Daniel Rosenberg

University of Oregon

Text for American Historical Association 2012

Data Before the Fact

Draft: Please do not quote without permission. dbr@uoregon.edu

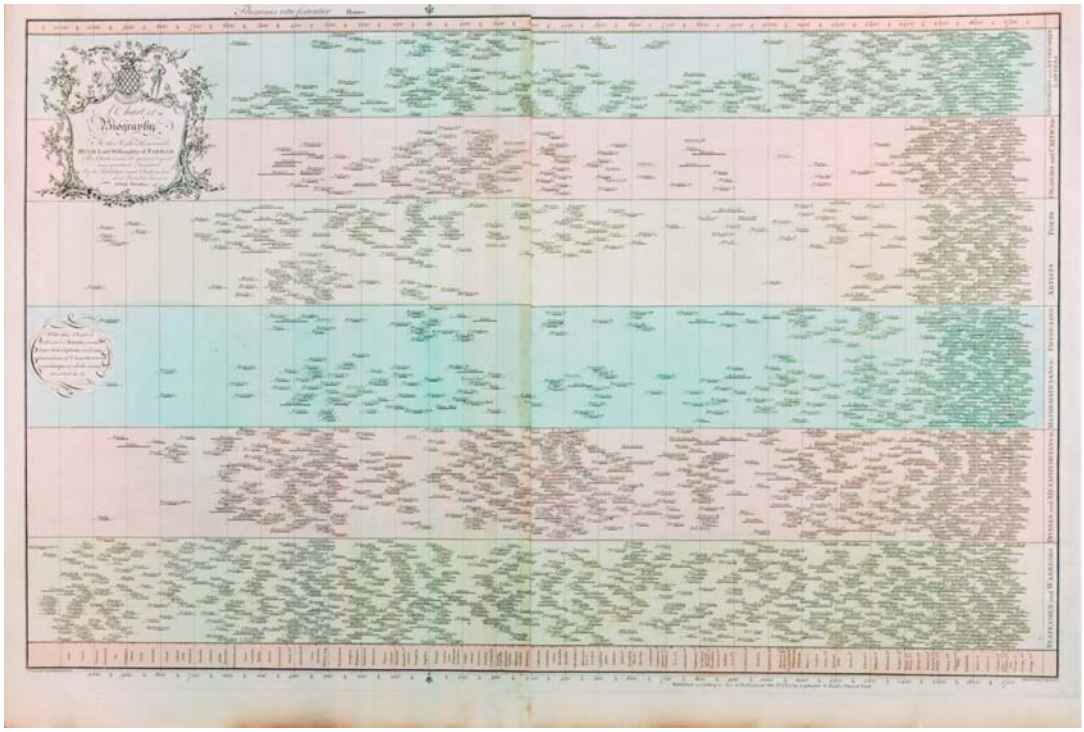


Image: Joseph Priestley, Chart of Biography, 1765. Densities of lines show patterns of achievement by category in different eras.

It is a great delight to be here to celebrate Peter Burke.

In the brief time that I have today, I'd like to talk about a project that I am just beginning on the history of the concept of "data."

My work on the concept of "data" began, as so many investigations do, with a happenstance textual encounter that eventually became a kind of irritation. In researching my last book, I ran across an odd passage in a work by the eighteenth-century natural philosopher and theologian, Joseph Priestley. In his 1788 *Lectures on History and General Policy*, Priestley refers to names and dates as the "data" we find in historians. The usage struck me as curiously modern.

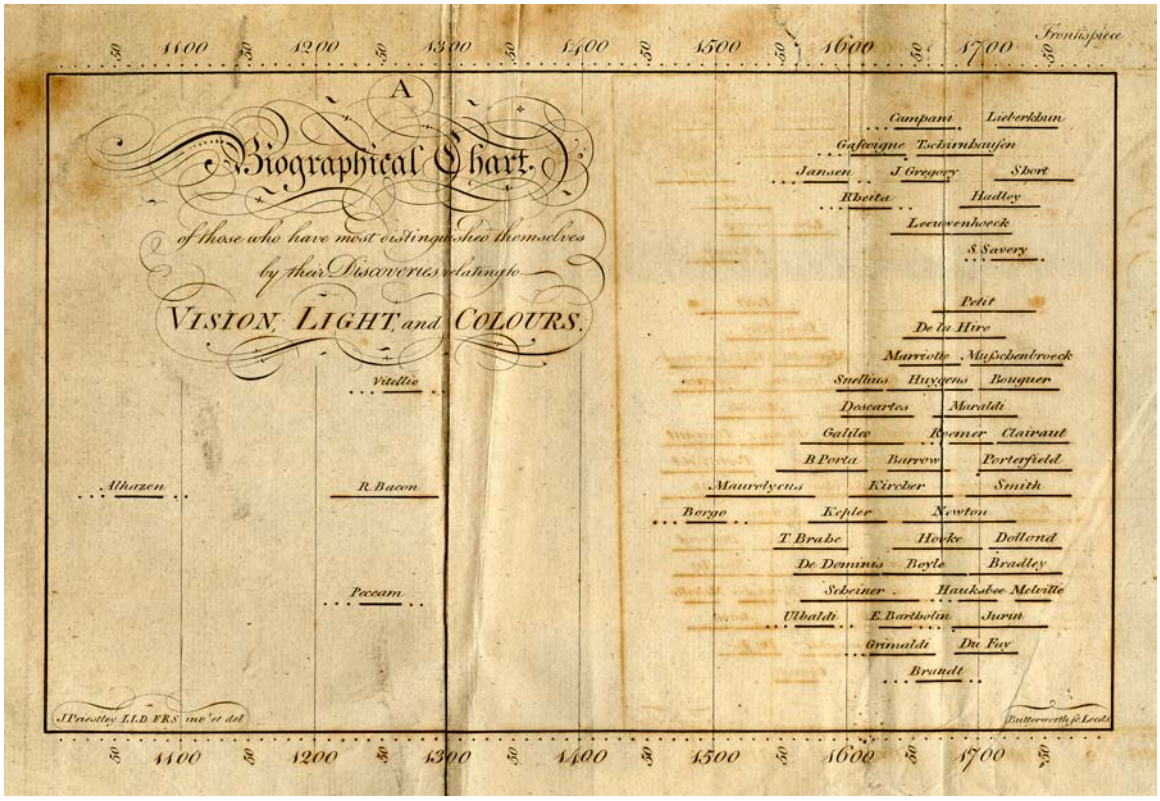


Image: Joseph Priestley, Biographical Chart from *History and Present State of Discoveries Relating to Vision, Light, and Colours*, 1772. Biographical information extracted from *Chart of Biography* showing lives of key figures in the history of optics.

Of course, if anyone in the eighteenth century was in a position to formulate a modern concept of historical data, it would have been Priestley. The image you see projected is Priestley's 1765 *Chart of Biography*, a giant double-folio graphic

representing the lives of approximately 2000 important historical figures over the course of 3000 years of world history categorized and laid out according to a linear measure. Priestley's chart is a monumental achievement in the history of data graphics, arguably the first modern timeline.

Still, Priestley's use of the term "data" bothered me. And as I continued my work on him, I noticed the term recurring. In his *Experiments and Observations on Different Kinds of Air* (1777), Priestley uses "data" to refer to measurements of volumes of air. In the *Evidences of Revealed Religion* (1794), Priestley says that scripture offers us no "data" on the physical nature of Christ's resurrected body.

Still, the passage seemed strange. Everything that I knew about data led me to associate the term with the bureaucratic and statistical revolutions of the nineteenth century and the technological revolutions of the twentieth.

Yet, having noticed data once in Priestley, I began to find it everywhere in the eighteenth-century corpus.

All of this raised questions: What was the history of the concept? What was the relationship between the emergent usage in the eighteenth century and familiar modern usages? And, if the term "data" did have an earlier importance, didn't **it** deserve a historiography equal to that received by sister terms such as "facts," "evidence," and "truth."

All of these questions, I think, are that much more compelling since, in the recent historiography, including foundational works by Lorraine Daston, Theodore Porter, and Mary Poovey, the term data appears frequently, even doing some very heavy lifting, yet is rarely, if ever, remarked upon.

Consider, for example, the first lines of Poovey's excellent book, *A History of the Modern Fact.* "What are facts?" Poovey asks. "Are they incontrovertible data that simply demonstrate what is true? Or are they bits of evidence marshaled to persuade others of the theory one sets out with?" In Poovey's construction, "facts" may be conceived either as theory-laden or as incontrovertible. We signal the latter case by calling them "data."

Of course, at this point, it would be very natural to attempt a little one-upsmanship. If "facts" can be deconstructed, surely "data" can be too. If facts can be shown to be theory-laden, why not data? Yet, in my view, there are good reasons to continue using "data" in precisely the unmarked, undeconstructed manner in which Poovey uses it. I'd just like to understand why it makes a plausible candidate for something we would *not* want to deconstruct.

To get there requires understanding what makes "data" different from other conceptual entities, in particular what makes it different from "facts."

So what was "data" prior to the nineteenth and twentieth centuries? How did "data" first acquire its pre-analytical, pre-factual status?

In this, the etymology of term is a good starting point.

The English word, "data," as you probably guess, is derived from Latin. It is the plural form of "datum," which itself is the neuter past participle of the verb *dare*, "to give." A datum is something given in an argument, something taken for granted.

This is in contrast to "fact," which derives from the neuter past participle of the Latin verb *facere*, to do, whence we have the "fact" as that which was done, occurred, or exists.

There is an important contrast here: facts are ontological; data is rhetorical.

In the influential formulation of Euclid, mathematical problems are structured around two basic elements, the *data* and *quaesita*, values that are given—let X=3—and values that are sought. And this Euclidean framework is one of the key conduits through which the Latin words "datum" and "data" first entered the English language.

In every language that I have examined, excepting Latin of course, the word "data" is recent, though it appears to emerge first in English. The Oxford English Dictionary finds its earliest usage in a 1646 theological tract that refers to "a heap of data." In seventeenth century English, "data" was used especially in mathematics, where it retained the technical sense given by Euclid, and in theology, where it referred to scriptural truths that were given and therefore not susceptible to question.

In the seventeenth century, then, historical data was information outside the realm of possible investigation that served the historian's pursuit of the *quaesita* of history. Similarly, the "heap of data" referred to in Henry Hammond's 1646 tract was not a pile of numbers but a list of theological propositions accepted as true for the sake of argument—that priests should be called to prayer, that the liturgy should be rigorously followed, and so forth.

So, this is where I was in my research not very long ago.

In any past situation, my next steps would almost certainly have been hermeneutic: my usual plan would have been to read Priestley more extensively and closely. And, of course, I did do plenty of that.

But, it occurred to me that in *this* case, with *this* subject matter, and at *this* historical juncture, it might also be appropriate to try to apply some quantitative tools, to take a stab at writing a *quantitative* history of "data."

My plan was to begin by collecting, categorizing, and counting occurrences of the term "data" in English in order to specify when the term came into use as a Latin loan word, when was it naturalized, when its achieved its various connotations, and when it became important in common usage—all the service of understanding both the historical problem and the historiographical opportunity offered by such an approach.

Now it happens that I performed my first round of tabulation just about a year ago, shortly before Google publicly debuted its Ngram Viewer, which provides a neat and easy way to do something very much like what I intended.

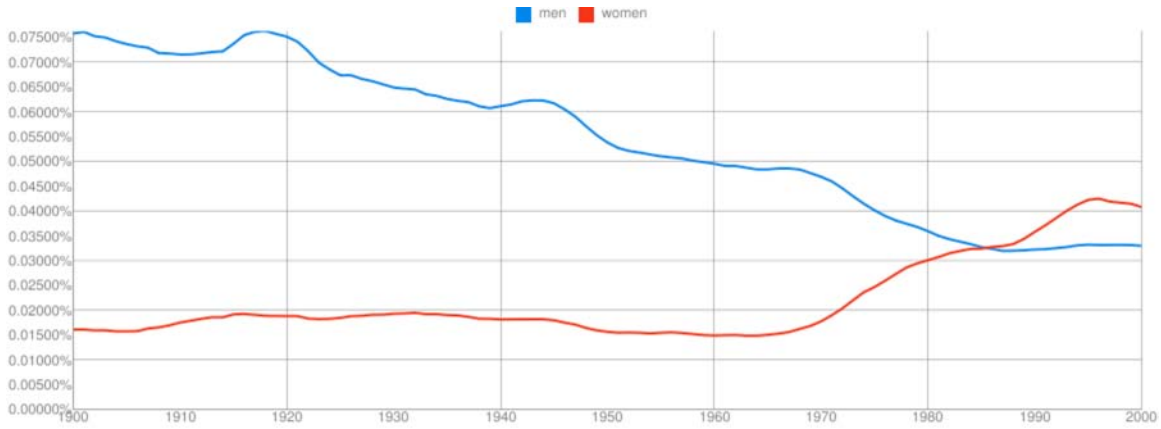


Image: Relative frequency of "men" vs. "women" in Google Books, 1900-2000, as conceived by Michel and Aiden, generated by Google Ngram Viewer.

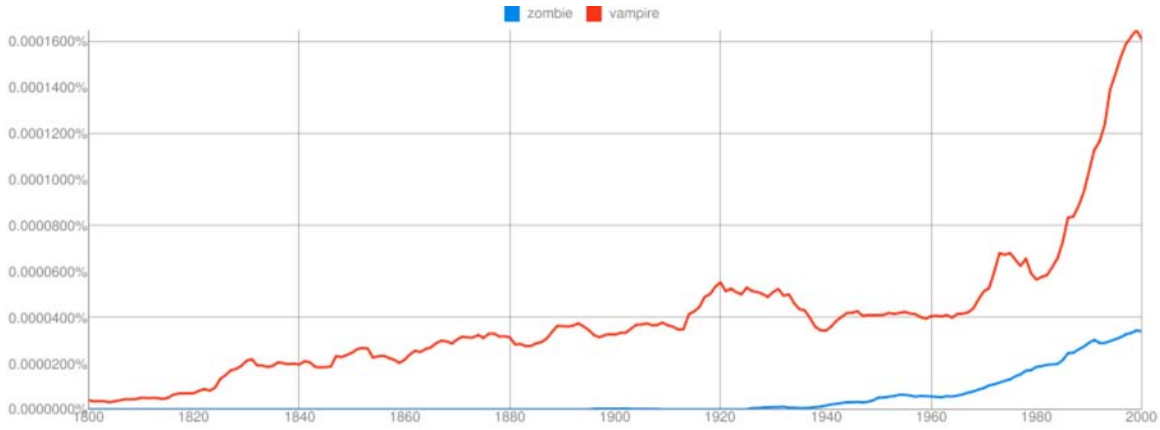


Image: Relative frequency of "zombie" vs. "vampire" in Google Books, 1800-2000, as conceived by theatlantic.com, generated by Google Ngram Viewer.

For those of you who have not yet played with the Ngram Viewer, I highly recommend it. It can instantaneously produce a whole variety of lovely historiographical artifacts—of varying significance—such as these.

In retrospect, I'm both a little sad and a little relieved that the timing of the release of the Ngram Viewer worked out the way it did. I'm sad because it could have saved me a good deal of work. I'm relieved because my labor, doing manually what Google can do automatically, turned out to be instructive in all sorts of ways.

So this is what one sees looking at the long history of "data" through the lens of the Ngram Viewer.

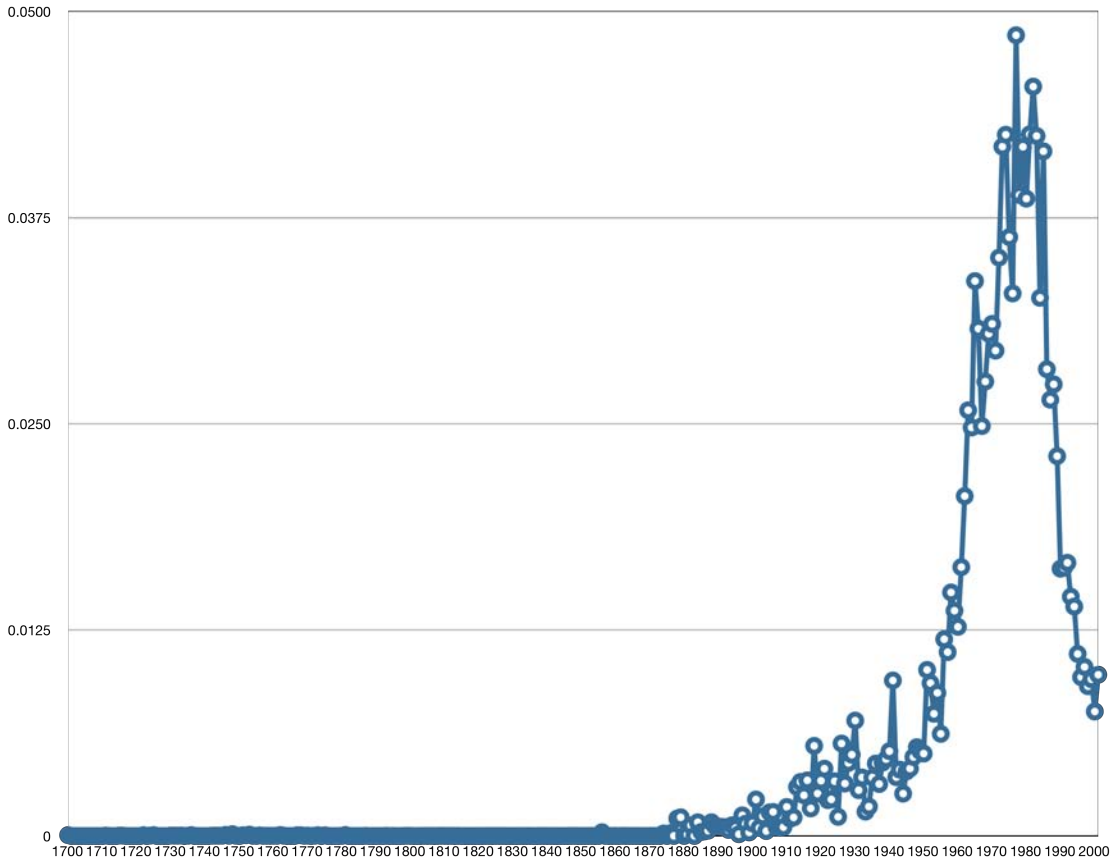


Image: Relative frequency of "data" in works in Google Books by year, 1700-2000, generated manually.

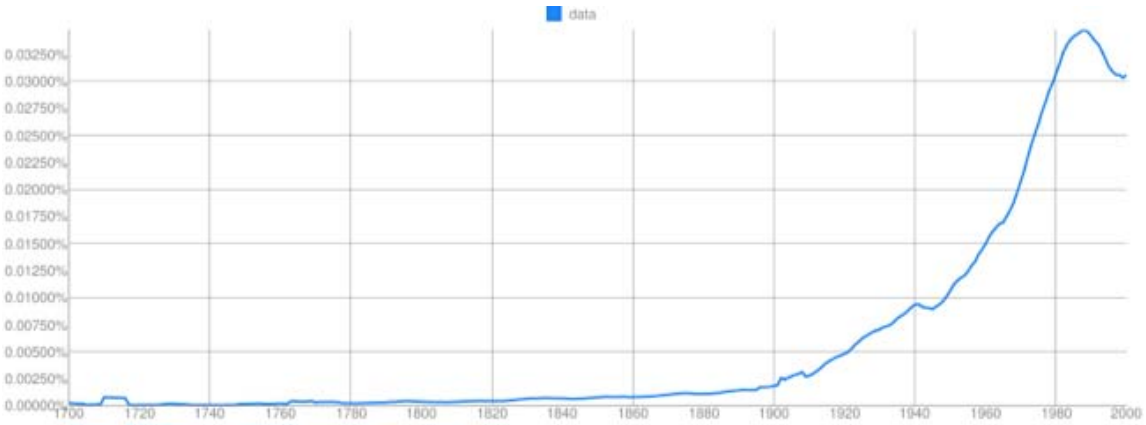


Image: Relative frequency of "data" in Google Books, by year, 1700-2000, generated by Google Ngram Viewer

There are a number of observations we might make and questions we might pose about this plot—particularly about the nosedive after 1980—but, in broad outlines, the story that it suggests is more or less what we might have expected, knowing nothing whatsoever about the quantitative facts of the matter.

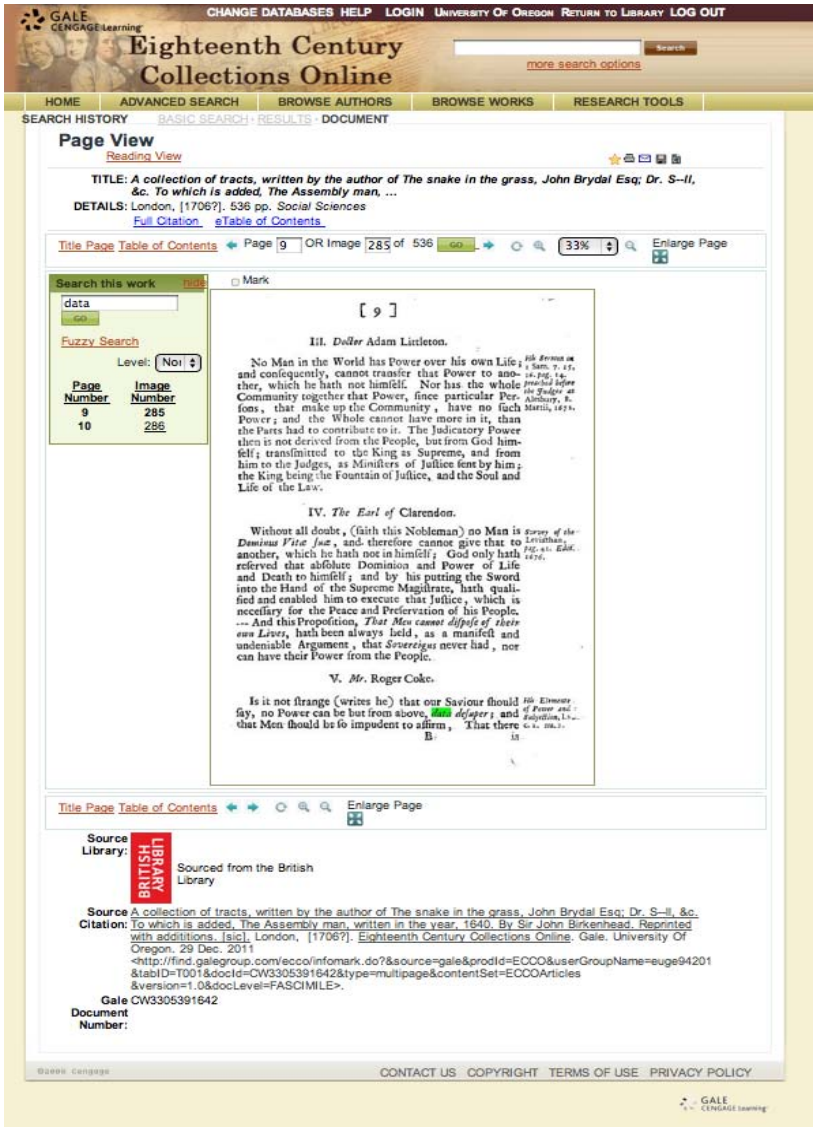
Broadly speaking, the big historical action appears to take place in the nineteenth and twentieth centuries, during which we see the rise of the concept.

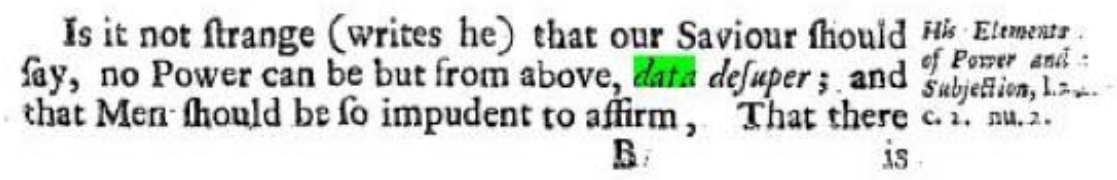
This is, of course, exactly what I imagined the history of data might look like before I first encountered that first quotation from Priestley. What is more, it's a good story, and probably a true story. Fortunately for me, I started my work just before the Ngram Viewer went public and therefore was unconstrained by self-evidence. I also began with a different system, the subscription database ECCO or Eighteenth-Century Collections Online.

ECCO is a primitive tool, and it suffers from many of the well-publicized faults of Google Books, particularly in scanning quality. (Incidentally, some recent work published in *Eighteenth-Century Studies* has shown just how problematic the scanning in ECCO turns out to be. What is more, the ECCO interface seems designed to thwart quantitative inquiry.) Yet ECCO has some notable advantages too. Its corpus, based on the English Short Title Catalogue, is well known, well defined, and relatively stable. ECCO provides a couple of clever proximity searching functions that are not available out of the box from Google. And ECCO has superb good book-level metadata.

In fact, a decade ago, one might have thought that ECCO would have had the revolutionary effect on historical scholarship that many now expect our interactions with

Google to produce. I remember a friend of mine in graduate school referring to the newly announced system as the "dissertation machine."





Images: ECCO screen shot and close up from "data" search

The first thing that has limited ECCO's effect, of course, is that it is not openly and freely available without an institutional subscription. Additionally, Thompson-Gale, the company that owns ECCO, treats its data as proprietary, and access is only readily available through the Thompson-Gale interface, which is limited in a number of important ways. Significantly, while ECCO users can view page images with the search terms highlighted, they cannot see or manipulate the OCR-coded text that underlies those images.

Interestingly, since they introduced the database about a decade ago, Thompson-Gale has loosened their rules on downloading page images. So it's now easy to save complete books from ECCO to your desktop computer in the form of page images. Yet you can't download a single page of OCR-coded text. Not even a line. Which suggests that over time Thompson-Gale has decided that there's no percentage in books, not even in digitized images of books, unless the books are already packaged as data.

The future is in data.

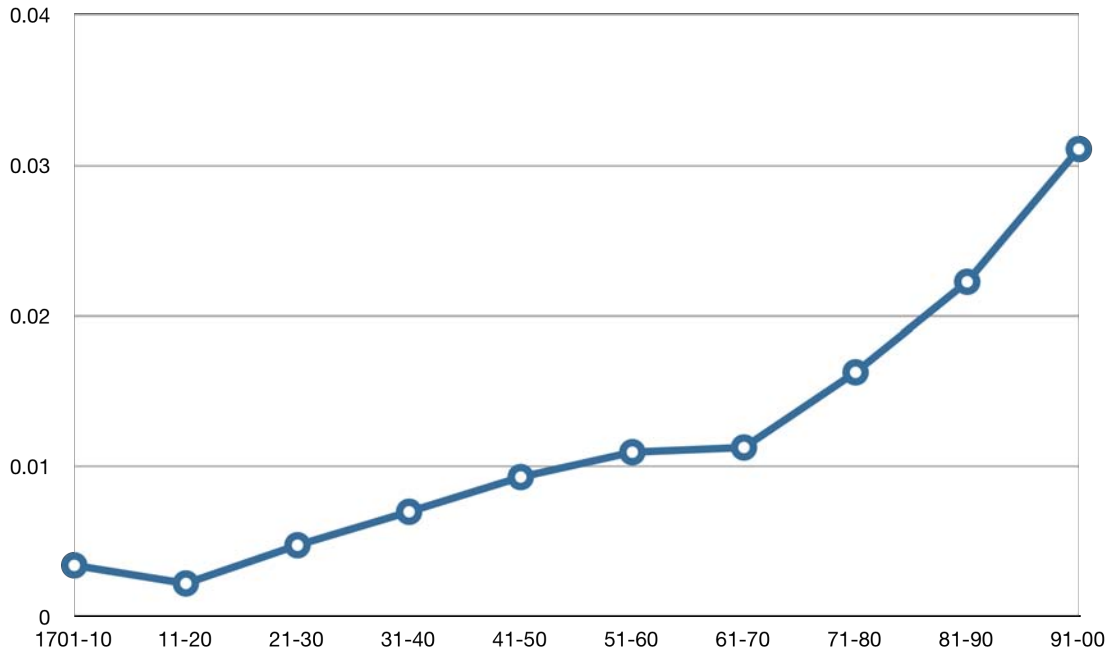


Image: Relative frequency of works including the word "data," 1701-1800, generated by analysis of ECCO I.

In my own work in ECCO, I began with a simple word search, identifying works that contained the word "data," year by year. Because ECCO is bounded and ultimately not that big—it contains only 136,000 books—it was practical, if time consuming, to examine every one of the approximately 10 thousand works in which the term "data" appears, and to apply a well-tested technique for analyzing and classifying them. I'll call this technique "reading."

There is much I could say already about this adventure. Looking closely at these usages revealed a good deal about what ECCO can and can't do. There were lots of scanning errors. Words such as "date" and "dare" were sometimes mistaken for "data." In many instances the word "data" was not read at all. Numerical calculations below the book level were very difficult.

Most importantly, ECCO (and this is true of Google too), does not distinguish between the Latin word "data" and the English. And this poses a problem when looking at frequencies. But once I separated Latin from English, usage trends emerged very nicely.

As I've said, my research in this area is still preliminary, but since it has already turned up some results that add nuance to the broader picture painted by Google, let me conclude by highlighting just a few:

First, the term "data" entered the English language in the seventeenth century and became naturalized in the eighteenth.

Based on results from ECCO, it appears that the term "data" appeared with increasing frequency during the eighteenth century relative to the total textbase. During the eighteenth century, "data" remained principally a term of art. Yet, by century's end, its range had been extended to a variety of new disciplines, and its use had become much more common.

Of course, as the Ngram we looked at earlier indicates, the term "data" would not receive a broad cultural application until later. In the last decade of the eighteenth century, less than 4 percent of total works included in ECCO employ it. By contrast, the term "fact" appears in about 28 percent of works. But, the trend for "data" is notable: over the course of the century, its relative use increases by about a factor of ten.

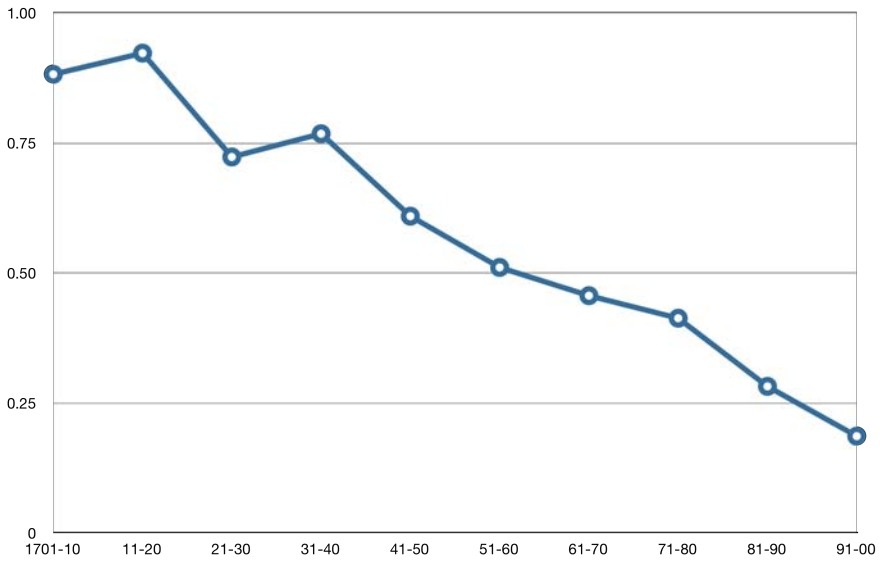


Image: Percentage of instances where term “data” is italicized.

Moreover, at the beginning of the eighteenth century, approximately 70% of published instances of the term “data” were italicized, suggesting that users still regarded it as a foreign word. By the end of the century, only about 20% of instances were italicized.

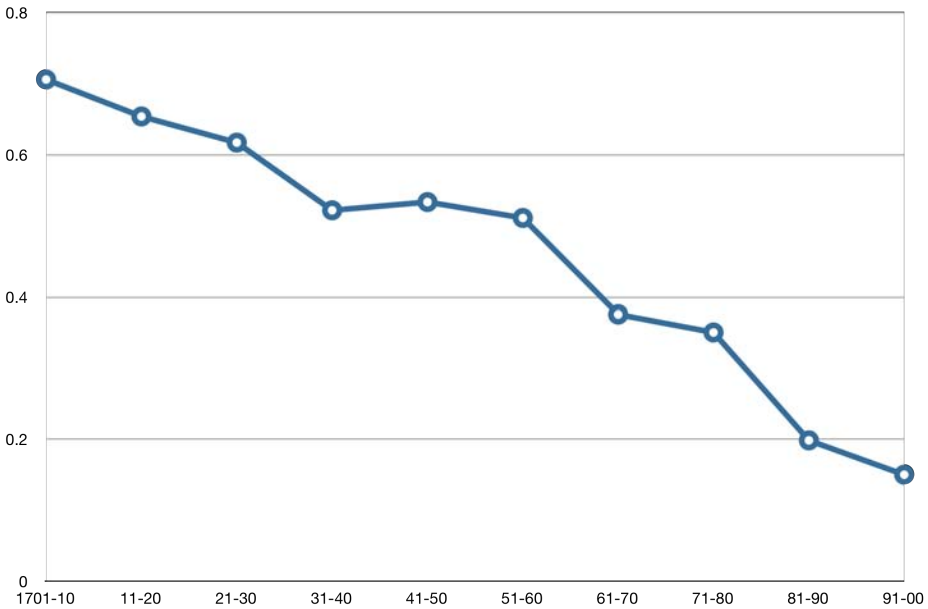


Image: Fraction of total usages of “data” in ECCO I pertaining to Mathematics and Theology.

Second, “data” came into English principally through discussions of mathematics and theology. By the end of the eighteenth century, dominant usages were in new and largely empirical areas of study including finance and natural history.

Third, over the course of the eighteenth century, the main sense of the term “data” shifted. At the beginning of the century, it usually referred to principles, facts, or values given and not susceptible to question. At the end of the century, the term typically referred to facts in evidence determined by experiment, experience, or collection. It had not only become possible but usual to think of data as the result of investigation.

This represents a near total semantic inversion. And while this inversion did not *produce* the twentieth-century meaning of data, it did provide one of its key enabling conditions.

In sum, the work so far has shown that there are definitive quantifiable trends in both the currency and usage of the term “data” in the eighteenth century. It took some fairly heavy manual work with the data derived from ECCO to get a good read on this, but having done it, it is clear that the very first tool that I employed in my pursuit of the history of the term, the *Oxford English Dictionary*, produced an account that fairly matches the quantitative results.

I suppose, in some respects this observation should be disappointing. After all, I did a lot of work creating a richly coded body of data on data only to find that nineteenth-century crowd sourcing had already discovered what my work confirms. But, to the contrary, I find it very interesting just how good the OED turns out to be on this matter.

For the moment, it's a win for nineteenth-century practices of reading, but don't expect this to hold up for long. If you follow the various strategies of the online OED, you know that even *that* venerable institution is moving to embrace a more data-driven model. And that fact alone suggests that we should all be ready to engage with the quantitative humanities in a strong, critical fashion.

In any event, I do think that my eventual results will be good news for reading even if they are not bad news for data.

What is more, as we have seen with Priestley, the techniques made possible by the data-fication of our archive are many ways consistent with ideas and writing native to the eighteenth century. In other words, at least in this corpus, there is a kind of pleasing echo of the material in the techniques.

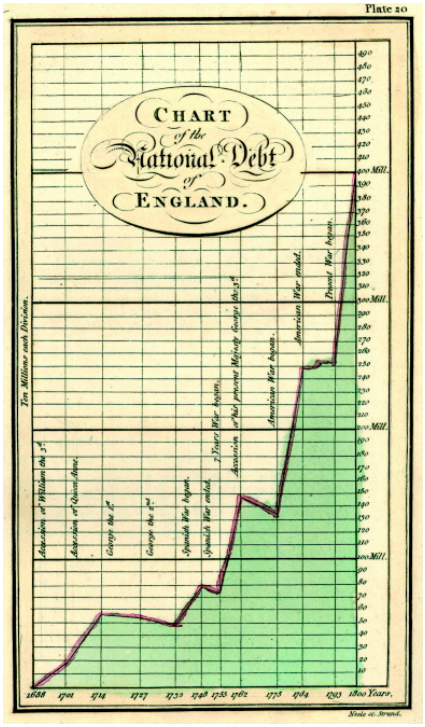


Image: William Playfair, Line graph from *Commercial and Political Atlas,* 1786. Playfair's *Atlas* was the first work to systematically employ the line graph.

In the end, what does the history of the term data have to tell us about data today? I think I've made a case for several possible answers, but to conclude, let me emphasize one that is supported by the numbers but not generated by them.

From the beginning, data was a rhetorical concept. "Data" means that which is given prior to argument. As a consequence, its sense always shifts with argumentative strategy and context—and with the history of both. The rise of modern natural and social science beginning in the eighteenth century created new conditions of argument and new assumptions about facts and evidence. But the pre-existing semantic structure of the term "data" gave it important flexibility in these changing conditions.

It is tempting to want to give data an essence, to define what exact kind of *fact* it is. But this misses important things about why the concept has proven so useful over these past several centuries and why it has emerged as a culturally central category in our own time. When we speak of "data," we make no assumptions about veracity. It may be that the electronic data we collect and transmit has no relation to truth beyond the reality that it constructs. This fact is essential to our current usage. It was no less so in the early modern period; but in our age of communication, it is this *rhetorical* aspect of the term that has made it indispensable.