THIS IS A PRE-PRINT CHAPTER (UNDER REVIEW WITH U MINNESOTA PRESS) FOR THE NETWORKED HUMANITIES COLLECTION (ED. BY MCNELY AND RICE).

The Boutique is Open: Data for Writing Studies

Cheryl E. Ball, Tarez Samra Graban, and Michelle Sidler[1]

Making Data Public

In February, 2013, the Presidential Office of Science and Technology announced a plan to increase access to federally funded research among major agencies that grant $100 million or more annually. The mandate describes two types of information, published journal articles and datasets, which require better public access through archiving and repositories. Providing public access to journal articles has received the lion's share of attention—both support and resistance—because of its impact on academic publishers, but it is our contention that access to data will have an equally strong impact on researchers' lives, including those in the humanities. Providing public access to data has the potential to speed the pace of research and discovery by creating data sets that can be aggregated, compared and analyzed. In addition, researchers who open their data can develop new collaborations that have the potential to create new lines of inquiry.

Although most humanities fields have not yet considered, in any systematic way, the idea of opening data to the public, a passionate community from fields of science is advocating greater access to research through a movement generally referred to as Open Data. Organizations like the Open Knowledge Foundation (OKFN) (http://okfn.org) perform many supportive activities to increase awareness, including sponsoring online projects that utilize public datasets, hosting hackathons to develop software tools, and composing open data guidelines. Recognizing researchers' potential concerns about making their data available, members of OKFN and other Open Data advocates composed the Panton Principles (http://pantonprinciples.org/), which serve as both an Open Data manifesto and a series of protocols for opening and licensing data.

Two major concerns for Open Data advocates are licensing and formatting. In order for data to be re-usable, researchers must explicitly grant permission rights to other users. The Panton Principles, in fact, strongly suggest using CC-Zero as the license, in effect putting data in the public domain. Other researchers will be re-aggregating and re-calculating data, so non-derivative licenses are particularly problematic, but even attribution and non-commercial licenses potentially slow down the research progress with copyright barriers. The format in which data is made available also has the potential to slow the progress of research. Currently, many journals require all Supporting Information like data tables to be in PDF format. However, this format is only human readable and not machine readable, essentially locking the data within the file rather than keeping it open and mobile. Advocates recommend using open spreadsheet or database software for quantitative data as well as indexing and tagging features in text data. The international Beyond the PDF Conference, held in 2011, created standards and researchers arguing for more open forms of public data and creating tools to more easily share it.

Many of the advances in open data for science happen as a result of research teams that collaborate on both research projects and the digital tools that facilitate those projects. As a

[1] This piece is a collaboratively authored article, and authors' names have been listed in ascending alphabetical order. Writing about rhetoric.io would not be possible without the other members of its advisory board, also in alphabetical order: Collin Brooke, Douglas Eyman, Derek Mueller, Michael Neal, and Karl Stolley.

result of these and other initiatives, we realize the need for a paradigmatically distinct conception of what it means *to make data public*--one that is committed more to core principles of the *collaboratory* than the *repository*. Unlike a repository that might be guided by an ethic of preservation (and in turn, organized according to assumptions about what should make entities stable), a collaboratory promotes an environment in which researchers can leverage the mutability of shared data and communication tools, facilitating networks of research teams and promoting cross-pollination of inquiry, data, and projects.

Forming Data Publics

One way that Open Data advocates can argue for new practices and standards is to conjoin what -- in analog spaces -- often refuses to be joined: entity relationships with multiple users. For networked humanities research, these entities are data sets that often go unseen outside of users' own preliminary research. Joining data sets with their users creates an activity system wherein the data entities and the publics (researchers, but also in open data the general public) are made available to each other; thus, *data publics*.

There is a plethora of this data waiting to be made public in writing studies, as Jenn Fishman and Joan Mullin discovered through their collaborative project to publish data and unpublished writing research. In 2006, Fishman and Mullin had the idea to build an online, open-access space to house data collected through writing research, with the aim to facilitate knowledge-making across the individualized studies we conduct in our classes and programs while also fostering more international collaborations in writing research, where quantitative and qualitative data collection is already the norm[2]. The principal problem was a lack of distribution venues -- repositories -- for these research entities and artifacts. Colleges and universities were only just beginning to set up data repositories or commit to open-access arrangements, both of which were usually reserved for post-publication copies of articles or accepted-for-publication pre-prints. Including data sets in these repositories was, and still is, an intellectual property issue for scientists, who are seen as the primary audiences for such collections (see Cohen; Heidorn; Waldrop).[3] However, wrapped up in the issue of audiences for such a repository, Mullin and Fishman also struggled to convince rhetoric and composition scholars in the United States that all of the surveys, interviews, writing samples, student projects, and other valuable corpora used for analysis in writing research were, in fact, *data*. Indeed, in 2013, many rhetoric and composition scholars still don't understand this fact; our field's slowness to adopt and adapt Institutional Review Board protocols for writing studies research is indicative of the difficulty the field has with accepting qualitative research, or even quantitative research collected for the purposes of program review, for instance, as data that *others* -- in and outside of our academic discipline as well as the general public -- might find value in. It may also point to an opacity surrounding the potential value and actual benefits of spending the time that it takes to justify and shape research designs according to these protocols.

Fishman and Mullin recognized that rhetoric and composition had decades of important data that was under threat of being lost due to neglect and disuse, and through a series of conference sessions, forums, and advisory discussions, the Research Exchange Index (REx)

---

[2] For a complete history of this project, see A Brief History of REx, http://researchexchange.colostate.edu/about/history.cfm.

[3] In 2013, it remains an issue for rhetoric and composition studies -- and for the humanities and social sciences more broadly -- even after notable advancements in the adoption of open-access agreements on university campuses; responses to the AHA's proposal to embargo students' intellectual work; and MLA President Marianne Hirsch's selection of "vulnerability" as the 2014 conference theme.

began. Hosted on rhetoric and composition's largest open-access publishing site, the WAC Clearinghouse,

> [t]he current goal of the Research Exchange Index is periodically to compile and publish a searchable database of research reports. The resulting publication will serve as both a directory of contemporary writing research and a means of aggregating data about it. The database remains true to its original intention, collecting studies that are ad hoc and local as well as those that are more regional, national and international. As a resource for everyone who is interested in writing and its study, REx will help researchers as well as scholars, teachers, and other writing stakeholders find models for new projects, put current projects in context, and review work in writing studies from 2000 to the present. (http://researchexchange.colostate.edu/about/history.cfm)

In theory, REx moves away from the concept of a repository for stable, completed scholarship and towards the concept of a collaboratory, housing in-progress and dynamic scholarship meant to facilitate ongoing research. It collects census-like information on current studies in writing research, including information on the principal investigators, project team, funding and support, study details with a specific focus on methodologies, reflections and outcomes, and related files users want to upload. Users submit this information through a form on the REx website, and the report that is generated from the submission is reviewed and developed (if necessary) by the editors and staff of REx. The current goal, now that REx has completed its first solicitation campaign (as of July 15, 2013), is to publish an annual, peer-reviewed database of the completed research reports with an introductory overview written by the editors. Because individual elements of each report are tagged, the database allows researchers to search within a report or across reports for relationships among tags.

The start-up history of REx is admittedly long, but this is due, in large part, to the editors' need to educate potential audiences and contributors about being stakeholders. In that process, one aspect of open data that REx decided it could not yet infrastructurally or intellectually support was the collection, storage, and distribution of the data these studies collected. Thus, it is a project a bit before its time, with a public of traditional rhetoric and composition scholars who are still becoming ready for open data. REx's refined scope -- to publish peer-reviewed summarizing reports *about* ongoing or completed, qualitative and quantitative, writing research studies (but not necessarily publishing the data sets that go with those studies) -- introduces rhetoric and composition scholars to the notion that their research is grounded in multivarious methodologies and methods which collectively produce all sorts of artifacts known as data sets, and that these data sets are always in progress. In this way, REx has laid the groundwork for traditional rhetoric and composition scholars to become *data publics.* Building from this valuable work, we argue that the time is right to move the field further, into a dynamic understanding of *data* themselves, wherein individual data sets can be combined, remixed, and leveraged for collaborative, multi-layered, or multi-authored research projects.

Big Data or Boutique Data?

Transdisciplinary fields of rhetoric and composition--including computers and writing and the digital humanities, which we will shorthand in this article as networked humanities fields--are already ripe as data publics. Non-data publics -- wherein potential users and data entities are not yet conjoined -- might particularly see this evidenced through the calls for "big data" projects in funding agencies such as the National Endowment for the Humanities' Office of Digital Humanities. **Big data**, as described in the NEH's Digging Into Data challenge, speaks to the breadth and depth of computational data researchers have access to:

> Now that we have massive databases of materials used by scholars in the humanities and social sciences -- ranging from digitized books, newspapers, and music to transactional data like web searches, sensor data, or cell phone records -- what new, computationally-based research methods might we apply? (http://www.neh.gov/grants/odh/digging-data-challenge)

The Digging Into Data challenge brings collaborative, international researchers together -- with funding supplied by 10 (as of 2013) international funding councils -- to ask researchers to dig into the research collections of an ever-growing collection of databases (more than 40 as of 2013) that house massive digital collections. A few examples of these databases that make their digital collections open for this data challenge include JSTOR, Project Muse, Hathi Trust, Internet Archives, and several major city and national libraries of funding-participant countries (New York Public Library, Library of Congress, National Archives, etc.) Some of the first round projects from 2009 included "Mining a Year of Speech," which focused "on large scale data analysis of audio -- specifically the spoken word" to "create tools to enable rapid and flexible access to over 9,000 hours of spoken audio files, containing a wide variety of speech, drawn from some of the leading British and American spoken word corpora, allowing for new kinds of linguistic analysis." Another project, "Digging Into Image Data," led in the U.S. by networked humanities scholar Dean Rehberger, analyzed authorship in manuscripts, maps, and quilts from the 15th through the 20th centuries. (http://www.diggingintodata.org/Home/AwardRecipientsRound12009/ConferenceforRound12009Awardees/tabid/184/Default.aspx).

These are a small sampling of the *huge* corpora -- big data in every sense -- that these Digging Into Data challenge projects use. But it is important to note the resources required for these projects: economically, they cost hundreds of thousands of dollars in funding supplied by two or more international funding agencies, coupled with the human resources of having international partnerships already in place; and physically, they require big computing -- supercomputing, in fact -- to algorithmically sort and analyze all that big data. Digging Into Data, and networked humanities' interest in big data, is a relatively recent entry (in 2009) to international funding patterns, and it's one that seems to require, for the most part, the staff and infrastructural support of an existing digital humanities center to carry out. This kind of research is beyond the reach of most networked humanities scholars, a critique that has been raised for years within digital humanities circles, as evidenced in some small ways by the move of THATCamp (The Humanities and Technology Camps) facilitators to take up these DIY technology unconferences at Small, Liberal Arts Colleges instead of solely at research-intensive institutions.

However, in the rush to embrace big data as a funding stream for large, networked humanities projects, researchers might miss the value in the thousands, perhaps hundreds of thousands, of uncounted data sets researchers in the networked humanities already have. These data sets come from what seem to be small-scale projects in relation to big data--small data sets such as assessments from our writing programs, student writing and projects from a class study, interviews from writers, syllabi collected under a single topic, discourse analyses from dissertation projects, data mining from job lists or journals in our field, and even data sets generated by historical questions that involve a defined, limited interaction of time to place. Library and information science professor, P. Bryan Heidorn (2008) calls this kind of data *dark data:* "Like dark matter, this dark data on the basis of volume may be more important than that which can be easily seen" (p. 281). It's the kind of data that networked humanities scholars (as well as traditional rhetoric and composition scholars) may already be interested in and already have access to, just as Heidorn explains scientists do:

> When asked, almost all scientists will quickly acknowledge that they are holding dark data, data that has never been published or otherwise made available to the rest of the scientific community. An example of dark data is the type of data that exists only in the bottom left-hand desk drawer of scientists on some media that is quickly aging and soon will be unreadable by commonly available devices. The data remains in this dark desk drawer, inaccessible to the scientific community until the scientist retires. At the point of retirement some scientists rush to find a more suitable home for their data be they in the form of slides, photographs, specimens or electronic media files. More often than not, even in a well planned retirement the desk drawer is eventually emptied into a dumpster because no one including the scientist knows exactly what the data is since it lacks adequate documentation. (p. 281)

Building on *Wired* editor Chris Andersen's (2004) notion of the long-tail phenomenon, Heidorn confirms that "There may only be a few scientists worldwide that would want to see a particular boutique data set but there are many thousands of these data sets" and access to these sets can have a huge impact on research in science--or the humanities, we contend. (p. 282) Thus the phrase **boutique data** seems apropos to describe the plethora of currently inaccessible sets of qualitative and quantitative data that exists behind much of humanities scholarship. These data sets are often small and built from local contexts, but when combined with other such sets, they are rich in potential new sources of inquiry and knowledge.

Arguing (from) Data

At such a critical juncture in data management and theorization, we need to try boutique data on for size in writing studies. We can learn from each other what has already been done, what is available, and how we might move forward in turning our pilot study into a longitudinal study, etc. Some researchers in the networked humanities have already begun to collect, share, and analyze boutique data. Several of the project examples mentioned above (e.g., interviews from writers, syllabi collected under a single topic, data mining from job lists or journals in our field) have already been started by writing studies researchers. For instance, in its five short years of existence, the Digital Archive of Literacy Narratives (DALN) has amassed a collection of over 5,000 multimedia stories in which people describe an important literacy event in their lives. Several of the projector directors and participants in the DALN, along with other writing and literacy researchers, have recently published a book exploring major themes and issues that have arisen in the collection (Ulman, DeWitt, & Selfe, 2013). In another example, Jeremy Tirrell (2012) has mined the locations and affiliations of all article authors and journal editors from the online journals in rhetoric and composition in order to map the geospatial locations of this disciplinary work. As he discusses in the methodology to his webtext on this work, not all journals include such information, so the data Tirrell has mined are unique, and researchers are fortunate that he has published data alongside his findings in an open-access format. (But we also know from Tirrell's work that many online journals are fleeting entities that cannot always be trusted with caretaking the field's scholarship.) Other scholars, such as Jim Ridolfo ("Rhetmap") and Derek Mueller (2012a, b) are working on similar disciplinary-based data-mining projects--the data for which live in random places such as a researcher-purchased domain name or, simply, on somebody's hard drive on their laptop. We, as researchers of our own studies and of others' scholarship, do not always know whether these data are available, who has them, how they are stored, or how they will be stored ten or twenty years in the future (or longer). Yet, these are all questions that scholars with boutique data sets should be able to definitively answer if we don't want to reinvent the wheels in our and our students' and collaborators' research.

Boutique data projects specifically require the *extraction* of data from ongoing research (rather than its *preservation* and *containment*), which effectively raises questions that are critical to the growth of certain disciplines. For rhetoric and writing studies, inherent questions about what defines text objects and how to measure their circulation make it a discipline necessarily committed to movement and access. Extraction of data sets allows for their reuse, providing researchers with materials that are beneficial to our disciplinary mission of understanding writing and pedagogy in the context of larger cultural, economic and historic trends.

> *What can we learn from open-data paradigms about the questions and methodologies that are intrinsic to our discipline?*

First and foremost, that our data needs are pan-historiographic and emergent in every sense -- that is, they demonstrate a renewing of rhetoric and composition's imperative to take up (and rewrite) disciplinary histories that help us to meta-theorize, and that are are not limited to a single temporal scope or geographic space (Hawhee and Olson 92). Open data paradigms remind us that although the field of rhetoric and writing is based on an assumption of social understandings and shared language, we as researchers often overlook the communal nature of our own knowledge-building. Sharing, aggregating, and comparing our individual data sets has the potential to grow our collective understandings of notions like pedagogical trends, student performance, and even our attitudes towards and participation in scholarly lines of inquiry.

Examination of existing data sets may also lead to a cultural shift in the ways we understand data collection. Shared data in the sciences, for example, has led to an expanded emphasis on crowdsourced data. An increasing number of scientific research projects employ large numbers of people-- many of whom are not scientists-- through such processes as the interpretation of images, census-taking of birds, and even the discovery of celestial bodies. What sorts of analogous projects might be possible through a collaboratory of rhetoric and writing? The DALN is perhaps the closest example we have in our field of large-scale community-generated data collection, but it is easy to imagine many other writing research projects that solicit participation from a large number of scholars and laypeople.

Second, that those needs require we mine data not as scientistic representations of what is there, but as *topoi* indicating what could be there. This challenges the idea of a "long tail" of inquiry into any field or discipline, much like Derek Mueller (2012a) poses challenges to "grasping the long tail" of rhetoric and composition (197). Even projects that aggregate widely from field journals and citation finders are capturing, at best, gaps or spaces between critical cultural moments in a discipline. Granted, Mueller shows that graphs of such activity can tell us a lot and at nuanced levels of detail (zooming in and out, seeing patterns of activity between topics and subject headings, for example); however, the data aggregated still represent imperfect and problematic measures of "citation events" and "citation activity."

Third, that our understanding of data relationships is not—and need not be limited to—simple object entity relationships (ER). An archival database supporting Library of Congress (LOC) subject relationships and the Dublin Core metadata tags can still be flexible and socially derived, but it will likely rely too much on flat relationships between texts and their users as objects, i.e., simple assumptions about what a "text" is, how it is created, and how it "moves." An archival metadatabase can employ more sophisticated tagging functions to account for searching activity as part of those entity relationships, but, as we describe in more detail below, rhetoric.io as a boutique metadatabase, offers a paradigm reconstruction altogether. It offers a limitless data interchange format that not only responds to but interacts with various data

agents, supporting multiple functions beyond searching and cataloging towards knowledge management, i.e., promoting inquiry and meta-inquiry at the same time. As a result, boutique data for the purposes of tracing such dynamic relationships as those put forward in rhetoric and composition research are uniquely distinct from other types of institutional research, which tend to employ archival paradigms in one of two ways: (1) either adapting analog methods and tools that digital historians make available; or (2) re-envisioning the database according to what extant historical questions require. Instead, boutique data define a third kind of participation, which is to consider who are all of the agents, and what are all of the agential functions involved in writing, distributing, historicizing, and theorizing pedagogical texts and their influence(s)?

Fourth, arguing from data (particularly through the functions enabled by rhetoric.io) is extremely important for rethinking the reading, researching, and examination practices of graduate programs focused in rhetoric and writing studies (see Eyman, Sheffield, & DeVoss, 2009). As (digital) rhetoric and composition scholars have long known, "imagin[ing] new couplings and scalings that are facilitated both by new models of research practice and by the availability of new tools and technologies" (Presner par. 11) is paramount to understanding writing as an activity system, especially when traditional definitions of texts as static and material often result in assumptions that our research is either highly quantitative based on retrieval or highly qualitative based on experiential. In truth, our reading and research practices evolve along with our tools—sometimes ahead of them—and so should our perceptions of how they work and why they evolve as well.

The Becoming of Rhetoric.io

It is in the spirit of this evolution within the networked humanities, in particular the context of the Networked Humanities conference at University of Kentucky (NHUK) during February 2013, that the rhetoric.io boutique data collaboratory was born. The NHUK conference, hosted by Jeff Rice, brought together a small group of humanities (mostly rhetoric and composition and digital writing studies) scholars who wanted the space of an informal conference to share research and ideas. Rice gave the conference a tone that was as much about the scholarship occurring in the sessions as it was about the spontaneous bourbon, beer, and hallway conversations that took place alongside. The combination proved fruitful in the case of rhetoric.io. As indicated on its website (http://rhetoric.io),

> Rhetoric.io is a boutique data repository for writing studies and related fields. The mission of this project is to provide an institutionally independent, centralized location for writing researchers to make their own datasets public. By publicizing our datasets, other researchers, as well as the non-academic public, can find and further writing research through analysis, remix, visualizations, and other uses of boutique data.

The idea for building a data repository for boutique[4] writing research actually prefaced the NHUK conference: In the winter of 2012-13, Joan Mullin had been discussing the technical and disciplinary challenges of REx with then-ISU colleague and *Kairos* editor, Cheryl Ball. Ball agreed with Mullin and Fishman that data preservation was an important avenue for writing researchers to focus on, but also agreed with their decision to not support data collection and storage as part of REx's immediate mission. REx needed to focus on supporting the visibility of (sometimes micro-) writing research and its methods and methodologies through publication of an open, peer-reviewed database, which left the data collection and storage issue still

[4] The terminology for calling it a *boutique* data repository came from Karl Stolley's NHUK presentation, "An API of Motives," in which he discussed boutique data, as opposed to big data, in writing studies. This concept fit perfectly with the intended purposes of REx and what was to become rhetoric.io.

unaccounted for[5]. With this idea in the air, Ball approached *Kairos* Senior Editor Douglas Eyman about starting such a repository as part of *Kairos*'s community outreach projects. The need had been presented, and who better (she thought) than the longest-continuously-running online journal in rhetoric and composition to manage such a project. Eyman responded wisely that it would be difficult to maintain such a database and to have it truly function as a long-term access and preservation node--to the point that this database could function as part of federal data management plans for future grant-seeking researchers--given the overload on *Kairos*' already fragile human resources. Indeed, both editors had previously decided not to take on other, equally administratively heavy *Kairos* projects so as not to tip the balance of their work from quality scholarly venue to something unwieldy or unsustainable. Additionally, this project needed to be much bigger than *Kairos*, and much more collaborative.

Enter NHUK a few weeks later. At the conference, Eyman and Ball had just finished a presentation about the social, scholarly, and technical infrastructures needed to maintain open-access, digital media publishing venues, such as *Kairos*. In other words, disciplines have to be ready for and willing to value digital scholarly publishing, and publishers have to have technologies that will support it. These are lessons long-learned through *Kairos*, which is an independent journal. Ball and Eyman then attended a session led by Michael Neal, in which he, Stephen McElroy, and Katherine Bridgman presented on the birth of the FSU Digital Postcard Archive, which is a pedagogical research archive that grew out of a Center for Everyday Writing Initiative, and that contains high-resolution scans of historical postcards, along with crowd-sourced tagging.[6] Neal explained in the presentation that the pedagogical component of the project made it difficult to collaborate with the FSU library, because, while the library was willing to host an Omeka installation for the archive, it wasn't able to accommodate a large number of students--under the guidance of an instructor in FSU's publishing studies program--to interact with and edit the archive. So, Neal hosts the archive, now filled with thousands of historical postcards and their metadata, where it continues to grow.

It is a boon to networked humanities scholars that our departments may now be willing to offer us such options so that we don't have to seek outside hosting, which we as individuals have to maintain outside of our academic responsibilities. However, hosting a valuable resource such as the FSU Card Archive on a personal or department site by a single researcher introduces a perhaps unnecessary bond between the project, the researcher, and the university. We won't speak to the contracts or agreements, if indeed there are any, set up between Neal and FSU's English department, but one has to ask: Who owns this data? Who is responsible for maintaining the data? What happens to it if the primary researcher moves to a different university? Are pedagogical applications of the archive so integrated into the department's curricula that the archive can withstand the absence of any single researcher? Does it matter? Yes. Because, at what point--even as soon as 5 or 10 years from now--does a systems administrator monitoring the webspace of departments decide that that folder hasn't seen any activity and must, therefore, not be needed? Or, at what point does the university change how its web architecture is structured so that the archive becomes inaccessible?

As well, most departments are still not equipped to offer networked humanities scholars the same support that Neal procured. For example, Tarez Samra Graban encountered some of the same hosting questions in developing the prototype for MDMP -- the Metadata Mapping Project,

[5] To be clear, REx can publish some filetypes associated with the research summaries as well as pointers to data published on other sites, but it is not in the business of summarily collecting the data sets used to write the reports, nor of focusing on the data sets as *the* scholarly work it publishes.

[6] http://english3.fsu.edu/mnealomeka/

which she presented at NHUK -- on an independent server and in conjunction with data designer Alli Crandell. As a project that occurs at the intersections of historical and quantitative research, MDMP attempts to re-portray (or portray for the first time) the migration of disciplinary activity in rhetoric and composition studies by mapping the terms, motivations, and uses associated with women pedagogues' curricular and teaching activities from North America's progressive era. It also aims to provide alternative constructions for topical relationships between terms that can be used to seek out women's intellectual activity in rhetoric and composition, and to note relationships between *terms* and *locations* of their work. This means that it requires both crowd-sourcing and gatekeeping functions that will allow a constant articulation and re-articulation of relationships between metadata, its users, and their queries so as to more critically account for the ways that women's pedagogical influences move into and through our disciplinary consciousness. Where Neal was able to procure departmental server space for the temporary construction and hosting of the growing Digital Postcard Archive, Graban found that MDMP had a more complex history as it first began as a concept during the 2011-2012 academic year at Indiana University, grew into a static prototype while she moved institutions from Indiana to Florida State, and then continued to grow as she sought a collaborator not affiliated with an institution or programming group. These circumstances should serve to better position MDMP as the flexible and crowd-sourced tool it strives to be, yet only make more urgent the questions of cost, sustainability, and -- surprisingly -- institutional *ethos* as she argues for its usefulness independently of a single academic unit, in her narratives on external grant applications.

So, in a kairotic face-to-face moment in February, those who were present at NHUK[7] were gathered in the hallway and dubbed the founding advisory board for what was to become rhetoric.io, the name later given to the project when Karl Stolley purchased the domain name in April 2013. Although .io is short for Indian Ocean, Stolley explained via exuberant text message that ".io (aka input/output) is the TLD [Top Level Domain] of fashion in the API-building/consuming parts of nerddom." Throughout the spring of 2013, the advisory board met via Google Hangout to discuss its agenda, the mission and vision for rhetoric.io, and the specs for its API--an application programming interface that would allow us, first, to collect data *about* the kinds of data sets researchers could eventually upload to rhetoric.io. But it was in those hangouts where the initial vision of rhetoric.io began to change from being a boutique data *repository* to being a boutique data *collaboratory*.

The initial idea was to create an institutionally independent repository to which researchers could upload their data for perpetual access and preservation, but that idea quickly morphed into one that was more flexible and manageable in the short run: to create a system-independent, federated data discovery tool that would point researchers to data stored already on faculty websites and institutional repositories. While this idea may have seemed too similar to what REx was doing, the key difference would be Stolley's suggestion to publish specs for researchers to mark up their data in the human-readable, flexible subset of the programming language Javascript, called JSON (javascript object notation). This mark up would allow data to be transferred into other programming languages, as needed, and would make the accessibility, circulation, and searching of data sets easier. Under this distributed model, rhetoric.io advisors (editors?) would publish pointers to the data sets marked up in JSON. This project, then, complements (and has plans to collaborate with) REx's mission to publish the research reports and will initially publish links to the JSON-tagged data.

[7] Those present at NHUK, and are part of the advisory board for rhetoric.io, include Collin Brooke, Derek Mueller, Karl Stolley, Doug Eyman, Cheryl Ball, Tarez Graban, Michael Neal, and (an addition later that spring) Michelle Sidler.

During this first phase of rhetoric.io, the advisory committee is set to complete three tasks:

1. **Creating an outreach API**: to invite participation from the networked humanities community regarding the kinds of data they would want to include
2. **Writing JSON spec**: based on known datasets, including those the advisory board already has access to as well as possible new data sets generated from the initial outreach API, write and test JSON spec for usability and flexibility
3. **Publish JSON spec**: distribute spec, train researchers to use it, help researchers collaborate with students who can use it

Below we offer a little more detail about each of these tasks.

Creating an outreach API

The purpose of building an application before creating a tagging specification for data is to ask the community of networked humanities scholars about the demographics of their data sets. We want to make sure, in other words, that the application we build that will eventually collect and distribute the tagged data will cover all the *kinds* of data researchers might want to include. This API includes the following types of questions, which require only very short answers:

- what is your project?
- what is it about?
- what motivated the data collection? (purpose)
- who do you think will find this data useful?
- what format is it in now?
- where is it located now?
- do you need a place to store your data?
- is the data set complete?
- if not, at what rate are you collecting more data?
- what do you want this data to do?
- who directs/runs/manages it?
- are you interested in working on this project?
- email contact

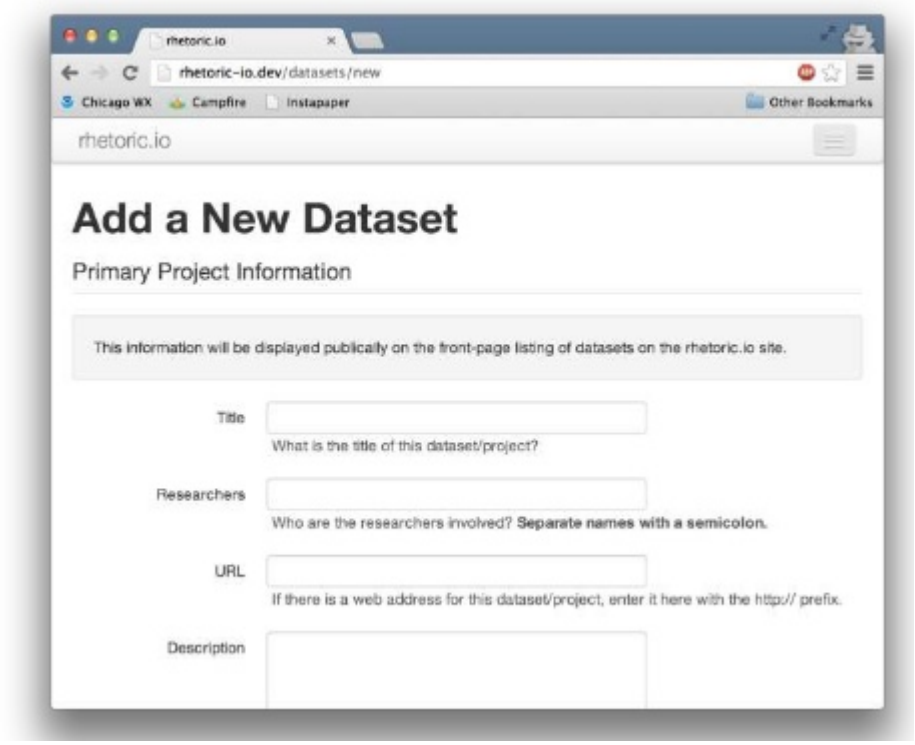

Figure 1: Partial screenshot of the in-development rhetoric.io dataset query API.

The question of formatting, in particular, is important to rhetoric.io's start-up, as researchers might have highly variant filetypes (videos, audio, word-processing docs, etc.) and mimetypes (.doc, .docx, .xls, .mov, .wmv, etc.). As we create the specifications for mark-up in JSON (described below), we need to accommodate as many of those file and mimetypes as possible. For instance, the tags used to mark up a video file (which other than its filetype, mimetype, and codec isn't otherwise searchable by most computers) may be much different than the tags used to mark up a web page (which is, by its native HTML language, already computer-readable).

Writing JSON spec

We settled on JSON (a derivation of the javascript programming language) because of its rendering speed in browsers and its flexibility and readability to both humans and machines. We believe JSON's readability will ease the learning curve for researchers less familiar with mark-up and programming languages, which is an important point even for some advisory board members who are still learning how to program in these languages. For instance, here is one of the many JSON examples Stolley has written for us to show the human-readability of this language:

```
{
  "title": "Our Amazing Research Project",
  "researchers": [
    "Alice Smith",
    "Bob Smith"
```

```
  ],
  "url": "http://example.com/amazing",
  "description": "Amazing research project is amazing."
}
```

To write the JSON spec, we will practice with known datasets and fine-tune the mark up (e.g., the tags that preface the data content, in the example above) that will be distributed as the tagging guideline for the repository. Some of these datasets that we have access to include are:

- Masters Degree in Rhetoric consortium survey data
- Metadata Mapping Project (MDMP) for tracing female pedagogues
- *CCC* archive
- FSU Card Archive
- *Kairos* webtext and media-element metadata
- Writing Studies Tree archive
- RikiWiki

In addition to these data sets, we will include examples (particularly additional multimedia ones) from other researchers' supplied datasets as part of the initial outreach API, or invite those researchers to join in the writing, if they are interested in collaborating with the advisory board.

Publishing the JSON spec

Through the rhetoric.io website, we'll distribute the JSON spec for any researcher who wants to test it out and use it. Education is a large part of this task, and it is the goal of the advisory board to collaborate with researchers and students to train them to use the spec, as well as make modifications of it, so that their data becomes valuable and usable outside of their own projects. As important as TEI (text-encoding initiative) has become to literary and linguistic fields within the networked humanities, making the marked-up copy of digitized corpora available to researchers worldwide, rhetoric.io's equally malleable specifications could help us collate decades worth of born-digital data for use and re-use. And, like TEI, we can teach each other through conference workshops, informal gatherings, online tutorials, and pedagogical projects to invest in our data as a continued research resource and to link to it in meaningful ways. This is another way that networked humanities scholars can take advantage of the National Endowment for the Humanities' grant offerings, specifically by proposing Institutes for Advanced Topics in Digital Humanities, which are typically week(s)-long summer workshops where participants are often compensated for their attendance.

One of the primary hurdles to this work succeeding, however, is the same hurdle that Fishman and Mullin encountered with data collection in their REx project: ethics. That is, writing studies has been slow as a discipline to take up their research as data-driven and generalizable in the sense that Institutional Review Boards define research. When, as this field knows, many IRB offices don't see our work as "research," it's easy to bypass the lengthy proposal forms that ask us to reenvision our qualitative research in quantitative terms--a process not amenable to some

types of writing research. But, a few minor changes in terminology--from "students" to "research participants", from "student assignments" to "participant data", from "classroom" to "study", for instance--and our (in this case, pedagogical) work fits easily within the forms of those oversight committees. Plus, as Will Banks and Michele Eble (2006) and Heidi McKee (2004) have argued, having to complete IRB forms makes us more ethical researchers, which can only be a good thing for us as scholars and as models for our students.

In addition to changing our terminology, however, we also need to change the way we think about the usefulness, security, and preservation of our data, as written up in our IRB proposals. For instance, at Illinois State University, the IRB proposal form has a whole section on data, which requires researchers to specify

- whether data are anonymous, confidential, or neither
- how and where data will be stored and secured (including the building and room number)
- who will have access to the data
- how the data will be used during and after the research has been conducted, including dissemination strategies, and
- how the data will be disposed of

Ball, in her work as department IRB representative, worked closely with faculty and graduate students to complete these forms in a way appropriate to the IRB office. During that experience, the answers for these questions rarely changed. In sum, humanities researchers would choose what they thought would be the least intrusive and least complicated means of tagging, storing, accessing, and deleting the data, which usually involved some random sentence about "this data will be destroyed 7 years from the date of collection," as if that number magically made the whole IRB proposal passable in the eyes of the IRB committee. These examples demonstrate an ongoing need for guidelines and standards that apply specifically to research in rhetoric and writing. Projects like REx and rhetoric.io make such standards even more important because they necessitate not just an understanding of basic IRB protocols, but also a critical awareness of how data must be collected and stored for long-term use and reuse.

What humanities scholars often fail to realize is that the data *are as important as* the published articles (or books) that use that data. For instance, several scholars have been trying to update the findings of a 2005 CCCC Research initiative grant on multimodal composition (see Anderson et al, 2006a), and to do so, they've looked to the 141 survey questions and answers from 45 scholars (i.e., loads of data) the research team published alongside the article in *Composition Studies* (Anderson et al, 2006b). Granted, the publication of that data is fairly messy, but it's at least open and available to future researchers. Making the change in disciplinary attitude from "destroy our data" to "publish and preserve our data" is part of the educational mission of rhetoric.io. In short, we should view our scholarly work in ways similar to other human subjects researchers. It includes participants, methods, data, and ethics. It builds on, and converses with, other research in the field. And, it involves data that may be reused by other researchers in the pursuit of other research questions.

Conclusion

In the second phase of rhetoric.io, once a more suitable and long-term social, scholarly, and technical infrastructure can be built (Eyman & Ball, forthcoming), the distributed system of links and specs will transform into a portal that can serve as a sustainable data storage venue suitable for long-term access and preservation. This is similar to WritingPro: Knowledge Center for Writing Process Research (http://www.writingpro.eu/share.php), which debuted during the summer of 2013. WritingPro, modeled on REx, collects research reports from scholars, but can also collect data from those studies (in private, semi-private, and public offerings).

As of this writing, WritingPro, like REx and rhetoric.io, has yet to publish any entries. However, our hoped-for outcome is that the rhetoric.io portal will embody and demonstrate a truly cross-disciplinary articulation of the ways we approach, value, and understand data relationships to function rhetorically and disciplinarily. The possibilities are there that these three archives (and hopefully others) will *make* data public, *form* data publics, help us argue by publicly *using* data, and open the boutique (data) of writing research *to the public*. These projects embody and perform more than technological advancements in data management. We believe they can enable a cultural sea change for our discipline analogous to that happening in many fields of science. We envision a time when our discipline comes to understand research—its inquiry, methods, and data—as a communal enterprise. And, we anticipate boutique data will lead to greater opportunities for research projects that are inclusive of many things, including data sets, researchers, and publics, both within and outside the academy.

REFERENCES

"A brief history of REx." *Research Exchange*: An Index of Contemporary Writing Research. http://researchexchange.colostate.edu/about/history.cfm

Anderson, Chris. 2004. "The Long Tail." *Wired* 12.10. http://www.wired.com/wired/archive/12.10/tail.html

Anderson, Daniel, Anthony Atkins, Cheryl E. Ball, Krista Homicz Millar, Cynthia Selfe, and Richard Selfe. (2006a). "Integrating multimodality in composition curricula: Survey methodology and results from a CCCC Research Initiative grant." *Composition Studies, 34*(2), 59–84.

Anderson, Daniel, Anthony Atkins, Cheryl E. Ball, Krista Homicz Millar, Cynthia Selfe, and Richard Selfe. Designed by Matt Bemer. (2006b). "Survey of Multimodal Pedagogies in Writing Programs." *Composition Studies*. http://www.compositionstudies.tcu.edu/archives/342/cccc-data/

Banks, Will, and Michele Eble. (2006). "Digital Writing Research: Technologies, Methodologies, and Ethical Issues" *Digital Writing Research* (ed. Danielle Devoss and Heidi McKee). Cresskill, NJ: Hampton Press.

Cohen, Daniel J. "The Social Contract of Scholarly Publishing." In Matthew Gold, ed. *Debates in the Digital Humanities*, 2013 open-access edition. http://dhdebates.gc.cuny.edu/debates/text/27

Eyman, Doug, Stephanie Sheffield, and Danielle Nicole DeVoss. 2009. "Developing Sustainable Research Networks in Graduate Education." *Computers and Composition, 26(1):* 49-57.

Fitzpatrick, Kathleen. "Beyond Metrics." In Matthew Gold, ed. *Debates in the Digital Humanities*, 2013 open-access edition. http://dhdebates.gc.cuny.edu/debates/text/7

Hawhee, Debra, and Christa J. Olson. "Pan-Historiography: The Challenges of Writing History Across Time and Space." In *Theorizing Histories of Rhetoric*, edited by Michelle Ballif. Carbondale: Southern Illinois University Press, 2013. 90-105.

Heidorn, P. Bryan. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends*, Volume 57, Number 2, Fall 2008, pp. 280-299.

Krause, Steven. 2007. "'Where do I list this on my CV?' Considering the values of self-published websites, version 2.0" *Kairos: A Journal of Rhetoric, Technology, and Pedagogy*, 12(1). http://kairos.technorhetoric.net/12.1/binder.html?topoi/krause/index.html

McKee, Heidi. "Examining the Process of Institutional Review Board Compliance." *College Composition and Communication* 54 (2003): 488-493.

Mueller, Derek. 2012a. "Grasping Rhetoric and Composition by Its Long Tail: What Graphs Can Tell Us About the Field's Changing Shape." *CCC* 64: 195-223.

Mueller, Derek. 2012b. "Views from a Distance: A Nephological Model of the CCCC Chairs' Addresses, 1977-2011." *Kairos: A Journal of Rhetoric, Technology, and Pedagogy, 16*(2). http://kairos.technorhetoric.net/16.2/topoi/mueller/index.html

Office of Digital Humanities. (n.d.) Digging Into Data Challenge. *National Endowment for the Humanities* [website]. http://www.neh.gov/grants/odh/digging-data-challenge

Presner, Todd. "The Digital Humanities 2.0: A Manifesto." *Humanitiesblast.com*. n.d. Accessed 1 January 2013. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf

Ridolfo, Jim. (n.d.) *Rhet Map: Mapping Rhetoric and Composition*. http://rhetmap.org/

Tirrell, Jeremy. (2012). "A Geographical History of Online Rhetoric and Composition Journals." *Kairos: A Journal of Rhetoric, Technology, and Pedagogy, 16*(3). http://kairos.technorhetoric.net/16.3/topoi/tirrell/index.html

UCLA Center for Digital Humanities. "A Digital Humanities Manifesto 2.0". http://manifesto.humanities.ucla.edu/2009/05/29/the-digital-humanities-manifesto-20/

Ulman, Louis, Scott Lloyd DeWitt, and Cynthia Selfe. 2013. *Stories that Speak to Us*. C&C Digital Press/Utah State University Press: http://ccdigitalpress.org/ebooks-and-projects/stories

Waldrop, M. Mitchell. "Science 2.0: Great New Tool, or Great Risk?" *Scientific American*, January 9 (2008), http://www.scientificamerican.com/article.cfm?id=science-2-point-0-great-new-tool-or-great-risk.