



Scholarly Networks and Collaborative Practices

Proceedings of the Colloquium held at Brown University on April 16-18, 2015

Massimo Riva

Abstract

The sections “Perspectives” and “Interventions” of this issue of *Humanist Studies & the Digital Age* present a selection from the proceedings of a colloquium held at Brown University in the spring of 2015. The event was hosted by the Virtual Humanities Lab in the Department of Italian Studies (VHL), in collaboration with the Center for Digital Scholarship in the Brown University Library (CDS), and DARIAH-Italy (Digital Research Infrastructure for the Arts and the Humanities). Its aim was to explore the new types of scholarly output produced when scholars use digital methods to collaborate on, annotate and visualize traditional materials.

Introduction

In the age of data mining, distant reading, topic modeling and cultural analytics, scholars and researchers in the humanities increasingly rely upon automated, algorithm-based procedures in order to parse their exponentially growing databases of digitized textual and visual resources. Yet, within this deeply networked and massively interactive environment, it is crucial to preserve the expert logic of primary and secondary sources, textual stability, citations, and other apparatus, which form the heritage and legacy of humanities scholarship. Digital editions, for example, must now live in the networked environment built within digital library repositories: emerging curatorial and editorial practices and the semantic act of interpretation are increasingly embedded together into the primary sources and such practices are also the conduit for training the next generation of digital humanists. In short, scholarly collaboration must problematize methodology, tools and interpretation at the same time.

Humanities researchers increasingly collaborate, in a laboratory mode, on shared platforms and in shared virtual environments, experimenting with open source tools often developed elsewhere, in the annotation and visualization of select corpora of primary sources. In the process, they produce new and yet unidentified typologies of scholarly objects (thoroughly embedded in library repositories) that incorporate curatorial and interpretive practices along with a new, and fully documented, technical instrumentation. This hybrid form of collaborative curation/publication is at the foundation of humanities scholarship in the digital age. In the Brown Colloquium, scholars from the U.S., the U.K.,

and Italy shared their ideas, diverse experience, and work-in-progress related to one or more of these aspects: new data-driven collaborative practices; the intersection of traditional scholarly methods and emerging computerized techniques; the new research infrastructure (digital library repositories and tools); the institutional framework within which computer-aided practices in the humanities may thrive.

The topics addressed in the colloquium also reflect the evolution of the collaborative activities conducted within the framework of the VHL since its creation in 2004 (thanks to a two-year grant from the National Endowment for the Humanities) by contributing scholars located in the U.S., Europe (Italy and the U.K.), Latin America (Mexico and Argentina), and Australia. In its original configuration, the VHL included a platform for the collaborative online editing, annotating and publishing of a variety of classic texts in Italian and Latin. It aimed to create an online environment in which scholars could train themselves in the new practices while attending at their editorial and scholarly work. More recently, in collaboration with the CDS at the Brown Library, the VHL has also embraced projects focused on the development of enriched digital library collections and archives, such as the Garibaldi Panorama & the Risorgimento Archive, and the Theater that Was Rome, both based on special collections of the Brown University library. These projects, and others under consideration, are particularly focused on visualization techniques and tools, including geospatial resources.

In his key-note address at the colloquium, entitled “The HathiTrust Research Center: Bringing you 4.7 billion pages of analytic opportunities!,” J. Stephen Downie, associate dean for research and professor at the School of Information Sciences of the University of Illinois, and the Illinois codirector of the HathiTrust Research Center, described the unprecedented opportunities offered to scholars across disciplinary boundaries by advanced digital library infrastructures, such as that built by the Hathi Trust consortium. This not only provides access to an ever-growing corpus of textual data but also creates an environment in which scholars can produce, store and share their annotated datasets, according to the topics of their research, thus contributing to generate what we may call collaborative curatorial practices that go beyond data mining into the realm of data sharing and perhaps also new forms of publication. This is entirely consistent with the original inspiration of the VHL, although clearly realized on a much larger scale, which poses the question: how can relatively small-scale projects such as the VHL take advantage of large-scale infrastructures such as Hathi Trust? Indeed, another aspect of the networked research environment in which scholars are increasingly able to collect, share and publish their data, along with their own analyses and conclusions based on those data, is the transformation of traditional forms of humanities publication, as we shall see also further on. These can be spearheaded by a synergy between large infrastructural digital library resources and highly focused, project-centered research “environments” which, like the VHL, aim to highlight special library collections, in what could be defined a “curatorial” mode.

With the dramatic increase in scope of digital resources, data-driven projects are a potentially new area of development for the VHL: in organizing this colloquium, we were particularly interested in exploring new collaborative possibilities in the area of text analysis and “distant reading,” including such emerging methodologies as topic modeling, text clustering, sentiment analysis, network analysis, but also geolocate and geospatial visualization techniques (the use of data-generated maps to parse a textual and visual corpus,

for example, as in a project recently developed within the framework of the Garibaldi & the Risorgimento digital archive, The Garibaldi Resource Explorer: <http://library.brown.edu/cds/garibaldi/resources/fullexplorer.php>).

A critical overview of these emerging methodologies was presented at the colloquium by Fabio Ciotti, of DARIAH, Italy (his paper could not be included in these proceedings having been already accepted for publication elsewhere). The methodologies listed above are central to the debate about data mining in, and for, the humanities and the related discussions about the “ontological statute” of humanistic data. In what is perhaps the most technical contribution to these proceedings, Dino Buzzetti (emeritus professor of the University of Bologna and a founding member of the Brown-Bologna Pico della Mirandola project) and Ernesto Priani (a researcher at the Universidad Nacional Autonoma de Mexico, and also a member of the Pico project) address these issues from a specific point of view: namely, how an annotated digital edition of Pico della Mirandola’s *Conclusiones Nongentae* (one of the ongoing projects in the VHL) could benefit from the integration of traditional editorial and publishing practices with the newly emerging computational methods.

A digital edition of, and commentary on, Pico’s 900 theses (as the *Conclusiones* are also known), Pico’s own critical compilation and theoretical “parsing” of hundreds of sources from all philosophical and theological textual traditions known to him, would indeed greatly benefit from the possibility of eliciting the help of automated techniques in order to reconstruct the intricate intertextual web of quotations and references often only implicitly embedded, or alluded to, in Pico’s text. Buzzetti and Priani discuss the merits and limitations of the new text analysis and data mining methods “introduced to inspect word contexts and vocabularies, through the classification of texts, the development of ontologies, and the identification of lexical similarities.” Their preliminary conclusion – “Intertextuality becomes then the theoretical foundation of further digital implementations” - is also the point of departure of their proposal of a topic modeling approach to the next phase of the Pico editorial and publishing project. In short, “finding the sources of Pico’s theses in an author’s corpus is a matter of context and intertextuality,” they write. From this point of view, they consider vector space modeling - a method to capture the relative importance (or weight) of a term in a document - promising. Equally worth consideration is, in their opinion, the approach of The Concept Lab, a research project at the University of Cambridge led by Peter de Bolla, because of its distinction between word senses and concepts: this is “particularly relevant in relation to Pico’s theses, since his vocabulary, depending as it does on the need to demonstrate the genuine concordance of all philosophical positions, is not necessarily the same as that of the authors they are ascribed to.”

Buzzetti and Priani then go on to describe an experiment conducted in collaboration with the Grupo de Ingeniería Lingüística of the Instituto de Ingeniería of the UNAM, in which an attempt is made to locate Pico’s “sixteen conclusions according to Albert the Great” in the latter’s corpus. The methods and tools thus tested (the Cosine Similarity method, the MALLET, and the word2vec packages) suggest several interesting analytical steps. Without delving into the technicalities addressed in the article, the benefit of a topic modeling approach consists for the authors in its integrative rather than alternative potential, in relation to more traditional scholarly methods: “Distant and close reading can then meet and in defiance of the common opinion they can effectively show their complementarity.”

This is particularly promising for a machine-aided annotated edition of Pico's work: "a controlled annotation language, obtained out of topic modeling and co-occurrence results, would possibly enable an intermediary course of action between manual and automated procedures, consisting for instance in a formalised annotation practice." The last section of the article takes into consideration another set of techniques and evaluates their potential use, namely how machine learning and artificial intelligence (AI) can contribute to process annotation data through the application of probabilistic soft logic (PSL). Again, skipping the most technical (and analytically useful) part and jumping to their conclusions, the authors' analysis shows "that both close and distant reading can play similar roles in respect of quantitative as opposed to qualitative, or factual as opposed to conceptual approaches, and that they can both be regarded as complementary practices."

A similar "critical" and experimental approach to emerging techniques and methodologies informs the contribution by Michael Papio, "Geospatial Visualizations for the Study of Boccaccio." Papio asks a very general question: what can mapmaking do to help us better understand the life and work of a medieval author such as Boccaccio? This question relates both to Papio's own engagement with Boccaccio's work as a "geographer" and "topographer" in *De Montibus*, and is not unrelated to pedagogical practices recently developed within the framework of the oldest among the VHL projects, the Decameron Web, that Papio contributed to create, back in the late 1990s. (These practices include collaborative class exercises in the mapping of the Decameron text according to geolative coordinates such as that designed by Nicole Gercke: the places mentioned, or alluded to, in the text, are located onto a historical map, the fifteenth century Catalan Atlas, layered over Google maps). In short, Papio sets out to test the idea that "geospatial representations of Boccaccio's world – a few pages taken from an unfinished Boccaccio atlas, so to speak – may... provide insights into additional things that had escaped our notice."

Simple visualization tools such a cluster map linked to word frequency may show what we already know: for instance, that "the *Decameron* is tightly moored to those areas its author knew from actual experience," or that "the *Decameron* is substantially Tuscan (which will come as a shock to no one)." Yet, the more "granular" our data parsing is, the more interesting are the potential analytical results: "in a collection of tales that seem to have been kept generally close to geographical areas known personally by their author, never is there mention of that handful of city blocks he knew as his own," that is the Oltrarno neighborhood around 1352. (Cross-referencing these data with other data, such as for example those retrievable from the Catasto and Tratte databases, also part of the VHL, could enhance even further our picture of Boccaccio's Florence). Referring to Bertrand Westphal (one of the driving forces of geocriticism) and his idea of the "unstable hierarchy of places" in the Middle Ages, Papio provides very useful considerations about the best way to take critical advantage of geographical visualizations: "The idea that geographical locations were unstable may seem to us, who have Google maps in our phones and a GPS in our cars, to be a curious concept." For a correct adjustment, we need first of all to be aware of the state of medieval cartography. (Layering a historical map such as the Catalan Atlas or medieval *portolani*, for example, over a geopositioned Google map can also provide an interesting measure of the distortion and instability of *both* systems). "The instability of geolocations across time and space is an especially productive notion for Boccaccio Studies," Papio writes, forcing us to look at the way "geopositioning" worked, in the late Middle Ages, based on textual traditions rather than computer-generated numeric coordinates; and making

us reflect on “the benefits of using maps to reconstruct a literary reality, even one distorted by time, that existed hundreds of years earlier.” A geocritical exercise thus requires a certain degree of historical imagination. In short, as the rest of Papio’s essay illustrate, it is precisely this multidimensionality that makes geocritical imagination tick: plotting or mapping Boccaccio’s world not by arbitrarily superimposing our own geographical representations on it, but by retrieving the depth of his own unstable, literary “geolocative” system. “Up to now -Papio writes - preference has been given to how geospatial approaches can open doors in literary studies; the reverse is of course just as true.” In conclusion, Papio suggests, “intelligent maps based on thoughtful criteria can lead to new areas of research, or at the very least can provide geospatial paratexts that will be useful to research that is already in full swing.” Again, as in the case of the Pico project, it is the cross-fertilization of traditional humanistic scholarship and computational visualization techniques that seems to produce the best results.

The other two contributions from the Brown colloquium presented here address the crucial question posed by the institutional framework of the new research infrastructure, from two different points of view: Guyda Armstrong and Marilyn Deegan summarize in their essay the results of a two-year investigation, sponsored by the Arts and Humanities Research Council in the U.K., about emerging forms of digital publication in the humanities, focusing in particular on the “digital monograph”; John Cayley (a professor of Literary Arts at Brown University, and an award-winning author of electronic literature) reflects on the way digital repositories could, and should, become the springboard for open, “creative” collaborative practices in the arts and the humanities. Let’s begin with the latter contribution, which appears here in a video format in the section “Interventions” of this issue of the journal *Digital Humanities & the Digital Age*. The long subtitle of Cayley’s presentation sums up the fundamental question he addresses: how can we persuade universities to own their responsibilities to the practice-based research that they patronize – while bringing new, fully-accredited methodologies and infrastructures to Arts and Humanities scholarship? The “we” in this title-sentence refers to scholars interested not only in the opportunities offered by the digital infrastructure being built but willing to actively contribute to the designing of this infrastructure. Cayley’s perspective is that of a “practitioner of digital literary arts,” interested in particular in the linguistic dimension of digital artistic practices. However, his argument applies also to humanities research more broadly conceived. With the shift toward the digital infrastructure, and the confusion which still surrounds definitions of the digital humanities, it seems that our attention as textual scholars has turned from the text to the indexing of textual resources. Massive (and according to Cayley “catastrophic”) restructuring of the publishing industry and unprecedented access to digitally represented (big) data, are the two main events we face, as scholars-practitioners in both the arts and humanities. The (academic) book is dead, according to Cayley, “students don’t read books anymore,” and the codex exists only as an archive (of individual research). Yet nothing has yet replaced the “book” as the corner stone of careers and, we may add, the main tool for sharing the results of research in the humanities (and perhaps also the non-digital “literary arts”).

The sweeping nature of these statements is intentionally provocative: his contribution, Cayley says, is partially a “manifesto.” His point of departure, certainly shared among those attending the colloquium is that scholars have to take active part in the “indexing” enterprise (not leaving it to Google & Co.) and universities must lead this effort, or better, enable scholars to lead this effort, in a participative and collaborative way. Yet, this

enterprise is more than “indexing” knowledge. As the digital data produced in research is feeding, or, in jargon, is being increasingly “ingested” into, universities’ repositories, scholars should be able to rely upon these repositories (rather than the “cloud,” for example) for archiving their work-in-progress: the datasets that are the equivalent of “field-notes,” drafts, compilations of annotated materials of various nature that they use in order to produce outcomes and “results” (still mostly fashioned or packaged as articles, books, etc.). This would not only guarantee against obsolescence (and dependence on market services) but would potentially change the nature of what “publication” is going to be in the future: new forms of repository-based publication will necessarily include the data collected and curated in the process of research, providing an account of both the “how” (the methodology) and “what” (the data) of a specific research. Institutional policies and practices should enable and encourage this process.

The future of the academic book in this changing landscape is the object of the final contribution presented here from the Brown colloquium. As the authors write: “Declining monograph sales, rising serials prices, funding problems, rapidly-changing new technologies, shifting policy landscapes, increasing pressure on academics to do more with less, all contribute to a sense of unease about the health of the academic book in the arts and humanities, and indeed about the health of the disciplines themselves.” In sketching out the results of the AHRC-sponsored survey, mostly centered on the U.K. but extending its reach also to North America, Armstrong and Deegan provide a more varied scenario than that envisioned by Cayley: “new developments...may point to diverse futures for different kinds of books – they write. Some of these are infrastructural and hold out promise of sustainable models; others are individual and experimental, and may point to some new and interesting possibilities. We need both, but we must bear in mind that some of the new models will not stand the test of time.” Proclaiming once again the “end of books,” as a famous *New York Times* article by the writer Robert Coover did already back in 1992, may be, again, premature (Coover, the founder of the program in digital literary arts at Brown left his position to Cayley). Yet, the cautionary character of the *Academic Book of the Future* (ABFR) report – “we have observed an enduring attachment to print that is neither sentimental nor habitual: print just happens to do some things particularly well, and will no doubt continue to do so” – may be also conditioned in turn by its specific point of view: the future of the *academic* book (Coover’s article mostly referred to the future of storytelling). This said, it should be added that while Cayley’s position is also a somewhat “militant” one, a manifesto in favor of an innovative approach to publishing methods and knowledge sharing in the humanities, the ABFR provides instead a snapshot of the current state of affairs, including the results of inquiries bridging the “old” and the “new.” Armstrong and Deegan for example cite “the Robb (2015) review of *Words Onscreen: The Fate of Reading in a Digital World* by Naomi Baron” which “in a survey of over 300 university students in the U.S., Japan, Germany, and Slovakia...found a near-universal preference for print, especially for serious reading” (which seems to straightforwardly contradict what Cayley stated in his presentation). Yet, these results should be probably adjusted to the relative percentage of time that the students surveyed actually spend reading books and reading online (two different kinds of reading, as we assume).

The snapshot of the current state of affairs is a very useful one, however, and the necessary point of departure for any attempt to predict, or envision, a possible future. The article touches upon many aspects of the problem, including: the technological side (or

platforms) available and under experimentation; the (broad) typology of academic genres which may or may not have a braver future in the new world of knowledge sharing in the humanities (digital editions, or e-journals, for example); the effect of emerging digital rhetoric on the printed page (the output may still be print but the process of researching and writing is increasingly digital); the issues of obsolescence and preservation, touched above and also Cayley's most urgent concern as an artist (the more innovative and experimental the emerging forms are, the greater the risk of obsolescence seems to be); the nature of the "enhanced monograph," the long- and single-author-form which is the cornerstone of careers in the humanities, and still the main carrier of symbolic value (with the imprint of a prestigious publisher), in the age of "open access"; and, last but not least, the economic sustainability of it all (the various private-public partnership models which make both the experimentation with new forms, and their establishment as models and standards, possible).

Many of these issues can be usefully considered from different points of view as well: that of the individual scholar, that of the scholarly community, that of a specific discipline, that of a specific institution (private and public universities, foundations, government agencies), that of publishers, even that of the "general audience," the humanities community at large etc. The complexity of these points of view, combined in what is sometimes referred to as the new "eco-system" of knowledge producing and sharing in the humanities, allows us to grasp the complexity involved in envisioning the role which "the academic book of the future," or what will eventually, possibly, replace the book, may have in it (including the Ph.D. dissertation which of the academic book is often the incunabulum).

The article goes on to review a number of current projects, of various nature and scale, which can provide at least a glimpse at the richness and variety of the experimentations in progress with both born-digital formats and/or the translation or hybridization of traditional scholarly outputs into the digital environment. Being myself currently involved in one of the many experiments with the "digital monograph," supported by the Mellon foundation in partnership with a number of North American academic presses and university libraries, I can see both the pros and cons of this enterprise to redesign the "book" as one of many "unidentified scholarly objects" on the horizon: this is not the venue to share my own insights based on this experience. Yet, I can perhaps conclude by saying that one of the fundamental issues for us today remains how to balance preservation and innovation: I mean "preservation," in both a specific sense and in the broader sense of the transmission of scholarly forms which have proven their value in the long past, including the printed book; and innovation, in both a specific sense, and in the broader sense of the comprehensive transformation of traditional scholarly forms and the emerging of new ones more adapted to thrive in the new knowledge ecology.

It could sometimes feel like a balancing act, and we can be tempted, from time to time, to tip that balance in one sense or the other, to abandon the "old" or resist the "new." In fact, what the Brown colloquium has confirmed is that the most productive attitude is an open, critical, pragmatic, and experimental one which sees "traditional" and "new" forms as cross-fertilizing and reshaping each other in a synergetic way. This has been the inspiration of the Virtual Humanities Lab, since its creation.