Books in Space: Adjacency, EEBO-TCP, and Early Modern Dramatists

Michael Witmore
Folger Shakespeare Library
mwitmore@folger.edu

Jonathan Hope
University of Strathclyde
jonathan.r.hope@strath.ac.uk

Imagine being given the keys to your own personal copyright library, containing every book printed in English since 1450 to the present day. You rush to the door and open it, keen to start exploring—but once inside the huge space you find there is a problem. There are miles of shelving, rooms and rooms of books, but there is no catalogue, and there seems to be no principle governing the arrangement of books on the shelves. You rush from floor to floor, scanning titles and authors you do not recognize, desperately looking for a familiar text. You have a Ph.D. in English literature, but none of these books were on the syllabus: there are hundreds of thousands of them, millions. You are lost.

This is the situation we are all about to be in—indeed, are all, to some extent, already in. A series of book digitization and transcription projects—EEBO-TCP, ECCO-TCP, HathiTrust, Google Books—is making almost every English printed book available.[1] Already, anyone can download 25,000 phase 1 TCP texts—with another 40,000 to follow. But "made available" is not the same as "made useable." Such an increase in the amount of available data has all kinds of effects: practical ones in terms of storage and processing; methodological ones in terms of how we manipulate it and measure it; and theoretical ones in terms of how it changes our subject or object of study. In this essay we will explore these effects, and suggest ways in which we can deal with them. We'll

[1] For EEBO-TCP (Early English Books Online), see http://www.textcreationpartnership.org/tcp-eebo/; for ECCO-TCP (Eighteenth Century Collections Online), see http://quod.lib.umich.edu/e/ecco/; for HathiTrust, see http://www.hathitrust.org; for Google Books, see https://books.google.com/intl/en/googlebooks/about/history.html—and also the alternative front-ends to the Google, and other, corpora built by Mark Davies: http://corpus.byu.edu. Text Creation Partnership to: "EEBO-TCP (Early English Books Online—Text Creation Partnership)" and "ECCO-TCP (Eighteenth-Century Collections Online—Text Creation Partnership)."

ISBN 978-0-86698-725-7 (online) ISBN 978-0-86698-557-4 (print)
New Technologies in Medieval and Renaissance Studies 6 (2016) 9–34

focus on EEBO-TCP for examples, as it is the data set most readers will know about or be familiar with (and it is the one we are most familiar with).

At the time of writing, anyone with an Internet connection can go to the following URL and download a file called TCP.csv:

> https://github.com/textcreationpartnership/Texts/blob/master/TCP.csv

This file lists all of the text files in EEBO-TCP—both those freely available as "Phase 1," and those limited to subscriber access. The csv file should open in any spreadsheet or statistics program as a structured spreadsheet with 61,315 rows (in the version available in April 2015)—corresponding to 61,315 text files, or "books."

Let's read *Hamlet*.

Text files in TCP are "named" numerically rather than with the titles of the books they contain. The TCP.csv file has the TCP number for each volume transcribed by TCP, as well as metadata such as "title," "date," "author," "terms" (roughly, subject or genre), and "pages" (i.e., length). So to read *Hamlet*, we need to find out the relevant TCP number from the csv file. Once we have that, we can download the right file from:

> https://github.com/textcreationpartnership

Let's search the csv file for "*Hamlet*."

If you have access to a networked computer, it might be instructive for you to try to do this: download the csv file, open it in a spreadsheet program, and do a search for the string "hamlet." When we did this with the csv file open in the software we use for data analysis (a commercial statistical package called JMP), we got a series of hits (depending on the software you use, and any updates to the TCP.csv file, you may get different results). The first hits are texts by authors with "hamlet" in their name. For example:

> A12788 Spenser, John, 1559–1614.; Marshall, Hamlett. 1615 A learned and gracious sermon preached at Paules Crosse by that famous and iudicious diuine, Iohn Spenser ... ; published for the benefite of Christs vineyard, by H.M. Bible. -- O.T. -- Isaiah V, 3–4 -- Sermons.; Sermons, English -- 17th century. 60

> A56269 Puleston, Hamlet, 1632–1662.1661 Monarchiæ Britannicæ singularis protectio, or, A brief historicall essay tending to prove God's especial providence over the Brittish monarchy and more particularly over the family that now enjoys the same / by Hamlett Puleston ... Monarchy -- Great Britain. 67

Then there's a hit in a title, but it's from a book about "villages and hamlets":

> A08306 Norden, John, 1548–1625?; Keere, Pieter van den, ca. 1571–ca. 1624, engraver. 1593 Speculum Britanniae. The first parte an historicall, & chorographicall discription of Middlesex. Wherin are also alphabeticallie sett downe, the names of the cyties, townes, parishes hamletes, howses of name &c. W.th direction spedelie to finde anie place desired in the mappe & the distance betwene place and place without compasses. Cum priuilegio. By the trauaile and vew of Iohn Norden. Anno 1593; Speculum Britanniae. Part 1 Middlesex (England) -- Description and travel -- Early works to 1800. 140

Then we get a play called *Hamlet*, but it is Davenant's adaptation:

> A59527 D'Avenant, William, Sir, 1606–1668.; Shakespeare, William, 1564–1616. Hamlet. 1676 The tragedy of Hamlet, Prince of Denmark as it is now acted at His Highness the Duke of York's Theatre / by William Shakespeare. 94

Only after a while do we find TCP number A11959, which is the file name for the TCP transcription of the 1603 quarto of *Hamlet*:

> A11959 Shakespeare, William, 1564–1616. 1603 The tragicall historie of Hamlet Prince of Denmarke by William Shake-speare. As it hath beene diuerse times acted by his Highnesse seruants in the cittie of London: as also in the two vniuersities of Cambridge and Oxford, and else-where; Hamlet 66

Great. Super.

Except this is Q1—the so-called "bad quarto," and very different from the text most of us are used to reading and seeing performed. It is nice to know that this important variant text is available in TCP, but we really wanted to read something closer to the "standard" text. We search on, but there is no sign of Q2 *Hamlet* (1604), on which most modern editors base their texts. We

have fallen afoul of one of the principles of TCP: the aim is to include one version of each text, not multiple editions, even if they are very different, as is the case with Q1 and Q2 *Hamlet*.

Then a great idea strikes us: surely Shakespeare's first folio is in? And of course, searching for "hamlet" will not find that, as *Hamlet* is not mentioned in the title of the collected volume. So, we remember that the title of the folio is *Mr William Shakespeares Comedies, Histories & Tragedies*, and we do a search on "mr william shakespeare." Nothing. Unable to believe that the first folio is not in TCP, we sort the spreadsheet by "date" and scroll to the section for 1623. There it is:

> A11954 Shakespeare, William, 1564–1616.; Heminge, John, ca. 1556–1630.; Condell, Henry, d. 1627. 1623 Mr. VVilliam Shakespeares comedies, histories, & tragedies Published according to the true originall copies.; Plays 916

Our initial search failed because "William" has been (accurately) transcribed "VVilliam."

So now we know that the first folio has the TCP number A11954. Before we download this file, though, let's reflect on this as a search: we were looking for something we knew existed; something we are very familiar with. Our familiarity with Shakespeare enabled us to know that A59527 was an adaptation, and that A11959, while interesting textually, was not exactly the text we wanted. It also enabled us to guess where to look to find the text, even though the title of the file gives no indication that it is present. This is fine for Shakespeare, but there are over 60,000 texts in TCP—the vast majority by writers of whom we have virtually no knowledge. What are the chances of us finding something by one of them? If we search, and get no results, how will we know if this is a true or a false negative?

Putting that uncomfortable thought aside, let's to GitHub, armed with our TCP number: A11954. In fact, to save time, you can use this URL with any TCP number at the end to take you to the relevant text file:

> https://github.com/textcreationpartnership/A11954

Now we are in business: because this text is in TCP phase 1, we get a link on the text's GitHub page to a nice HTML edition:

> http://tei.it.ox.ac.uk/tcp/Texts-HTML/free/A11/A11954.html

as well as a live download link for an EPUB version, and a link to the page images on JISC Historical Texts (for UK users at subscribing academic institutions only—other users at subscribing institutions may have access to the page images via EEBO). If we had tried this with a phase 2 text, however, we might have found ourselves trying to open an XML-coded file—and if we had been able to do that, we might well have found a text marked frequently with encoding errors. Not only is everything not there; what is there is not perfect.[2]

It is worth spending some time playing with the TCP texts, if only to temper the excitement that has surrounded their release. For any collection of books to be usable, or understandable, we need paths through it—paths that allow us to do at least two very different things. First, we want to be able to find texts we already know about: "I want to read *Hamlet*." And even this apparently simple request, as we have seen, can be tricky. Secondly, and perhaps more importantly, we want to be shown texts we *don't* currently know about,

---

[2] TCP transcribers were told not to spend too much time trying to work out illegible sections of text. As they were working from the same images online subscribers to EEBO get, they were using digital images of microfilms of early modern books. EEBO users are well aware of the many problems with reading these images. Consequently, the TCP files have regular marked gaps where transcribers could not read the image. Martin Mueller is leading AnnoLex (http://annolex.at.northwestern.edu/about/), a project to enable the collaborative correction of errors and gaps.

In addition to the gaps in individual TCP texts, we should remind ourselves that TCP does not contain "the whole" of the early modern print record (although it is tempting, and easy, to fall into such rhetoric). TCP, of course, can contain only surviving printed texts—so we must bear in mind likely survival rates when using it to represent early modern culture (Alan Farmer has a forthcoming essay on survival rates of printed material). We must also remember not only lost manuscript material, but the huge amount of surviving manuscript material not in TCP—although we can also welcome projects like Early Modern Manuscripts Online: http://folgerpedia.folger.edu/Early_Modern_Manuscripts_Online_(EMMO). But even taking these materials out of consideration, TCP does not contain "everything." The aim was to have one copy of each *text*, not a copy of each *book*, so texts are usually included in one edition only. Anupam Basu's graph of TCP text counts against ESTC entries (http://earlyprint.wustl.edu/tooleeboestctexts.html) is a striking visual reminder of the difference this makes. Finally, we should remember that TCP is ongoing: texts are being added constantly. In a recent comparison between drama texts recorded in the Database of Early English Playbooks (http://deep.sas.upenn.edu/index.html), Beth Ralston found a number of dramatic texts not currently included in TCP (the data from this study, funded as part of the Visualizing English Print project, is available from http://winedarksea.org).

but which are relevant to our interests: "Out of these tens of thousands of texts, I want to read texts with a relationship to *Hamlet*."

Typically, in the past, such relational paths through collections of books have been based on humans making high-level, subjective comparisons between texts. The pathway of traditional literary history, for example, was made by clearing away the vast majority of books, leaving a narrative formed out of just a chosen few (Seneca leads to *Hamlet*, which leads to ...). More comprehensively, library science developed a subject-based arrangement, which placed books on similar topics close to each other, allowing the apparent serendipities of open-stack research: effectively projecting and visualizing the results of human-based content analysis in the three-dimensional spaces of the library. Crucially, both literary history and librarianship rely on metadata (title, author, date) and comparison-based sorting (this text is drama, that one is religious prose; this text is "good," that one is not).

How do we deal with this within the new digital collections? Increasing numbers of books have always been a problem for scholars, so perhaps we can learn from the past. Let's move away from GitHub and csv files for a moment, and think about medieval libraries and the impact of Renaissance humanism on the employment rates of carpenters.

Imagine you are standing in a medieval library.[3] There are a couple of things to note. There are not many books, and those that are there are stored in fixed positions, chained to desks. When in use they are placed open on a reading surface. When not in use, the books are generally stored lying horizontally, either on top of the desk, or on a shelf beneath the reading surface. This tells us a lot about the tenor of the medieval intellectual world, and the lack of climate control in medieval libraries. It is not too much of a caricature to say that medieval scholasticism meant that intellectual life was focused on a small number of authoritative texts. The main job of a library was to hold copies of that small number of texts, and replace them with new copies as they wore out with use and the depredations of damp, mould, cold, and heat. Books were stored horizontally in piles of one because (a) there were not very many, so space was not an issue; and (b) the clasps and decorative metalwork affixed to the covers of books would damage other books if they were piled on top of each other.

[3] Our discussion of libraries, storage, and cataloguing methods draws on the following sources (complete information can be found in the Works Cited): Balsamo 1990; Campbell and Pryce 2015; Leedham-Green and Webber 2006, especially the essays by Gameson, Sargent, and McKitterick; Norris 1939; Petroski 1999; and Webster 2015.

With the rise of Renaissance humanism, however, things changed. The manuscript hunters scoured Europe and beyond for "lost" works which were added to the store of classics that libraries might be expected to hold. Humanists started to write new books—some commenting on the old ones, and some even introducing new ideas. And of course the invention of the printing press allowed books to be reproduced more cheaply and in greater numbers than had previously been possible.

One consequence of this intellectual revolution was a storage crisis. Libraries had to find space for more books. In the Renaissance, librarians responded in two ways: they employed carpenters, and they rotated books through 90 degrees. The carpenters added shelves to the wooden reading desks of medieval libraries, both above the reading surfaces and in the hollows below them where seated scholars had previously been able to put their knees. At the same time, librarians began to store books standing vertically, rather than horizontally, since this was more space-efficient. The Renaissance can be thought of as a ballet of books, thousands of them spinning gracefully through 90 degrees so that they become upright in space.

People realized pretty early on in the Renaissance that having more and more books was not an unmitigated good. Not only was there nowhere to put your knees when reading, but the stalls carpenters were building above reading desks now blocked out the light from the small, low windows usual in medieval libraries. It is a rather nice paradox that the new books brought into being by the light of the Renaissance quite literally blocked out the light required to read them.

And there was another worry: how could you possibly read them all? The restricted medieval canon had something going for it in terms of removing the stress of the unread. As book production gathered momentum, however, scholars became uneasy in the face of all the new knowledge: how could any one person master it? Very soon, the first attempts at listing books, and significantly organizing and excerpting them, appeared—because, of course, it *is* a good thing to have more books, as long as you can have some kind of meaningful access to them.[4]

[4] The pioneer in this field was Conrad Gessner, whose work is frequently cited in several landmark studies of information management in the Renaissance and beyond: see Blair 2010, Krajewski 2011, and Rosenberg 2013. For specific work on Gessner, see Blair 2003; Nelles 2009; Rissoan 2014; and Rosenberg 2003.

Returning to our medieval library for a moment, if you wanted to find a book, you asked the librarian, whose job it was to know where the books were in physical space—on which table or shelf (and if the librarian died suddenly, libraries would become less usable). In the Renaissance, we get the rise of the shelfmark. Those interested in digital humanities would do well to read up on the history of libraries, bibliographies, and cataloguing techniques: this is a large part of what we do, and librarians are pretty good at it. Originally, of course, the shelfmark was exactly that: generally a three-part mark consisting of letters and numbers, which typically identified the press or bookcase ("A"), the shelf ("3"), and the book's position on the shelf ("11"). In our example, then, the book is the eleventh book on the third shelf on bookcase A. The book's position is fixed in physical space. And the position of the book is very likely to be decided by the librarian's assessment of its subject matter: bookcases contain all the books relating to a certain subject area.

This may seem pretty obvious, but in fact its implications are conceptually substantial. Libraries that organize books by subject are effectively three-dimensional search engines, physical instantiations of the Amazon algorithm that generates those "If you liked X you may also like Y" messages.

Scholars of our generation like to gush nostalgically about the serendipities of the open-stack library. Geoffrey Hill did it recently in one of his Oxford poetry lectures in the midst of a brag about never going online, claiming that going to libraries was better because it allowed for the serendipitous discovery of the book you needed, but did not know you wanted.[5] But in fact Hill is wrong—there's nothing serendipitous about these finds, despite the air of self-congratulation that usually accompanies narratives of such "discoveries"—"Wasn't I clever to spot this?" Well, no. The librarians and the catalogue were clever to place similar books proximately in physical space. Otherwise you'd have had to wander the stacks pulling books randomly off shelves.

Those "serendipitous" discoveries are thanks to the invention of relational cataloguing systems like Dewey, which constitute an advance on literal shelfmarks—one again occasioned by the needs of storage. Relational systems number books relative to each other and their subject areas, but have no necessary or fixed relation to the physical space of the library. They are meant to

[5] Hill 2013. This lecture was not published, but the audio is available at the link provided in the Works Cited. We are grateful to Mary Erica Zimmer for this reference, and apologetic to Professor Hill for singling him out. We have all made similar claims about open-stack research.

allow books to be shifted around in the library space as new books, shelves, rooms, floors, even buildings, are added. The introduction of such numbers is another significant event in intellectual history, severing the organization of cultural materials from the organization of the buildings they are stored in—it is driven by a practical need (the maximal utilization of space), and perhaps a theoretical development (the multiplication of categories of intellectual culture).

So the physical organization (and cataloguing system) of a library is one way of combating the information overload of the Renaissance, the Enlightenment, and the present. But of course, modern libraries have to deal with exponential rates of book publication, and the biggest new libraries deal with this by an even more radical severing of the subject–location relationship, and even by (actually or potentially) severing the relational links between books noted above.

Book storage hangers like the new Bodleian book depository, or the British Library facility at Boston Spa, store books in largely human-free zones. Books tend to be fetched by robot. They are identified by bar code. This may sound clinical, anti-human. Perhaps it is. But perhaps we should remember that humans are bad for books, with their coughs and sneezes and greasy fingers, and a liking for high temperatures and humid spaces. These new book depositories are atmospherically, environmentally kinder to books than human-ridden open-stack libraries. And there's another payoff. With books identified by bar code in vast hangars patrolled by robots, we can break the final link between subject and physical space. In these spaces, books can sit beside any other books. Geoffrey Hill is never going to be allowed in to wander randomly—and even if he were, he wouldn't fit between the towering shelves, which are not accessible to humans, however slim, and are far too tall for safe browsing. This gives us the potential of "organizing" these books in any number of ways—in the actual physical space of the facility, they may be added, and clustered, by date of cataloguing, or size. But our access to these books, since it can't be physical, is virtual—opening up all the possibilities of modern search engines. And here we have a parallel with the opportunities and methodologies afforded by the digitization of collections such as TCP: by digitizing all of our books, we enable multiple reorganizations—either radical or traditional.

So you can think of the digital humanities as a million different ways to organize the books on a shelf, allowing you to make "serendipitous" discoveries more frequently, and more mind-bendingly, than in any open-stack library.

Indeed, digital tools allow you to have your books on a hundred different shelves simultaneously—rather like the high-dimensional library imagined at the end of the film *Interstellar*. This ability to reorganize enables us to map out multiple pathways through the book collection. If we count the right things, we can recover texts, and relationships between texts, currently lost to literary history: but everything depends on us counting the right thing, and being able to interpret the results.[6]

So let's have a look at some of the things that can happen when you have easily countable texts. You can learn things about texts by counting pretty simple things, although the things you learn tend to be quite simple in themselves. But we will begin with some simple examples because they give us the basic principles. Let's count the frequencies of a few words in Shakespeare's plays, starting with the word "king."[7]

Table 1 arranges Shakespeare's plays by relative frequency of the word "king," from highest to lowest. When we do this, the results are impressive, if predictable: all the histories go to the top. Only one non-history play gets amongst them: *King Lear* (which Shakespeareans will remember is actually called a history in its quarto publication). But this is telling us something really very obvious: there are lots of kings in the history plays, so the word king gets used a lot. Behold the golden new dawn of digital humanities!

---

[6] Here the paradigm shift turns back to literary scholars: computer science and statistics (and corpus linguistics) have an array of well-established techniques for counting and analyzing—but the decision of what to count, and the analysis of the results, can only sensibly be made by literature specialists.

[7] Before we count, let's note that an apparently simple phrase like "We'll count the word *king* in each Shakespeare play" makes a whole host of assumptions and covers up a lot of work and thinking: which plays do we mean by "each Shakespeare play"? Are we including *The Two Noble Kinsmen, Pericles, Sir Thomas More*? Which texts will we use? Paper, electronic—Q1 or folio? Edited? Unedited? Which parts of the text will we count? All of the text, or just those parts spoken on stage (i.e., not speech prefixes or stage directions, or running titles at the top of pages)? How will we define "king"? As just the character string < _king_> or as the "lemma" *king(n)*, including *kings*, *king's*, but not including *king (v)*? How will we report our results? As raw totals for each play, or as standardized relative frequencies adjusted for length to allow direct comparison? In this case, we "simply" counted the string <_king_>, using our project's web-based text tagger Ubiqu+Ity (http://vep.cs.wisc.edu/ubiq/)—which automatically reports results in relative frequencies. The edition of Shakespeare used was that of the Folger Digital Texts (http://www.folgerdigitaltexts.org).

Table 1. Shakespeare's plays arranged in order by relative frequency of the word "king" (highest to lowest)

Play	Genre
Richard II	History
3 Henry VI	History
2 Henry VI	History
Henry VIII	History
Henry V	History
1 Henry VI	History
Richard III	History
King John	History
1 Henry IV	History
King Lear	Tragedy
2 Henry IV	History
Hamlet	Tragedy
Pericles	Late
Winter's Tale	Late
Macbeth	Tragedy
Tempest	Late
Love's Labour's Lost	Comedy
All's Well	Comedy
Cymbeline	Late
Titus Andronicus	Tragedy
Two Noble Kinsmen	Late
Antony & Cleopatra	Tragedy
Midsummer Night's Dream	Comedy
Measure for Measure	Comedy
Troilus & Cressida	Tragedy
Julius Caesar	Tragedy
Two Gentlemen of Verona	Comedy
Twelfth Night	Comedy
Merchant of Venice	Comedy
Merry Wives	Comedy
Romeo & Juliet	Tragedy
Taming of the Shrew	Comedy
Much Ado about Nothing	Comedy
As You Like It	Comedy
Othello	Tragedy
Coriolanus	Late
Comedy of Errors	Comedy
Timon of Athens	Tragedy

But note the principle of what we've done here: we have rearranged our books on the shelf, using an unusual method (frequency of one word, rather than alphabetical order of title, or date of writing). This has had the effect of isolating a single group normally identified by other methods, and has picked out a further play with a potentially interesting generic relationship to that group. Strolling round the library of serendipity, we have apparently stumbled on an idea for a scholarly essay (albeit not a very original one).

We have also established a quantitative test for identifying the genre of Shakespeare's plays. Based on this evidence, if someone discovered a previously lost play by Shakespeare, and we wanted to find out if it was a history or not, all we'd need to do would be to count the frequency of the word "king" in it. Above a certain level, we'd be happy saying it was a history; below a certain level, we'd start to think it was something else.

Of course, that's a bit of a daft way to decide if a newly discovered play is a history or not. What we'd really do is get humans to read it and argue about the genre for a bit—and it would probably be a short argument because humans are pretty good at ascribing texts to genres. We have a good idea of what makes a play a history play, and it is not the frequency of the word "king": we ascribe a play to the genre "history" on the basis of relatively high-level features such as its relationship to its source material, and its thematic concerns. We can see, given this definition of a history play, why Shakespeare's history plays have a high frequency of the word "king." This is a statistical fact about history plays, but not, we would suggest, a very interesting fact. It is not very interesting because it is not surprising—it doesn't tell us anything we did not already know, or challenge our assumptions about history plays.

What happens if we count a different word? Let's try "love." Table 2 shows Shakespeare's plays arranged in order of frequency of the word "love." The comedies now come to the top, with an extra added tragedy, *Romeo and Juliet*, hardly a surprise.

Table 2. Shakespeare's plays arranged in order by relative frequency of the word "love" (highest to lowest)

Play	Genre
Two Gentlemen of Verona	Comedy
Midsummer Night's Dream	Comedy
Romeo & Juliet	Tragedy
As You Like It	Comedy
Love's Labour's Lost	Comedy
Much Ado about Nothing	Comedy
Twelfth Night	Comedy
Two Noble Kinsmen	Late
Taming of the Shrew	Comedy
Othello	Tragedy
Merchant of Venice	Comedy
All's Well	Comedy
Troilus & Cressida	Tragedy
Richard III	History
Hamlet	Tragedy
Merry Wives	Comedy
King Lear	Tragedy
King John	History
Timon of Athens	Tragedy
Julius Caesar	Tragedy
3 Henry VI	History
Antony & Cleopatra	Tragedy
Henry V	History
Richard II	History
Pericles	Late
Comedy of Errors	Comedy
Measure for Measure	Comedy
1 Henry VI	History
1 Henry IV	History
Macbeth	Tragedy
Titus Andronicus	Tragedy
Cymbeline	Late
Coriolanus	Late
Winter's Tale	Late
Henry VIII	History
2 Henry IV	History
Tempest	Late
2 Henry VI	History

The separation here isn't as good as with histories, but it is not bad—and note a couple of "interesting" results:

1. two comedies are very low on "love": *Comedy of Errors* and *Measure for Measure*;
2. four of the late plays are right at the bottom of the "love" ladder.

No doubt literary scholars could offer several theories about these findings, and this illustrates one of the benefits of counting things and rearranging books on the shelf. We don't really do it to allow us to identify newly discovered histories or comedies: we do it to suggest new questions about texts we already know. What is it about *Comedy of Errors* and *Measure for Measure* that makes them behave differently in this case? What shifts between the main comic group and these plays? Why are so many of the late plays so low in "love"? One of their typifying features is supposed to be the redemption of parents through the love of lost and rediscovered children, so this is a surprising result.

Note here that we are not only generating questions for humanities scholars; we are observing absence. Humans are pretty good at seeing things that are present in texts—especially things that happen relatively infrequently—but we are not very good at spotting things that aren't there, or are there relatively less frequently. Computers, being undiscriminating, are really good at this.

For our final word, let's pick "might"—the results of our count are in Table 3.[8]

[8] In choosing "might," we were inspired by the work of Lynne Magnusson on modal verbs in Shakespeare (for example, Magnusson 2009).

Table 3. Shakespeare's plays arranged in order by relative frequency of the word "might" (highest to lowest)

Play	Genre
Twelfth Night	Comedy
Hamlet	Tragedy
All's Well	Comedy
Pericles	Late
Two Noble Kinsmen	Late
2 Henry IV	History
Antony & Cleopatra	Tragedy
Timon of Athens	Tragedy
Measure for Measure	Comedy
Winter's Tale	Late
Tempest	Late
Julius Caesar	Tragedy
As You Like It	Comedy
Cymbeline	Late
Othello	Tragedy
Coriolanus	Late
King John	History
Richard III	History
Henry VIII	History
1 Henry VI	History
Love's Labour's Lost	Comedy
Merry Wives	Comedy
King Lear	Tragedy
2 Henry VI	History
Comedy of Errors	Comedy
Two Gentlemen of Verona	Comedy
Troilus & Cressida	Tragedy
Midsummer Night's Dream	Comedy
3 Henry VI	History
Macbeth	Tragedy
Much Ado about Nothing	Comedy
Henry V	History
Merchant of Venice	Comedy
Titus Andronicus	Tragedy
Richard II	History
1 Henry IV	History
Romeo & Juliet	Tragedy
Taming of the Shrew	Comedy

Look at how the late plays all come to the top—the only genre completely in the top half of the table. This result makes us want to investigate hypothetical and speculative language across Shakespeare's career—and it makes us wonder if he writes more definitive language early in his career, and gradually becomes less certain, more indefinite, as he develops.[9]

So far, we've been reorganizing our books along one shelf—one dimension—at a time, using just one frequency count. We could do the same reordering for anything we can count: nouns, verbs, murders, marriages, references to the Bible, scenes with more than three speaking characters. But we don't have to use just one dimension all the time. We can look at two frequency counts at once, arranging our books on a two-dimensional plane, as we do in Figure 1.

Here we plot Shakespeare's plays in a two-dimensional space, with the position of each play fixed by a pair of coordinates: the frequency of "king" on the vertical axis, and the frequency of "love" on the horizontal one. Now we can see something of the relationship between the two features. The "L"-shaped distribution of the dots, with a blank space in the top right of the plot, tells us that no play has high frequencies of both words—which might lead us to the republican hypothesis that where there are many kings there is no love—although more prosaically it shows that at high frequencies the words are negatively correlated, while at low frequencies there is no necessary relationship between them (a high value of one word is a reliable predictor of a low value for the other, but a low value of one does not reliably predict the value of the other).

[9] We have found some support for this notion in other studies of Shakespeare's texts: see Hope and Witmore 2014.

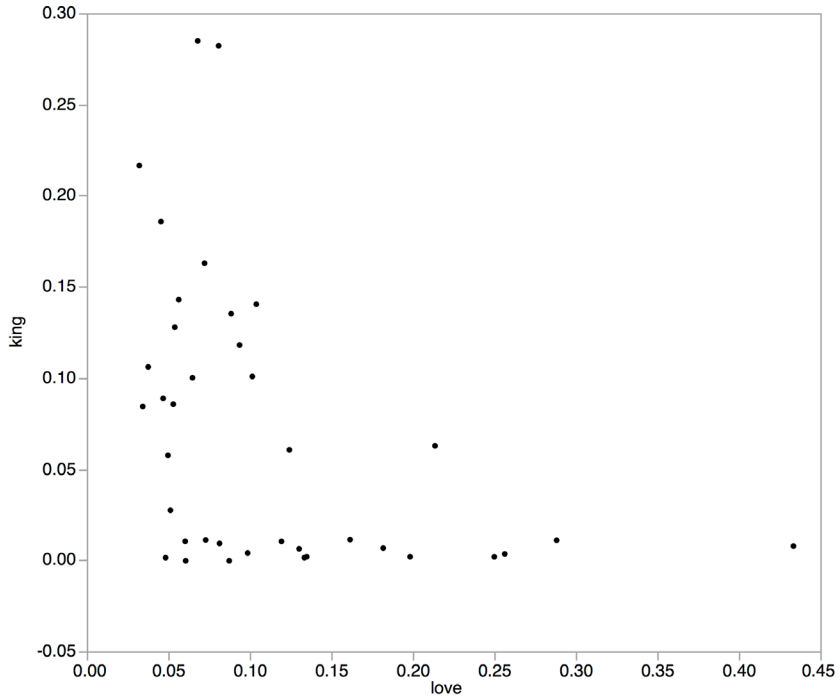

Figure 1. Shakespeare's plays plotted according to the relative frequencies of the words "king" and "love."

And there is no need for us to stop at two dimensions. We can further explore the relationships between our texts by adding a third feature to the plot, such as the frequency of the word "might," and projecting the texts into a three-dimensional space. But the things we've counted here—the frequencies of just three words, two of which are pretty obvious—restrict the new questions we are generating, and they do not capitalize on the other great advantage of computers: their ability to count lots of things across hundreds of texts. It would just about be possible for a human to count every instance of "king," "love," and "might" across all of Shakespeare's plays, but soon we will all have access to tens of thousands of texts. Digital tools allow us to work at scales beyond those that limit human readers, so let's shift up from three words in 38 Shakespeare plays, and start counting over 72 linguistic dimensions in 554 early modern plays. We are currently working on a project funded by Mellon to produce software tools, and methodologies, to allow humanities scholars to access and work with large corpora like EEBO-TCP and

ECCO. We are trying to find ways to allow humanities researchers to make sensible use of all of this data: how can they find what they want, and how can they find new things?

One obvious way to come to grips with an expanding data set is to begin with what you know and work outwards. So we began with Shakespeare, and we are working out to "the whole of" early modern printed drama—about 554 texts, depending, of course, on how you define "drama."[10] Rather than counting individual words, we've been using a piece of linguistic analysis software called DocuScope.[11] DocuScope counts functional language units and sorts them into groups called "Language Action Types" (LATs). Each LAT consists of words and phrases that have the same function—marking first person, for example, or encoding anger, or introducing turns in rapid dialogue. Because LATs are more sophisticated linguistically than individual words, counting them gives us a more complex, nuanced picture of what's going on linguistically in the texts we're comparing. DocuScope was designed to pick up the different things writers do when they try to achieve different things with their texts—and because it was designed using the OED as a base, it is surprisingly good at "reading" early modern English.

[10] Our current definition is "plays" printed before 1660 which either were performed, or were intended to be performed, or look as though they could have been performed (note the vagueness of the last category—some of our texts are closet dramas). A corpus of less tightly defined "dramatic" texts including masques, entertainments, and so on, would run to more than 700 texts. Including dialogues (a form often employed in philosophy and instructional texts) in the corpus would push that over the thousand mark—but dialogues were not intended for performance. The point here is that there is no single "right" corpus of early modern drama: our corpus of 554 attempts to be inclusive, but necessarily lacks any plays known only in manuscript, all lost plays, and several plays not yet transcribed by TCP. The needs of any one researcher are likely to differ from ours (we have already spoken to scholars who want to include Peele's Lord Mayor entertainments in "the" corpus)—and one of the aims of Visualizing English Print is to give scholars tools that allow them to construct their own corpus from TCP easily. Doing this kind of work is very good for the soul: it makes you define your object of study very precisely!

[11] For DocuScope, see https://www.cmu.edu/hss/english/research/docuscope.html. The language theory underpinning DocuScope, and the categories it sets up, are detailed in Kaufer et al. 2004. A number of studies illustrating its use in the classroom, and authorship work, are listed at http://wiki.mla.org/index.php/Docuscope.

Now, we just moved from one, to two, to three-dimensional analysis, by adding an extra feature at each stage—an extra frequency score. Effectively, we arranged our books on a line, a flat plane, and then in a cube, allowing us to see the relationships between the books, and in the case of two- and three-dimensional representations, between the features themselves. When we get beyond three dimensions, our puny human brains seize up: we can't imagine adding a fourth axis to the three-dimensional graph (or not without a lot of difficulty). But mathematicians have long known how to describe spaces with more than three dimensions, and once you regard these spaces as mathematical objects, there is no limit to the number of dimensions you can project data points, or books, into.

So we can continue adding dimensions to our virtual library right up to the number of features DocuScope counts—which happens to be 72.[12] The resulting spreadsheet has 72 values for each of the 554 plays: 72 coordinates fixing them precisely, if unimaginably, in 72-dimensional space. All well and good—but how can we look at this space to see where the books are on the shelves, which ones are serendipitously next to each other? The answer is that, having built up a space too complex for us to see, we use statistics to simplify it—effectively to throw away information we hope isn't important, allowing us to visualize the space in a form we can read. Now, there are lots of ways of doing this: statistics is an art, not a science. The method we have been using, not without qualms and unease, is principal component analysis (PCA)—a well-known method for looking at relationships between features in complex data sets.[13]

To cut a long and complex story short (which is, effectively, what PCA itself attempts to do), PCA collapses the axes making up our 72-dimensional space into a series of super-axes, each one of which attempts to summarize some of the variation, or space, of our original space. If we take the two super-axes (called "principal components") that together summarize the largest amount of the original space, we can use them to plot a two-dimensional graph, showing something of the relationships between our texts.

---

[12] In fact, the version of DocuScope we use here counts 113 LATs, but we use only 72 in the analysis because we discard LATs with very low frequencies.

[13] We give a fuller account of PCA in Basu, Hope, and Witmore (forthcoming). Most standard statistics textbooks cover PCA (and factor analysis, to which it is closely related); we have found Field (2013) useful. Literary scholars will probably get most out of Alt (1990), which is a brief and very clear conceptual account of what the statistical procedures are trying to achieve.

Figure 2 is a plot of the 554 plays which make up early modern drama in what we can call "PCA space." When you look at it, remember that you are looking at a huge simplification: 72 dimensions have been reduced to two. We have thrown away all but 26 per cent of the information in our 72-dimensional space. Let's hope we kept the important bits.

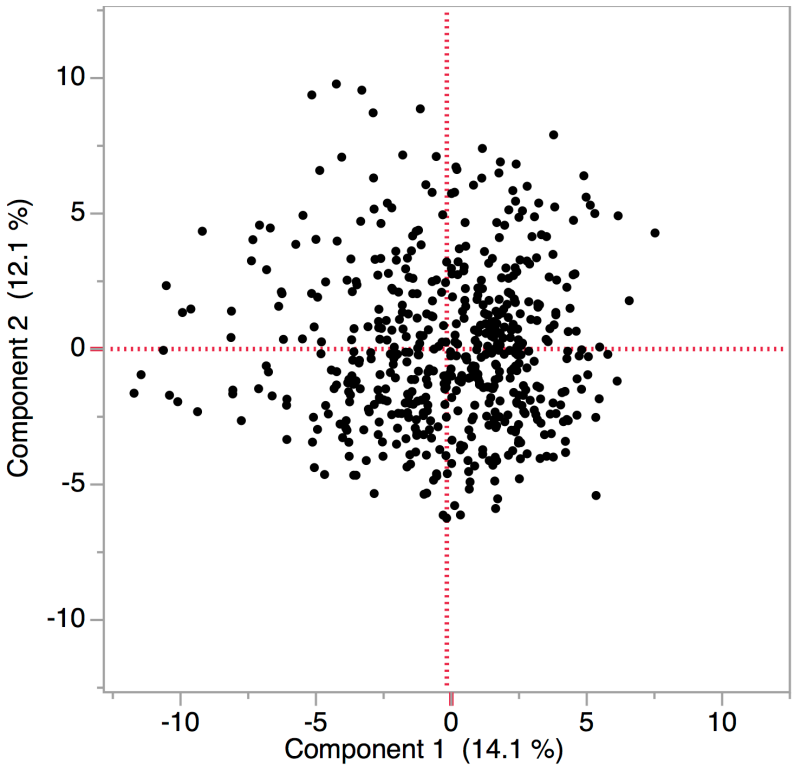

Figure 2. The corpus of 554 early modern plays visualized in PCA space.

So Figure 2 is our visual summary of a set of linguistic relationships in the corpus of early modern drama. Each dot is a play, and we can say, roughly speaking, that plays appearing next to each other use similar types of linguistic features at similar rates (and avoid similar groups of features). Conversely, plays a long way from each other will differ linguistically (we are hedging this because our statistical simplification may warp space, so we need to check things, but let's pretend all this is true for now). We now have an overview of early modern drama in linguistic space, and we can start to "read" it. One thing is clear: the dots are not evenly distributed across the

space. There are empty areas, especially in the bottom half, free of dots. So there are combinations of language that just don't get used in drama—why not? One interesting avenue of further research would be to add texts from different genres to this corpus to see if they occupy the spaces drama avoids: perhaps sermons, or a nascent genre like the novel would show up here—or perhaps these combinations are simply not used in English writing.

Another thing: the dots seem to clump in a circle roughly centred on the point of origin, with a sparse cloud of less densely-packed plays to the upper left. It looks as though most plays are broadly similar to each other, using the identified set of linguistic features at more or less similar rates. When playwrights vary from this norm, they do so by moving into only certain areas. Again, why should this be?

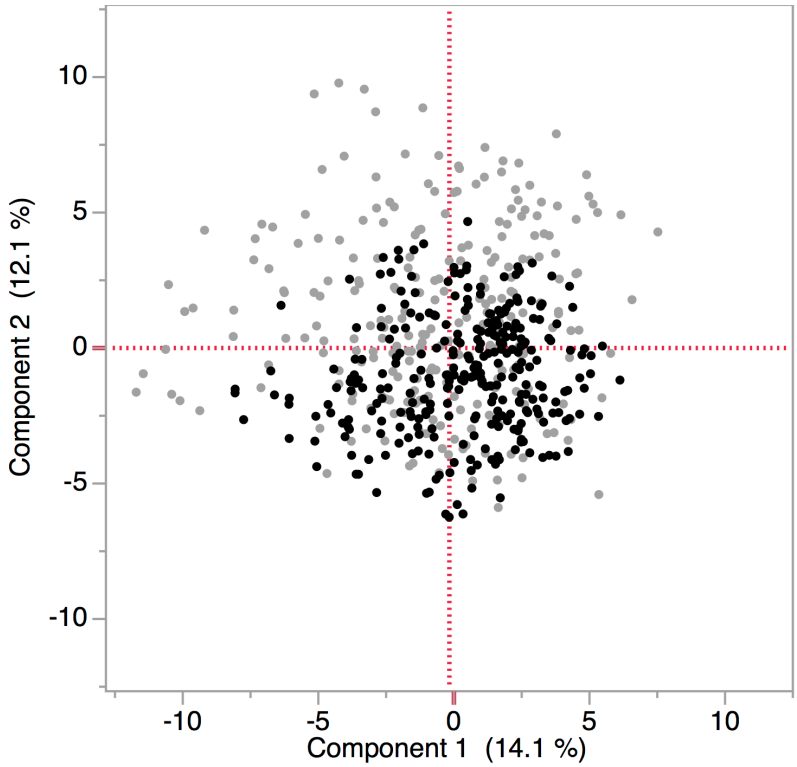

Figure 3: Early modern drama with 299 "career plays" highlighted.

We have been playing around with this data, and one thing we did was to pick out the plays in this sample written by "career" playwrights. These are writers who have written large numbers of plays, five or more—"professional" playwrights in the true sense. The big names of the period—Shakespeare, Fletcher, Dekker, Massinger, Middleton, and so on. When you add up the plays by these men, they total about 299. The rest of the plays in the sample are written by people who have written only one or two—or they are translations not intended for the professional stage. Figure 3 shows those 299 professional plays highlighted in black, with the rest in grey.

We found this result surprising—the clustering effect we noted above becomes even more pronounced. What this result suggests to us is that professional early modern playwrights were exactly that: professional. Professional in the sense that they knew what a professional play sounded like, and could hit the target every time. They are not, on this evidence, a set of daring experimenters, despite including some of the most celebrated names in English literature. The experimenters are the grey dots out in the upper left regions of the graph—these turn out to be translations from the classics, and mavericks either not writing for the stage at all, or writing once, and never being employed again.

Once again, this overview suggests several paths for future research. What is the relationship between date of writing and position on the graph? (We suspect that many outlier plays are either very early or very late.) What about genre? (We begin this discussion in a forthcoming paper.[14]) Can we learn anything about "successful" playwrights by looking at the outlier plays? Do certain types of play group in the outlier areas?

The answer to that final question is, in some cases, yes. For example, 12 of the outlier plays grouped along the horizontal axis to the left turn out to be translations of Seneca's tragedies (see Figure 4). Any history of the early modern stage will tell you that these plays are the foundation of, the key influence on, early modern drama as a whole, and early modern tragedy in particular. What do we make of the linguistic space separating Seneca from the professional dramas he is universally held to have influenced? Does the extreme language of the Senecan plays mean we need to reassess these claims for "influence"? Or can we read the distance from the Senecan region to that of the "core" early modern tragedies as "influence" or adaptation?

[14] Witmore, Hope, and Gleicher (forthcoming).

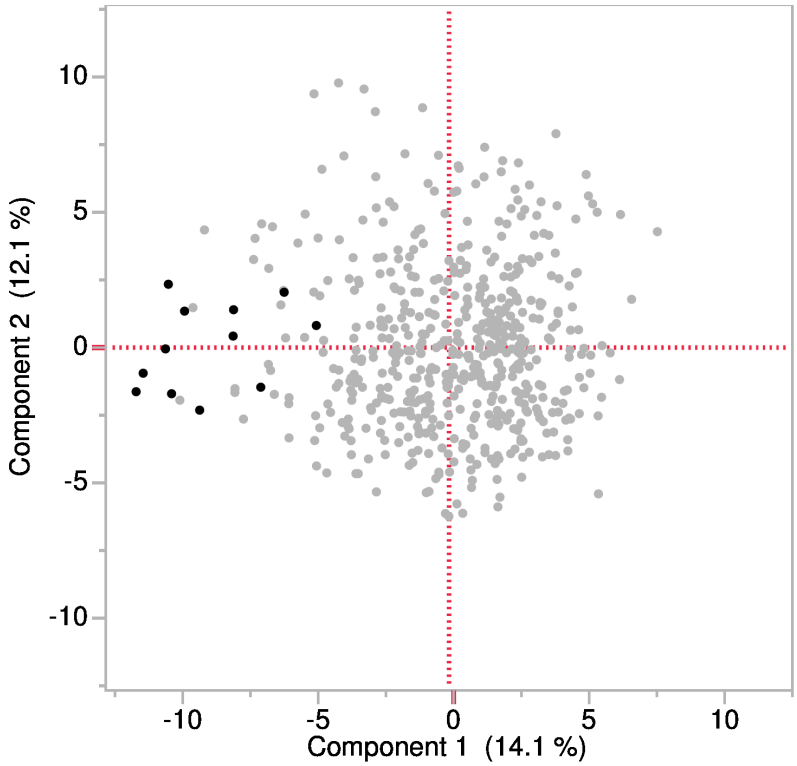

Figure 4. Early modern drama with translations of Seneca highlighted.

And what about Shakespeare? Shakespeare clusters with his professional colleagues in the central mass of dots—resolutely average, doing similar things to them, at similar rates. How do we reconcile this with our notion that Shakespeare is, by about as many orders of magnitude as you care to name, "better" than everyone else? Either our sense of Shakespeare as exceptional is wrong, or, whatever it is that makes Shakespeare great, we ain't counting it yet.

These two findings—that early modern professional dramatists stick together, and that Shakespeare sticks with them—chime with something other digital scholars have found in other periods. Ted Underwood, who works on nineteenth-century literature, has noted (2013) that the narratives of traditional literary history focus on rupture and revolution: break points triggered by the emergence of radically new individual genius. When he looks

for shifts in literary history in large digital corpora, however, he does not see sudden shifts. Instead, he sees similarity and continuity.

> Viewed through a large-scale lens, the subject itself becomes a different thing—a series of slow developments, incremental changes—and the story of genres slowly emerges from other types of text over decades, rather than springing fully-formed from the brow of one exemplary genius. Where does the novel really come from? What about scientific writing? How do sermons, by far the largest text-type for most of the period, but read by almost no one now, fit in? Up to now, literary history has worked a bit like PCA, by cutting away the unnecessary and hoping to leave behind the important. We had to do that because we didn't have the time, or the brain capacity, to read everything. But if you chuck most of the books away, it is hardly surprising that the ones remaining start to look like exceptional peaks rising above the plain, appearing without preparation. It is as if we had gone round the library pulling most of the books off the shelves, leaving only the odd one to represent whole subject areas or periods.
>
> Now, as the paradigm shifts, we are going round again, putting the books back—thousands of them, hundreds of thousands of them, most unread since they were published, containing who knows what. The only thing we have to do is learn how to find them, and then read them again.

WORKS CITED

Alt, Mick. 1990. *Exploring Hyperspace: A Non-Mathematical Explanation of Multivariate Analysis.* London: McGraw-Hill.

Balsamo, Luigi. 1990. *Bibliography: History of a Tradition.* Translated by William A. Pettas. Berkeley, CA: Bernard M. Rosenthal.

Basu, Anupam, Jonathan Hope, and Michael Witmore. Forthcoming. "The Professional and Linguistic Communities of Early Modern Dramatists." In *Community-Making in Early Stuart Theatres: Stage and Audience*, edited by Roger D. Sell, Anthony W. Johnson, and Helen Wilcox. Farnham, UK: Ashgate.

Blair, Ann. 2003. "Reading Strategies for Coping with Information Overload ca.1550–1700." *Journal of the History of Ideas* 64 (1): 11–28.

_____ 2010. *Too Much to Know: Managing Scholarly Information before the Modern Age*. New Haven, CT: Yale University Press.

Campbell, James W.P., and Will Pryce. 2015. *The Library: A World History.* London: Thames and Hudson.

Field, Andy. 2013. *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll*. 4th ed. London: Sage.

Gameson, Richard. 2006. "The Medieval Library (to c. 1450)." In *The Cambridge History of Libraries in Britain and Ireland*, edited by Elisabeth Leedham-Green and Teresa Webber, 13–50. Cambridge: Cambridge University Press.

Hill, Geoffrey. 2013. "'Legal Fiction' and Legal Fiction." Oxford Professor of Poetry Lecture, March 5. http://media.podcasts.ox.ac.uk/engfac/poetry/2013-03-21-engfac-poetry-hill-2.mp3 (audio only).

Hope, Jonathan, and Michael Witmore. 2014. "Quantification and the Language of Later Shakespeare." *Actes du congrès de la Société française Shakespeare* 31: 123–49. https://shakespeare.revues.org/2830.

Kaufer, David, Suguru Ishizaki, Brian Butler, and Jeff Collins. 2004. *The Power of Words: Unveiling the Speaker and Writer's Hidden Craft.* London: Routledge.

Krajewski, Markus. 2011. *Paper Machines: About Cards and Catalogues, 1548–1929.* Cambridge, MA: MIT Press.

Leedham-Green, Elisabeth, and Teresa Webber, eds. 2006. *The Cambridge History of Libraries in Britain and Ireland.* Vol. 1, *To 1640*. Cambridge: Cambridge University Press.

Magnusson, Lynne. 2009. "A Play of Modals: Grammar and Potential Action in early Shakespeare." *Shakespeare Survey* 62: 69–80.

McKitterick, David. 2006. "Libraries and the Organisation of Knowledge." In *The Cambridge History of Libraries in Britain and Ireland*, edited by Elisabeth Leedham-Green and Teresa Webber, 592–615. Cambridge: Cambridge University Press.

Nelles, Paul. 2009. "Reading and Memory in the Universal Library: Conrad Gessner and the Renaissance Book." In *Ars reminiscendi: Mind and Memory in Renaissance Culture*, edited by Donald Beecher and Grant

Williams, 147–69. Toronto: Centre for Reformation and Renaissance Studies.

Norris, Dorothy May. 1939. *A History of Cataloguing and Cataloguing Methods 1100–1850: With an Introductory Survey of Ancient Times.* London: Grafton and Co.

Petroski, Henry. 1999. *The Book on the Bookshelf.* New York: Alfred A. Knopf.

Rissoan, Bastien. 2014. "La gestion de l'information au XVIe siècle (2/2): La bibliographie universelle de Conrad Gesner." *Interfaces/Livres anciens de l'université de Lyon* blog post, July 24. https://bibulyon.hypotheses.org/4825.

Rosenberg, Daniel. 2003. "Early Modern Information Overload." *Journal of the History of Ideas* 64 (1): 1–9.

_____ 2013. "Data before the Fact." In *"Raw Data" is an Oxymoron*, edited by Lisa Gitelman, 15–40. Cambridge, MA: MIT Press.

Sargent, Clare. 2006. "The Early Modern Library (to c. 1640)." In *The Cambridge History of Libraries in Britain and Ireland*, edited by Elisabeth Leedham-Green and Teresa Webber, 51–65. Cambridge: Cambridge University Press.

Underwood, Ted. 2013. *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies.* Stanford, CA: Stanford University Press.

Webster, Keith. 2015. "Redefining the Academic Library—Revisiting a Landmark Report." *Library of the Future* blog post, April 13. http://www.libraryofthefuture.org/blog/2015/4/13/redefining-the-academic-library-revisiting-a-landmark-report.

Witmore, Michael, Jonathan Hope, and Michael Gleicher. (Forthcoming) "Digital Approaches to the Language of Shakespearean Tragedy." In *The Oxford Handbook of Shakespearean Tragedy*, edited by Michael Neill and David Schalkwyk. Oxford: Oxford University Press.