ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Words, Words, Words: How the Digital Humanities Are Integrating Diverse Research Fields to Study People

Chad Gaffield

Department of History, University of Ottawa, Ottawa K1N 6N5, Ontario, Canada; email: gaffield@uOttawa.ca

Annu. Rev. Stat. Appl. 2018. 5:119–39

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031017-100547

Keywords

text, interdisciplinary, history, statistics, society, culture

Abstract

The rapidly developing field of digital humanities (DH) is showing how unprecedented volumes of data such as written expression can be studied to reveal new insights into humans and, therefore, into individual and collective experiences within and across societies. Scholars from disciplines such as literature and history are collaborating with scientists from disciplines such as statistics and computer science. Moreover, these interdisciplinary teams often reach beyond campuses to companies as well as local, national, and international public and nonprofit institutions. Surprisingly, the computational research that began in the humanities in the 1950s did not develop an important presence within mainstream scholarship until half a century later. The DH experiences thus far reflect the complexity of both human expression and research collaborations across diverse fields and sectors. Learning from past successes and failures will help meet today's data analytic challenges and prepare us for opportunities in statistical applications ranging from literary studies and cybersecurity to business intelligence and health indicators.

INTRODUCTION

Researchers have long recognized that what people write and read provide key clues to human thought and behavior. Until recently, the dominant research approach in the text-based disciplines, now commonly known as the humanities, involved close reading of well-known official documents and the writings of leading figures. Characteristically, such reading falls within qualitative research and has contributed to the common association of humanities research with words, in contrast to the association of scientific studies with numbers. Beginning in the 1950s, however, changing perspectives on humans have been combining with new digital technologies to transform the ways in which researchers study what humans write and express to others personally and publicly.

Along the way, well-established distinctions between scholarly interpretative studies and scientific empirical research are being disrupted in surprising and unpredictable ways. Increasing numbers of scholars from disciplines such as literature and history have begun collaborating with those in disciplines such as statistics and computer science (Schreibman et al. 2004). Moreover, interdisciplinary teams now often include partners beyond campus, including companies such as IBM, Microsoft, and Google as well as local and national governments. Most notably, the field now known as the digital humanities (DH) that began emerging during the 1950s and then matured in fits and starts in the following decades is finally blossoming as a leading research area today (Hockey 2004, Bonnett & Kee 2009, Gouglas et al. 2013). This long, uneven, and contested emergence reflects not only the complexity of textual analysis but also, and more importantly, the challenges of research collaborations across diverse fields and sectors. Accordingly, DH projects have been impressive and disappointing, exhilarating and frustrating.

The following discussion distills key themes from past successes and failures to explain why computational research in the humanities took a half century to develop an important presence within mainstream scholarship and thereby to help accelerate the digitally enabled study of the human experience. Particular attention is paid to the ways in which diverse research perspectives and approaches have been colliding, converging, and evolving to resolve data analytic challenges in research on the human experience. The discussion draws especially on examples from historical research to illustrate key aspects of the multifaceted and rapidly changing research landscape of the social sciences as well as humanities (Siemens & Moorman 2006). While firmly based in the humanities, the discipline of history is also at home in the social sciences and thus offers a strategic window on research debates and controversies as well as innovative ways in which researchers have been examining text (Thomas 2004).

It should be emphasized that, in focusing on written expression, many features of DH research are left for future discussion, including well-established and continuing development in the study of images, ranging from prehistoric to those on YouTube, and in the study of sounds, such as music and speech (Klein & Gold 2016). The key premise of DH is a conviction that unprecedented volumes of all kinds of human expression can be captured in digital form and systematically studied along with other evidence to reveal profound insights into the human experience. In addition to the long-standing importance of writing in the making of the modern world, text merits our specific attention for many reasons including the surprising ways in which digital technologies are helping redefine who can be an author (and what that means) and whose written expressions deserve study, by whom and for what purpose (Jockers 2013).

In addressing the continually changing answers to such questions, I am drawing upon my experience as a historian whose research since the 1970s has explored digitally enabled approaches to historical documents; this experience is evident in the examples I have chosen to highlight. In keeping with this approach, the article's presupposition is that our ability to advance today depends upon knowledge of past successes, failures, and missed opportunities (Nyhan & Welsh

2013). While the past does not offer directly applicable lessons, such knowledge provides a firm foundation for understanding and thinking through next steps (Ayers 1999). In addition, the following discussion reflects my years beyond campus, most notably during 2006–2014 when I served as President and CEO of the Social Sciences and Humanities Research Council of Canada, the federal agency that funds most campus-based research in these fields including student fellowships, faculty research projects, and institutional initiatives. In this sense, my perspective should be understood as that of a participant observer who, despite many reasons for pessimism, remains optimistic about how scholars, scientists, and diverse research partners can integrate their expertise to advance the study of the human experience (Gaffield 2016b).

DEFINING TEXT WITHIN DIVERSE RESEARCH INITIATIVES

Text has many forms. Centuries-old as well as recent research has combined to identify a variety of research questions that can be posed about the form and content of written expression (DeRose et al. 1990). In one schema (Ide & Véronis 1995), seven ways of viewing text have been identified: (*a*) physical objects ranging from clay tablets and stones to papyrus and paper; (*b*) typographical objects such as fonts; (*c*) linguistic objects such as words; (*d*) formal objects such as paragraphs and chapters; (*e*) rhetorical objects such as tropes; (*f*) propositional objects such as names or events; and (*g*) historical or cultural objects of commentary that are used in various ways. In each of these distinct, overlapping, and complementary categories, research fields have been expanding and evolving rapidly in light of new thinking and new technologies.

In the case of historical research but more generally as well, profound changes converged by the 1950s to ignite research interest in the digitally enabled study of human written expression. First, researchers began greatly expanding their assumptions about what types of text were worthy of research attention. While the writings of those in positions of official and unofficial authority had long been a focus of research activity, new efforts to take seriously the expressions of less well-known individuals began multiplying as researchers increasingly suspected that societal patterns were determined not only from the top-down, as was remarked by the 1950s, but also from the bottom-up, as a result of distinct perspectives and preferences across society.

To study such economic, social, political, and cultural change, researchers turned to what Ian Winchester (1970) termed routinely-generated sources that selectively documented the lives of most if not all residents in specific jurisdictions. Such sources went beyond censuses to include church records of births, marriages, and deaths as well as many other types of administrative documents created by expanding government and institutional structures especially after the early nineteenth century. Thus, researchers placed a new emphasis on studying evidence about as many people as possible rather than only certain writing by some individuals or groups; the new ambition became known as total history. In this way, changed research assumptions greatly expanded the realm of relevant text for systematic study, thereby posing serious methodological challenges as researchers began grappling with increasingly vast quantities of documents created in diverse ways.

The expanded definition of text for study includes all three kinds of research data identified by Sallie Keller and her colleagues (2017): designed data, administrative data, and opportunity data (daily life). The result helps explain the widespread attention to big data; indeed, as recently observed, "Much of the big data now available involves information about people—this is one of the aspects of big data that contributes to its popularity" (Franke et al. 2016, section 3.9). This point deserves emphasis since just a few years ago, many readers of *Science* were surprised to learn that the corpus of books already digitized by Google in 2010 included 500 billion words comprising a sequence of text 1,000 times longer than the human genome (Michel et al. 2011). The US State

Department's Office of the Historian recently estimated that the department produces 2 billion emails a year (Connelly & Immerman 2015).

Researchers are now focusing on two kinds of digital textual evidence: digitized and born-digital sources. The written expressions being digitally transformed include those from the earliest known literate societies as well as those created in the rapidly expanding print cultures that followed the development of the printing press (Milligan 2013). While private and public repositories have preserved over the centuries only a fraction of written expression, the result still provides a seemingly infinite evidentiary basis for studying individuals and societies. The new relevance attributed to all such evidence has rapidly increased academic initiatives including major research projects that have often joined government institutions in digitizing archival, library, and museum holdings. The increasing number of these projects in recent decades has opened up multiple new avenues of inquiry as well as highlighted key epistemological issues at every stage of the research process.

FROM DATA CREATION TO ANALYSIS PREPARATION

While the ambition to study increasingly large volumes of text was quite widespread after the 1950s, the extent and character of digitally enabled projects were limited, uneven, and often contested on campus. In fact, the experiences of the initial research efforts foreshadowed key issues of enduring importance today. One of the first digitally enabled text-based research projects was undertaken in the 1950s by the historian Merle Curti, who became interested in the possibilities for historical research of new approaches in the social sciences. At a time when history was clearly identified as a humanities discipline, Curti's reputation as a highly respected humanist ensured that his efforts would be taken seriously by students as well as by other scholars. In particular, Curti began imagining how computers could facilitate the study of the documents created during census enumerations. His ambition was to examine the frontier experience in the United States by systematically examining economic, social, and demographic patterns at the level of individuals, families, and households. Taking an approach that would later be associated with the so-called new social history, Curti chose Trempealeau County in Wisconsin for detailed study based on the censuses of 1850, 1860, 1870, and 1880 (Curti 1959).

In turning to computer applications, Curti was benefitting from a long history of census work, especially since the mid-nineteenth century, when enumerations and tabulations became a major government activity of countries around the world and therefore of the emerging statistical sciences (Cohen 1999). The character and content of modern enumerations have been remarkably similar over the decades, with each enumeration characteristically including questions about individual and family identity, as well as economic status and activity. In the case of countries like the United States and Canada, the similarity of the census questionnaires reflects the consistent rationale for the decennial enumerations that began with the objective of allotting congressional and parliamentary seats but also included an extensive effort to enhance the government's knowledge of social and economic patterns (Citro 2016).

Rather than simply study the published tables of frequencies and cross-tabulations as done previously, Curti focused on the original manuscript census schedules to create individual-level data that supported new multivariate analyses involving data on individual ages, birthplaces, occupations, household contexts, and other characteristics included in the enumerations. Going far beyond the tabulations pioneered by census officials, Curti and a team of assistants studied the ways in which personal characteristics such as birthplace corresponded to occupation.

Curti's seemingly straightforward ambition highlighted what only recently has been fully embraced by domain specialists as a major theoretical as well as a technical step in the statistical

sciences: creating and preparing data for analysis. Indeed, one of the most important contributions of humanists to appropriate statistical analysis has been an unrelenting critical interest in the interpretive assumptions of each aspect of digitally enabled research beginning with data creation. To study occupation, for example, Curti had to group the many individual written entries on the enumeration forms into a reasonable number of analytic categories. Today, the continuing challenges of such work explain the vibrancy of international debate about occupational classification along with ongoing proposals for revised approaches by international associations dedicated to promoting standards and guidelines that facilitate comparison across societies and over time (Roberts et al. 2003). While such work has become located within the social sciences rather than DH for reasons discussed below, it aptly illustrates the complexity of analyzing even short textual expressions in relationship to other such short expressions.

Another important feature of Curti's research was how it illustrated the significance of technological capabilities in data creation and analysis. Not surprisingly, in hindsight, census enumerations became the first use of early tabulating machines especially following the late nineteenth-century inventions of Herman Hollerith, who helped develop the punched card electrical machines chosen to handle the 1890 enumeration in the United States. While embracing the benefits of Hollerith's inventions, census officials always had to confront the challenges of accommodating the written responses to enumeration questions on the punched cards, given the number of columns available. These responses often exceeded the expected letter and word length and thus made aggregation into categories increasingly difficult, especially as the number of census questions rose from dozens to hundreds by the early twentieth century (Coats 1946, Dom. Bur. Stat. 1955).

The challenge of data creation resembles part of today's efforts to appropriately handle text in data analysis even for quite controlled survey-type administrative census documents that often have text of widely varied (and sometimes unpredictable) length. While the width of the punched cards familiar until the later twentieth century varied from 25 to 80 columns, enumerators were only physically constrained by how much writing they could squeeze in the allotted space on the forms. This kind of constraint is certainly no longer as relevant today but creators of data from documents must still be aware of how the transformation of text by intended or inadvertent simplification during initial data capture not only may facilitate analysis but also may have profound research implications. The current popularity of Twitter-focused data research is undoubtedly connected to the advantage of its 140-character limitation for each textual submission.

Overall, as recently reported (Franke et al. 2016, p. 375), researchers today recognize that "a very large hurdle in the analysis of data, which is especially difficult for massive data sets, is manipulating the data for use in analysis." They emphasize that "big data is often very 'raw': considerable preprocessing, such as extracting, parsing, and storing, is required before it can be considered in an analysis" (p. 375). To such work, humanists would add the key step of addressing a question highlighted a half century earlier for readers of *Computers and the Humanities* by Robert Zemsky: "Just how did our evidence come into being?" (Zemsky 1969, p. 31).

Among the humanists who took significant steps in preprocessing text were literary scholars who saw the connection between the early scientific uses of computers and the labor-intensive work needed to study the content of books. As Curti and his research team were undertaking research on Trempealeau County, Roberto Busa was developing an initiative that humanists today consider a founding project in the DH. Busa's focus was the voluminous writings of Thomas Aquinas (Busa 1980).

His research ambition was to enable various analyses based on the uses of certain words both in isolation (including all forms of the chosen word) and in the context of other words. In this ambition, Busa was taking from print culture a well-developed tradition of concordances, in which manual ways had been developed to create alphabetical lists of words used in the writings of key

individuals such as religious figures. While starting with the goal of automating such work, Busa began realizing that computer technology offered not only better ways of completing familiar tasks but also unforeseen possibilities for addressing new questions (Winter 1999).

According to Keller et al.'s (2017) categories, Busa's project involved opportunity data, defined as data arising from the unfolding of daily life. In the other research projects that were beginning to explore computer applications for documents, investigators such as Curti were able to fit their needs into the existing technological expectations for rectangular data based on sources that could fairly easily be codified into numeric symbols to form part of a conventional data file. In contrast, Busa had to face the challenge of creating data from the entire Aquinas corpus that was estimated to contain about nine million words in Latin. For their part, IBM staff prepared a report on Busa's proposal that explained to their president, Thomas Watson, why the current technology was not capable of meeting such a request (such as the limitations of 80-column formatting) (Tasman 1957). Surprisingly, Busa was eventually able (while also pursuing other options) to convince Watson that IBM's slogan of "the difficult we do right away; the impossible takes a little longer," meant that the company should accept the challenge of finding ways to handle large quantities of unstructured text (Jones 2016). As a result, IBM researchers began coming to grips with creating a seemingly endless string of data that could not obviously be imagined as observations and variables with clear conceptual differences between the two.

In the following years, Aquinas's writings were duly transformed at IBM into 0s and 1s (beginning with poetry to accommodate initial technological limitations) and were then subjected to innovative computations to identify and analyze word usage in ways that came to be well-known, such as KWIC (key words in context) and KWOC (key words out of context) (Luhn 1960). During the 1960s, a small group of diverse researchers realized that, at the heart of textual analysis, was the need for text encoding and classification, and their work became crucial to later developments. The first major conference on computing in the humanities was held in Cambridge, United Kingdom, in 1970 and then continued in Europe and North America, alternating location during the following years. The subtly changing titles of the published conference proceedings reflect the changing ways in which the researchers were viewing their objectives. Articulated initially in 1971 as *The Computer in Literary and Linguistic Research* (Wisbey 1971), the proceedings became *The Computer and Literary Studies* (Aitken et al. 1973) and then *Computers in the Humanities* (Mitchell 1974) before settling on *Computing in the Humanities* (Lusignan & North 1977, Bailey 1982) for the remaining two volumes of this period. These were among the first steps taken toward today's intensive work to harvest insight from statistical analysis of written expression as well as to build digital tools such as search engines (Burrows 1992).

In his editor's preface to the 1974 volume, J.L. Mitchell (1974) emphasized that computer-based humanities projects were often collaborative initiatives involving researchers from diverse disciplines. Overall, Mitchell listed 12 distinct academic affiliations among the authors, who came from the social and natural sciences as well as the humanities.

He noted one article written by an electrical engineer as well as one by a psychologist. The expectation was that the volume would be used in the courses on computers in the humanities that were said to be offered in many major universities in the United States and Europe (Mitchell 1974). Additional steps forward were also taken during these years at IBM's Thomas J. Watson Research Center, where innovative thinking through the 1970s and later led to language breakthroughs in areas such as speech recognition.

The largest collaborative international effort that eventually grew from such early projects was the Text Encoding Initiative (TEI) that, on the thirtieth anniversary of its launch, was awarded the Antonio Zampolli prize for 2017 by the Alliance of Digital Humanities Organizations at its annual conference. Over the decades, members of this consortium have been collectively developing

guidelines for encoding text that have become the standard for use not only in research and publishing but also in cultural institutions to guide current digitization as well as curation and preservation. Established as a nonprofit organization, the TEI Consortium has proven to be the most impressive example of how humanities expertise can enable the international and society-wide character of digital to be matched to the similarly pervasive importance of text as a key form of human expression.

STEPS AND MISSTEPS IN STATISTICAL ANALYSIS

If judged in comparison with successful DH research projects today, the initiatives of Curti, Busa, and certain others would suggest that a new research field based on conceiving and implementing digitally enabled projects was poised for takeoff (Gouglas et al. 2013). However, as the first researchers began embracing computer-assisted textual analysis, their work became embroiled in heated controversies that worked against both development and integration into mainstream academic activities. Despite the success of the TEI Consortium and certain other initiatives over the years, these controversies have had enduring consequences that help explain both the belated blossoming of DH and its continuing challenges in integrating expertise from distinct disciplinary cultures. As recently as 2007, specialists admitted that "quantitative analysis has not had much impact on traditional literary studies. Its practitioners bear some of the responsibility for this lack of impact because all too often quantitative studies fail to address problems of real literary significance, ignore the subject-specific background, or concentrate too heavily on technology or software" (Hoover 2013, p. 518).

In a study (Swierenga 1970) of how historians were developing what became known as "the new social history" or "cliometrics," Robert Swierenga found that research publications during the 1960s often included descriptive and inferential statistics such as frequency distributions, scaling, correlations, so-called likeness indexes, and, in certain cases, factoring. Various studies of legislative voting patterns made use of Guttman scaling, cluster bloc analyses, and cohesion indices. Swierenga emphasized that such work was accelerated by the availability of new facilities such as the "marvelous machine" at the University of Iowa Computer Center that was able to provide 6,441 fourfold tables or paired comparisons between the individual voters being examined in one research project. Swierenga felt convinced that the future was bright, as "more historians are becoming statistically oriented and computer proficient" (Swierenga 1970, pp. 19–20).

Such optimism soon proved unjustified. The first salvo was Carl Bridenbaugh's Presidential Address to the American Historical Association in 1962, in which he offered a wide-ranging assessment of the state and future of the discipline. Bridenbaugh perceived that "like nearly any activity in the Western world, historical scholarship has undergone a technological revolution, and we now possess, and probably will add to and improve, remarkable techniques for handling our raw materials, advantages of which previous historians never dreamed. Among other ways, bigness has struck us by proliferating sources and editing, thereby deluging us with an overwhelming mass of data for the study of the last one and a half centuries of history. The new age has built up a stock pile of sources and forced us to resort ever more frequently to statistics" (Bridenbaugh 1963). While admitting some advantages, Bridenbaugh focused on the negative consequences of this emerging "quantitative history." He emphasized that the "realization that historical facts are unique in character, space, and time restrains the historian from trying to fit them into a rigid theory or fixed pattern—and here he can render emergency yeoman service to his unhistorical colleagues in other disciplines." In his view, a historian must not ever "worship at the shrine of that Bitch-goddess, QUANTIFICATION, [*sic*] History offers radically different values and methods" (Bridenbaugh 1963).

While it may be tempting to dismiss Carl Bridenbaugh's address, the thrust of his remarks became characteristic in continuing controversy that gained traction during the 1960s and took flight during the 1970s and 1980s (Bogue 1983). The controversy was regularly fueled by probing methodological criticism of the data analysis undertaken by early practitioners such as Curti. This criticism exposed repeatedly the characteristic limits of statistical expertise among historians as well as the inadequate domain knowledge among the collaborating statisticians (Smith 1984).

One reaction focused on the tendency of researchers to state confident conclusions that went far beyond the evidentiary data under examination. For humanists, research like Curti's study of a county in Wisconsin was hardly appropriate for addressing general questions about the frontier experience. The humanist privileging of the individual and the specific context discouraged the use of data to study large-scale historical change. Moreover, historians could always be criticized for not studying all potentially relevant data even by converts to statistically informed historical research. In 1969, John J. McCusker (1969) claimed Theodore Rabb had committed a fundamental statistical error by downplaying missing data in his empirical study of investment in England during the late sixteenth and early seventeenth centuries. Such critiques soon became limited to those who gravitated toward the new social science historical groups, while most historians disconnected from the increasingly sophisticated debates about statistical methods.

Two examples illustrate the methodological issues that unfortunately worked against the spread of data analysis in historical research even among early adopters: statistically explained variance and multicollinearity. As historians increasingly saw the value of multiple regression for addressing questions about the relative importance of characteristics such as gender and birthplace in influencing other attributes such as migration and school enrolment, they were initially quite disappointed to find that their conceptualizations often translated into relatively modest statistical explanations of their data. Such disappointment sometimes reflected expectations based on studies in psychology and sociology especially those using experimental designs. The conclusion that, for example, the independent variables under study could explain, for example, only 20% of the variation in the dependent variable seemed to raise questions more than provide historical insight.

While some researchers lowered their expectations that their own estimates of historical significance would be matched by their data's statistical significance, some stayed with basic descriptive statistics and others lost all enthusiasm for data analysis and rejoined the qualitative majority. Similarly, historians only belatedly recognized that multiple linear regression can be unreliable when there is a high degree of statistical dependence among the variables chosen for systematic analysis. In the case of census data, for example, variables such as birthplace, ethnicity, and religion were often closely related both conceptually and statistically; in fact, enumerators sometimes used the response to one variable to help determine the value of another (Fitch 1984).

While specialists in statistics were focusing on issues like multicollinearity by the mid-1960s, historians were still catching up years later in recognizing problems such as variable interactions and confounding. Over the years, a small minority of historians did develop increased sophistication in conventional data analysis such that criticism over methods gave way to new norms for historical data analysis, most notably, the use of logistic regression. Although the label "cliometrics" came and went in the early 1970s, this minority of historians bolstered the continuing success of the Social Science History Association (SSHA); however, their association with the group also emphasized their distance from the humanities.

After the 1960s, researchers increasingly embarked on individual research projects, usually borrowing from social sciences such as psychology, sociology, and economics, especially as statistical software packages such as the Statistical Package for the Social Sciences became familiar on campus. As a result, historical data analysis increasingly moved away from its intellectual origins in the humanities. During the 1960s, early key developments were regularly presented and debated

by historians and diverse other researchers in the journal *Computers and the Humanities*. During the 1970s, more historians joined with political scientists and economists to bolster the SSHA that soon published *Social Science History* as well as organized an annual conference that quickly became the go-to event of the year. Along with other initiatives such as the transformation of the newsletter, *Historical Methods*, into a full research journal, historians interested in data soon had little contact with other humanists.

While fragmenting humanist engagement with data, developments such as the establishment of SSHA did deepen considerably the sophistication with which a small number of historians began coming to grips with data analysis. Over time, what is now termed data wrangling came to require a sophisticated understanding of the relationship between the creation of the data and its value for providing insight into the human experience (Darroch & Soltow 1991). Unlike earlier researchers like Curti who viewed the census enumerations, for example, as providing quite objective evidence about the residents of Trempealeau County, historians developed over the years a far more robust appreciation of the underlying mindsets, motivations, and consequences of the ways in which humans express their perspectives on themselves and the world around them (Novak 1988). For Western societies, scholars (Porter 1986, Hacking 1990) have shown how a particular type of thinking—namely, linearity—underpinned many of the documents from modern times that are studied by researchers.

The census enumeration forms illustrate clearly this so-called statistical mind since they were crafted within a worldview determined to impose linearity on understandings of change and the construction of social hierarchies. Behind each census question was the anticipation of singular responses that were more or less close to an ideal. For this reason, for example, enumerators in countries like Canada struggled to apply the dominant-culture questions to indigenous peoples, especially those living in rural areas who did not occupy the expected "dwelling unit" or have a "household status" such as "head." In that sense, each census enumeration was not benignly counting and classifying social realities but was, in fact, constructing hierarchies in which individuals were singularly lined up from "best" to "worst" according to often implicit criteria that reflected the dominant mindset of the era. Whereas researchers had initially focused on the taking of the census, the revised approach emphasized the making of the census (Curtis 1994, Dunae 1998). And rather than counting responses to census questions, researchers began studying the questions themselves as illustrations of the ways in which governments were attempting to increase and solidify their power.

More recently, researchers have begun embracing a quite robust understanding of how text reflects multiple authors, if defined in terms of who influenced the actual form and content of documentary evidence (Baskerville & Sager 1998, Gaffield 2005). From this perspective, complementary contextual data can help researchers interpret how documents like enumerations provide evidence of individual lives that were both linguistically constructed and materially based. For this reason, the Canadian Century Research Infrastructure (CCRI) includes, for example, databases of related newspaper articles as well as census microdata to study how familiar and new words and concepts have been used and understood (Gaffield 2007).

The ways in which humanist thinking is enriching the interpretation of data is worth emphasizing since it is crucial that the emerging field of data science avoid assumptions that the facts speak for themselves. Expertise is essential in applying effectively the realizations that humans decide what the "facts" are and that humans make data speak "facts" in specific ways just as computers reflect societal context both explicitly and implicitly (Golumbia 2009). DH makes clear that such expertise depends upon integrated transdisciplinary approaches based on both specialized knowledge and well-developed familiarity with related research fields. Humanists need not become experts in statistical sciences and vice versa, but each research collaborator requires appropriately

broad as well as deep expertise to fully engage in data analysis. The case of historical evidence related to indigenous peoples clearly illustrates this requirement since even descriptive statistics must recognize the complex ways in which different worldviews may have collided in the creation of specific documents. In this spirit, official statistical agencies now provide researchers with fulsome quality guidelines that do not go as far as desired by DH researchers but nonetheless reflect the increasing attention to the interplay of linguistic and material realities (Stat. Can. 2009).

The example of historical research illustrates a larger trend in which increased sophistication in textual analysis after the 1960s occurred within specialized groups that established their own niches, often in isolation from related efforts in nearby disciplines (Lancashire 2005). For example, the Association for Computational Linguistics that was founded in 1962 increasingly moved away from interactions with other groups of humanities scholars and, instead, pursued their own activities that continued to grow impressively; their publicly accessible digital repository now includes 41,024 articles on the study of computational linguistics and natural language processing. Such work has had a limited relationship to other research endeavors in the humanities, even closely related work such as the TEI that computational linguists often saw as unnecessarily complicating research projects rather than offering simple codes in a single standard for use by all investigators. Illustrative of the enduring legacy of the formative 1970s and 1980s, computational linguists have only recently begun seeing themselves as members of a big tent DH community.

TOWARD INTEGRATED CROSS-CAMPUS DIGITAL HUMANITIES RESEARCH COLLABORATION

In surprising ways, the humanities' sensibilities and insights that underpinned early critiques of computational research have subsequently become a focus in the statistical sciences. Not only is there greater and greater knowledge about the interface of linear-based statistical models and specific disciplines like history, but researchers in statistical sciences are increasingly concerned about the use and misuse of specific measures in data analysis. In 2016, for example, the American Statistical Association felt compelled to emphasize that statistical significance is not equivalent to scientific, human, or economic significance (Wasserstein & Lazar 2016).

This point resonated with humanities scholars who consistently define significance in light of their own reading of various texts. For them, statistical analysis is valued as a way to think through the meaning of text rather than to prove what the text might or might not mean in a fixed sense. Humanists use computation to inspire new interpretations and insights rather than to justify a hypothesis (Bod 2013). Historians, for example, sometimes focus on the ways in which statistical insignificance may nonetheless have societal significance (e.g., election results) or may reflect an overall empirical result that does not reveal noteworthy patterns among one or more subgroups. In the same way, humanists have begun using topic modeling not to uncover the so-called real themes emphasized in a text corpus but rather as a way to stimulate thinking about the focus and meaning of the work under study with a view toward invigorating rather than ending discussion. Some scholars are also becoming interested in Bayesian thinking and other innovative ways to develop interpretations of text that are informed by, rather than based on, statistical analysis (Fenton et al. 2016).

From this perspective, DH humanists are now joining with scientists who have become similarly interested in statistical features at multiple and unanticipated levels and among unrecognized clusters (Sawyer 2005, Li 2016). Just as economists have abandoned ideas of "the rational man" (to use the term from decades ago) and sociologists no longer refer to "normal" behavior as in years past, the emphasis of humanists on contingency and context is now converging with scientific thinking about complexity including ideas of emergence and nonlinear change (Miller & Page

2007). Viewed in this way, all quantitative analysis is also inherently qualitative, just as qualitative analysis usually includes quantitative statements about size and scale (Perron et al. 2000).

By recognizing these converging perspectives across disciplines, we can understand the importance of collaboration between humanists and those in statistical sciences in developing statistical methods appropriate for textual data. As recently observed, "For many big data problems, the variety of data types is closely related to an increase in complexity. Much of classical statistical modelling assumes a 'tall and skinny' data structure with many more rows, indexing observational or experimental units n, than columns, indexing variables p" (Franke et al. 2016, p. 373). And research at the level of millions of words is now seen as compatible with close study of certain passages (Clement 2013, Sinclair & Rockwell 2014).

A stunning example of the value of carefully crafted cross-campus collaboration involved Ian Lancashire, a literature scholar and founding figure in DH, and Graeme Hirst, known for his research in artificial intelligence and natural language processing. Their project revealed for the first time what had been previously suspected: Agatha Christie's later novels reflect the onset of dementia. By systematically analyzing words and groups of words in keeping with Busa's early study of Aquinas's writing, Lancashire and Hirst showed a statistically significant drop in vocabulary as well as an increased use of certain phrases during her career. Along with other changes, this evidence showed that Christie was not only aging but also dealing with Alzheimer's disease during her later years (Lancashire & Hirst 2009). Lancashire and Hirst's study topped the "Health" category in the *New York Times Magazine's* "9th Annual Year in Ideas" and soon inspired new diagnostic approaches to early detection in clinical settings (Fortini 2009). At the heart of the success of this collaboration was the required partial overlap in expertise that this scholar and scientist shared as well as their distinct backgrounds that enriched the entire research process from data capture to interpretation (Lancashire 2010).

The potential of this type of collaboration within large international research initiatives was pursued over a decade during the building of the CCRI by a research team that included historians, sociologists, geographers, demographers, archivists, and computer scientists. From the start in 2001, this team embraced what might be characterized as T-shaped engagement between members, with the vertical axis representing distinct expertise and the horizontal axis the knowledge overlap that connected the various individuals. Since such engagement is easier in theory than in practice, special commitment was required for team members to learn new vocabularies and different ways of working and sharing results throughout the collaboration, beginning with writing grant applications (Gaffield 2016b).

Overall, CCRI team members had to become familiar with what Michèle Lamont describes as the various cultural scripts that distinguish professors on different parts of the campus (Green & Gutmann 2007, Lamont 2010). Of course, the CCRI project found that there were limitations to how well each member was able to fully understand the implications of the diverse scholarly and scientific traditions but, nonetheless, it was key that everyone appreciated and embraced the challenge. Serious commitment to understanding different cultural scripts was also needed in CCRI's partnerships beyond campus, such as those with public institutions and the private sector. The limited number of collaborations like CCRI reflects the continuing challenges of updating entrenched disciplinary structures and practices to facilitate DH (Green & Gutmann 2007).

THE ROLE OF RESEARCH GRANTING COUNCILS

The experience of DH since the 1960s emphasizes the extent to which dedicated efforts must be made to deal with the complexity of the collaborative, campus-wide, cross-sectoral initiatives that first emerged in the 1960s but failed to mature successfully until recent years. The diverse parts of

such initiatives in terms of either individuals or structures cannot simply be added together; rather, they must be integrated appropriately. Toward this end, national research funding agencies have played key roles especially during the past 15 years. Experience has shown that such agencies can enable positive change in multiple ways from grant program design and selection criteria to the conditions of award holding including accommodation of domestic policies affecting international collaborations.

During the later twentieth century, there were certainly some examples of granting agency support for digitally enabled textual study. The most notable was the foundational partnership among the US National Endowment for the Humanities (NEH), the Directorate XIII of the Commission of the European Communities, the Andrew W. Mellon Foundation, and the Social Sciences and Humanities Research Council (SSHRC) of Canada to mount the TEI. Among the first dedicated efforts was SSHRC's Image, Text, Sound and Technology (ITST) funding opportunity that followed consultations during 2001–2002. This initiative was based on the perception that "to examine and interpret individuals and their cultures, researchers currently use three fundamental kinds of digital information: images, text and sound. These digital forms of information are, however, very sensitive to changes in the technologies through which they are created, analyzed, published and preserved. In recent decades, innovative technologies have transformed the very definition of text and its relationship to image and sound. To benefit fully from these new technologies, researchers must not only be aware of technological developments, but also be directly involved in them" (SSHRC 2009).

In this way, SSHRC defined scholars in the social sciences and humanities as cocreators of digital technologies rather than simply users. The funding opportunity gave many examples of possible research: electronic editing and publishing; web programming; immersive and virtual environments in multimedia research; textual analysis; 3D imaging technology; creativity, culture, and computing; digital image design; information aesthetics; and computer gaming. On the basis of the success of this funding opportunity, SSHRC began focusing on ways for national granting agencies to transcend their domestic mandates by developing straightforward mechanisms for international DH research collaboration.

In 2008, SSHRC built on the domestic success of ITST by partnering with its counterparts in the United States and the United Kingdom to create an ingenious way to efficiently support international research collaboration. Launched the following year under the leadership of NEH's Office of Digital Humanities (ODH), the Digging into Data Challenge asked applicants to develop responses to the unprecedented availability for research of millions of books, millions of newspaper pages, millions of photographs of artwork, and other massive repositories of digitized data that simply could not be studied conventionally in many lifetimes. The key logistical feature of this funding opportunity was the agreement proposed by ODH's creative and innovative team, led by Brett Bobley, for each national granting agency to respect the decisions of a single application and adjudication process by funding their "own" members of the winning research teams. In this way, the various agencies (initially from the United States, United Kingdom, and Canada) were able to combine their national mandates with recognition of the increasing internationalization of research in the digital age. The quality and originality of these teams attracted worldwide media attention and inspired concerted efforts to combine domestic and international support for research collaborations that crossed jurisdictional boundaries (Williford & Henry 2012).

The single-process approach of the Digging into Data Challenge proved so successful that it inspired by 2013 the start of construction on a global platform for research funding. Led by the SSHRC, a consortium of national funding agencies involving ten partners from Europe and eight from North America and South America built the TransAtlantic Platform with support from the European Union's Framework Program for Research and Technological Development. The first

funding opportunity was a Digging into Data competition open to researchers from 16 countries across the Atlantic world. The long-term ambition is to extend the platform to enable seamless global research teams.

Moreover, while currently composed mostly of national funding agencies that focus on the social sciences and humanities, the logic of the new research paradigm calls for full inclusion of all the ways of knowing, including those based in the larger society. One encouraging example is the Canadian participation in the first TransAtlantic Platform call for proposals that was supported not only by SSHRC but also by the Natural Sciences and Engineering Research Council and the Canada Foundation for Innovation as well as provincial partners. In such ways, DH is now helping lead the way in taking seriously the importance of integrated support for those across campus and beyond to benefit exponentially from their distinct strengths in major research projects.

CAPACITY-BUILDING IN DIGITAL HUMANITIES

The decades since the mid-twentieth century have been surprisingly stable in academic programming, especially when compared with the rapidly changing research landscape. Only recently have serious efforts begun to prepare students for handling complex data, especially in the humanities. In his contemporary analysis of nascent work in "computing and history" as it had developed by the end of the 1960s, Robert P. Swierenga remarked presciently on the importance of scholars creating appropriate digitally enabled methods. "Borrowing from other disciplines is not the solution," he emphasized, since historians deal with different kinds of evidence that call for different statistics and computer programs (Swierenga 1970, p. 20). Swierenga cited Robert Zemsky's earlier call for historians to "invent a methodology—including computer programs—of our own, a methodology designed to cope with the peculiar kinds of evidence with which we deal" (Zemsky 1969, p. 39). Writing in 1970, Swierenga defined this need as the vital task of the next generation (Swierenga 1970).

The following years included some impressive examples of work on this vital task, but it is also clear that more needed to be done to help students and researchers develop the different competencies required for inventing digitally enabled approaches to the peculiar character of research on human expression (Gaffield 1988). Some courses to enhance computational skills for humanities students were introduced during the 1970s and 1980s but most fell by the wayside before becoming routine features of the curriculum. In fact, major steps forward have been taken only in the past 15 years and, characteristically, these steps have involved additions to, rather than changes to, established policies and practices including academic programming.

One positive step has been the creation of stand-alone research groups and centers that offer undergraduate and graduate students as well as professors the chance to acquire skills in data creation, analysis, and curation. For example, the University of Virginia established the Institute for Advanced Technology in the Humanities in 1992 to provide "consulting, technical support, applications development, and networked publishing facilities" in support of "a unique collaboration between humanities and computer science research faculty, computer professionals, student assistants and project managers, and library faculty and staff" (Inst. Adv. Technol. Humanit. 2017). Six years later, the University of Virginia added to this initiative by launching the Virginia Center for Digital History to further support digitally enabled scholarship. A key ambition was to offer students opportunities that were not available elsewhere "to explore the ways that technology can enhance their own scholarly work, while teaching them the skills that will enable them to produce their own digital projects" (Inst. Adv. Technol. Humanit. 2017). Moreover, the Virginia Center for Digital History offered workshops to elementary and high school students to help them integrate digital resources into their curriculum (Va. Cent. Digit. Hist. 2017).

While groundbreaking in many ways, even the positive initiatives at the University of Virginia illustrate the extent to which academic structures and practices have not kept up with the changing landscape of research. In 2005, a report on *Digital Humanities at the Crossroads: The University of Virginia ECAR Case Study 6* described the preceding years as "an exceptionally successful garage phase" that was, however, not maturing into enduring institutional change (Blustain & Spicer 2005, p. 2). The report concluded that the University of Virginia's efforts composed a "cautionary tale" for digital scholarship in which "its practitioners have been challenged by traditional academic structures, values, and incentives." Specifically, for example, those involved in the various initiatives were found to encounter "issues related to promotion and tenure, the limits of decentralized autonomy, and even the concept of scholarship itself."

In contrast, there are certainly a few examples of multiyear major initiatives dedicated to digitally enabled humanities' scholarship such as George Mason University's impressive Roy Rosenzweig Center for History and New Media and the University of Maryland's innovative Maryland Institute of Technology for the Humanities. In addition, research groups such as the University of Ottawa's Textual Analysis and Machine Learning Group have been expanding to include researchers from diverse fields; one early achievement in this context was an innovative study of individual census enumeration responses using decision trees to identify historical subgroups of specific interest (Drummond et al. 2006). Overall, however, DH has developed without full, sustained support in resources and expertise on most campuses.

Another example of how the emergence of DH has depended upon capacity building beyond established academic programming is the rapid growth of the Digital Humanities Summer Institute (DHSI) launched at the University of Victoria in British Columbia in 2001 by a group of early career scholars who wished to build a supportive community of practice for computational applications in the arts and humanities. Most known for its annual spring weeklong instructional sessions, DHSI now attracts nearly one thousand participants from around the world each year. The institute's creator, Ray Siemens, was awarded the Antonio Zampolli Prize by the Alliance of Digital Humanities Organizations in 2014. In addition to learning both introductory and advanced skills, emerging and reorienting researchers develop networks as well as explore collaborative possibilities with other researchers who are often from quite different backgrounds. An important recent development has been the participation in, and support of, DHSI by the Canadian Statistical Sciences Initiative, which is becoming a leader in promoting and enabling cross-campus engagement to enhance sophisticated data analysis.

While truly impressive, the success of DHSI also reflects the fact that scholarly and scientific academic programming across many institutions continues to lag behind the rapidly changing expectations for digitally enabled learning (Bonnett & Kee 2009). There are certainly a growing number of important exceptions such as new undergraduate and graduate programs like the DHSI-sponsored Graduate Certificate in Digital Humanities offered by the University of Victoria. Similarly, it is also encouraging that campuses can now benefit from the recently published "Curriculum Guidelines for Undergraduate Programs in Data Science" (De Veaux et al. 2017). For example, Lynne Siemens gave reasons for optimism in her 2013 report on DH capacity overall in Canada, where she found a "growing acceptance of digital methods, resources and tools" across the social sciences and humanities, although she also noted that "issues for funding both initial development and ongoing sustainability and relevance of digital resources remain unresolved and may become more critical over time as more digital resources and tools are created" (Siemens 2013).

The current observation that much more should be done immediately to normalize support for DH was similarly emphasized by the British Academy in its position statement made in 2012. The rationale for this statement was the academy's conviction that "to understand social dynamics,

cultural phenomena and human behavior in the round, researchers have to be able to deploy a broad range of skills and techniques." As a result, the "critical need for quantitative research is not confined to any particular field but, rather, applies to disciplines drawn from the full spectrum of the social sciences—and increasingly, of the humanities." Despite this need, however, the British Academy expressed deep concern that "the UK is weak in quantitative skills, in particular but not exclusively in the social sciences and humanities" with "serious implications for the future of the UK's status as a world leader in research and higher education, for the employability of our graduates, and for the competitiveness of the UK's economy" (Br. Acad. 2012, p. 1). Taken together with similar observations elsewhere, this perspective emphasizes that there is a long way to go in conceptualizing and implementing digitally appropriate policies and practices.

One concern for humanists is the possibility that DH leadership will be taken over by researchers from other academic traditions and without cross-campus engagement. An op-ed in the *New York Times* claimed that the humanities risk being sidelined by "biologists, economists and physicists" who had begun analyzing text within a new "mathematical theory of culture." In "One Republic of Learning: Digitizing the Humanities," Armand Marie Leroi, a professor of evolutionary developmental biology at Imperial College, London, described how a "code-capable graduate student" could now trump a "traditional, analog scholar" in examining a "claim about the origin, fate or significance of some word, image, trope or theme." She wondered whether or not all humanists, "weakened by their own interminable, internecine Theory Wars," would "gratefully accept the peace imposed by science" (Leroi 2015).

Within the DH community as in many other areas, specific attention is also being paid to the importance of inclusive capacity building, especially to attract women and visible minorities. While the humanities are often considered welcoming disciplines for women as well as men, some scholars note that DH conferences characteristically include women as less than one-third of the presenters. In comparison with trends in the sciences, however, this proportion seems encouraging. Observers have repeatedly focused on the limited extent to which women enroll in computer programming courses, and there are indications that the gender gap has been getting worse. The American Association of University Women reported in 2016 that the percentage of women among computer science graduates in the United States fell from 37% in 1984 to 18% in 2016 (Amer. Assoc. Univ. Women 2016). When compared with this trend, the more balanced gender distribution in DH has attracted positive attention from administrators and public policy makers. Sandra Collins of the Digital Repository of Ireland has emphasized that "digital humanities researchers are working at the cutting edge of technology in a field that has significant potential for Ireland. In the context of government policies to increase female participation in STEM, and lots of positive action in the area, here is a field that is evolving gender balance organically. We should look to the digital humanities sector for clues as to how to address imbalance in other areas of STEM" (Faller 2016).

More intensive studies would undoubtedly help balance the established scholarly portrayals of men leading in the creation of digitally enabled research. For example, close reading of the records of Merle Curti's innovative census-based study indicate that his wife, Margaret, who had an advanced degree in psychology, participated in the project and may have influenced her husband in applying research approaches familiar in her own work. Scholars have already shown that it was Lotte Bailyn, also trained in psychology, who convinced her husband, the historian Bernard Bailyn, to abandon his hand tabulation of ship records in his study of trade in favor of new data processing facilities. He later confirmed that her advice had been essential to the success of his project that "would have been impossible without the use of tabulating machines" (Swierenga 1970, p. 3). Both Margaret Curti and Lotte Bailyn were published researchers experienced with data processing and statistical analysis, as was Myrtle Kitchell Aydelotte, who published quantitative studies in nursing

education while her husband, William O. Aydelotte, undertook computer-assisted analysis of voting in the British Parliament in the 1840s (Swierenga 1970, p. 4).

CONCLUSIONS

In 1964, Marshall McLuhan launched a major collaborative research project with "awesome implications" in the words of the *Globe and Mail*, Canada's leading English-language newspaper (Munro 1964, p. 2). By bringing together researchers in medicine, architecture, engineering, political science, psychiatry, museology, anthropology, and English, this project sought to discover the "impacts of culture and technology on each other" with a view toward developing a "surefire method of planning the future and making a world free from large scale social mistakes." McLuhan and his cross-campus team at the University of Toronto aimed to confront the "failure in communications" across society "with the aid of such sophisticated machines as the computer and the head camera" that would allow researchers "to see what a man is really looking at, not what he thinks or says he is looking at." One of the researchers, Daniel Cappon, a psychiatrist, anticipated that "with a model built on the relationships between technology and culture and perceptual typology, the enhanced ability to predict and control would bring light to the darkness of the future" (Munro 1964, p. 2).

In hindsight, we can see why the ambition of McLuhan's too-far-ahead-of-its-time "awesome" project remained unfulfilled. The good news is that current DH initiatives often replicate McLuhan's inclusion of diverse scholars and scientists while also partnering with companies and public sector institutions (Irvine 2015). The bad news is that DH depends upon continuing changes to research policies and structures as well as academic mindsets. Peter Baskerville has recently reminded us of fellow historian Michael Rothberg's frank admission in 2010 that "when I first read the phrase 'quantitative cultural history,' I was tempted to reach for my revolver" (Rothberg 2010, p. 319; Baskerville 2015). While more welcoming in his comments, also made in 2010, the leading historian Anthony Grafton still assumed that counting could be separated from interpretation. "The digital humanities do fantastic things. I'm a believer in quantification. But I don't believe quantification can do everything. So much of humanistic scholarship is about interpretation" (Cohen 2010, p. C1).

The research developments of the past half century, however, make clear that it is misleading to see DH as continuing rather than transforming scholarly and scientific traditions. While Roberto Busa could be considered an important originator of DH for his efforts to use computers for handling words rather than numbers, this juxtaposition does not capture Busa's breakthrough work that, in fact, foreshowed many of the computational features of twenty-first-century textual analysis enabled by hypertext. This is why, for example, Roberto Busa has also been praised for helping pave the way for societal transformation today. At the time of Busa's death in 2011, the journalist Stefano Lorenzetto aptly noted that "if you surf the internet, you owe it to him and if you use a PC to write emails and documents, you owe it to him" (Priego 2011).

In the coming years, it seems likely that diverse research traditions will continue converging on the challenges and opportunities of textual analysis and related studies of human expression. Fortunately, changes across campus and beyond are now converging in the belated blossoming of digitally enabled research (Berry 2011). While some may say that the humanities are becoming more science-like, key long-standing features of disciplines like literature and history are now resonating in fields as diverse as genomics and computer science. The result is meaningful engagement in discipline-based interdisciplinarity in which historians learn technical skills and computer scientists learn humanistic skills (Graham et al. 2016). The example of historical research is only one of many dramatic changes now underway that highlight the fading distinctions

between scientific and scholarly disciplines and highlight the importance of integrated discussion about scientific objectivity, quantitative and qualitative research, and other perspectives that have propped up C.P. Snow's concept of two cultures of arts and sciences (Graff 2015, Gaffield 2016a).

The example of DH historical research also illustrates the general need to pay greater attention to fundamental issues in enhanced statistical analysis ranging from data creation through appropriate computation to preservation and capacity building. These issues include the increasing challenge of defining and assessing data quality as well as the possibilities of combining well-established and emergent statistical approaches to study unstructured and deeply complex data. Perhaps the most important issue is the continuing and indeed increasing importance of robust and readily accessible documentation describing the complete research workflow from data creation to all stages of computation. Such documentation is key to data preservation and curation as well as follow-up research projects, especially given increasing concern about reproducibility. Critically discussing DH as part of this larger context makes the foundational and, in fact, increasing role that humanities scholars have played in data analysis of the human experience more understandable.

Finally, recent DH developments also illustrate why the multiplying benefits from new insights through textual analysis are reorienting—or disrupting as often said—many aspects of social organization from business strategies and public policies to governance expectations and health practices. To some extent, the results so far demonstrate that this impact can help create a better future with enhanced well-being both individually and collectively. However, headlines often remind us of increasing digitally enabled crime such as identity theft and election sabotage as well as unintended negative consequences that result from misleading or inappropriate statistical analysis or interpretation (Liu & Meng 2016). The rapid growth in data about people raises not only complex ethical questions with long research traditions in the humanities but also new concerns such as how linked data can reveal personal identities in unintended and sometimes illegal ways. These examples reinforce the urgency for scholars and scientists to learn from initial and recent experiences to pursue integrated DH research approaches that draw upon the full range of expertise relevant to appropriate data creation, analysis, and preservation.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Aitken AJ, Bailey RW, Hamilton-Smith N. 1973. *The Computer and Literary Studies*. Edinburgh, UK: Edinburgh Univ. Press

Amer. Assoc. Univ. Women. 2016. *Solving the equation: the variables for women's success in engineering and computing*. Washington, DC: Amer. Assoc. Univ. Women. **http://www.aauw.org/research/solving-the-equation/**

Ayers EL. 1999. *The pasts and futures of digital history*. Charlottesville: Va. Cent. Digit. Hist. **http://www.vcdh.virginia.edu/PastsFutures.html**

Bailey RW, ed. 1982. *Computing in the Humanities*. Amsterdam: North Holland

Baskerville P. 2015. *Big Data: So What?!* Presented at McMaster Univ., Dep. Hist. Annu. Grad. Colloq., Big History, March 26, Hamilton, Ont., Can. **https://www.youtube.com/watch?v=WRtcyaJ6CqI&index=7&list=PLzLUWMt2NZLQIvHuAQZ3IYCLCWDD0FAiv**

Baskerville P, Sager EW. 1998. The census as historical source. In *Unwilling Idlers: The Urban Unemployed and Their Families in Late Victorian*, pp. 195–216. Toronto: Univ. Toronto Press

Berry DM. 2011. The computational turn: thinking about the digital humanities. *Cult. Mach.* 12:1–22
Blustain H, Spicer D. 2005. *Digital humanities at the crossroads: the University of Virginia ECAR case study 6*. Boulder, CO: EDUCAUSE. **https://library.educause.edu/~/media/files/library/2005/7/ecs0506-pdf.pdf**
Bod R. 2013. Who's afraid of patterns? The particular versus the universal and the meaning of humanities 3.0. *BMGN: Low Ctries. Hist. Rev.* 128(4):171–80
Bogue AG. 1983. *Clio and the Bitch Goddess: Quantification in American Political History*. Beverly Hills, CA: Sage
Bonnett J, Kee K. 2009. Transitions: a prologue and preview of digital humanities research in Canada. *Digit. Stud./Le Champ Numér.* 1(2). **https://www.digitalstudies.org/articles/10.16995/dscn.106/**
Br. Acad. 2012. *Society counts—quantitative studies in the social sciences and humanities: a British Academy position statement*. October. London: Br. Acad. **https://www.britac.ac.uk/publications/society-counts-quantitative-studies-social-sciences-and-humanities**
Bridenbaugh C. 1963. The great mutation. **https://www.historians.org/about-aha-and-membership/aha-history-and-archives/presidential-addresses/carl-bridenbaugh**
Burrows JF. 1992. Computers and the study of literature. In *Computers and Written Texts*, ed. CS Butler, pp. 167–204. Oxford: Blackwell
Busa R. 1980. The annals of humanities computing: the Index Thomisticus. *Comput. Humanit.* 14(2):83–90
Citro CF. 2016. The US federal statistical system's past, present, and future. *Annu. Rev. Stat. Appl.* 3:347–73
Clement T. 2013. Text analysis, data mining, and visualizations in literary scholarship. In *Literary Studies in the Digital Age: An Evolving Anthology*, ed. KM Price, R Siemens. New York: Mod. Lang. Assoc. Am. **https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/**
Coats RH. 1946. Beginnings in Canadian statistics. *Can. Hist. Rev.* 37:109–30
Cohen P. 2010. Digital keys for unlocking the humanities' riches. *New York Times*. Nov. 16, p. C1. **http://www.nytimes.com/2010/11/17/arts/17digital.html**
Cohen PC. 1999. *A Calculating People: The Spread of Numeracy in Early America*. New York: Routledge
Connelly M, Immerman RH. 2015. What Hillary Clinton's emails really reveal. *New York Times*. March 4, p. A25
Curti M. 1959. *The Making of an American Community: A Case Study of Democracy in a Frontier County*. Stanford, CA: Stanford Univ. Press
Curtis B. 1994. On the local construction of statistical knowledge: making up the 1861 census of the Canadas. *J. Hist. Sociol.* 7(4):416–34
Darroch G, Soltow L. 1991. *Property and Inequality in Victorian Ontario: Structural Patterns and Cultural Communities in the 1871 Census*. Toronto: Univ. Toronto
DeRose SJ, Durand DG, Mylonas E, Renear AH. 1990. What is text, really? *J. Comput. High. Educ.* 1(2):3–26
De Veaux RD, Agarwal M, Averett M, Baumer BS, Bray A, et al. 2017. Curriculum guidelines for undergraduate programs in data science. *Annu. Rev. Stat. Appl.* 4:15–30
Dom. Bur. Stat. 1955. Mechanical techniques. In *Ninth Census of Canada 1951*, Vol. 11: *Administrative Report*, pp. 125–52. Ottawa, Can.: Queen's Printer and Controller of Stationery
Drummond C, Matwin S, Gaffield C. 2006. Inferring and revising theories with confidence: analyzing bilingualism in the 1901 Canadian census. *Appl. Artif. Intel.* 20(1):1–33
Dunae PA. 1998. Making the 1891 census in British Columbia. *Hist. Soc./Soc. Hist.* 31(62):22–39
Faller G. 2016. *Women in STEM winning in digital humanities: press release*. News Release, May 19, 2015. **https://www.insight-centre.org/content/women-stem-winning-digital-humanities**
Fenton N, Neil M, Berger D. 2016. Bayes and the law. *Annu. Rev. Stat. Appl.* 3:51–77
Fitch N. 1984. Statistical fantasies and historical facts: history in crisis and its methodological implications. *Hist. Meth.* 17:239–54
Fortini A. 2009. Literary Alzheimer's. *New York Times*. Dec. 13
Franke B, Plante J-F, Roscher R, Lee EA, Smyth C, et al. 2016. Statistical inference, learning and models in big data. *Int. Stat. Rev.* 84(3):371–89
Gaffield C. 1988. Machines and minds: historians and the emerging collaboration. *Hist. Soc./Soc. Hist.* 21(42):312–17

Gaffield C. 2005. Evidence of what? Changing answers to the question of historical sources as illustrated by research using the census. In *Building New Bridges: Sources, Methods, and Interdisciplinarity*, ed. J Keshen, S Perrier, pp. 265–74. Ottawa, Can.: Univ. Ottawa Press

Gaffield C. 2007. Conceptualizing and constructing the Canadian Century Research Infrastructure. *Hist. Meth.* 40(2):54–64

Gaffield C. 2016a. Mindset and guidelines: insights to enhance collaborative, campus-wide, cross-sectoral digital humanities initiatives. *Int. J. Humanit. Arts Comput.* 10(1):8–21

Gaffield C. 2016b. The surprising ascendance of digital humanities: and some suggestions for an uncertain future. *Digit. Stud./Le Champ Numér.* 9. **http://doi.org/10.16995/dscn.2**

Graff H. 2015. *Undisciplining Knowledge: Interdisciplinarity in the Twentieth Century*. Baltimore, MA: Johns Hopkins Univ. Press

Graham S, Milligan I, Weingart S. 2016. *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press

Golumbia D. 2009. *The Cultural Logic of Computation*. Cambridge, MA: Harvard Univ. Press

Gouglas S, Rockwell G, Smith V, Hoosein S, Quamen H. 2013. Before the beginning: the formation of humanities computing as a discipline in Canada. *Digit. Stud./Le Champ Numér.* 3(1). **https://www.digitalstudies.org/articles/10.16995/dscn.244/**

Green AG, Gutmann MP. 2007. Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Syst. Serv. Int. Digit. Libr. Perspect.* 23(1):35–53

Hacking I. 1990. *The Taming of Chance*. Cambridge, UK: Cambridge Univ. Press

Hockey S. 2004. The history of humanities computing. In *A Companion to Digital Humanities*, ed. S Schreibman, R Siemens, J Unsworth, pp. 3–19. Oxford, UK: Blackwell. **https://dx.doi.org/10.1002/9780470999875.ch1**

Hoover D. 2013. Quantitative analysis and literary studies. In *A Companion to Digital Literary Studies*, ed. R Siemens, S Schreibman, pp. 517–33. London: Blackwell

Ide N, Véronis J, eds. 1995. *Text Encoding Initiative: Background and Context*. Dordrecht, Ger.: Kluwer

Inst. Adv. Technol. Humanit. 2017. About IATH. Charlottesville, VA: Inst. Adv. Technol. Humanit. **http://www.iath.virginia.edu/about_iath.html**

Irvine D. 2015. From angel to agile: the business of the digital humanities. *Schol. Res. Commun.* 6(4). **https://doi.org/10.22230/src.2015v6n4a208**

Jockers ML. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: Univ. Ill. Press

Jones SE. 2016. *Roberto Busa, S.J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York: Taylor and Francis

Keller S, Korkmaz G, Orr M, Schroeder A, Shipp S. 2017. The evolution of data quality: understanding the transdisciplinary origins of data quality concepts and approaches. *Annu. Rev. Stat. Appl.* 4:85–108

Klein LF, Gold MK. 2016. Digital humanities: the expanded field. In *Debates in the Digital Humanities 2016*, ed. MK Gold, LF Klein, p. ix. Minneapolis, MN: Univ. Minn. Press

Lamont M. 2010. *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard Univ. Press

Lancashire I. 2005. Computers in the linguistic humanities: an overview. In *Encyclopedia of Language and Linguistics*, ed. K Brown, pp. 789–809. Amsterdam: Elsevier. 2nd ed.

Lancashire I. 2010. *Forgetful Muses: Reading the Author in the Text*. Toronto: Univ. Toronto Press

Lancashire I, Hirst G. 2009. *Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: a case study*. Presented at the 19th Annu. Rotman Res. Inst. Conf., Cognitive Aging: Research and Practice, March 8–10, 2009, Toronto. **ftp://ftp.cs.toronto.edu/pub/gh/Lancashire+Hirst-extabs-2009.pdf**

Leroi AM. 2015. One republic of learning: digitizing the humanities. *New York Times*, Feb. 13. **https://www.nytimes.com/2015/02/14/opinion/digitizing-the-humanities.html?_r=0**

Li J. 2016. Exploring the logic and landscape of the knowledge system: multilevel structures, each multiscaled with complexity at the mesoscale. *Engineering* 2:276–85

Liu K, Meng X-L. 2016. There is individualized treatment. Why not individualized inference? *Annu. Rev. Stat. Appl.* 3:79–111

Luhn HP. 1960. Key word-in-context index for technical literature (KWIC index). *Am. Doc.* 11:288–95

Lusignan S, North JS. 1977. *Computing in the Humanities*. Waterloo, Can.: Waterloo Univ. Press
McCusker JJ. 1969. Book review: *Enterprise & Empire: Merchant and Gentry Investments in the Expansion of England, 1575–1630*. *Hist. Meth. Newsl.* 2(3):14–18
Michel J-B, Shen YK, Aiden AV, Veres A, Gray MK, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331:176–82
Miller JH, Page SE. 2007. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton, NJ: Princeton Univ. Press
Milligan I. 2013. Illusionary order: online databases, optical character recognition, and Canadian history, 1997–2010. *Can. Hist. Rev.* 94(4):540–69. **https://doi.org/10.3138/chr.694**
Mitchell JL. 1974. *Computers in the Humanities*. Minneapolis: Univ. Minn. Press
Munro H. 1964. The impacts of culture, technology on each other: research project with awesome implications. *Globe Mail*, May 7, p. 2
Novak P. 1988. *That Noble Dream: The "Objectivity" Question and the American Historical Profession*. Cambridge, UK: Cambridge Univ. Press
Nyhan J, Welsh A. 2013. Uncovering the "hidden histories" of computing in the humanities 1949–1980: findings and reflections on the pilot project. *Digital Humanities 2013: Conference Abstracts*, pp. 326–29. Univ. Neb., Lincoln, July 16–19. **http://dh2013.unl.edu**
Perron P, Sbrocchi LG, Colilli P, Danesi M, eds. 2000. *Semiotics as a Bridge Between the Humanities and the Sciences*. Toronto: LEGAS
Porter TM. 1986. *The Rise of Statistical Thinking, 1820–1900*. Princeton, NJ: Princeton Univ. Press
Priego E. 2011. "Father Roberto Busa: one academic's impact on HE and my career." *Higher Education Network Blog, The Guardian*, Aug. 12. **https://www.theguardian.com/higher-education-network/blog/2011/aug/12/father-roberto-busa-academic-impact**
Roberts E, Woollard M, Ronnander C, Dillon LY, Thorvaldsen G. 2003. Occupational classification in the North Atlantic population project. *Hist. Meth.* 36:89–96
Rothberg M. 2010. Quantifying culture? A response to Eric Slauter. *Am. Lit. Hist.* 22(2):319–24. **https://academic.oup.com/alh/article-abstract/22/2/341/172183?redirectedFrom=fulltext**
Sawyer RK. 2005. *Social Emergence: Societies as Complex Systems*. Cambridge, UK: Cambridge Univ. Press
Schreibman S, Siemens R, Unsworth J. 2004. *A Companion to Digital Humanities*. Oxford, UK: Blackwell
Siemens L. 2013. Developing academic capacity in digital humanities: thoughts from the Canadian community. *Digit. Humanit. Q.* 7(1). **https://dspace.library.uvic.ca/bitstream/handle/1828/8201/Siemens_Lynne_DHQ_2013.pdf?sequence=1**
Siemens R, Moorman D, eds. 2006. *Mind Technologies: Humanities Computing and the Canadian Academic Community*. Calgary, Can.: Univ. Calgary Press
Sinclair S, Rockwell G. 2014. Towards an archaeology of text analysis tools. *Digital Humanities 2014: Conference Abstracts*, pp. 359–60. École Polytechnique Fédérale de Lausanne and Univ. de Lausanne, Switz., July 7–12. **http://dharchive.org/paper/DH2014/Paper-778.xml**
Smith PH. 1984. Statistics, epistemology and history. *Hist. Meth.* 17:3
Social Sci. Humanit. Res. Counc. (SSHRC). 2009. *Image, text, sound and technology*. Summer 2009 competition. **http://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/itst/workshops-ateliers-eng.aspx**
Stat. Can. 2009. *Statistics Canada Quality Guidelines*. Ottawa, Can.: Stat. Can. 5th ed.
Swierenga RP. 1970. Clio and computers: a survey of computerized research in history. *Comput. Humanit.* 5(1):1–21
Tasman P. 1957. Literary data processing. *IBM J. Res. Dev.* 1(3):249–56
Thomas WG. 2004. Computing and the historical imagination. In *A Companion to Digital Humanities*, ed. S Schreibman, R Siemens, J Unsworth, pp. 56–68. Oxford, UK: Blackwell
Va. Cent. Digit. Hist. 2017. History and highlights. *Outreach*. Univ. Va., Charlottesville, VA. **http://www.vcdh.virginia.edu/index.php?page=Outreach**
Wasserstein RL, Lazar NA. 2016. The ASA's statement on *p*-values: context, process, and purpose. *Amer. Stat.* 70(2):129–33
Williford C, Henry C. 2012. *One Culture. Computationally Intensive Research in the Humanities and Social Sciences.* Washington, DC: Counc. Libr. Info. Resour.

Winchester I. 1970. The linkage of historical records by man and computer: techniques and problems. *J. Interdiscip. Hist.* 1:107–24

Winter TN. 1999. *Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance*. Faculty Publications, Classics and Religious Studies Department. Paper 70. **http://digitalcommons.unl.edu/classicsfacpub/70**

Wisbey RE, ed. 1971. *The Computer in Literary and Linguistic Research*. Cambridge, UK: Cambridge Univ. Press

Zemsky RM. 1969. Numbers and history: the dilemma of measurement. *Comput. Humanit.* 4:31–40