Research Note

Possibilities for a Public-Facing Digital Writing Program Archive in the Age of Analytics

Kyle Oddis, Avery Blankenship, Brice Lanham, and Neal Lerner
Northeastern University

Structured Abstract

- **Aim:** The aim of this research note is to demonstrate the great potential of digital writing program archives in expanding our knowledge of writing program history and encouraging similar efforts at multiple institutions. The project we describe is a digital public archive of one institution's writing program—Northeastern University in Boston, Massachusetts—intended to add to the cumulative history of the "archival turn" in writing studies scholarship by combining the archival turn with the "public turn" through digital access. In this research note, we discuss possibilities for the use of analytics on our digital archive—the first of its kind that we know of—and outline the methodological, legal, ethical, and practical complexities of collecting, sorting, ingesting, and organizing these data with the ultimate goal of making them accessible to the public. We also discuss the implications of using analytics tools on this archive and what this could mean for future institutional assessment, classroom practices, and writing program administration.
- **Problem Formation:** Beyond using analytics to understand the learning processes of student writers, a digital writing program archive also offers us a chance to use writing analytics as means to understand the processes, products, ethos, values, and goals of writing instructors, administrators, and the institutions that house writing programs. Where a broader set of tools is being used in learning analytics, a digital writing program archive offers us a

new set of texts as data that reach beyond students' texts and into the texts that administrators and instructors use to teach each other, learn from one another, and improve and (re)form writing programs to better serve their institutional and broader communities. Such an archive also asks us to consider the ways in which we narrativize institutional histories.

- **Information Collection:** The work of developing a workflow for our archive has meant confronting and organizing the informal archives that exist in a variety of spaces in Northeastern University's English Department. Much of the materials of our archive, before it was formalized as an archive, existed in filing cabinets and abandoned boxes in storage closets. The work of categorizing these materials has largely involved sorting, tagging, and making judgment calls based on the perceived relevance of the material to our project. In this research note, we present a potential workflow for researchers dealing with similar informal writing program archives and suggest that some of the steps towards formalization involve attending to the politics of data collection and metadata development. Metadata, or data that describes data, can be developed and shaped to suit the needs of a particular project, but crucially, it must still belong to a wider ecosystem of metadata formatting and patterns. The ability of a particular project's metadata to be legible and useful to anyone who is not familiar with the project is a significant design consideration, and metadata and the data categorization decisions behind its development are one of the many ways that a writing program digital archive can begin to "talk back" to the institutions from which it came. In this section of the research note, we outline our process and discuss our practices for ingesting and describing our texts as data.
- **Programs of Research:** As we proceed with our archive in consideration of the four programs of research outlined in the writing analytics taxonomy, a discussion of values and tools becomes vital: we cannot understand how various constructs (represented in an institution's values and as reflected through things like mission and outcomes statements and learning goals) operate in context if we do not consider context on the local level. An institutional archive allows us to more closely examine our local context, but we are also aware that tools and technologies are not neutral objects—they reflect their designers' values and interests. As such, tools and technologies are always situated in historical contexts: that is, sites. This section of our research note examines technologies situated within local writing sites as a reflection of institutional values.
- **Conclusions:** Our project to this point grapples with many unanswered questions. The material realities of creating and curating a digital archive are

laborious; we recognize the importance of institutional resources and discuss the potential for partnerships that might help us better understand our own institutional history as situated within the vast landscape of writing studies. We also discuss challenges to narrativizing our history and presenting that history in a public digital space. Ultimately, we conclude that the future of our archive must consider not only the ways in which we curate and manage data, but also the ways in which technologies, as the vehicles of the archive, also require the stewardship of human hands. Therefore, it is important to consider the ways in which designing the archive's digital space might enable us to be better stewards of the information it houses while considering the ways in which its information might be taken up by future generations of writing studies and writing analytics researchers.

Keywords: archives, archival studies, digital archives, historiography, project management, writing analytics, writing assessment, writing curriculum, writing pedagogy, writing program, writing program administration, writing studies research, texts as data

1.0 Aim

Expanding the body of knowledge in writing analytics and in writing studies more broadly requires deep consideration of how and why we collect, store, manage, and use the data we have. It also prompts the necessary question: *What do we have?* This research note stems from realizing that development of new tools and technologies for dealing with data opens new doors for us to (re)consider what counts as data—and in that reconsideration, rethink how we tell not only the stories of our students and their learning of writing, but also how we tell the stories of the institutions where we research and teach writing through texts collected over years (and often decades) that change hands or sit in boxes in writing program, writing center, and department offices. Enter: the potential of the archive.

The ubiquity of digital environments for learning, teaching, and researching in higher education prompts us to consider how we might use archival materials to broaden our understanding of our communities of practice, methods of assessment, and curricular and pedagogical developments. Necessary, too, is considering the ethical implications of unboxing and transforming archival texts into usable, living data for researchers and practitioners seeking to identify and improve upon the insights garnered from ink-stained spectres of the past. Within sometimes long-forgotten boxes, we find new avenues for expanding knowledge-making; within the landscapes of our digital environments, we find hope for rewriting our shared history and potential for remaking our practices to respond to the challenges of our changing world.

The aim of this research note is to demonstrate the great potential of the creation of digital writing program archives in expanding our knowledge of writing program history and encouraging similar efforts at multiple institutions, endeavoring to tell the stories of our pasts in a new era: one capitalizing on the power and ubiquity of digital access. Our digital public archive

of one institution—Northeastern University in Boston, Massachusetts—has great potential, also, to expand the body of knowledge in writing analytics by rethinking what it means not only to curate and store texts as data, but also to assess historical texts comprehensively using digital tools and platforms in hopes of improving student learning and program efficacy in writing.

Central to the difficulty of imagining our archive and how it might contribute to the creation of similar future archives is reckoning with the fact that part of what makes our archive unique and potentially powerful is precisely what also gives it the most potential for harm. Biesecker (2016) claims that "the archive may best be understood as the scene of a doubled invention rather than as a site of a singular discovery," (p. 156) and we feel it is our responsibility as scholars talking back to our institution and to our field to interrogate our "under-interrogated relationship to the archive" (p. 157) as one which necessitates not only active awareness of, but active participation in how we position ourselves through its contents. A major aim of this work, then, is to make visible how processes of classification (Bowker & Star, 2000) not only inform but influence our decisions and thereby, perhaps, the future of our program and our work within the larger field. We recognize here, also, the need for more extensive digital, accessible archives in the field that represent both individual institutions and their broader contexts (e.g., conferences, institutes, and external organizations); the archive we describe building here is only one of many steps we might take to better understand the work we already do and the work we hope to do in the future.

1.1 The Archival Turn in Writing Studies

The past 15–20 years have constituted an "archival turn" of sorts in writing studies scholarship, given the volume of publications that have drawn on archival sources to tell the histories of writing courses and writing programs in U.S. higher education. Recent publications have focused attention on individual institutions (e.g., Ostergaard & Wood, 2015; Ritter, 2009, 2012b), individual programs (e.g., Fleming, 2011; Gold, 2008; Lamos, 2011), and individual figures (e.g., L'Eplattenier & Mastrangelo, 2004; Lerner, 2001; Varnum, 1996;) as a complement—and at times counter—to the grand sweeping historical narratives of writing studies (e.g., Berlin, 1987; Connors, 1997; Kitzhaber, 1990). These studies complicate what the field knows about its history and challenge long-held beliefs that all first-year writing grew out of Harvard's implementation of English A in the 1870s (Donahue & Moon, 2007). This work also demonstrates that writing instruction has long been a feature at most, if not all, colleges and universities (Gannett, Brereton, & Tirabassi, 2010), and that this instruction—its purposes, values, and methods—has been as diverse and situated as all instruction needs to be.

While these published research narratives of institutional and program history draw on archival materials, there have also been several efforts to create accessible archives in writing studies, ones we see as complementary to what we are building at Northeastern, as well as examples to learn from when it comes to access, availability, and usability. These efforts include the Digital Archive of Literacy Narratives (DALN; https://www.thedaln.org/), a wide-ranging attempt to capture brief perspectives on the reading, writing, and communicating experiences of

its contributors. The DALN is also a public and freely accessible archive, thus offering material for teaching or further study as we hope to do with the Northeastern Writing Program Digital Archive. Another long-standing effort is the National Archives of Composition Rhetoric, a joint project between the University of New Hampshire and the University of Rhode Island that houses a great deal of material but has struggled with issues of access and availability. The Writing Centers Research Project, originally housed at the University of Louisville and now at the University of Arkansas Little Rock, also was able to curate a great deal of material (including oral history interviews with over 25 early figures in the field; see Smitherman, 2007), but very little of that material is available online, and its current physical status is unknown. A related and rich project is a section of the larger Documenting the American South archive at the University of North Carolina (UNC): "True and Candid Compositions: The Lives and Writings of Antebellum Students at the University of North Carolina" offers readers access to writing from students and administrators at UNC between 1795 and 1868, including letters, reports, and essays (see https://docsouth.unc.edu/true/), all transcribed, edited, and annotated. Finally, the online search engine CompPile (https://wac.colostate.edu/comppile/) is a digital archive of sorts, both in its online and full-text holdings of composition studies materials between 1939 and 1999. As one reviewer of a draft of this manuscript suggested, the possibilities for integrating CompPile or other search engines (what digital archivists described as "linked data") with the Northeastern Writing Program Digital Archive are exciting, offering multiple ways to achieve our purpose of easy and broad access. One more example is the Open Syllabus Project (https://opensyllabus.org/), currently offering public digital access to over seven million syllabi from colleges and universities in 80 countries, largely culled or scraped from materials posted online. While it is not specific to writing classes, the top two most frequently assigned texts are William Strunk's *The Elements of Style* and Diana Hacker's *A Writer's Reference*, offering some evidence that writing program courses do make up a good deal of that database.

In sum, the raw data of the writing program and writing center histories—the written materials that any writing program/center collects over the course of a semester, an academic year, a decade, or more—with the few exceptions we just described, are rarely accessible for research (or any) purposes other than for an individual scholar, usually on site in a physical archive. However, these syllabi, memos, emails, reports, and meeting minutes offer powerful possibilities for analysis, particularly with the use of digital analytical tools.

1.2 Our Digital Archival Turn

The history of the Northeastern University Writing Program (NUWP) is one that many writing studies scholars and practitioners can easily relate to. In the early 1980s, led by Richard Bullock, the program grappled with assessing and placing incoming students unprepared for college-level writing; its long-standing writing center, designed to serve the entire university, and an upper-division-required disciplinary writing course (AWD) have adapted over time to meet the ever-evolving needs of students, the university, and the surrounding community. Finally, the program has shifted as undergraduate student demographics have shifted from a largely local, urban

student population to a much more diverse, highly-prepared, and global student body. Through these changes, the activities of the NUWP—and the English department more generally—were documented in a relatively large collection of paper materials, including meeting minutes, memoranda, reports (including two writing program administration external evaluator reports), news clippings, readings, student writing, syllabi, and assessment prompts from the early 1980s to the present.

The project we describe in this research note is intended to add to this cumulative history, but not simply by crafting a narrative based on archival materials. Instead, our intent is to combine the "archival turn" with the "public turn" in our creation of the Northeastern University Writing Program Digital Archive and to suggest some ways in which an archive like this might create unique opportunities in the growing field of writing analytics to develop and utilize data methodologies/practices while considering and challenging long-held beliefs surrounding the ways in which institutions curate and narrativize their histories.

In this research note, we discuss possibilities for the use of analytics on a digital archive like this—the first of its kind that we know of—and outline the methodological, legal, ethical, and practical complexities of collecting, sorting, ingesting, and organizing this data with the ultimate goal of making these data accessible to the public. We also discuss the implications of using analytics tools with these data and what this could mean for future program, classroom, and institutional assessment and administration, including but not limited to the following:

- Writing assessment: Researchers can track the attention to writing placement over time.
- Writing products: Researchers can conduct a variety of text-based or coded analysis of the collection of student writing in our archive.
- Writing curriculum: Our deep collection of syllabi for required writing courses offers a variety of angles on the changing nature of curriculum.
- Writing pedagogy: Our collection of teaching materials lends itself to analysis of how teaching practices are influenced by particular required texts, student populations, and university initiatives.
- Writing program administration (WPA): The memos, reports, and other documents on the day-to-day functioning and larger strategic planning of writing program goals, values, and practices could be explored for what they say about the changing nature of WPA work.

2.0 Problem Formation

2.1 The Archive is Not Merely a Collection

Archival research draws on multiple disciplinary perspectives, including history/historiography, library science, writing studies, digital humanities, and information sciences. Important to all of these traditions is the idea that an archive is not a static collection of materials; instead, as Cook

(2001) describes, current understanding of the archive is "a shift away from looking at records as the passive products of human or administrative activity and towards considering records as active agents themselves in the formation of human and organizational memory" (p. 4). The materials in an archive, in other words, are in a dynamic relationship with the context of their creation, the creators themselves, and the beliefs, values, and goals of anyone examining those materials. It is perhaps a statement of the obvious, but not every document generated ends up in the archive (as our project surely attests), not every document is chosen to remain in the archive or be retrievable, and not every document in the archive itself might be viewed or used with equal weight by an eventual user. Archival research is thus a social process, one permeated with human decision making—both implicit and explicit—and always, then, an incomplete capturing of some phenomenon and only a partial epistemological product of the one who attempts to discover what the archive contains.

Add to this complexity the digital nature of contemporary archives. While digitization offers the promise of increased access (i.e., rather than an individual researcher needing to travel to a physical archive, access is granted to anyone with a computer and an internet connection), digital archives create their own issues of retrieval and partial knowing, whether a limited set of scripted tags or keywords for information retrieval (Sternfeld, 2011), a separation of materials from their original context (Zhang, 2012), or simply a structure selected by the creator(s) that tells a limited story of the materials themselves (Zhang, 2012) or one with a clear political agenda (White & Gilliland, 2010). Still, the move to digital archives and the resulting increased access promise to fulfill the intent of the archive as a dynamic, socially constituted rhetorical space, as well as an educational resource (Hayden, 2017; White & Gilliland, 2010).

In writing studies more specifically, recent attention to archival research has stressed its importance to the field, while deepening our understanding of methods (L'Eplattenier, 2009), the need for transparency and the researcher's relationship to the materials themselves (Ritter, 2012a), as well as the power and authority that archival creation and maintenance will always encounter (Cushman, 2013). Hayden (2017) extends these understandings to a curriculum for an undergraduate writing course, using digital archival research to offer students opportunities to explore communities, educational contexts and histories, and to "recover lost voices" (p. 143). In sum, a writing program digital archive is potentially a means for understanding our history, offering access to that history to others, and achieving our learning and teaching goals for our students.

Achieving the goals we just described is not as simple as collecting and sorting lots of paper materials, scanning them into PDF files, and making that collection available via web interface—though those were all necessary steps. As we learned in this project and as we describe in this research note, the process from inception to current status (an incomplete digital archive) was filled with choices—ethical, strategic, uncertain, and collaborative.

2.2 The Archive's Place in Writing Analytics

Before addressing the methodological and practical issues involved in the building of an archive like ours, it is important to discuss the place of a digital archive in the emerging field of writing analytics. In the most recent issue of *The Journal of Writing Analytics*, Palmquist (2019) draws a valuable distinction between learning analytics and writing analytics, highlighting the "disconnect" that exists between the two emerging fields. Palmquist points out that the "two areas of analytics research might productively inform each other" (p. 1), and we believe that the creation and curation of a digital archive and development and implementation of tools to engage with it is one way we might begin to bridge this disconnect.

In a larger context, learning analytics primarily consists of sets of tools turned into products offered by organizations that are external to the institutions at which these tools are utilized; for instance, Canvas and Blackboard are examples of learning management systems (LMS) which offer adaptive assessment tools that reveal student users' learning behaviors (Palmquist, 2019). Learning management systems generate a range of reports that are becoming more common as data sets that programs and institutions can use to evaluate student "success," retention, and achievement; however, as Palmquist also points out, skepticism toward these tools is warranted given that they are both 1) still in development and 2) consequential, since they "have the potential to shape the academic paths taken by a large number of students" (p. 4). As of writing this piece, a global pandemic is changing the face of higher education; we must be mindful of our willingness to, on the one hand, embrace the reality of the growing ubiquity of learning analytics tools, and, on the other hand, be critical of the potential consequences of their use as we (as writing studies scholars) help shape the fields of learning analytics and writing analytics moving forward.

We take care to revisit Palmquist's differentiation of learning analytics and writing analytics in order to address one of the root problems of the perceived divide: that some scholars have "embraced the term 'writing analytics,' often without an awareness of the origin of the term or its connection to efforts in learning analytics" (p. 4). Norbert Elliot, Palmquist notes, has argued that the key difference between learning analytics and writing analytics is that "writing analytics is construct-based, while learning analytics is not" (2019). In Elliot's understanding, writing is seen as "a network of reading, speaking, and listening" as "understood within four domains: cognitive, intrapersonal, interpersonal, and neurological," and it is "fine-grained construct articulation" in writing analytics which "allows an actionable framework for our research" (Palmquist, 2019, p. 5).

To take up Elliot's definition and Palmquist's call to employ a "larger set of methodological tools and a more robust set of research goals," we present the potential of our archive. Even beyond using analytics to understand the learning processes of student writers, a digital writing program archive also offers us a chance to use analytics tools as means to understand the processes and ethos of writing instructors, administrators, and the institutions that house writing programs. Where a broader set of tools is being used "in learning analytics to understand how our students learn to write, the points in a course at which they encounter difficulties, and

instructional strategies we might employ to improve their learning and success" (Palmquist, 2019, p. 6), a digital writing program archive offers us a new set of texts as data that reaches beyond students' texts and into the texts that administrators and instructors use to teach each other, learn from one another, and improve and (re)form programs to better serve their institutional and broader communities.

2.3 Texts as Data

Understanding texts as data means reimagining what texts are and what happens when humans read texts differently from the ways in which machines read texts. Because writing analytics is a field that uses both qualitative and quantitative methodologies (the latter of which often involves transforming qualitative observations into counts and scores that can then be used to run statistical analysis), it is important to outline what it means to engage digital and analytics tools in processing and interpreting texts created by humans. This becomes particularly important in considering possibilities for a digital archive, given that it is a hybrid environment that combines human and machine labor.

Gentzkow, Kelly, and Taddy (2017) discuss representing text as data in economics, articulating the importance of making a distinction between human and machine readers in that "much of the text analysis in machine learning more generally, ignores the lion's share of [interpretive and linguistic] complexity" (p. 4), which means that representing text as data often involves a process of "cleaning" to "reduce the number of unique language elements we must consider and thus the dimensionality of the data," which can provide "massive computational benefit" but requires "careful decisions about the elements likely to carry meaning in a particular application" (p. 6). Writing analytics engages in this practice through use of automated writing evaluation (AWE) and corpus analysis, for instance. But what does this mean for our archive if we understand our texts as data that might be mined and cleaned for computational benefit? First, it means being deliberate about the decisions we make regarding the texts we include in the archive to become available for machine uptake. Second, it means understanding that, as Grimmer and Stewart (2013) note, "automated content analysis methods will never replace careful and close [human] reading of texts," but machines may be used to amplify and augment "careful reading and thoughtful analysis" (p. 268). Third, use of analytics tools opens doors into discovering new methods of classification and scaling of texts one might find in a digital archive in order to "discover new ways of organizing" those texts (Grimmer & Stewart, 2013, p. 269), which might reveal insights into curricular and programmatic realities that are invisible to the careful reading and thoughtful analysis enacted by human agents.

Research in writing analytics (and writing curriculum and pedagogy more broadly) frequently uses student texts as data, but there is a major opportunity to apply the same analytics tools and methodologies to curricular course and program texts (as well as potentially to institutional communications) in order to discern and better understand, put simply, where we come from and where we want to go. Indeed, many departments and programs are sitting on a "gold mine" of data from intentionally or unintentionally collected documents that span years or

decades—Digication and Blackboard repositories and Google Drive folders of syllabi, assignment sheets, rubrics, etc., for example—so the questions that surface are these: What can we do with all that data? What *should* we do with it, and why? The fact, for instance, that the materials collected and ingested into our archive sat in boxes for decades *untouched* speaks to the potential to learn from what we have as a way to help inform where we might go. The potential is also expansive, which is why—as we will discuss below—documentation becomes increasingly important as we look toward the future to better equip those that might take up this work in subsequent generations of writing studies and writing analytics scholars. Hovering in the background, however, is a sobering realization that some of what we might find by applying analytics methods to our archive will reveal truths about our past and current efficacy that we might be reluctant to face. For instance, how do we make sense of a history of basic writing that focused on sorting and classifying students based on far-from-perfect placement instruments? What might we say about syllabi that are seemingly well intentioned but simply replicate structural inequities and treat students as mere consumers of knowledge, not knowledge creators? These legacies of likely most college and university writing programs in the United States are revealed by the archive; a digital public archive renders them visible for all.

What consequences for analytics research might this visible rendering have? In imagining the work we might be able to do with a digital archive, we envision a number of potential avenues for future scholarship. Following the initial steps of information collection and ingestion (described below), taking steps to render machine-readable text files will allow researchers to apply corpus linguistics tools to sets of texts; for example, a researcher might want to trace the usage of a specific textbook in a writing program over time, and opt to plot that data on a visual or graphic (say, a timeline) or create a digital cluster model of texts that appear frequently in a first-year writing curriculum. Researchers might also choose to focus on texts like meeting minutes to quantify how much time is spent discussing or revisiting functional departmental issues. Even a simple word frequency count in certain text genres might render interesting information for program administrators seeking to grasp a deeper understanding of the ways in which a writing program articulates its values and priorities in training manuals, which could lead to valuable insights for future program assessment, creation of new faculty development tools, or syllabi. We also imagine our archive as a usable tool for professional development of students (undergraduate and graduate) who might want to practice applying new qualitative or quantitative research methodologies to texts as data. As we will discuss at length, creating files that are usable for writing analytics work requires multiple steps and deep consideration of how we decide to tag, sort, and classify the data we have, since this will affect the kinds of work that researchers are able to do. Outlined in the next sections are the practical steps we have already taken to lay groundwork for making our materials usable for this type of work in the future.

3.0 Information Collection

3.1 Developing a Metadata Format

In this section, we present a methodology for gathering, sorting, and developing a metadata format for the NUWP Digital Archive. As we have previously discussed, writing programs, sometimes intentionally and unintentionally, collect the ephemera of their programs. This ephemera includes syllabi, meeting notes, memos, and many other documents which—when pieced together—detail the history of a writing program. While arguably, the most interesting step in the process of developing a digital archive involves the set of decisions a researcher makes once the archive is preparing to be launched, attention to metadata as well as collection strategies is both essential and fascinating. Metadata can be understood as data that describes other data. In the case of a digital archive, metadata is the set of data which describes and categorizes the material. With limited similar projects[1] upon which to base our metadata format, the development of this format was a crucial first step which required a detailed first pass through the physical material with careful attention paid to the format, intention, and perceived audience of each piece.

In order for data to be useful to users of a digital archive, it is crucial that metadata for the collection is descriptive and coherent. Developing useful and thoughtful metadata for our project was a crucial step in our workflow because metadata can often determine what parts of a historical record are able to be searched for and retrieved by users of the archive. The effectiveness of search and retrieve ultimately impacts what aspects of the archive are considered to be possible avenues for study and encourages particular interactions with archival records that are informed by social, political, and technological powers. According to the National Information Standards Organization (NISO; 2016), "the existence of searchable descriptive metadata increases the likelihood that digital content will be discovered and used." Since increased use and accessibility of writing program historical materials is a priority of our project, this description of metadata is particularly salient.

In the case of a digital archive, the transformation of a material record into a digital one requires a level of separation from context, which makes reliable metadata even more crucial. Because our archive is composed of documents created by many students and instructors who are no longer a part of our institution, the context of our records is already fragile; the transformation of a folder of syllabi into individual scanned files further isolates these records from their inherent sense of connection to other records.

Metadata is typically grouped under three distinct categories: descriptive metadata, structural metadata, and administrative metadata (Bantin, 2016, p. 120). For the purposes of our archive, our metadata is primarily descriptive, documenting the bibliographic and material attributes of a

[1] Although there are no predecessors to our Writing Program Digital Archive, we acknowledge and give credit to the other digital archive projects housed by Northeastern University whose existent work provided us a base for developing meaningful metadata categories. Special thanks to the work of the Early Caribbean Digital Archive and the Thoreau Drawings Archive, which are both housed by Northeastern.

record and uniting these various elements under a unique ID. However, because much of the process of sorting and cataloging our records was an act of discovery (when we began this project, we had some sense that the material we had acquired would be connected to the writing program, but not all of it), it was necessary that our metadata be mutable and editable. We paid particular attention to NISO's guideline that "different types of metadata can be added by different people at various stages of an information object's life cycle."

3.1.1 Material Acquisition and Intake

This project began with a total of seven very large boxes of paper materials. Most had been contributed by a previous writing program administrator who decided to clean out her office; others represented materials collected by previous academic administrators who had since left the university. The oldest materials were found in a filing cabinet in the then-writing center director's office, though that involved tracking down a key that had been lost years prior. Each individual sheet of paper in these boxes would eventually be checked over for its relevance to the project, and then materials would be categorized and logged with appropriate metadata—but before this process could begin, key observations had to be made regarding how the item could be understood in an archival setting. Additionally, as is often the case with unsorted and unlogged archival material, not everything in each of the seven boxes would necessarily be useful for the purposes of our archive, though they may be of use to Northeastern University's Special Collections, which is a much larger archive.[2] For this reason, it was determined that metadata should not only be recorded for the items of interest, but that items which were to be excluded from our particular project should be marked as such and appropriate metadata recorded for these items as well.

The metadata record for each item was recorded in a spreadsheet using Google Sheets. Google Sheets' ability to allow team members to work on the document simultaneously was an essential element of our workflow since metadata and document records would be recorded at various junctures of the project's timeline and not necessarily always in chronological order. The ability for the spreadsheet to be arranged by date or title was especially important for keeping track of timelines. In addition to metadata, for each record, a column was devoted in the spreadsheet to marking the document as a "yes" (to be included), a "no" (not to be included), and a "maybe" (possibly to be included).

The heuristic we selected for determining whether or not an item would be included in the archive was to question whether or not each item could be connected to the English department's writing program at Northeastern. If an item was definitely of interest, that item was tagged "yes," and if that item was determined to not have a connection to the writing program, it was tagged "no" with accompanying metadata. Those items of which we were uncertain were tagged "maybe," which indicated that they should be reevaluated (a process that involved discussion and follow-up regarding legalities, permissions, and relevance). Each of these documents was then catalogued in a spreadsheet with metadata columns, including document titles, document

[2] https://library.northeastern.edu/archives-special-collections

creators, intended recipients, dates, document type tags, and a brief descriptive abstract. These metadata columns in the spreadsheet were populated as the information was present and left blank in cases where data was not explicit on the document itself. For example, if no author name was clear on the document itself, the author column was left blank for that particular document. We chose to specifically not make assumptions regarding authorship or any other identifying metadata categories since to do so opens up the possibilities that our own politics and biases as archivists would be inscribed through the process of digitization. The documents were also filtered for any sensitive personal information such as student names, addresses, etc., and for the time being, those documents were either removed, or in some cases, this information was redacted.

Using Google Sheets' filtering functions, we were able to separate the "yes" files from the "no" and "maybe" files when we eventually made a second pass through all of the documents. During this second pass, file names were created and added to the metadata for each "yes" file so that anyone scanning the documents would be able to faithfully name the resulting PDF files. File names are an essential part of the metadata record since the file name is what links the PDF file to its corresponding metadata once the file is ingested into the archive. While other projects may choose to match metadata to archival materials using other metadata, there must be some form of unique ID that will connect the metadata to the material. In developing a metadata format for this project, team members met with members of Northeastern University's Digital Scholarship Group (DSG) to discuss how the metadata format we had developed—which was specific to our project and our particular needs—could be legible to the larger digital repository that our institution maintains for digital archival material.

Table 1

Metadata Format

Note. *Our metadata format can be seen above as well as how the fields are populated. The rows that are highlighted indicate that those documents have been ingested into Northeastern's DRS, though the collection still remains private.*

Date	Title/Subject	File Name	Document Type	Author	Recipients	Content Tags	Tagged For Interest	Description
	Computers & Writing	Proposal_Computers&Writing_NoDate_0002.pdf	Proposal/ Essay		The english department	Proposal	Yes	A description of a proposal for a hybrid program in the english department using technology. The program started in the Freshman Composition program
03/26/03	Talk to TLTR	Outline_TalktoTLTR_2003_0003.pdf	Outline			Outline	Yes	An outline of the history of hybrid courses in the english department
2006	ENG U311 Advanced Writing for Pre-Law	Syllabus_ENGU311_2006_0004.pdf	Syllabus	Elizabeth Britt		Syllabus, Courses, Writing Instruction	Yes	A syllabus for an advanced writing course
	English 309: Advanced Writing in the Humanities	Syllabus_English309_NoDate_0005.pdf	Syllabus	Ben Leubner		Syllabus, Courses, Writing Instruction	Yes	A syllabus for an advanced writing course
	Advanced Writing in the Disciplines 309: Writing for the Humanities	Syllabus_WritingintheDisciplines309_NoDate_0006.pdf	Syllabus	James Weiss		Syllabus, Courses, Writing Instruction	Yes	A syllabus for an advanced writing course
2006	Eng U309.2: Advanced Writing in the Humanities	Syllabus_U309.2_2006_0007.pdf	Syllabus	Christopher Weinmann		Syllabus, Courses, Writing Instruction	Yes	A syllabus for an advanced writing course
	English U307: Writing for Careers in the Natural Sciences	Syllabus_U307_NoDate_0008.pdf	Syllabus	Suzanne Richard		Syllabus, Courses, Writing Instruction	Yes	A syllabus for an advanced writing course
2006	English 307: Advanced Writing in the Sciences	Syllabus_English307_2006_0009.pdf	Syllabus	Cecelia Musselman		Syllabus, Courses, Writing Instruction	Yes	A syllabus for an advanced writing course

3.1.2 Ingesting a Sample Set of Documents

The metadata format was written to intentionally map directly to an ingestion format for Northeastern University's Digital Repository Services (DRS). The DRS is a digital archive that hosts archival material online on behalf of Northeastern community members. Choosing to upload our material into the DRS guarantees that the material is both securely preserved according to library and archival standards and guarantees that the material will be maintained digitally, even if our website were to malfunction. The DRS also comes equipped with useful workflow technologies that can limit the number of team members who can edit documents as well as close the collection to the public. This latter feature was especially important as we began to discuss the look of the archive's website as well as its function. Finally, storing the material in the DRS means that we will be able to use The CERES Exhibit Toolkit,[3] a tool that allows for simple builds of curated exhibits and digital collections in WordPress. Northeastern University created CERES as a specialized WordPress plugin which uses the DRS' API to pull metadata and digital objects from the DRS into WordPress.

In order to test that the scanned files would be able to match up with the metadata entries in the metadata fields successfully, we selected a sample set of one hundred documents that were relatively representative of the variety of documents that would eventually be uploaded onto the archive's website. With the help of Northeastern University's Digital Repository Services, the sample set of documents were queued and uploaded online, with their respective metadata attached. Once the sample set of files had successfully been ingested into the DRS, the ingested files were marked to indicate that they were ingested into the DRS. Initially, there were a total of 1,852 documents, and the eventual "yes" set of documents totaled 659. This sample set of documents remains closed to the public as of the writing of this research note, pending future decisions on access and analytics.

3.2 Naming, Organizing, and Optimizing Files

Before we could send a representative sample from the 659 documents to the DRS at Northeastern, we had to turn our physical collection into a digital one—and before we could do that, we needed to ensure that the collected metadata could be linked to both the physical and, eventually, digitally scanned documents. In order to do this, we developed a file naming convention. This file naming was done retroactively (once all the metadata had been collected) instead of simultaneously during the metadata collection process. In retrospect, it would have been more efficient to do this synchronously.

In order to ensure that file names were both representative of the content contained within the documents, as well as distinct so that one document could be identified from another, we used criteria from our metadata to develop a file naming convention. The three criteria we chose were document type, document title, and year so that the contents of any given document were clear from the file name, while a distinct four-digit number based on the row number of our populated

[3] https://dsg.northeastern.edu/ceres/

metadata spreadsheet was assigned to each document so that all other metadata for each document was linked and could be easily found. Knowing that the current collection of documents within the archive amounted to a triple-digit number (659), we specifically chose to include a four-digit number in the file name to ensure that the file naming convention allowed for future growth of the archive. The resulting naming convention is shown in Table 2.

Table 2

A Single Row in Our Metadata, Populated by a Computers and Writing Proposal

Note. *The third column contains the file name, which is developed by aggregating the Document Type, Document Title, Date, and Row Number columns. If a particular column entry is blank, for instance Date, that category is still marked; for instance, rather than the date, this file name reads NoDate.*

Date	Title/Subject	File Name	Document Type	Author	Recipients	Content Tags	Tagged For Interest	Description
	Computers & Writing	Proposal_Computers&Writing_NoDate_0002.pdf	Proposal/ Essay		The english department	Proposal	Yes	A description of a proposal for a hybrid program in the english department using technology. The program started in the Freshman Composition program

From the example file name given (Proposal_Computers&Writing_NoDate_0002.pdf), a user can discern that the document is a proposal, possibly for a course on writing and digital technology, that was created at an unknown date. Finally, the number 0002 indicates that any additional available metadata can be found on row 02 of the metadata spreadsheet. In cases where information required by the file naming convention was unavailable, such as the document title or the year, we added "NoTitle" or "NoDate" to the file name, as can be seen in the above example.

3.2.1 The Importance of Naming Conventions

When creating our file naming convention, we considered it important that it had the potential to be adopted as a model so that users interested in archival work might be encouraged to develop their own sub-collections or conduct their own archival research. However, while we did not have any particular method, theory, or user-interaction end goals in mind when naming files initially, we have since come to understand that file names are just as important for identifying the human elements contained within digital archives as they are for understanding the contents and structure of them.

In general, a file name should make it clear to a user what is contained in a file while simultaneously outlining the unique features of that file. The Princeton University Library states that, as a best practice, file names should:

- be consistent
- be short but descriptive
- avoid special characters or spaces
- use capitals and underscores rather than periods and spaces
- use the YYYYMMDD format for dates
- include version numbers
- be written down with a clear convention in the data management plan.[4]

Our naming convention adheres to each of these suggestions.

Additionally, and specific to our digital archive, file names hint at the metadata contained within the artifacts and, as a result, hint at the structuring tenants of our archive. Many archival scholars—including Jonathan Alexander and Jacqueline Rhodes, Jean Bessette, Barbara Biesecker, Cara Finnegan, and K. J. Rawson, to name a few—understand that, "because [archives] are created in time and space by human beings who make decisions about the selection, preservation, and presentation of materials, and each of these decisions (and more) shapes in important ways the kinds of meanings that can emerge from the sites," archives are rhetorical—in their creation and in their reception (Enoch & VanHaitsma, 2015, p. 217). We find that ensuring that file names transparently reveal to users any underlying assumptions about the structure of the archive is important because doing so provides insight into how an archivist rhetorically understands their metadata, and therefore how they believe users will interact with the archive. No archive is without bias, and such transparency in file names allows users to find the gaps and biases of an artifact, which a user can then complicate or provide greater insight to through a particular lens. As researchers, we understand that meaning making in archives is a social activity, requiring active participation from both the user and the archivist, and while an archivist may provide an environment (physically and digitally) for research and contextual information for any given artifact, a user cannot simply accept what the archivist provides as constituting a whole picture. A user must also contribute their own knowledge to an artifact's understanding—and it may be easier to do so if a user can see these biases more transparently.

In the future, processes of automation might come to exist which could relieve archivists of the burden of analyzing each individual document and entering its characteristics into a spreadsheet. As one reviewer of this manuscript pointed out, there is potential for algorithms to "learn" from processes of human-created data entry (e.g., our metadata fields) to create more efficient input and data organization methods going forward. However, we want to be careful to

[4] https://libguides.princeton.edu/c.php?g=102546&p=930626

emphasize that awareness of our biases when initially inputting this information is of utmost importance should such technologies become automated and tasked with this labor as this project evolves. Put simply, we are currently limited by the technological constraints of our archive's platform given that we only have a certain set of tools available to us. Should other institutions follow our lead and create digital archives of their own programs, they might not be limited to the same constraints. Again, we want to stress that when we think about what our platforms are and are not capable of doing, we closely and actively consider what existing biases our tools might already have based on the decisions their creators made when building them.

3.2.2 The Material Labor of Scanning and Organizing

As individual file names were created for each document, we placed them in two locations: within the metadata spreadsheet in a new column and on the physical documents themselves. To ensure the original physical documents were not altered by this process, file names were hand-written on Post-it notes and attached to the front of each document. It is important that the physical documents remain unaltered because any permanent alteration of the document means that that particular aspect of the data is lost forever. In some cases, losing this data might be necessary, such as when a student's personal information is revealed. However, in most instances, a best practice involves ensuring that our contemporary judgement calls do not influence what may be considered part of the historical record in the future. The physical documents were also organized into stacks ordered from 0001 to 0659. These processes ensured that the document's name could be easily added to our scanned PDF files once we began scanning, and that anyone interested in seeing a specific physical document would be able to do so in the future.

We should note that materials were converted to PDF files only because the scanner we used generates PDFs. When the scanner is scanning a document, it is really just taking a high-quality photograph of the page. This is an important limitation to consider in doing this work: much depends upon the technology that is available. In our English department at Northeastern, our scanner options render either PDFs or JPEGs. To generate other formats (plain text, Word docs, etc.), we have to hand-transcribe the documents ourselves or scan them into an image format that would enable Optical Character Recognition (OCR). In our case, when documents get uploaded to the DRS at Northeastern, they are uploaded as PDF files and then OCR is used at some later point.[5] Whether or not the step to use OCR is taken in the future (which we intend to do), we must ensure that scanned documents first become either a PDF or a picture of the page; generally, OCR will then take that picture, conduct a layout analysis where it identifies where the text is positioned, and determine what the words on the page are. When dealing with messy or handwritten documents, this task becomes more difficult given that the letters are not as easily distinguishable as they are in a clean, printed copy, and the lines are not perfectly vertical or

[5] For more on how the DRS conducts work on text analysis and named entity recognition and the complexities of how computers work with text analysis tools, see https://bostonresearchcenter.org/using-named-entitiy-recognition-on-the-east-boston-community-news/.

horizontal. OCR or transcription is an important task for the kinds of text analysis we eventually want to be able to do; however, nearly every text analysis tool that can do this work depends on machine-readable text.

While there are some useful tools out there that conduct automated data analysis (Voyant, Palladio, Tableau, etc.), these tools all depend on machine-readable data. Automatically generating a machine-readable version of our texts would be possible for all of the typed and printed material, but when it comes to handwritten material, we are aware that much could be lost in processes of automated transcription. Perhaps the best and most efficient way to tackle this unique problem will be to OCR typed material and then proofread, or even hand code, the handwritten bits while also providing images of the original handwritten documents for display to archive users. OCR continues to move and develop at an incredible rate, and some of the technologies out there are incredibly precise, so while this problem of OCRing handwritten documents exists for us today, it may well cease to be an obstacle in the near future. We should note, too, that while Intelligent Character Recognition (ICR) for processing hand-printed documents does exist (it is used, for example, in recognizing hand-printed characters on hospital or DMV forms where each character is written in its own dedicated box), ICR for true handwriting (which, in our case, is still very distinct from hand-printed text) is still not very precise. We will need to use a combination of OCR and ICR in order to achieve the kind of usability we envision for our archive.

This is a prime example of where material meets machine: there are many steps required to build an archive like ours, and while we do have many tools available to aid us in our work, there are very real material constraints in this process that require time and labor.

The scanning process itself required many, many hours of physical labor[6], and was often hampered by the need to remove paper clips and staples, manipulate objects with bindings, and ensure that the scanner did not get jammed, which would have resulted in damaged artifacts. To ensure that the file name could be found on the physical document, but was not scanned onto the document itself, each Post-it note containing the pre-planned filename had to be moved to a blank white sheet of paper and placed at the front of a scan to provide a reference when later re-naming the scanned files. For the majority of our work, we used a sheetfed scanner, which did add a layer of convenience, but which was especially finicky if too many documents were queued at once. In rare cases, some documents were too large to fit into the sheetfed scanner and had to be done on a flatbed scanner. Additionally, a number of the tagged documents represented entire packets or folders of similarly combined items (some exceeding 100 pages), in which cases the documents were scanned in groups and split into multiple PDF files, as the scanner could only handle so many documents at once. In cases where documents were scanned in groups and split into multiple PDFs, they were later re-combined digitally. Finally, to ensure consistency, all scanned documents were correctly oriented (landscape or portrait, depending on the document), and all blank pages within the PDF files were removed to improve the digital user experience.

[6] We would like to thank Cameron Barone and Alex Shad for their contributions of time and labor toward this effort.

The files are currently stored in a dedicated Google Drive that is not attached to any particular team member; this decision was made with future team member access in mind, given that we see this project as an ongoing effort.

3.2.3 Considerations for Optimization

Once everything was scanned and re-named with the previously tagged file name, two steps remained: optimization of the document type metadata criteria and one final pass through the documents in search of sensitive or personal information. The first of these efforts was done to improve the usability of the contents of the archive. In order to optimize the document type criteria, we created a "Code Book" in which we retroactively defined each of the document types in an effort to help users better understand how one document type was distinct from another. We also hoped this would provide insight into our rationale when placing any given document into a particular document type category (see Appendix). In doing so, we found that some of our document types had overlapping definitions where distinctions were insignificant or essentially non-existent. In these cases, we agreed upon which of the two document type definitions was more inclusive and consolidated the two types into one.

Additionally, in one or two cases, some of the document type categories contained only one or two artifacts, which we made note of and re-examined to discern whether this was due to the documents being truly distinct in a significant way—in which cases we kept that document type, despite the low number of representative files—or whether those distinctions were based upon an insignificant feature (for example, we had a "Chart" document type but found that its distinction from "Outline" and "Statistics" was difficult to determine). Once this process was completed, we were sure to change any physical and digital file names as necessary.

Finally, we filtered each of the 659 documents in our archive once more: both to ensure the privacy and safety of the individuals represented in the archive, and also to ensure the legality and safety of the archive itself. Given the digital nature of the archive, and the accessibility that digital mediums provide, it is of great importance to adamantly protect those represented within it. An archive is not simply a collection of documents filled with texts and images, it is a collection of people's thoughts, experiences, conclusions—and, subsequently, information. For example, some of our documents dated in the 1970s and 1980s contained Social Security numbers, which Northeastern once used similarly to their unique student ID numbers today. Other examples include home addresses, phone numbers, and even salary information.

3.2.4 Legalities and Permissions

Since our archive is specific to Northeastern University, many of the people represented in it are present or former employees and students of Northeastern University. Because of this, special care must be given to documents that might be considered the intellectual property of the author, in the case of assignment sheets, or that are protected by laws and regulations such as the Family Educational Rights and Privacy Act (FERPA), in the case of student writing. As we discuss below, institutions also, in many cases, consider syllabi the intellectual property of the

university, but granting access to student writing is complex given various state and national laws. Knowing what to be aware of when going through the review process can be challenging, as data protection laws in the United States are supported and enforced by a variety of regulators (Markus, 2016, p. 152), and even making proper inquiries to institutions and lawyers can leave you with unsatisfactory answers, as we experienced first-hand.

As a result, we found it pertinent to review the items in our archive many, many times. We have also set aside a number of documents that we are still uncertain about, as we believe it better to choose to not include those items in order to ensure the safety of individuals and our archive until further notice. In the event that any personal information is made publicly accessible through our archive, it's possible that legal action could be taken—and the last thing we want to see is two years' worth of work get shut down due to carelessness. It is also important to note that while, legally, permission is not required to post institutional documents (e.g., syllabi, memoranda, emails), getting authors' permission—if possible—is an act of good faith and one that we feel is important for those documents that might put the author or recipient in an unflattering light.

Despite our intent to follow legal requirements, we still needed official sign off from Northeastern's Office of General Counsel (OGC). In our initial meeting with that office, its representative was not quite sure what precedent was for including student materials in a public archive. For example, could we include placement essays with students' names redacted? Could we have an inward-facing version of some aspects of the NUWP archive, accessible only by those with Northeastern credentials (which would be contrary to our intent to create a public digital archive, but could still be very valuable for classroom/instructional use)? We still await follow-up on this and other permissions issues.

4.0 Programs of Research

4.1 Understanding the Archive Within the Writing Analytics Taxonomy

Lang, Aull, and Marcellino (2019) created a writing analytics (WA) taxonomy "inspired by efforts to map out practice areas in learning analytics" (p. 15), which deals with "applying WA (e.g., gathering data, assessment) and accounting for context (privacy and data safety, implementation paradigms)" (pp.14–15) and is useful in helping us to understand how our project fits within the larger ecosystem of this field.

In the taxonomy, four potential programs of research are outlined: 1) educational measurements, 2) massive data analysis, 3) digital learning ecologies, and 4) ethical philosophy. We see a clear connection to these research programs and feel there is potential to tap them even further if our digital archive can eventually become one node within a much larger network of digital archives across multiple institutions.

As we proceed with our archive in consideration of these four programs of research, a discussion of values and tools becomes vital: we cannot understand how various constructs (represented in an institution's values and as reflected through things like mission and outcomes

statements and learning goals) operate in context if we do not consider context on the local level. An institutional archive allows us to more closely examine our local context, but we are also aware that tools and technologies are not neutral objects—they reflect their designers' values and interests. As such, tools and technologies are always situated in historical contexts: that is, sites.

4.1.1 Using Analytics to Discern Values Across Writing Sites

Traditionally, the field of writing studies clusters research in three major sites: writing programs, writing classrooms, and writing centers—all of which are historically embattled spaces that have seen their borders (and therefore, to some extent, their possibilities) redrawn by shifting institutional power dynamics, globalization, and burgeoning desire for efficiency, replicability, and production. All of these sites are changing, largely due to new tools. For example, no longer is educational measurement and assessment chiefly about scoring of student essays for placement testing or the use of anti-plagiarism software. Today, analytics can predict what makes writing "good" and extend that prediction to shape longitudinal definitions of "success" through supporting data like grades and retention. Digital tools can trace every move of a student's writing process with the click of a button and offer language replacement suggestions that potentially erase or reinforce predominant values of identity and culture (Leki, 2007; Young et al., 2014). The questions of the consequences of educational measurement (gesturing toward the taxonomy's first program of research) no longer exist on a linear timeline; analytics reveals the ways in which writing is iterative and recursive in its processes and practices. The development of new digital tools to assess these processes and practices is also increasingly automated and instantaneous in generating data. This can leave little room for negotiation of values (encapsulating research programs two, three, and four), which is troublesome given the historical lack of agreement in values observed across writing sites. We see our archive addressing all four programs of research in that by examining the materials contained within through these lenses, we might better come to discern where values across sites at a specific institution *do* overlap even while they infrequently align.

Complicating this outcome is an understanding that some institutions do not have writing programs that allow them to design or assess writing curriculum; therefore, the question of "what belongs in a writing program archive" becomes extremely valuable. For example, Haswell (2001) noted in *Beyond Outcomes* that Washington State University only recently (at the time) established a program that contained many of the elements we would like to identify as belonging to a writing program: course sequences, a writing across the curriculum (WAC) program, an assessment office, or a writing center. Nineteen years later, there are still plenty of colleges and universities that lack these elements. In fact, what writing programs "are" and what they consist of varies from institution to institution, if they exist within a university's ecosystem at all. The process of creating and curating a writing program archive can also help scholars in our field more deeply understand the precariousness of what a writing program "is" and (re)consider what it might be.

The programs of research outlined in the writing analytics taxonomy, when applied to an institutional archive, can help us stand before the technological advances of 21st-century tools not from a place of fear or ignorance, but from a place that recognizes the very real consequences these tools might have on the future of writing across institutional writing sites. Analytics can show us what is going on in our institutional contexts and highlight to us precisely the ways in which our situated practices are being separated from their contexts, especially given that "data production technologies . . . are able to detach situated practices from their context and make them into calculable goods for use in policy and economic decision-making" (O'Keeffe, 2017, p. 126). Sharer et al. (2016) encourage writing studies scholars to consider the ways in which assessment and accreditation might support scholarship and the building of writing programs, arguing that companies are inspired to create and use tools that harness the power of big data to enable better comparability; the key here is that WPAs need to be more involved in not only deciding the goals of assessment but also in influencing how these companies collect and use that data.

An institutional archive is one way we imagine we might be able to make situational contexts not only more visible but more *usable* through applying writing analytics methods within the taxonomy's programs of research. Writing analytics, much like writing studies, is working to establish disciplinarity; institutional archives provide an opportunity to better understand situated local contexts for purposes of assessment, curriculum design and pedagogical practice, and program administration, which can aid quests for disciplinarity in both overlapping fields simultaneously.

4.2 Writing Analytics Meets the Archive

If our archive is to be useful to the field of writing analytics, we must consider the work that is already being done as a way to consider where we might go. Given that writing programs, where they exist, serve the entirety of an institution across colleges of study, a discussion of program assessment alone quickly gets complex. Dryer (2013), for example, describes the complexity of writing program assessment through analysis of 83 scoring rubrics and grade definitions at public U.S. research universities, and problematizes the assumptions and methods used to measure student writing proficiency while highlighting the difficulty of doing so. Dryer uses corpus analysis to show that traits used to score writing are unstable and applied speculatively, while critiquing "missed opportunities to emphasize the *situatedness* of the students' writing" (p. 27). Dryer asserts that it is important to be aware of this situatedness to the extent that we recognize how any assessor's appraisal of a program is embedded in cultural and material contexts; however, program assessments often fail to comprehensively take these into account. For instance, assessment studies that focus largely on demographic data collected from students run the risk of obfuscating other situational factors that impact student learning and place undue burden on undertrained and underpaid contingent faculty that teach writing courses. Add this to the observation that proactive planning for accessible and usable curricular texts has been found to be inconsistent at best (Scott et al., 2017), and program assessment becomes fraught with

challenges that traditional tools fail to comprehensively address. We imagine that by making our program's artifacts publicly accessible to researchers, we might offer a new means of observing situational factors that otherwise might have remained invisible.

Moreover, little work has been done to map use of natural language processing (NLP)-based automated writing evaluation tools onto situated outcomes-based assessments that can vary widely by institution. Burstein et al. (2017) describe the "real-time, dynamic nature of NLP-based AWE" as affording researchers the ability to "explore linguistic features and skill relationships across a range of writing genres in postsecondary education" that can "provide educational analytics that could be informative for various stakeholders" (p. 101). This lack of situated application arguably results in rejection of use of these tools by many WPAs; however, these tools *can* help writing programs understand what they are actually doing and how they can do it better—*if* they are able to clearly articulate outcomes and use these tools to assess whether or not their programs are enacting them, perhaps by mapping features of texts observed in archival corpora across time or across specific courses. This is where writing analytics has true potential when applied to a writing program archive.

Collecting and reviewing the materials of a writing program archive might also bring more squarely into view a disconcerting chicken-or-egg problem: If we use writing assessments to place students into writing classes, and those assessments do not account for what is actually being delivered in curriculum or taught in classes, do we need to look at curriculum more closely to discern what *should* be assessed, or do we need to restructure our assessments in order to determine what we *really value* in designing curriculum? Lerner (2019) argues that poorly defined curriculum in composition results in the conflation of curriculum with pedagogy. The notion that curriculum is often not clearly defined on a local program level prompts this question: If we don't know *what* we are actually teaching, how can we understand our effectiveness in teaching it? Applying analytics tools to our archive also offers ways of better understanding not only what students write, but how we teach them to write it. Most of the research done on classroom assessment to this point focuses on students' grades and how instructors determine those grades as reflections of competence; little work has been done to assess the instructor's role in clearly framing the tasks that students are being asked to complete in, for example, assignment sheets, and what has been done has been limited to study of rhetorical situations, genres, and discourse communities of college writing without considering additional factors in assignment design beyond disciplinary content (Melzer, 2014).

A collection of program materials in a digital archive offers an opportunity to apply writing analytics programs of research to tackle these complex questions in more comprehensive, data-driven ways, particularly if we are able to take artifacts from our archives and use text analysis tools to identify patterns in syllabi, assignment sheets, and program assessments not only to more closely examine trends in our own institution, but perhaps to map those onto trends we might find in artifacts at other institutions where writing programs exist. Our archive might also offer a way for us to observe how the use of technologies in classrooms has shifted over time, for instance, and how that has impacted writing program curriculum. If our archive reflects the kinds

of materials that many WPAs and department staff members collect over the course of any given academic year, then building a searchable database of these materials provides one way of actually *using* those materials in consequential ways, rather than simply storing them in separate folders in physical and digital spaces that exist to be forgotten.

Finally, adding writing center texts to our program archive contributes to an even richer space for research potential in analytics. Northeastern's writing program also houses its writing center, and while this is not true for every institution that might seek to create an archive, adding writing center materials such as tutor training manuals, workshop materials, and records of utilization data might help writing programs demonstrate the function of a writing center within disciplinary and institutional contexts. Johanek (2000) called for writing centers and writing studies more broadly to use qualitative methods alongside quantitative methods to help us make better sense of the work writing centers do in the larger landscape of writing in higher education settings. Examining survey and utilization data, however, is not enough to understand the writing center's role in context, and we believe that if situated in an institutional archive like ours, analytics can do some of the work of tracing patterns that more clearly reveal the value of writing centers in relation to writing programs and the larger institutions where they live.

One promising starting point in the research potential we see for our archive is the use of corpus analysis tools that can process a higher volume of texts at one time than a human can manually; we can use analytics to show us patterns in our institutional texts across sites that might not otherwise be visible—and we can also use them to trace program development over time in terms of these patterns. Because, as mentioned previously, digitization and OCR are separate labor processes, we don't yet have searchable PDF files, but this is a goal for our developing archive, and one that we hope will make research programs in writing analytics more possible.

5.0 Conclusions

5.1 Material Realities

It is important to note that the process of developing a digital archive is not linear—it happens in stops and starts, and decisions that might have been made at the project's genesis are often revisited as the realities of digitization demand an uncovering of a variety of technical apparatuses. Among the many tasks associated with creating a digital archive are questions of legality, collaboration with existing institutional resources and archives, the politics of the internet, and many others. However, at its most basic level, digitization is the transformation of physical material to pixels and binary code. Digitization, though seemingly separating the physical from the digital, is a very physical type of labor. The reality is that before the archive as we know it could exist in a single folder in an online file sharing platform, it existed in seven boxes, only some of which would contain relevant material.

The process of creating a digital archive, though we often associate process and processors with computation, involves exposing oneself to the material realities of documentation. Though

the form of a PDF file flattens the material, in some ways, in the process of making this shift from the physical to the digital, our hands were dotted with papercuts and twenty-year-old dust.[7] Some of that dust might have even been transferred to our keyboards as we meticulously entered in metadata for each object. In this way, the physical material of the archive is transformed, true, but the digital object that emerges from this transformation is necessarily infused with the material.

"Digital" is not the antithesis of "material." While our seven boxes may now live in a single file, the archive is still very much physical material. Each PDF file represents an object that still continues to exist, and those objects must also be given a place to live, whether that be in the institution's larger archive or even in a department storage closet. Even the bytes that have been clustered and drawn together to form a single file on a computer exist physically in the form of hard drives and servers. Behind each of these material realities are people who are required to maintain and watch over them. In almost every way, the digital archive cannot exist without the material. Digitization, while concerned with questions of preservation, is also a process that concerns itself with revealing material in new ways—providing access to history and putting documents together that may not exist in the same place without the digital archive to facilitate this connection.

The digital archive is about expanding the possibilities and connections, bringing the archive into one's home rather than hidden away. In a somewhat alchemic exchange for this access, as researchers, our backs ached as we lugged boxes throughout our department, moving them back and forth between scanner and storage closet like Sisyphus pushing his boulder; our hands were dirtied with rust and dirt as we carefully removed every staple from every piece of paper in those seven boxes, and our eyes strained to take in the tiny font of a spreadsheet. This experience only remarks upon our particular exposure to the labor of digitization, yet there are many other people who in many ways contributed labor to this project: the members of the library who maintain the institution's digital repository service, the digital scholars who maintain websites for projects affiliated with our institution, and perhaps, most importantly, the people in our department who created (and saved) these documents to begin with. Each of these forms of labor work together to form what appears only on the surface as a flattened PDF file.

5.2 Complexities of Narrativizing Our History

One issue that we grappled with throughout the process that we have described is how much we wanted a final site that was curated for readers, how much we wanted simply to provide access to raw data, and how much we wanted a combination of those two. Any curation, of course, is a narrativizing that would involve deliberate steps on our part: decisions to include some materials and exclude others, summaries and analyses that might reflect our biases and "good intentions" but that are as subjective as any historiographical account. On the other hand, if we leave our data "raw," how might we prevent users from narrativizing with our data in ways that we feel misrepresent the materials themselves and the rich histories they contain? Certainly, including

[7] We tip our hats to Carolyn Steedman's *Dust* (2001).

disclaimers cleared by our Office of Legal Counsel is one practical approach, but perhaps another is simply to let go, akin to any archive—physical or digital—and the uses an archival researcher might make of its materials. Framing this project as a potential source of data for researchers in writing analytics recognizes this impetus to "let go" while acknowledging our fears of what might be done with our data. Our initial intention to create a *public* digital archive is one we strongly adhere to, and with that public intent comes a certain degree of trusting the public to make ethical and fair use of the materials that we provide—just as we would hope that those who encounter our archive would trust us to make ethical and fair choices of what we included and how we provided access.

5.3 A Note on Resources and Partnerships

Much like a physical space, a digital space is not maintained by itself. Key to the development and maintenance of digital spaces is knowing what resources are available. Familiarizing with these resources at an institution (e.g., partnering with the library/DRS), and forging and cultivating relationships with these partners, is important for sustainability purposes so that future generations will be there to maintain it. But what happens when an institution does not have the kinds of resources that Northeastern has to do this kind of work?

In a keynote address at the 2020 Writing Analytics Conference, Jessica Nastal discussed the importance of the role of writing analytics in two-year colleges while noting the many challenges inherent in conducting the kind of research that would be necessary to broaden the field to include institutions of all types:

> . . . many two-year college faculty lack access to disciplinary research outside the open-access journals, like *Journal of Writing Analytics*, and books and resources, including those on the WAC Clearinghouse. College libraries, however, have limited academic databases, which makes it difficult to find and access current and foundational scholarship. . . . Institutional Review Boards do not exist at all two-year colleges. Each IRB has a different interpretation of the federal standards. And, even when IRBs are approved, it may be difficult to access the data. Partnerships with Institutional Research are critical for our work to succeed; however, at two-year colleges and even for our state-wide bodies, the offices are often understaffed and may be unprepared to assist faculty pursuing research and institutions that require access to more sophisticated metrics than they have previously needed . . .

Nastal's keynote address highlights the need to more critically consider the ways in which institutions might partner and offer resources to take up the kind of work that is necessary to create and maintain a truly inclusive network of institutional archives that might support future research. In order to make truly data-informed decisions, our developing field(s) are called to take stock of what resources are available and what can be done while also realizing that an archive is a living space—one which is never "finished," and one that is necessarily collaborative and inter-generational. The project we have described here was a consequence of the labor of

two MA students, one PhD student, one tenured professor, and two undergraduates with federal work study funds. It was possible because of the availability of internal funding (to pay for that labor) and institutional resources. We recognize the privilege in our particular circumstance, but we also challenge the field of writing studies/writing analytics to pursue partnerships, to share resources, and to expand digital work beyond privileged institutions.

How might we expand our digital spaces to include the important work done at two-year institutions and historically Black colleges or universities (HBCUs) in order to amplify systemically silenced voices? The creation of our digital archive is a potential first step in thinking about the labor processes and practices required for critical institutional inquiry, and we hope that our work offers a bridge to a broader discussion about how we might, as Nastal urged, "transform what we know about writing and writing instruction." While we enter this phase of our project with, in many ways, more questions than we have answers, we hope to have outlined our approach in ways that make visible the kind of labor and considerations required for taking up the call to more meaningfully approach archival texts as data, and in doing so, urge other institutions to consider the possibilities for partnership that exist within digital spaces.

5.4 The Future of Our Archive

The interdisciplinary field of writing analytics offers potential ways of pinpointing and reconsidering what it is we're asking students to do, how we're asking students to do it, and how we can provide feedback that translates (in tangible and calculable data) into skills-based improvement in situated writing tasks. An institution is a multidisciplinary ecosystem that requires critical participation from all of its organisms in order to not only survive, but to thrive. Archives, on a very basic level, allow us to trace the origins of our ecosystems and evolve. The question for us after *What do we have?* becomes *What do we want?* More specifically: *How do we want to evolve?*

What exists in our archive already, and what doesn't exist that we might like to be able to make visible in order to address the above questions? What would we want to try to add in the future that isn't there now? The future of our archive must consider not only the ways in which we curate and manage data, but also the ways in which technologies, as the vehicles of the archive, also require the stewardship of human hands. Therefore, it is important to consider the ways in which designing the archive's digital space might enable us to be better stewards of the information it houses. We see the future of our archive evolving to include oral histories, timelines, and lists of institutional actors for which we already have metadata—but considering the ways in which this data is presented to the public presents a plethora of challenges we have yet to address. The next phase of our project will require deep consideration of not only the ways in which we choose to narrativize our history, but also the ways in which we present that history for digital user experience.

Since so many of the documents that are featured in this archive are documents which have been annotated and revised by other readers, much of the thought process surrounding how we might think about features the website might incorporate, such as data visualization, filtering,

and searching, has involved figuring out how to deal with the messiness of these documents first. The next phase of our project will require extensive consideration of user experience in how we choose to actually present this material now that it has been collected, digitized, tagged, and sorted. Going forward, we face new challenges in terms of our presentation of our data; we must think about how we want to display our materials (Do we want to make downloadable corpora available? If so, what documents should and can we include? Do we want to create exhibits containing research projects of our own which use our data and demonstrate the kind of work that is possible? How can we encourage new and veteran scholars to learn from what our archive has to teach?). We must also decide what levels of access we can provide to users within and beyond our institution (both ethically and legally), and we must also determine the degree to which we want to narrativize our own history to provide context for the users that will engage with our materials for practical and scholarly purposes. We plan to make many of these decisions during the next phase of our project, and we are confident that the challenges and questions this next phase will raise will be crucial to the longevity and vitality of our archive.

The picture we have of our writing program right now is messy, as was the process to this point of capturing it. But we believe this messiness matters for our institution, for writing studies as a field, and for the future of writing analytics scholarship.

Author Biographies

Kyle Oddis is the Project Manager of the Writing Program Digital Archive and a PhD student in English writing studies at Northeastern University. Her programs of research are writing assessment, writing curriculum and pedagogy, and community literacy at intersections of data ethics, analytics, and institutional/organizational cultures.

Avery Blankenship is a PhD student in the English Department at Northeastern University. Her areas of research are digital humanities, metadata design and implementation, digital archives, nineteenth-century American cookbooks and domestic manuals, and the politics of consumption.

Brice Lanham holds a Master's Degree in English from Northeastern University. His research interests include rhetoric of space and publics, community literacy, writing center studies, digital archives, and Black American literature.

Neal Lerner is Professor and Chair of the English Department at Northeastern University. His scholarship has focused on writing centers, writing across the curriculum, writing pedagogy and curriculum, citation analysis, and qualitative and quantitative research methods in writing studies.

References

Bantin, P. C. (2016). *Building trustworthy digital repositories: Theory and implementation*. Rowman & Littlefield.

Berlin, J. (1987). *Rhetoric and reality: Writing instruction in American colleges, 1900–1985*. Southern Illinois University Press.

Biesecker, B. (2016). Of historicity, rhetoric: The archive as scene of invention. In L. L. Gaillet, H. D. Eidson, & D. Gammill, Jr. (Eds.), *Landmark essays on archival research* (pp. 156–162). Routledge.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. The MIT Press.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2017). Exploring relationships between writing and broader outcomes with automated writing evaluation. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), EMNLP 2017, Copenhagen, Denmark.

Connors, R. J. (1997). *Composition-Rhetoric: Backgrounds, theory, and pedagogy*. University of Pittsburgh Press.

Cook, T. (2001). Archival science and postmodernism: New formulations for old concepts. *Archival Science, 1*, 3–24.

Donahue, P., & Moon, G. F. (Eds.). (2007). *Local histories: Reading the archives of composition*. University of Pittsburgh Press.

Dryer, D. B. (2013). Scaling writing ability: A corpus-driven inquiry. *Written Communication, 30*(1), 3–35.

Cushman, E. (2013). Wampum, Sequoyan, and story: Decolonizing the digital archive. *College English, 76*(2), 115–135.

Enoch, J., & VanHaitsma, P. (2015). Archival literacy: Reading the rhetoric of digital archives in the undergraduate classroom. *College Composition and Communication, 67*(2), 216-242.

Fleming, D. (2011). *From form to meaning: Freshman composition and the long sixties, 1957–1974*. University of Pittsburgh Press.

Gannett, C., Brereton, J. C., & Tirabassi, K. E. (2010). "We all got history": Process and product in the history of composition. *Pedagogy, 10*(2), 425–450.

Gentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as data* (NBER Working Paper No. 23276). National Bureau of Economic Research. http://www.nber.org/papers/w23276

Gold, D. (2008). *Rhetoric at the margins: Revising the history of writing instruction in American colleges, 1873–1947*. Southern Illinois University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*, 267–297.

Haswell, R. (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Hayden, W. (2017). And gladly teach: The archival turn's pedagogical turn. *College English, 80*(2), 133–158.

Johanek, C. (2000). *Composing research: A contextualist paradigm for rhetoric and composition*. Utah State University Press.

Kitzhaber, A. R. (1990). *Rhetoric in American colleges, 1850-1900*. Southern Methodist University Press.

Lamos, S. (2011). *Interests and opportunities: Race, racism, and university writing instruction in the post-civil rights era*. University of Pittsburgh Press.

Lang, S., Aull, L., & Marcellino, W. (2019). A taxonomy for writing analytics. *The Journal of Writing Analytics*, *3*, 13–36.

Leki, I. (2007). *Undergraduates in a second language: Challenges and complexities of academic literacy development*. Lawrence Erlbaum.

L'Eplattenier, B. E. (2009). An argument for archival research methods: Thinking beyond methodology. *College English*, *72*(1), 67–79.

L'Eplattenier, B. E., & Mastrangelo, L. (Eds.). (2004). *Historical studies of writing program administration: Individuals, communities, and the formation of a discipline*. Parlor Press.

Lerner, N. (2001). Searching for Robert Moore. *The Writing Center Journal*, *22*(1), 9–32.

Lerner, N. (2019). *Reformers, teachers, writers: Curricular and pedagogical inquiries*. Utah State University Press.

Markus, L. M. (2016). Obstacles on the road to corporate data responsibility. In C. R. Sugimoto, H. R. Ekbia, & M. Mattioli (Eds.), *Big data is not a monolith* (pp. 143–161). The MIT Press.

Nastal, J. (2020, February 6–8). Writing Analytics and Pedagogy in the Two-Year College [Keynote Address]. The 9th International Conference on Writing Analytics, St. Petersburg, FL, United States.

Melzer, D. (2014). *Assignment across the curriculum: A national study of college writing*. Utah State University Press.

National Information Standards Organization (2016). *Metadata*. http://framework.niso.org/24.html

O'Keeffe, C. (2017). Economizing education: Assessment algorithms and calculative agencies. *E-Learning and Digital Media*, *14*(3), 123–137.

Ostergaard, L., & Wood, H. R. (2015). *In the archives of composition: Writing and rhetoric in high schools and normal schools*. University of Pittsburgh Press.

Palmquist, M. (2019). Directions in writing analytics: Some suggestions. *The Journal of Writing Analytics*, *3*, 1–12.

Ritter, K. (2009). *Before Shaughnessy: Basic writing at Yale and Harvard, 1920–1960*. SIU Press/Studies in Writing and Rhetoric.

Ritter, K. (2012a). Archival research in composition studies: Re-Imagining the historian's role. *Rhetoric Review*, *31*(4), 461–478.

Ritter, K. (2012b). *To know her own history: Writing at the woman's college, 1943–1963*. University of Pittsburgh Press.

Scott, L. A., Thoma, C., Puglia, L., Temple, P., & D'Aguilar, A. (2017). Implementing a UDL framework: A study of current personnel preparation practices. *Intellectual and Developmental Disabilities*, *55*(1), 25–36.

Sharer, W., Morse, T. A., Eble, M., & Banks, W. (2016). *Reclaiming accountability: Improving writing programs through accreditation and large-scale assessments*. Utah State University Press.

Smitherman, C. (2007). *Oral histories of the National Writing Centers Association: A look at group dynamics* [Unpublished doctoral dissertation]. University of Louisville.

Steedman, C. (2001). *Dust: The Archive and Cultural History*. Manchester University Press.

Sternfeld, J. (2011). Archival theory and digital historiography: Selection, search, and metadata as archival processes for assessing historical contextualization. *The American Archivist*, *74*, 544–575.

Varnum, R. (1996). *Fencing with words: A history of writing instruction at Amherst College during the era of Theodore Baird, 1938–1966*. National Council of Teachers of English.

White, K., & Gilliland, A. (2010). Promoting reflexivity and inclusivity in archival education, research, and practice. *The Library Quarterly: Information, Community, Policy*, *80*(3), 231–248.

Young, V. A., Barrett, R., Young-Rivera, Y., & Lovejoy, K. B. (2014). *Other people's English: Code-Meshing, code-switching, and African American literacy*. Teachers College Press.

Zhang, J. (2012). Archival context, digital content, and the ethics of digital archival representation. *Knowledge Organization*, *39*(5), 332–339.

Appendix

"Code Book" Defining Document Types

- **Agenda:** a list of to-do items, likely in a meeting or short period of time
- **Assessment:** placement test material/assessment of student
- **Checklist:** a checklist-format document of things to do
- **Committee Description:** a description of what a committee does/will do
- **Course Description:** a description of what a course entails/will entail
- **Departmental Documents**: documents pertaining to department administration
- **Email:** electronic communication between one or more individuals, usually marked by a user's email address
- **Evaluation:** the determination of a course's, instructor's, or the department's success
- **Exam:** a course exam/test
- **Fax:** a faxed document, clearly indicated on the document itself
- **Floorplan:** a top-down view of a building
- **Flyer:** an advertisement of information or an event on campus/in the department
- **Form:** a document that provides blanks to be filled (in some cases have been filled)

- **Guidebook:** a book-length description of how to produce/do something
- **Guidelines:** a description of how to produce/do something
- **Handbook:** a collection of rules or requirements
- **Handout:** a sheet with pertinent information for a specific event
- **Image:** a document that is entirely an image
- **Letter:** a correspondence between one or more people that was not sent as a memo or email
- **List:** items organized into a column
- **Job Listing:** a listing for a job
- **Job Description:** a description for a job
- **Memo:** an informal correspondence that provides information about events; more often than not, says "memo" at the top.
- **Minutes:** notes/information about a department meeting
- **Newsletter:** an informal publication describing updates or news in/on the department/campus
- **Notes:** supplemental handwritten annotations, often regarding another document or event
- **Outline:** an example of the end product
- **Overview:** a summary review or examination of courses, requirements, and the writing program itself
- **Policy:** a department- or university-wide rule
- **PowerPoint:** a collection of slides that are often used to present information
- **Prompt:** directions for an assignment
- **Proposal:** a document that proposes a plan, course, or intention
- **Publication:** a piece of writing that was published (by someone from the department . . . that we know of)
- **Report:** a write-up of findings following an examination of something
- **Resolution:** a determined plan of action
- **Schedule:** a plan applicable to one or more days, often over a greater range
- **Statement of Role:** a statement about the role of a department's project
- **Statistics:** any numerical data about students/employees/the university, etc.

- **Student Writing:** writing completed by a student
- **Syllabus:** a course description and schedule for any given course
- **Webpage:** a printed version of a then-internet-accessible page