind of attention is imperative if humanities researchers are to find the menu data valuable for their scholarship.

As we considered methods for preserving diversity within our large dataset, the work of anthropologist Anna Tsing offered us a valuable theoretical framework through which to approach these issues. In "On Nonscalability: The Living World Is Not Amenable to Precision-Nested Scales," Tsing critiques scalability as an overarching paradigm for organizing systems (whether world trade, scientific research, or colonial economies). By scalability, Tsing means the quality that allows things to be traded out for each other in a totalizing system without regard to the unique or individual qualities of those things—like many stalks of sugarcane (which are biological clones of one another) or, subsequently, workers in a factory. From this definition of scalability, she goes on to argue for a theory of nonscalability. Tsing

writes, "The definition of nonscalability is in the negative: scalability is a distinctive design feature; nonscalability refers to everything that is without that feature. . . . Nonscalability theory is an analytic apparatus that helps us notice nonscalable phenomena" (509). While scalable design creates only one relationship between elements of a system (what Tsing calls "precision nesting"), nonscalable phenomena are enmeshed in multiple relationships, outside or in tension with the nesting frame. "Scales jostle and contest each other. Because relationships are encounters across difference, they have a quality of indeterminacy. Relationships are transformative, and one is not sure of the outcome. Thus diversity-in-the-making is always part of the mix" (510).

Currently, the imagination of the cultural heritage world has been captured by crowdsourced information production on the one hand and large-scale institutional aggregation on the other: the *What's on the Menu?* project exemplifies both of these trends. Our difficulties working with the open data from this project suggest that it is a vital moment to consider the virtues of nonscalability theory in relation to digital scholarship. Engineering crowdsourced cultural heritage projects usually involves making object transcription, identification, and the development of metadata scalable. For example, the makers of the *What's on the Menu?* project designed their system to divide the work into parcels that could be completed quickly by users while reducing the friction that arise from differences in the menus; for example, the organization of the information on the page or other evidence of physical manifestations like handwriting and typeface variations (Lascarides and Vershbow). The images of menus and the metadata about them are also being republished through projects like the Digital Public Library of America (DPLA), another example of how things get shaped and parsed for purposes of scaling up to ever wider distribution. Tsing reminds us, "At best, scalable projects are articulations between scalable and nonscalable elements, in which nonscalable effects can be hidden" (515). She argues that the question is not whether we do or do not engage in scalable or nonscalable processes. To explore the articulations between scalable and nonscalable, Tsing tells the story of the contemporary matsutake industry, which encompasses both foraging (by immigrant harvesters in the ruins of large-scale industrial forestry in the U.S. Pacific Northwest) and global supply chains serving Japanese markets. Tsing's account focuses our attention on how "scales . . . arise from the relationships that inform particular projects, scenes, or events" (509). The elements of nonscalable systems enter into "transformative relationships," and these "contact[s] across difference can produce new agendas" (510). Following Tsing, we came to see points of articulation that had previously been invisible to us as would-be consumers of scaled data. Beginning from the creation of the original, physical menus and tracing the development of the crowd-created data, we identify and account for "nonscalable elements"—and consequently, edge further and further from the terminology of "cleaning."

Seeing Nonscalability in NYPL's Crowdsourced Menus Project

Making menus is a scalable process. Although menus are sometimes handwritten or elaborately printed on ribbon-sewn silk, the format of a menu is designed to be scalable. Menus are an efficient typographical vehicle for communicating a set of offerings for often low-margin dining enterprises. Part of the way that we know that menus are scalable is how alike they appear. "Eggs Benedict" or "caviar," with their accompanying prices, may fit interchangeably into the "slots" of the menu's layout. Within the menus themselves, we also see evidence of the nexus of printing scalability, dish scalability, and cost in, for example, the use of ellipses to express different options: eggs with . . . cheese, . . . ham, . . . tomatoes, etc. The visual evidence of *What's on the Menu?* shows us how headings, cover images, and epigraphs—for all their surface variations—follow recognizable patterns. These strong genre conventions and the mass production of the menus as physical objects allow us to see and treat them as scaled and scalable.[4]

However, the menus also express nonscalable elements—historical contingencies and encounters across difference. Some of these nonscalable elements are revealed by the kind of questions we find ourselves asking about the experience of ordering from these menus. How were they understood as part of the interactions among purveyors, customers, and diners? How did diners navigate elements like the pervasive use of French in the late nineteenth and early twentieth centuries? How did they interpret the particular style and content of cover images or quotations? Evidence for these questions manifests in the menus as objects, but does not fit within the scalable frames of menu production nor the menu data we have at hand. Yet the nonscalable elements cannot be disregarded and have the potential to affect how we interpret and handle the scalable data. Nonscalability theory encourages us to grapple with this dynamic at each point of articulation in the process of making scalable objects.

The collection of these menus was also scalable. The system set up for their accession and processing treated the menus not only as interchangeable objects but also like the many other paper materials that entered the collections of the NYPL in the twentieth century. Perhaps the clearest evidence of this is in their cataloging. The catalog cards fit each menu into the same frame—with fields for the dining establishment, the date of creation and date of accession, and the sponsor of the meal, if available. Cataloging is a way of suppressing or ignoring the great differences in the menus, choosing one type of data to attend to. The cards index the collection so that the institution has a record of its holdings and a user can find a particular object. The menus, with their scalable and nonscalable features, become scalable library inventory through this process (cf. Tsing, 519).

Cataloging's aim is to find a way to make items at least interchangeable enough not to break the system. The practice is rife with examples of catalogers navigating encounters with difference. Catalogers practice nonscalability theory constantly.

Sometimes the answer is institutionally specific fields in machine-readable cataloging (MARC) records; sometimes the solution is overhauling subject headings or creating a new way of making records (like the BIBFRAME initiative). However, the answer is almost never to treat each object as a unique form; instead, the object is to find a way to keep the important and usable information while continuing to use socially and technologically embedded forms of classifying and finding materials.

Digitization is also a process designed for scalability. As long as an object can fit into the focal area of an imaging device, its size, texture, and other material features are reduced to a digital image file. The zooming enabled by high-resolution digital images is one of Tsing's prime examples of design and engineering for scalability. In the distribution of digitized images, the properties of the digital surrogate that are suited to scalability are perpetuated, while the properties of the original that are nonscalable (the feel of the paper, its heft or daintiness) are lost.

The point at which certain objects are selected for digitization is one of the moments of articulation Tsing describes between the scalable and nonscalable. Digitization transforms diverse physical materials—brittle, acidic paper or animal parchment, large wooden covers or handstitched bindings, leaves or inserts—into standardized grids of pixels. From the point of digitization forward, the logic of scalability permeates projects like *What's on the Menu?* The transcription platform is constructed to nest precisely within the framework of how cultural heritage organizations like NYPL create digital objects from their original materials.

Paul Beaudoin from the NYPL Labs team discusses some of the logic behind their approach to these kind of projects in a blog post announcing Scribe, an open-source software platform released in 2015, but derived from the library's experience with crowdsourced transcription projects. Beaudoin describes how the Scribe platform is based on "simplification [that] allows us to reduce complex document transcription to a series of smaller decisions that can be tackled individually . . . the atomicity of tasks makes projects less daunting for volunteers to begin and easier to continue." For example, *What's on the Menu?* presents visitors with a segment of a digitized image of a menu and a single text input box to record what they see.[5]

The NYPL Labs team is explicit about its commitments to designing for scalability. We know from work in the domain of scholarly editing that what comprises "transcription" is not self-evident.[6] It could be modeled and implemented in software in a number of ways. The menus project uses optical character recognition (OCR) software to generate bounding boxes that determine what human volunteers will transcribe. In this, we can see the precision nesting of scales at work. OCR software is designed to scalably produce machine-readable, machine-processable digital text from images of printed material. In the case of the menus, the software can detect regions of text within the digital images; however, due to the variation in typefaces, the aging of inks and paper, and other nonscalable elements, the output of the OCR algorithm is not a legible set of words. Using the bounding boxes but discarding the OCR text in favor of text supplied by human volunteers is a clever and

elegant design. It constructs the act of transcription in such a way that it matches the scalable process of digitization and ways of understanding the content of a menu that privilege scalable data.

Yet, even here, now that we know to look for them, the nonscalable effects cannot be completely hidden. The controls allow users to zoom through three levels of an image, a feature that evidences slippage in the segmentation algorithm. This element of the tool acknowledges that someone might need to zoom out to complete a transcription—often because the name of a dish has to be understood through the relation between the line of text in the bounding box and other nearby text, like a heading. Further, the text box on the transcription screen is unadorned, implying that what to do is self-evident; however, the help page is full of lengthy instructions for how to "navigate some of the more commonly encountered issues," revealing the ways that transcription is not a self-evident, scalable process.

In addition, the project was designed so that people did not have to create accounts or sign in to submit transcriptions. Creators and managers of such projects are working to make more data and to allow more people to participate in making data. This entails creating systems that treat volunteers as interchangeable in the service of allowing more work to get done more quickly. However, what is construed as a scalable workforce is, in fact, made up of people who have different levels of understanding or adherence to the guidelines and different perceptions or interpretations of the materials. When we understand this workforce as a collection of individuals, we can see how any crowd as large as the one that has worked on the menus project will contain such diversity. The analytic apparatus of Tsing's nonscalability theory makes all these design choices visible and allows us to see the transcription task, as framed within *What's on the Menu?*, as another moment of articulation between scalable and nonscalable elements.

When we download the open data from the *What's on the Menu?* site and open up the files, we are presented with the results of all this activity: menu collection and digitization and transcription. Instead of seeing mess, we see the ways in which diversity has seeped or broken into what were designed to be smoothly scaling systems. Now we are better prepared to envision how our work—creating new data organized around concepts of historical food practices—begins from what the NYPL has released, which is transcription data: words volunteers typed into the boxes at *What's on the Menu?* linked to metadata from their digital library systems. In both of these datasets there is something called "dish." In NYPL's data, "dish" is the name of the field in which a transcribed string from a menu is stored in the project's database. In *Curating Menus*' data, "dish" is a representation created to reflect and name an arrangement of foods and culinary practices in a particular historical moment. This is an example of, as Tsing puts it, the ways that "scales jostle and contest." We know that the response to this friction is not to retreat from working at scale. Instead we have to find ways of working while being aware that *precision* nesting hides diversity and that there are consequences to diversity being hidden.

Indexes: Making Scalability Explicit and Preserving Diversity

Our answer to this challenge is an index. We are suggesting that indexing is a more precise replacement for some of the work that travels under the name of "cleaning." An index is an information structure designed to serve as a system of pointers between two bodies of information, one of which is organized to provide access to concepts in the other. The list of terms and associated page numbers from the back of a book is one familiar example. An array of other terms that people use alongside "cleaning" (wrangling, munging, normalizing, casting) name other important parts of working with data, but indexing best captures the crucial interplay of scalability and diversity that we are trying to trace in this chapter.

We began to think of the work we were doing as building something like a back-of-the-book index for the *What's on the Menu?* data. We aimed to create additional data structures around and atop the existing data, generating links between a set of names or categories we created and the larger and more heterogeneous set of data from NYPL. Ultimately, we decided to build two interconnected indexes: one focused on historical food concepts and one on the organizations connected to the menus (businesses, community organizations, fraternal societies, etc.). We began with the food index and are developing a framework that echoes cookbook indexes to structure our data: ingredients, cooking methods, courses, cuisines.

If we had felt no unease continuing the lineage of precision nesting that links the scales of digitization and crowdsourced transcription, we could have proceeded with a completely algorithmic approach: "cleaning" our data using scripts, linguistic rules, and even machine learning. These methods yield results by predictively circumscribing variants in language so as to aggregate and enable analysis at another level of abstraction. We can see the usefulness of these algorithmic approaches, but we also know that they would hide diversity in the menus. However, to understand the implications of these diversity-hiding methods, we needed to create a grounded approach to making a data model for our index.

Today, when we look at a list of variations on "Potatoes au Gratin" or some other group of transcriptions, we focus on the task of choosing a label that will be a node in our dataset and that will serve as a pointer to all the other transcribed values in the NYPL dataset. We are building the set of labels from the set of data, rather than beginning by writing some large, hierarchical domain model. We want to represent the concept of two eggs and bacon, not caring if it was written "bacon and two eggs" or "2 Eggs and Bacon."

To get from transcription to concept, we began with a set of simple rules: spell out numbers and use lowercase letters. Actually engaging with the menu transcriptions quickly raised other questions. For example, on the question of place names, we decided to apply capitalization rules (in accord with style guides like *The Chicago Manual of Style*) that say that you capitalize when the reference to place is literal, but not when the reference makes a generic association: yes to Virginia

ham or Blue Point oysters but no to swiss cheese or scotch. We also found many single transcriptions containing multiple concepts, like "steak, sauce béarnaise." Since we wanted a way to be able to systematically find multiple components of a dish, we opted to standardize how we labeled the addition of sauces, garnishes, and other added ingredients. Here is one instance where we plan to use algorithmic tools to help us analyze some of this big data, since it is grounded in a specific data model.

In building our index, we are conscientiously creating scalability. We know that scalability is a process of articulations between different scales; however, Tsing suggests—and we believe—that those articulations are often hidden. Conversely, indexes are tools of scalability that make these articulations explicit.

Our index is about ingredients, meal structures, and cooking techniques. Someone else could re-index the menu material in a different way. Variations might involve attending to the species of the plants and animals that are in foods or taking a nutritional approach that classifies food based on calories, vitamins, and carbohydrates. We can also imagine projects that attend to language use in ways that our index suppresses. However an index is conceived, it allows us to build up explicit and flexible bases of knowledge that people can continue to access and understand.

Sharing Control of Authority

One of the mechanisms that librarians and archivists have used to build and maintain large, distributed information systems is a set of practices referred to as authority control. In brief, these practices involve creating defined and agreed-on taxonomies, as well as guidelines for the application of such arrangements of terms. The Library of Congress Subject Headings represent one instance of authority control. Maintaining such a system is labor intensive and has been used only for supporting core library activities like managing collections and supporting patrons in finding materials. Libraries and archives are trying to take advantage of technological developments in linked data—merging their centuries-old authority control practices with the affordances of the World Wide Web. However, what relatively few have seized on are new opportunities to apply the practices of authority control outside the original core needs of collection organization and wayfinding.

These new opportunities fall somewhere between digital library practices and digital humanities research, but the gap is one that more projects should embrace the opportunity to fill. There is a need for projects that take "authority work" as an invitation to encourage creativity, an invitation for making and building. In such a model, multiple regimes of authorities might be built up from critically aware and engaged intellectual communities to meet their own specific needs while also participating in larger technological and information systems.

We imagine those communities will contain librarians and scholars. Though librarians and humanities scholars have frequently intersected, they have rarely interacted in this way. Simplifying to the point of caricature, these existing interactions go

something like this: humanities scholars point out that the structure and content of a specific archive or collection represent and even re-create certain cultural logics. For example, the systems of describing collections, such as the widely used Library of Congress Subject Headings, reify concepts about persons or cultures that should be interrogated more closely or perhaps discredited and dismantled altogether. For the most part, librarians, archivists, and information scientists acknowledge these flaws and perhaps even work to remedy them in the course of maintaining systems that preserve whatever partial archives do exist or helping patrons find information they need.

We are looking for new forms of collective action that can be expressed through the professional work of humanities scholars and librarians. This is not simply a call for the production of more data—attempting to subvert the work of categorization and classification through the production of ever more local, finely wrought distinctions, details, and qualifications. Our aim is to develop ways of working that validate local experiences of data without removing them from a more global network of information exchange.[7] These practices, as we imagine them, resonate with Bethany Nowviskie's interpretation of the Afrofuturist philosophy of Sun Ra (as expressed by Shabaka Hutchings), which claims, "Communities that have agency are able to form their own philosophical structures"(Nowviskie, "Everywhere") The transition to working in a linked data paradigm should be valued not principally for the ways in which it might make large-scale information systems operate more smoothly, but rather for the ways in which it can create localized communities of authority, within which people can take control of the construction of data and the contexts in which it lives. In a keynote presentation at the 2015 LITA Forum, Mx (Mark) A. Matienzo articulated a parallel version of this view:

> We need to begin having some serious conversations about how we can best serve our communities not only as repositories of authoritative knowledge or mere individuals who work within them. We should be examining the way in which we can best serve our communities to support their need to tell stories, to heal, and to work in the process of naming.

Discussions of cleaning data fail to capture this need. The cleaning paradigm assumes an underlying, "correct" order. However tidy values may look when grouped into rows or columns or neatly delimited records, this tidiness privileges the structure of a container, rather than the data inside it. This is the same diversity-hiding trick that nonscalability theory encourages us to recognize.

But it is not enough to recognize how cleaning suppresses diversity; we must also envision alternate ways of working. The first thing we must do with our datasets, rather than normalizing them, is to find the communities within which our data matters. With those communities in mind and even in dialogue with them, we must ask these questions: What are the concepts that structure this data? And how can

this data, structured in this way, point to other people's data? This way of thinking will allow us to see the messiness of data not as a block to scalability, but as a vital feature of the world that our data represents and from which it emerges.

NOTES

1. The fact that these communities have developed discourses for describing the boundaries of the claims they make does not inoculate them from critique about the costs and shortcomings of their methods (cf. Latour).

2. For a variety of reasons, we would not recommend this course of action to others without serious deliberation. There is a reason why applications like OpenRefine are so popular and useful. If you would like to know more, contact the authors.

3. If we had a dictionary to compare these materials to, the process may had been more automatable; however, from what we have found thus far, that particular language resource—Wordnet for Oysters!—does not exist.

4. The NYPL menu collector Frank E. Buttolph's acquisition practices reinforce the role and scale of printers in menu production in the twentieth century. In addition to restaurants and customers, she went straight to the menu source—printers—to fill out her collection.

5. Early versions of the project interface featured a social-media-style call-to-action below the image snippet ("What does this say?"), as well as brief instructions below the text input box: "Please type the text of the indicated dish EXACTLY as it appears. Don't worry about accents" (see for example, https://web.archive.org/web/20120102212103/http://menus.nypl.org/menu_items/664698/edit). This accompanying text was quickly dropped—presumably because the task seemed self-evident enough from the layout of the transcription screen.

6. See https://datasymposium.wordpress.com/sahle/.

7. Compare with the "speculative guidelines" in Loukissas.

BIBLIOGRAPHY

Beaudoin, Paul. "Scribe: Toward a General Framework for Community Transcription." New York Public Library (blog). November 23, 2016, https://www.nypl.org/blog/2015/11/23/scribe-framework-community-transcription.

Lascarides, Michael, and Ben Vershbow. "What's on the Menu?: Crowdsourcing at the New York Public Library." In *Crowdsourcing our Cultural Heritage,* edited by Mia Ridge, 113–38. Surrey, UK: Ashgate, 2014.

Latour, Bruno. "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry* 30, no. 2 (2004): 225–48.

Loukissas, Yanni Alexander. "Taking Big Data Apart: Local Readings of Composite Media Collections." *Information, Communication & Society* 20, no. 5 (May 4, 2017): 651–64. doi:10.1080/1369118X.2016.1211722.

Matienzo, Mx (Mark) A. "To Hell with Good Intentions: Linked Data, Community and the Power to Name." Mx A. Matienzo (blog). February 11, 2016. http://matienzo.org/2016/to-hell-with-good-intentions/.

Muñoz, Trevor. "Using Pandas to Curate Data from the New York Public Library's What's On the Menu? Project." http://nbviewer.jupyter.org/gist/trevormunoz/8358810. Accessed January 10, 2014.

Newman, Harvey B., Mark H. Ellisman, and John A. Orcutt. "Data-Intensive E-Science Frontier Research." *Communications ACM* 46, no. 11 (November 2003): 68–77. doi:10.1145/948383.948411.

Nowviskie, Bethany. "Everywhere, Every When." April 29, 2016, http://nowviskie.org/2016/everywhere-every-when/.

Tsing, Anna Lowenhaupt. "On Nonscalability: The Living World Is Not Amenable to Precision-Nested Scales." *Common Knowledge* 18, no. 3 (2012): 505–24.

What's on the Menu? New York Public Library. http://menus.nypl.org/. Accessed July 29, 2016.

Wickham, Hadley. "Tidy Data." *Journal of Statistical Software* 59, no. 10 (2014): 1–23.