Paper Machines: A Text Analysis and Visualization Toolkit for Zotero Libraries

by Jo Guldi | Brown University; Harvard Society of Fellows | guldi@fas.harvard.edu

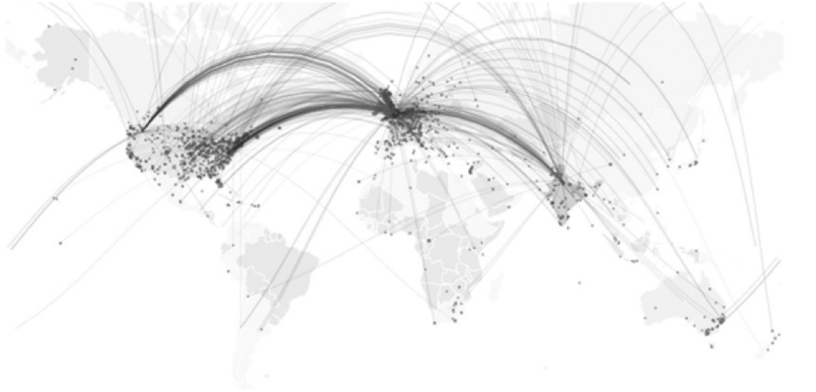
From the 1790s on, official government minutes and reports were printed for most Western governments. Printing was expensive and occasioned only by meetings of official bodies. From the 1820s on, these state papers were supplemented by occasional propaganda from nongovernment bodies like the Society for the Promotion of Christian Knowledge. By the 1880s and the rise of Fabian socialism, this form of para-governmental reporting had become a standard apparatus of professional groups, each of which, from doctors to lawyers to property surveyors, had its own professional journal, routinely distributing professional reports. Modern history is the moment of too much paper to read, a problem that traps historians in a game of characterizing the institutions of power even when their social and disciplinary commitments lead them to critical stances toward those institutions.

This summer, my team of researchers released Paper Machines, a digital toolkit designed to help scholars like myself parse the massive amounts of paper involved in any comprehensive, international look at the over-documented twentieth century. Its purpose is to make state-of-the-art text mining accessible to scholars across a variety of disciplines in the humanities and social sciences who lack extensive technical knowledge or immense computational resources.

Paper Machines was designed with the range of historians' textual sources in mind. While tool sets like Google Books Ngram Viewer utilize preset corpora from Google Book Search that automatically emphasize the Anglo-American tradition, Paper Machines works with the individual researcher's own hand-tailored collections of texts, whether mined from digital sources like newspapers and chat rooms or scanned and saved through optical character recognition (OCR) from paper sources like government archives. Paper Machines is an open-source Zotero extension; Zotero is a program that allows users to create bibliographies and build their own hand-curated libraries in an online database. These libraries may be large or small—as minute as a particular literary author or as large as the discipline of economics in the nineteenth century—and the libraries may overlap (for example, novels about economic life could appear in an economics library). Zotero also provides for scholarly collaboration. Nested inside Zotero, Paper Machines works with Zotero's provision for multiuser text collections, allowing a class, a group of scholars, or scholars and activists together to collect and share archives of texts. These group libraries can be set as public or private depending on the sensitivity and copyright restrictions of the material being collected. Indeed, I have heard reports of historians of Panama using a Zotero group library to collect and share the texts of government libraries for which no official finding aid exists. The scholars themselves are thus engaged in preserving, annotating, and making discoverable historical resources that otherwise risk neglect, decay, or even intentional damage.

In Paper Machines, these hand-tailored libraries are visualized with data-mining techniques like geoparsing, named entity recognition, and latent Dirichlet allocation (LDA). The results of these processes can then be translated into a variety of categorical, chronological, and geographical visualizations to show the distinctive features of textual corpora. With Paper Machines, scholars can create visual representations of a multitude of patterns within a text corpus using a simple, easy-to-use graphical interface. One may use the tool to generalize about a wide body of thought—for instance, things historians have said in a particular journal over the last ten years. Or one may visualize libraries against each other—say, novels about nineteenth-century London set against novels about nineteenth-century Paris. Using this tool, a multitude of patterns in text can be rendered visible

Figure 1: Geoparser map showing mentions of place-names in a corpus of texts about land law, with time slider at top left.



through a simple graphical interface. For example, Figure 1 is a map of place-names mentioned in a corpus of texts on land law. In the interactive version, a time slider would allow the scholar to watch the appearance of new place-names over time.

Paper Machines itself, as a piece of infrastructure, is both a work in progress and a community initiative. As the grant writer behind a collaborative effort involving several people, I designed Paper Machines to help scholars like myself by capitalizing upon the masses of texts already available online, digitized by thoughtful digital archivists, and also drawing upon the work already developed by our colleagues in computer science. To this end, it provides an easily extensible framework for combining data sources—including JSTOR Data for Research and the user's own corpus of texts—with techniques such as geoparsing, named entity recognition, and latent Dirichlet allocation (LDA), each of them the product of graduate work by computer scientists. For instance, David Mimno, a Princeton postdoctoral student, wrote Mallet, the timeline used by Paper Machines for visualizing topic modeling. Figure 2 shows a timeline designed with Mallet wherein the computer identifies particular topics (sets of words that probabilistically appear near each other in a book) over time. Designed to work with English, it was released as an early open-source alpha version of code this summer, with the idea that scholars interested in retooling the code to work with Spanish-corpus place-names or alternative historical gazetteers might do so, hiring computational research assistants of their own to adjust the tool to their specific needs.

Applying Paper Machines to text corpora allows scholars to accumulate hypotheses about longue-durée patterns in the influence of ideas, individuals, and professional cohorts. By measuring trends, ideas, and institutions against each other over time, scholars will be able to take on a much larger body of texts than they normally do. In my own work I have begun applying Paper Machines to a text corpus that I hand-curated for my *Long Land War* monograph project. It is already proving useful to the problem of "distant reading" large numbers of bureaucratic texts from the twentieth century, which is a major problem for scholars asking large-scale questions in my field. However, scholars from other fields will be able to imagine new and innovative uses for Paper Machines' functions—or new and innovative functions altogether.

Working with Paper Machines allows me to trace the conversations in British history from the local stories at their points of origin forward, leaping from microhistorical research in the British archive into longue-durée synthesis of policy trends on a worldwide scale. That digitally enabled research operates through a threefold process: digitally synthesizing broad swaths of time, critically inquiring into the microhistorical archive with digitally informed discernment about which archives to choose, and reading more broadly in secondary literatures from adjacent fields. For example, in Figure 2, the topic-modeling algorithm Mallet has been run on a corpus of scholarly texts about land law. The resulting image is a computer-guided timeline of the relative prominence of ideas—some mentioning Ireland and some mentioning India—that can then be changed and fine-tuned. This visualization of changing concepts over time guides me to look more closely in my corpus at the 1970s, when the intellectual memory of land struggles in Ireland was helping to guide contemporary policy in Latin America.

The Long Land War tells the story of the global progress of land reform movements, tracing ideas about worker allotments and food security, participatory governance, and rent control from the end of British Empire to the present. British lawyers wrestled with ideas of peasant proprietorship in India, Ireland, and Scotland, inventing patterns of land reform that appealed to administrators at the United Nations and the postcolonial governments of the global south. Throughout my reading in the history of British property law, I have used Paper Machines to synthetically characterize the nature of particular debates and their geographic referents, making for instance timelines and spatial maps of topics and place-names associated with rent control, land reform, and allotment gardening.

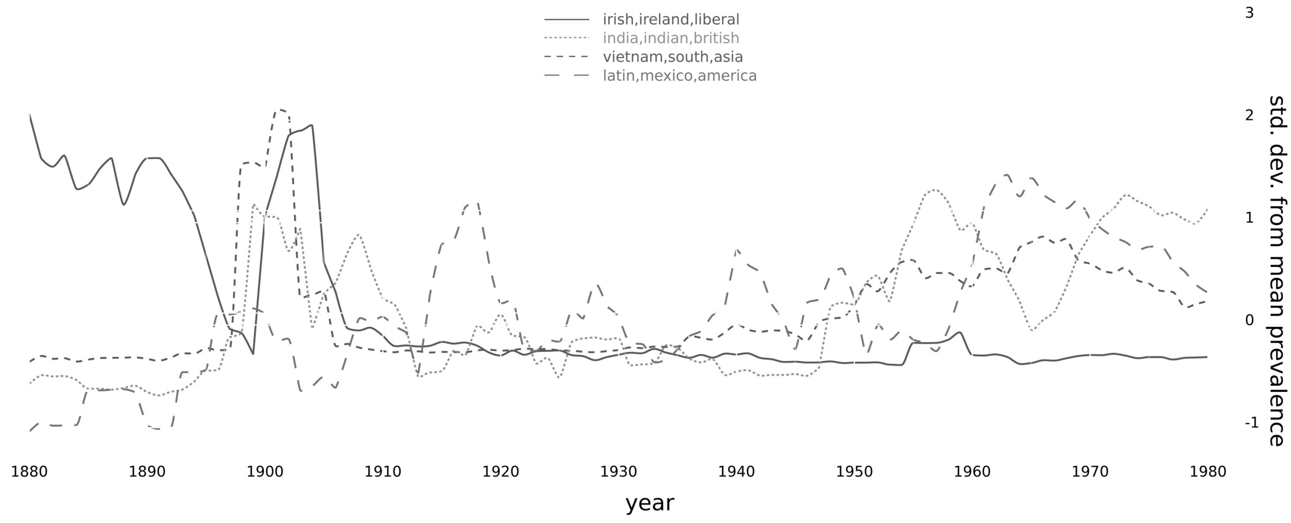
Paper Machines also makes the scholar better able to discern which archives to choose and in which parts of archives to invest her reading time. I came back from visiting Rome this summer with 4,600 OCR documents from the archive of the Food and Agriculture Organization of the United Nations, one of the chief offices to preside over midcentury land reform. Paper Machines was designed as a tool for hacking those bureaucracies, for forming an instant portrait of their workings, giving an immediate context to documents from the archive that is more substantial than the high-profile events, schools of thought, or individuals—Milton Friedman, John F. Kennedy, and so on—who tend to dominate our understanding of history. Instead, the user of Paper Machines can afford to pay attention to the field agents, branch heads, and directors-general of UN offices, or indeed to the intermediate faculty of the University of Wisconsin and the University of Sussex who offered so much advice to both heads of state and generations of undergraduates on their way to the civil service. Paper Machines allows us to instantly take the DNA of each of these organs, identifying the ways in which they diverge and converge. All of these fields spoke a common language of modernization theory: of national governments, democratic reform, government-provided extension, training and management and the provision of new equipment that resulted in quantitatively verifiable increased production.

Reading with a digital finding aid renders possible intentional reading, as for instance in reading for dissent. Traditional research, limited by the sheer breadth of the non-digitized archive and the time necessary to sort through it, becomes easily shackled to histories of institutions and actors in power, for instance characterizing universal trends in the American Empire from the Ford and Rockefeller Foundations' investments in pesticides, as some historians have done. By identifying vying topics over time, Paper Machines allows the reader to identify and pursue particular moments of dissent, schism, and utopianism—zeroing in on conflicts between the pesticide industry and the Appropriate Technology movement or between the World Bank and the Liberation Theology movement over exploitative practices, for example. Digitally structured reading means giving more time to counterfactuals and suppressed voices, realigning the archive to the intentions of history from below.

For the scholar writing as a British historian, that gesture implies an end to scholastic isolation of Anglo-American traditions from debates in other corners of the world. My work with Paper Machines has required me to spend more and more of my time reading contemporary literature on Latin American and South Asian history, coming to grips with the way land stories unfold on a global scale, and thus seriously addressing the import of my subject through dialogue with colleagues in neighboring fields. Digital synthesis, by allowing me to look at the grander sweep of time, has made me, as a historian of Britain, better able to write as a citizen of the world.

My hope is that tools like these can teach historians how to take our own questions more seriously. I also hope that open-source, reusable tools like Paper Machines, building upon existing resources, will encourage historians and indeed the public to look at events in their deep contexts, drawing out the most important narratives possible for a history of the present. ■

Figure 2: Mallet, topic modeling software by David Mimno, modeling the relevant prominence of mentions of India, Ireland, and other topics in relationship to each other over time.


irish,ireland,liberal
india,indian,british
vietnam,south,asia
latin,mexico,america
std. dev. from mean prevalence
year