How to do digital philosophy of science

Charles H. Pence
Department of Philosophy and Religious Studies
Louisiana State University
Baton Rouge, LA, USA
charles@charlespence.net
https://charlespence.net

Grant Ramsey
Institute of Philosophy
KU Leuven
Leuven, Belgium
grant@theramseylab.org
http://www.theramseylab.org

Abstract

Philosophy of science is beginning to be expanded via the introduction of new digital resources—both data and tools for its analysis. The data comprise digitized published books and journal articles, as well as heretofore unpublished and recently digitized material, such as images, archival text, notebooks, meeting notes, and programs. This growing bounty of data would be of little use, however, without quality tools with which to analyze it. Fortunately, the growth in available data is matched by the extensive development of automated analysis tools. For the beginner, this wide variety of data sources and tools can be overwhelming. In this essay, we survey the state of digital work in the philosophy of science, showing what kinds of questions can be answered and how one can go about answering them.

1. Introduction. Our understanding of science is being broadened by the digitization and automated analysis of the various outputs of the scientific process, such as scientific literature, archival data, and networks of collaboration and correspondence. These technological changes are laying the foundation for new types of problems and solutions in the philosophy of science. The purpose of this article is to provide an overview and guide to some of the novel capabilities of digital philosophy of science.

To best understand the reasons why digital philosophy of science lets us ask a new class of questions, let's consider how it differs from more traditional approaches. For example, consider how we might draw conclusions about articles in the journal *Nature.* It has published over 360,000 articles since its founding in 1869, meaning that one would have to read ten articles a day for one hundred years to work through the complete archives of this journal alone. Of course, the standard response in the philosophy of science is to favor depth over breadth, and closely read a much smaller number of articles. While there is certainly much we can learn about science in this way, some broad questions about the nature and history of science—questions, for example, about how theories arise and become established in the literature as a whole—would remain unanswerable without a way to glean information from hundreds of thousands or even millions of journal articles. Much the same argument holds for scientific images, or information about the collaboration, communication, training, or citation connections between researchers.

The question, then, is to what degree we can learn from the vast scientific literature without having to read every article closely—to instead do what is called *distant reading* (Moretti 2013). With distant reading, we input a large body of literature into a computer, and use it to do the "reading" for us, extracting large-scale patterns that would be invisible or impractical to find otherwise. In the philosophy of science in particular, this process has been aided by a

number of large digitization efforts targeted at the outputs of the scientific process. One crowning achievement of these efforts is the nearly complete digitization of the academic journal literature. This content is thus now accessible in ways that it has never been before.

Digital approaches to the philosophy of science contrast with traditional methods involving *close reading*—intently reading a narrow body of literature within a focal area. With close reading, a philosopher will have an impressive command over a limited domain. He or she closely reads a select set of documents from the scientific literature, or analyzes the experimental, training, or collaborative records of a small group of researchers to attempt to extract the structure of a scientific theory, or to understand the meaning of its terms.

We should stress that the close reading-based traditional philosophy of science and distant reading-based digital philosophy of science are not in competition. Instead, they are complementary. If, for example, a researcher wants to know how the meaning of a particular term has changed over time, he or she could use automated textual analysis tools to locate instances of the term, find hot spots in which the term is used frequently, quickly see which words it is associated with, and how these word associations have changed over time. In conjunction with digital analysis, performing close reading of key texts will be invaluable. The close reading may then spur further digital inquiries, and so on. Thus, traditional and digital philosophy of science work in tandem, each supporting the other.

The remainder of this article will canvass a number of significant issues that must be dealt with in order to develop a digital philosophy of science research program. We hope that this overview will be helpful to researchers who are interested in moving forward with digital tools but are not certain where or how to begin.

2. Getting started. Because digital philosophy of science is a relatively new field, not only is there no set of standard tools, it is often unclear what sorts of questions can be answered by the extant tools. Thus, let's begin by considering some of the new kinds of questions one can address.

One of the most significant advantages of distant reading comes from the ability to engage with corpora significantly larger than those usually treated by philosophers and historians of science. For example, Murdock, Allen, and DeDeo (2017) were able to analyze large-scale patterns in Darwin's reading by accessing the full text of every book that we know him to have read over a period of decades. These kinds of analyses simply would not be possible without the aid of technology. Answering research questions that leverage broad (yet still circumscribed; see section 4) sets of data are thus likely to be a fruitful use of digital tools. For example, one could track concepts over the entire print run of a journal, the collections of books published in the Biodiversity Heritage Library (Gwinn and Rinaldo 2009), or the PubMed Open Access Subset of contemporary biomedical journal articles (Roberts 2001). These kinds of investigations allow us to explore the conceptual landscape of a field through distant reading, by offering (at least in some cases) an exhaustive analysis of an area.

Another advantage comes from the ability of analytical algorithms to parse texts in ways that even well trained close readers cannot. For example, fine-grained patterns of language usage, such as the shift in a term from a noun use to a verb use, or a shift from referring to science as a one-person activity to a group activity, could be traced in the literature with a level of exhaustiveness, objectivity, and care that would simply be impossible for a single reader. Automated tools can analyze sentence structure, word order, or parts-of-speech usage in a way that would try the patience of any scholar (Manning et al. 2014).

The ability of digital tools to increase the breadth of a research question is also important. If one has a hypothesis drawn from a particular domain (maximization or optimality inferences in biology, for example), this hypothesis could be tested in other, separate domains (economics, psychology, sociology) with only a modest further investment of resources.

While digital tools can aid in answering existing research questions, these tools also open the possibility of framing new questions without a clear analogue in the pre-digital world. For instance, work by Manfred Laubichler and colleagues applies dynamic network analysis to our understanding of scientific conceptual development (Miller et al. 2015). The questions they ask arise in conjunction with the digital tools, and in dialogue with digital humanities researchers in other disciplines.

3. Choosing the right tools. Now that we have a sense of the advantages of digital analysis, let's consider the currently available tools and corpora of data relevant to the philosophy of science.

To begin, we should draw attention to the central repository of digital humanities tools, known as the DiRT Directory, accessible at <https://dirtdirectory.org/> (for more on its construction and predecessors, see Dombrowski 2014). There are nearly as many digital humanities tools as there are digital humanities researchers, and the landscape of contemporary software changes rapidly. For nearly any kind of analysis, the directory will include some tool which performs it—the most important question will be whether the data available can efficiently be converted into the format required by that tool.

3.1. Basic tools. There are a number of tools that may be used immediately by researchers, as they do not require that one collate a set of documents of interest in advance.

Perhaps the most famous of these is the Google Ngram corpus (Brants and Franz 2006; Michel et al. 2011), accessible at <https://books.google.com/ngrams>. This corpus contains the entirety of the scanned Google Books project, current as of 2012, with frequency data for single words as well as pairs and longer sequences (so called bigrams, trigrams, and, more generally, n-grams).

Obviously, the Ngrams project does not exclusively contain scientific or philosophical content, and hence a number of queries that might interest philosophers of science will simply not be meaningful when queried against the Ngram Viewer. For example, the scientific usage of the term "evolution" will be completely masked by the broader cultural use of the term, and hence philosophers interested in the use of this term are unlikely to be able to uncover interesting data. There are also a number of worries about the statistical representativeness of the Google Ngram corpus, even when judged as a measure of broader cultural usage or popularity (Morse-Gagné 2013; Pechenick, Danforth, and Dodds 2015).

Much more precise search and analysis may be performed by using JSTOR's Data for Research project (Burns et al. 2009), available at <http://dfr.jstor.org/>. This tool allows users to perform searches and analyses against the entire corpus of JSTOR journals. Researchers may search for articles by journal, publication date, author, subject, and more, allowing for careful control over the set of articles to be analyzed. These articles may then be queried for word frequencies (and ngram frequencies), as well as automatically extracted "key terms," which are words common in the selected articles but uncommon in the corpus as a whole (computed using the *tf-idf* score). The frequency scores from JSTOR DFR may also be used as an input to a variety of the tools described below.

3.2. Gathering a corpus. The more advanced tools are set apart primarily by not coming with a pre-loaded corpus of material to study. This means that the challenge of obtaining data falls to individual researchers. As mentioned above, we find ourselves in a particularly fertile period for data availability in the philosophy of science. Much of the journal literature, in some cases back into the nineteenth century, is available online in PDF or HTML form. Comprehensive online projects are available that focus on the works, life, and correspondence of figures like Darwin (Secord 1974; van Whye 2002), Newton (Iliffe and Mandelbrote 1998), Poincaré (Walter, Nabonnand, and Rollet 2002), Einstein (Mendelsson 2003), and others (Pouyllau et al. 2005; Beccaloni 2008; Mills 2011). A number of discipline-specific archives have also been constructed, such as the Embryo Project Encyclopedia, an open access, digital repository covering the history of embryology and developmental biology (Maienschein et al. 2007). To this may be added the digital collections now increasingly available from a wide variety of museums and libraries. With an appropriate collection of data obtained for a researcher's private use, it becomes possible to leverage a much wider variety of analytical tools. (These data must also be carefully curated and safely preserved; we will return to the question of data management in the next section.)

A researcher gathering a corpus must consider how and to what extent the data should be annotated. Minimal annotation—for example, leaving content as plain text with only bibliographic data for tracking—allows for the rapid creation of a large corpus, and lowers the future burden of maintaining and updating the annotations. But more significant annotation—such as marking up textual data in a format like that described by the Text Encoding Initiative (Ide and Véronis 1995)—allows for more complex, fine-grained, and accurate analyses. This annotation can take a variety of forms. For textual data, TEI allows users to indicate the locations

of various parts of the document (pages, paragraphs, chapters, indexes, figures, or tables), or the various kinds of references made by a piece of text (dates, citations, abbreviations, names of persons or institutions, etc.).

This process of cross-referencing documents may be aided by the use of external ontologies—in the sense (not the one common in philosophy) of collections of standardized verbs and concepts that allow for the same term to refer unambiguously across multiple documents. In philosophy, the Indiana Philosophy Ontology project, or InPhO (Buckner, Niepert, and Allen 2007), available at <https://inpho.cogs.indiana.edu/>, allows standardized reference to concepts such as "sociobiology," or to particular philosophers. A number of such ontologies also appear in other areas of the sciences, and a document may be marked up with multiple ontologies to add further semantic richness.

With a heavily annotated document, significantly more complex analysis may be applied, as the computer now "knows" where particular concepts are mentioned, how they are used, and how they relate to other ideas. While the use of such methods is relatively untested in philosophy, the biomedical field has made significant strides in this direction in recent years—for example, analysis of the usage of gene and chemical concepts in the scientific literature has actually enabled the extraction of novel relationships (previously unpublished by researchers, but discernible from the body of literature as a whole), and even the generation of novel hypotheses about future drug development (A. M. Cohen and Hersh 2005).

The question of the representativeness of one's sample of data is also a significant one with which researchers must engage. As we noted above, even in the largest corpora, such as Google's Ngram collection, there are still problems with the statistical significance of the sample (Morse-Gagné 2013), with biases in temporal availability of data (more data tends to be available

closer to the present, as the relevant outputs were "born digital"; Michel et al. 2011) and systematic sources of error such as that introduced by optical character recognition (Hoover 2012). These concerns are somewhat alleviated when using a curated corpus known to be complete (such as databases of historical correspondence), but even in these instances, researchers must remain constantly vigilant against statistical bias.

3.3 Advanced tools. With a corpus in place, there is a variety of options for users interested in performing analyses impossible with the basic tools described above.

First, there are a number of tools designed to aid researchers in presenting their material as an easily navigable, searchable, categorized public resource—a public digital archive or museum. The most popular of these is Omeka (D. Cohen 2008), available at <https://omeka.org/>. Omeka is a free, open-source software product that allows users to construct online archives and museum exhibitions, to add catalog information and metadata to digital items, and to attractively present all of this material to the public at large. Deploying a website such as this is a nice way to garner some immediate, public-facing payoff from the difficult work of obtaining and curating a digital collection.

One alluring feature of large digital data sets is the possibility of analyzing the networks found within them—whether these are networks of collaboration drawn from experimental archives or lab notebooks, networks of correspondence drawn from digitized letters, or citation networks extracted from the journal literature. Such network analysis can often allow us to see patterns in the overall structure of a field that would be otherwise difficult to discern. One of the most user-friendly network analysis tools available is Gephi (Bastian, Heymann, and Jacomy 2009), available at <https://gephi.org/>. Gephi allows users to import graphs in a number of

formats (including ones as simple as CSV spreadsheet data), and to perform a variety of analyses and visualizations. The network may be broken into clusters (using a standard measure known as modularity; Blondel et al. 2008), the degree of connectivity of individual nodes may be easily explored, and the results can then be rendered graphically for presentation.

If the data to be analyzed is text, a popular choice is Voyant Tools (Sinclair and Rockwell 2016), available at <http://voyant-tools.org/>. Once a corpus of text is uploaded to Voyant, the user is immediately presented with a wide variety of options: a word cloud, a cross-corpus reader, a tool for tracking word trends through the text, and a short snippet concordance are among the immediately available tools, and a variety of other, more complex analyses and attractive visualizations may be performed using plugins. Voyant may also be used to save online corpora for future use, which facilitates classroom usage of textual analysis.

Another challenging problem likely to be faced by philosophers of science interested in the scientific literature is the analysis of a large number of journal articles, a kind of analysis not often performed in traditional digital humanities, which often focuses on book-length source material. To solve these problems, one of us has created a software package, RLetters (Pence 2016), available at <http://www.rletters.net>. (One public installation of this software, containing a corpus of journals in evolutionary biology, is available at <http://www.evotext.org>, and described in (Ramsey and Pence 2016).) This is a web application, backed by a search engine and database, which may be deployed by anyone wishing to analyze a corpus of academic journal articles. It includes a variety of analysis methods (sharing many of those described for Voyant), including an especially powerful word frequency analyzer.

Finally, should all of these tools fall short, the statistical computing language R (R Core Team 2017, available at <http://www.r-project.org/>) has become a very popular base for

constructing novel analyses in the digital humanities (Jockers 2014). R combines a comprehensive set of standard statistical analyses (such as principal component analysis and dendrogram or tree clustering) with an extensive collection of user-contributed packages which may be utilized to perform complex tasks such as querying Google Scholar or Web of Science. This power comes at the cost of significant complexity, however, as R operates like a programming language rather than a graphical application.

3.4. Copyright issues. One of the most common pitfalls that users are likely to encounter when building corpora of digital data is copyright and licensing issues. While much material pertaining to figures like Newton or Darwin is available in the public domain, a confusing legal landscape besets all work created after 1923 (the date of "public domain" for published works in the United States). A number of recent court decisions (most significantly Authors Guild v. HathiTrust; Bayer 2012) have begun to clear the legal landscape in the United States, indicating that scholarly textual analysis and other sorts of digital-humanities work are likely to fall under the U.S. "fair use" provision. This, however, does nothing to simplify *obtaining* copyrighted materials, nor does it help scholars in other countries, many of which lack an analogue to fair use. It also may well be cold comfort to litigation-sensitive universities.

Increasingly, however, publishers are recognizing the demand for digital analyses of their materials. Elsevier has deployed a text and data mining policy that applies to all of their journals, and will allow researchers to access and analyze articles as part of any institutional subscription (Elsevier 2014). Under the auspices of JSTOR's DFR project, researchers may request access to full-text articles, if their university subscribes to the appropriate JSTOR collections. We also have had some degree of personal success negotiating access contracts for closed-access journal

articles with their publishers, including with Nature Publishing Group, who were very receptive to the possibilities opened by digital analyses. We anticipate that this trend toward increased ease of access will only continue.

4. Data. The academic process relies on the ability of other researchers to access, verify, and reproduce the results of analyses such as these. We will next consider how to publish and archive data, and how make public the tools and techniques used to achieve the results.

4.1. Data management. Philosophers are not, as a rule, accustomed to producing large amounts of data as part of our research. When using digital tools, we find ourselves faced with many of the same questions our scientific colleagues have dealt with for some time—how do we document, store, and preserve the data that our research generates? We cannot offer comprehensive answers to these questions here; we raise them only to emphasize that problems of metadata, documentation, and archiving have been discussed extensively in other contexts and should not be neglected. Early engagement with these resources will prevent significant problems from arising in the long term (York 2009; Michener 2015).

4.2. Reproducibility. If digital analyses are to serve as elements of the permanent research record along with journal articles, then we must take care to make those analyses reproducible in the future. This is a multifaceted problem that has, in recent years, received significant attention from the scientific community (Munafò et al. 2017). For most digital philosophy projects, there are three key components to reproducibility.

First, software must be reproducible—that is, easily installed and run by those with the relevant technical expertise. To that end, the development and use of open source software is laudable, as is using a readily accessible distribution platform such as GitHub.

Second, corpora must be reproducible. This can be a difficult challenge, particularly if one has negotiated access to a body of copyrighted materials for analysis. It is often possible to negotiate access not just for an individual researcher or research team, but also for any researchers accessing a public resource (Ramsey and Pence 2016 successfully negotiated such contracts for evoText). We encourage researchers to think very seriously about this challenge as they develop corpora.

Finally, the original forms of data must be—and remain—available. Open data repositories such as figshare (figshare Team 2012; Kraker et al. 2015) or Zenodo (CERN 2013) will accept raw data and make it citable. Researchers should also take care to upload data into these repositories in formats that are likely to remain readable indefinitely into the future, such as comma-separated value (CSV) format for spreadsheets, or plain Unicode text or XML for textual data.

5. Integrating digital results into philosophy of science. The digital tools are powerful and they have great potential for the philosophy of science. But digital results do not automatically translate into philosophical results. We therefore must consider how to integrate them with broader answers to philosophical questions.

5.1 Justifying digital results. A recurring problem with digital humanities results consists in how we can be certain that we have obtained genuine information supporting the conclusions

we hope to draw. We can in part resolve this by proceeding in an "hypothesis-first" manner—forming clear hypotheses prior to performing analyses. All datasets are apt to contain chance patterns, and we should not be led astray by these patterns by basing our conclusions upon them. And when we formulate hypotheses, we should attempt to be open to a range of possible conclusions, since approaching a statistical analysis system with an answer to one's question already in mind tends to result in the cherry-picking of tools and methods to produce the desired result (Ioannidis 2005).

That said, it can be difficult, even having carefully formulated and tested an hypothesis, to be certain that one has in fact demonstrated it conclusively. Many analyses in the digital humanities lack statistical validation, and have only a history of successful use as evidence in their favor (see, e.g., the discussion of validation in Koppel, Schler, and Argamon 2009). Others require collaboration between experts in philosophy and statistics, computer science, or even electrical engineering (Miller et al. 2015). An important step in developing a digital research program, therefore, is to consider how to assess whether a project has succeeded or failed. This may involve validating the methods, producing standard kinds of analysis outputs, or, as we now consider, using digital research methods only as a first step in a broader program of philosophical research.

5.2 Digital humanities as research generator. Because digital tools give us significantly increased breadth and depth, we have found that they are useful not just as research tools in and of themselves, but as a compass, directing us toward questions that would be answered by traditional methods in philosophy of science. For example, Pence has recently combined existing work on an episode in the history of biology (Pence 2011) with digital tools (Ramsey and Pence

2016), to produce a more general hypothesis about debates over paradigm change, which is now ripe for a non-digital analysis (Pence in preparation). We anticipate that this workflow will, in fact, be quite common. As a digital tool shows us a provocative but not fully theorized result, this can provide us with an excellent working hypothesis, case study, or set of sample data for developing a philosophical thesis.

6. Conclusion. As scholars interested in studying the natural sciences, we cannot ignore the availability of digital data that might assist us in our research. It was once the case that the body of scientific literature was modest in size and represented only a narrow distillation of and reflection upon the world. Now the literature has become so massive, complex, and diverse that it constitutes a world unto itself, one poised for scientific and philosophical analysis. Adding to this all of the digital traces of work not heretofore published—archival images, notebooks, and so on—we are confronted with an overwhelming, but incredibly rich, world of information. Philosophers are beginning to see how this information can bear on questions in the philosophy of science, and can inspire new ones. But the profusion of sources and formats of data, on top of the assortment of available tools, some of which require considerable technical savvy, provides a barrier to the philosopher. In this essay, we have attempted to provide a window into digital philosophy of science, with both an overview of what is possible and some guidance in seeking data and analysis tools. We are excited about the prospects for future work in this field, and hope that this article will help to spread our excitement.

References

Bastian, Mathieu, Sebastian Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In *Third International AAAI Conference on Weblogs and Social Media*, 361–62. AAAI Publications.

Bayer, Harold, Jr. 2012. The Authors Guild, Inc., et al., v. HathiTrust, et al., 11 CV 6351 (HB). United States District Court, Southern District of New York.

Beccaloni, George. 2008. "The Alfred Russel Wallace Correspondence Project." http://wallaceletters.info.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. doi:10.1088/1742-5468/2008/10/P10008.

Brants, Thorsten, and Alex Franz. 2006. *The Google Web 1T 5-Gram Corpus Version 1.1 (LDC2006T13)*. Philadelphia, PA: Linguistic Data Consortium.

Buckner, Cameron, Mathias Niepert, and Colin Allen. 2007. "InPhO: The Indiana Philosophy Ontology <http://inpho.cogs.indiana.edu/>." *APA Newsletter* 7 (1): 26–28.

Burns, John, Alan Brenner, Keith Kiser, Michael Krot, Clare Llewellyn, and Ronald Snyder. 2009. "JSTOR - Data for Research." In *Research and Advanced Technology for Digital Libraries*, 416–19. Lecture Notes in Computer Science 5714. Berlin: Springer.

CERN. 2013. *Zenodo*. Geneva. https://zenodo.org/.

Cohen, Aaron M., and William R. Hersh. 2005. "A Survey of Current Work in Biomedical Text Mining." *Briefings in Bioinformatics* 6 (1): 57–71.

Cohen, Dan. 2008. "Introducing Omeka." http://hdl.handle.net/1920/6089.

Dombrowski, Quinn. 2014. "What Ever Happened to Project Bamboo?" *Literary and Linguistic Computing* 29 (3): 326–39. doi:10.1093/llc/fqu026.

Elsevier. 2014. "Text and Data Mining." https://www.elsevier.com/about/our-business/policies/text-and-data-mining.

figshare Team. 2012. *Figshare*. London. https://figshare.com/.

Gwinn, Nancy E., and Constance Rinaldo. 2009. "The Biodiversity Heritage Library: Sharing Biodiversity Literature with the World." *IFLA Journal* 35 (1): 25–34. doi:10.1177/0340035208102032.

Hoover, David L. 2012. "Textual Analysis." In *Literary Studies in the Digital Age: An Evolving Anthology*. Modern Language Association. http://dlsanthology.commons.mla.org/textual-analysis/.

Ide, Nancy, and Jean Véronis, eds. 1995. *Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer.

Iliffe, Rob, and Scott Mandelbrote. 1998. "The Newton Project." http://www.newtonproject.sussex.ac.uk/.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. doi:10.1371/journal.pmed.0020124.

Jockers, Matthew. 2014. *Text Analysis with R for Students of Literature.* Cham, Switzerland: Springer.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. "Computational Methods in Authorship Attribution." *Journal of the American Society for Information Science and Technology* 60 (1): 9–26. doi:10.1002/asi.20961.

Kraker, Peter, Elisabeth Lex, Juan Gorraiz, Christian Gumpenberger, and Isabella Peters. 2015. "Research Data Explored II: The Anatomy and Reception of Figshare." http://arxiv.org/abs/1503.01298.

Maienschein, Jane, Manfred D. Laubichler, Jessica Ranney, Kate MacCord, Steve Elliott, and Federica Turriziani Colonna. 2007. "The Embryo Project Encyclopedia." https://embryo.asu.edu.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Baltimore, MD: Association for Computational Linguistics.

Mendelsson, Dalia. 2003. "Einstein Archives Online." http://www.alberteinstein.info.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014): 176–82. doi:10.1126/science.1199644.

Michener, William K. 2015. "Ten Simple Rules for Creating a Good Data Management Plan." *PLoS Computational Biology* 11 (10): e1004525. doi:10.1371/journal.pcbi.1004525.

Miller, B. A., M. S. Beard, M. D. Laubichler, and N. T. Bliss. 2015. "Temporal and Multi-Source Fusion for Detection of Innovation in Collaboration Networks." In *2015 18th International Conference on Information Fusion (Fusion)*, 659–65.

Mills, Virginia. 2011. "The Joseph Dalton Hooker Project." http://www.sussex.ac.uk/cweh/research/josephhooker.

Moretti, Franco. 2013. *Distant Reading.* London: Verso.

Morse-Gagné, Elise E. 2011. "Culturomics: Statistical Traps Muddy the Data." *Science* 332: 35–36.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1: 21. doi:10.1038/s41562-016-0021.

Murdock, Jaimie, Colin Allen, and Simon DeDeo. 2017. "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks." *Cognition* 159: 117–26. doi:10.1016/j.cognition.2016.11.012.

Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PLOS ONE* 10 (10): e0137041. doi:10.1371/journal.pone.0137041.

Pence, Charles H. in preparation. "How Not to Fight about Theory: The Debate between Biometry and Mendelism in *Nature*, 1890–1915." In *The Evolution of Science*, edited by Andreas De Block and Grant Ramsey.

———. 2011. "'Describing Our Whole Experience': The Statistical Philosophies of W. F. R. Weldon and Karl Pearson." *Studies in History and Philosophy of Biological and Biomedical Sciences* 42 (4): 475–485. doi:10.1016/j.shpsc.2011.07.011.

———. 2016. "RLetters: A Web-Based Application for Text Analysis of Journal Articles." *PLoS ONE* 11 (1): e0146004. doi:10.1371/journal.pone.0146004.

Pouyllau, Stephane, Christine Blondel, Marie-Helene Wronecki, Bertrand Wolff, and Delphine Usal. 2005. "Ampère et l'histoire de l'électricité." http://www.ampere.cnrs.fr.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Ramsey, Grant, and Charles H. Pence. 2016. "evoText: A New Tool for Analyzing the Biological Sciences." *Studies in History and Philosophy of Biological and Biomedical Sciences* 57: 83–87. doi:10.1016/j.shpsc.2016.04.003.

Roberts, Richard J. 2001. "PubMed Central: The GenBank of the Published Literature." *Proceedings of the National Academy of Sciences* 98 (2): 381–82. doi:10.1073/pnas.98.2.381.

Secord, James. 1974. "The Darwin Correspondence Project." http://www.darwinproject.ac.uk.

Sinclair, Stéfan, and Geoffrey Rockwell. 2016. *Voyant Tools*. http://voyant-tools.org/.

Walter, S. A., Ph. Nabonnand, and L. Rollet. 2002. "Henri Poincaré papers." http://henripoincarepapers.univ-nantes.fr.

Whye, John van. 2002. "The Complete Work of Charles Darwin Online." http://darwin-online.org.uk/.

York, Jeremy. 2009. "This Library Never Forgets: Preservation, Cooperation, and the Making of HathiTrust Digital Library." *Archiving Conference* 2009 (1): 5–10.