EDDAC or The Book of English:

Towards Digital Intertextuality and a Second-Generation Digital Library

By Martin Mueller
[Draft April 2009]

1 Introduction and Summary 2

1.1 An English Diachronic Digital Annotated Corpus (EDDAC) 2
1.2 Textkeeping or Distributed Collaborative Data Curation 2
1.3 The library as laboratory and provider of tools and services 2

2 EDDAC: English Diachronic Digitally Annotated Corpus 3

2.1 Legal obstacles to digital intertextuality 3
2.2 Half a loaf of EDDAC 3
2.3 Conditions for Digital Intertextuality: the good enough edition 5
2.4 Data curation to maximize digital intertextuality 6
2.5 A prototype of EDDAC: the text corpus of the Monk Project 8

2.5.1 Shared baseline encoding in the Monk Project 9
2.5.2 Linguistic annotation in the Monk Project 9

2.6 Adding more texts to EDDAC 9

3 Textkeeping or Distributed Collaborative Data Curation 11

3.1.1 Correcting orthographic and similar errors 11
3.1.2 Creating digital editions in a collaborative fashion 13
3.1.3 Correcting morphosyntactic errors 13
3.1.4 Adding new levels of metadata: identifying spoken language 13

4 EDDAC, Digital Intertextuality, and the Role of the Library 14

4.1 Libraries as the natural institutional home for EDDAC as a cultural genome 15

5 Works Cited 18

1 Introduction and Summary

1.1 An English Diachronic Digital Annotated Corpus (EDDAC)

In this essay I make a case for a project that consists of three distinct but overlapping components. The first is an English Diachronic Digital Annotated Corpus (EDDAC), in which

1. each individual text is an accurate transcription of an edition of some standing, is explicit about its provenance, and wherever possible is linked to a digital facsimile of its print source
2. the texts exist in the public domain, which in practice and for the foreseeable future limits such a corpus to texts published before 1923
3. the texts are linguistically annotated
4. the texts are consistently encoded and support a high level of 'digital intertextuality' in the sense that any subset of texts from this archive can be readily compared with any subset or the whole archive for a variety of literary, linguistic, historical, philosophical, or rhetorical purposes, whether directly or through the metadata associated with them.

Think of such a corpus as a Book of English or 'cultural genome', a metaphor to which I will return from a variety of perspectives. Between 5,000 and 10,000 texts would constitute a sufficient seed corpus to begin reaping the benefits of digital intertextuality. Whether such a corpus would benefit from growing beyond a range of 25,000 to 50,000 is an open question, which need not be answered until the archive has grown to that dimension.

1.2 Textkeeping or Distributed Collaborative Data Curation

The second component is a scholarly user community that is actively engaged in the task of building and keeping this corpus. User contributors should be 'textkeepers'. I coin this term on the analogy of housekeeping as an activity that goes on all the time in a humble, invisible, but essential manner. Distribute collaborative data curation or DCDC is a more technical name for this component.

Any inquiry is constrained by the quality of the data on which it rests. It is a formidable task to build and maintain a large diachronic and fully intertextual corpus sufficiently complex and accurate to meet high scholarly standards. Who has a greater stake in the quality of the data than the scholars whose work depends on them? We are here in the world of 'Wikinomics' or 'crowdsourcing'. Central to collaborative digital data curation is the idea that beyond the assembly of an initial seed corpus the scope and direction of further growth will result from the choices of users who want to add this or that text for this or that purpose. User-driven growth will provide the best direction over time.

1.3 The library as laboratory and provider of tools and services

The third component is consortial activity by academic libraries --for instance, the CIC libraries--to provide the logistical and technical framework for EDDAC and DCDC. This framework will also support the analysis tools needed to explore the textual resources created by a diachronic and fully intertextual digital corpus.

This will blur the traditionally clear distinction between libraries and publishers. It also involves involve substantial renegotiations of the implicit contracts that have governed the relationships of librarians and their patrons. Librarians are comfortable with the motto "More books for more readers." But with digital technology libraries need to think about 'enhanced' as well as 'extended' access. The distinction is fuzzy at the edges but clear in many contexts. It is one thing to grow by 'extending' access to more materials and more readers. It is another to grow by enhancing access to the materials you have. When Ranganathan formulated the fifth and final law of library science as "The Library is a growing organism" I take it that he had both means of 'growing' in mind.

Extended access uses digital technology in a emulatory mode and thinks of it as bringing more books to more readers. Enhanced access thinks of digital technology as providing new tools for a more sophisticated analysis of available materials. With enhanced access the distinction between catalogue information 'about' the book and information 'in' the books becomes increasingly blurred. With regard to the primary sources that constitute the evidentiary basis for text-centric scholarship, the concept of the 'finding aid' will increasingly involve tools that go beyond the catalogue record of a given book and help users look inside the book or across many books.

Extended and enhanced access are not in conflict, and the digital library of the future must deal with both. But it would be a mistake to sequence them and think that there is no need to enhance anything until you have extended everything. You could certainly make a strong case for the position that doing more things with the stuff you have has a greater pay-off than getting more stuff to do things with.

2 EDDAC: English Diachronic Digitally Annotated Corpus

2.1 Legal obstacles to digital intertextuality

As a corpus of fully interoperable primary texts, EDDAC should allow scholars to use digital texts and tools without constraint. The model here is the open-stack library in which researchers walk among shelves and are free to choose and analyze any combination of books for any purpose, subject only to the constraints of human feet, hands, and eyes.

The chief obstacles to exploring the affordances of digital texts in a single document space are legal rather than technical. The original intent of copyright legislation was clearly to protect intellectual property rights for a shorter than a longer time. Recent legislation has gone the other way (Danton 2009). For the foreseeable future, the benefits of full digital intertextuality will not be available to literary scholars whose work is anchored in literary texts since the 1920's, because commercially available digital texts are typically tied to particular access tools that severely constrain their use outside of the parameters envisaged by the vendor.

2.2 Half a loaf of EDDAC

Half a loaf is better than none, especially if it is large in its own right. More than half of the colleagues in my Department of English have their scholarly centre of gravity in texts before 1923. The percentage is about the same for graduate students, and a quick survey of my colleagues at the University of Chicago suggests similar proportions. I es-

timate that a comprehensive version of EDDAC would constitute a basic research tool for approximately half the faculty in research universities. For pedagogical work, the percentage is probably lower.

The intertextual affordances of EDDAC reach far beyond Literary Studies. The traditional range of "Letters" includes texts that lend themselves to forms of rhetorical, linguistic, philosophical, historical, political, social, or cultural analysis across a wide range of disciplines.

It might be instructive to use JStor as the basis for a study that looks at this range of texts and asks how many of them are in the public domain and have been cited more than twice in the secondary literature of the last fifty years. The result would probably identify a core group of texts and authors measured in the low thousands. It is a reasonable assumption that a Book of English, consisting of such an initial core collection and supplemented over time by user-contributed texts would meet important needs of a global scholarly community well into the middle of this century. Whether anybody thereafter will read anything written before 2000 is a question about which future scholars will vote 'with their feet'.

Another way of measuring an initial size of EDDAC points in a similar direction. Consider a collection of '1001 novels' from Sidney's *Arcadia* or Wroth's *Urania* to *Ulysses* compiled eclectically from various bibliographies. How often would this collection fail you if you wanted to follow up references from scholarly articles? Not very often. Now consider other genres, again taking a broad view of 'Letters'. How many books would it take to construct a library that covers other genres at the same density that 1001 texts achieve for fiction from the late 1500's to the early 1900's?

A collection of 10,000 books would include all the memorable and quite a few not so memorable texts. For the part of English literature that is no longer subject to copyright, a collection of that size constructed on the principle of a very high plateau of digital intertextuality would amount to a resource sufficiently comprehensive for many scholarly purposes. There is, however, one caveat. In any collection, however thoughtfully selected by somebody else, there will always books that are missing from the perspective of my particular project. From 'my perspective' therefore a successful EDDAC will need two components:

1. enough texts to make digital intertextuality a working reality for me
2. a procedure that lets me add additional texts I need, preferably in a manner that will be helpful to others as well.

Growth beyond the size required for an initial seed corpus should be driven by the needs of particular users who care enough to spend some of their own time and energy to add to the collection.

The Life Sciences provide a useful model. Evolutionary biologists carefully extract DNA sequences from specimens and contribute them to Gen Bank, an 'annotated collection of all publicly available DNA sequences'. GenBank is part of the International Nucleotide Sequence Database Collaboration. In this enterprise, the immense phenotypical variety of life is reduced to systematic description at the level of the genotype. Think of it as a Book of Nature, written in a four-letter alphabet, with collaboration and reduction as

both the cause and cost of scientific insight. The DNA sequences individual researchers contribute in a standardized format acquire much of their meaning from their incorporation into a large gene bank that support different forms of contextualization and analysis. One by one, the contributions of hundreds or thousands of biologists enrich the query potential of this resource. The Book of Nature and the Book of English, the biological and the cultural genome, both support exercises in digital intertextuality of a kind beyond the dreams of earlier scholars and scientists.

2.3 Conditions for Digital Intertextuality: the good enough edition

People often talk about 'digitization' as if it were the same thing, but digitization has many affordances and needs to vary with the purposes of the user. Robert Whaling is engaged in a digital edition of the manuscripts of George Herbert, a small but exquisite corpus. He asks why in one version of given poem a particular word is capitalized and whether the choice was the poet's or the printer's. He uses the affordances of the digital medium to draw the reader's attention to the minutiae of intra-textual variance, and like other scholarly editors, who have chafed under the constraints of a print-based *apparatus criticus* (read about as often as manuals of your computer), he is delighted by a technological tool that makes readers 'see' complex textual relationships (Isn't it poets that 'make you see'?). The side-by-side display of textual variance between the Ellesmere and Hengwrt texts of the *Canterbury Tales* in Estelle Stubbs' edition is driven by a similar delight in the power of digital tools to make you see textual variance.

At the other end of the scale, there is Google Books and the Hathi trust with its slogan "There is an elephant in the library." Here you are in a world of search engines that will find a needle in the haystack of millions of books and billions of other documents. The requirements of data curation in such an environment are completely different from those of a scholarly edition. This is not a matter of better or worse, but of different purposes. There are of course many more people who use Google than people who read Herbert or worry about the poetics of typographical choices. But the world would be a much poorer place were it not for the myriad of very small 'interest groups' that care passionately about things that few others care about.

The digital intertextuality of which I speak sits somewhere between the microscopic scale of intratextual variance and the astronomical, global, and transdisciplinary scale of Google Books. The objects are books from the past that for one reason or another are worth remembering. The purpose is to use digital technology to make these books talk to each other and to you. Literary scholarship is largely a matter of an endless conversation about the relationships of past authors to each other and to us, and like Michael Oakeshott's ship of state it has "neither starting-place nor appointed destination."

How can digital technology further this conversation and what standards of data curation are appropriate to such an enterprise? This question divides into two parts. What standards of data curation are appropriate to a particular text considered by itself, and what is required to maximize is query potential in a space of intertextual inquiry?

As for the first, a digital text must be a 'good enough'. I borrow the term from Winnicott's 'good enough' mother to define a level that is dangerous to drop below, while rising above it may for many purposes not add a whole lot. A 'good enough' edition is first of all an orthographically accurate transcription of a print source of some standing. It must be explicit about its provenance, and it must be citable.

From the perspective of a critical scholarly edition these are very modest goals, but they are typically not met by texts in Project Gutenberg, which are orthographically clean but more often than not bibliographically opaque. They are typically met by digital texts that have been encoded according to the Level 4 Guidelines by projects housed at Michigan, Virginia, Indiana, and North Carolina.

As for intertextual inquiry, texts from these collections are typically not easy to compare with each other. This does not matter so much for human readers, who are used to negotiate a great deal of stylistic and typographical variance and read everything on the level playing field of their understanding. But if you want to explore the power of what Gregory Crane calls 'machine-actionable' texts, encoding practices in different projects create hurdles that machines stumble over although human readers manage them effortlessly. A simple thing like the treatment of hyphenated words at the end of a line or page is a good example of the difference between man and machine in that regard.

From a theoretical perspective, it is possible to imagine a set of tools so 'smart' and comprehensive that they can take in arbitrary textual data and 'on the fly' perform the curatorial tasks that will guarantee a high plateau of digital intertextuality. Such tools would combine the smart but slow skills of human readers with the fast but dumb routines of computers. In practice, this is still Utopian. Text processing programs of any kind depend on a case logic: if . . else if . . else if ... ..else. The variety of typographical and text encoding practices is such that the construction of an adequate case logic for all kinds of texts is not an achievable goal. It is not only a matter of too many cases interacting in too many unpredictable ways. In any collection of digitized texts there are likely to be cases that do not yield to algorithmic treatment of any kind but require some human editorial intervention.

There are two choices. Either you take texts as they come, model them at the most primitive level as sequences of spellings, and see what you can do on that minimal level of interoperability. Or you move texts through curatorial processes that raise them to a plateau of digital intertextuality that supports more complex forms of analysis. While data curation differs from traditional scholarly editing in many ways, both involve intrinsically labor-intensive procedures. It takes ingenuity and patience of one kind to write and test the scripts that do the algorithmic part of data curation. It takes ingenuity and patience of another kind to remedy the cases that resist algorithmic treatment.

With digital data curation as with scholarly editing there is always a speculative element. Will the labor justify itself over time by the insights supported by the data in a new and enhanced format? Martin West and his students at Oxford spent years on the Teubner edition of the *Iliad,* which in its detail of textual witnesses and testimonia from later sources is much superior to any previous edition (West 1998). If the cost benefit analysis of this project measures the benefits in terms of what the edition does for the average reader of Homer in Greek the costs may seem excessive. The scholarly cost/benefit calculus runs differently. There may be lessons here for making similar calculations in the field of digital data curation.

2.4 Data curation to maximize digital intertextuality

The card catalogue of a library is the guarantor of intertextuality in a world of printed books. Think of the difference between books on shelves accessible through a model of

their order in the card catalogue and the same books scattered across the floor. The catalog defines the book as an 'object' and assigns it a place in a hierarchy of other objects.

'Object' is a big word in digital discourse. Programmers may speak of a 'book object' or a 'page object'. You might ask "why don't they just say 'book' or 'page'? The answer is that a book on the floor is just a thing. But a catalogued book is a 'book object' that is clearly defined through a set of relationships. Scholars read books rather than 'book objects'. But without the 'book objects' that are created and maintained through the cataloguer's activity, their work would grind to a halt.

When computers came into general use in the sixties, it was both an exciting and a difficult achievement to convert the catalog records of large library -- a million books or more -- into digital objects. Difficult because it strained the storage capabilities of the computers of the time. Exciting because it held out the promise of much more sophisticated manipulation of bibliographical data.

Today it is possible to extend the cataloguing of books to the word level. Think of EDDAC as a library of 'word objects' with something like a MARC record for each of them. In the sixties Senator Dirksen wryly remarked of the Federal budget: "a billion here and a billion there; pretty soon you're going to talk about real money." At least at the Federal level, a billion dollars has just become a rounding error. Similarly a billion 'word objects' or word occurrences with catalog records attached to them is a much smaller programming task today than cataloguing a million books was fifty years ago.

The transformation of texts into catalogs of word objects has been a centerpiece of the sub-discipline of corpus linguistics. The linguists call it 'annotation'. It can be done automatically with tolerable levels of accuracy (~97%), and it transforms the opening words of *Emma* into something like

Emma_name Woodhouse_name, handsome_adj, clever_adj, and_conj rich_adj

This does not tell human readers anything they do not know already, but that is not the point. Through the tedious process of 'explicitation' that injects some rudiments of readerly knowledge into the text the machine acquires a very pale simulacrum of human understanding. More importantly, it acquires powers that humans lack. If you have a large body of annotated texts the machine can at lightning speed retrieve all cases of three adjectives following each other. If each file searched by the machine is associated with 'metadata' about its author, date, genre, etc. the machine will dutifully report those association. In almost the twinkling of an eye you have the materials for the analysis of the 'three-adjective rule' on which Jane Austen consciously drew in the opening sentence of her novel and which she expected her readers to recognize. If there is an interesting story to be told about who uses three adjectives in what combinations and where, it is a story that, given a sufficiently large corpus, has moved within the grasp of a bright undergraduate.

Linguists, who are interested in low-level linguistic phenomena for their own sake, discovered the query potential of linguistically annotated corpora fifty years ago, and invested an extraordinary amount of data curation into the original Brown corpus of a million words of American English. Literary scholars and other humanists typically do not share this interest. On the other hand, there is very often an interesting path from low-

level observation of verbal usage to larger thematic or narrative patterns. Thus a linguistically annotated corpus is a powerful resource for many scholars who would not describe themselves as linguistically or philologically oriented.

Linguistic annotation of a particularly comprehensive kind underwrites most of the affordances of digital intertextuality. The German project DDD (DeutschDiachronDigital) makes this point very well (Lüdeling 2004). Linguistic annotation creates a descriptive framework that lets you describe 'word objects' or the molecular components of a text in a metalanguage that bridges orthographical or morphological variance due to differences in time, place, genre, social status, or other factors. The point is not to erase, but to articulate difference: words, phrases, sentences become comparable across large data sets. Readers do this for the few texts they can hold in their memory. Machines can help readers extend their memory in new and powerful ways.

EDDAC thus is a digital library of specially curated texts that are catalogued at the highest level of the 'book object' and the lowest level of the 'word object'. It is much harder to extend such cataloguing to the internal structural articulation of a text. You can successfully model just about any text as a sequence of sentence, but beyond the level of the sentence, the variance of internal structure among texts poses almost insuperable challenges to a structural metalanguage -- except for plays with their conventional division into speeches, scenes, and acts.

2.5 A prototype of EDDAC: the text corpus of the Monk Project

A fairly substantial prototype of EDDAC exists in the corpus of ~2,000 linguistically annotated texts (~150 million words) that were prepared from existing digital texts for the Monk Project (http://monkproject.org/). This corpus consists of

1. ~650 texts from the EEBO collection with special emphasis on plays, sermons, a heterogeneous collection of works from the birth of Elizabeth to the death of James I, and witchcraft texts
2. ~1,100 18th century from the ECCO collection
3. ~100 American novels from the public domain texts of the Early American fiction collection at the University of Virginia
4. ~300 American novels from the Wright archive of American novels 1851-75
5. The 'Library of Southern Literature,' a subset of 120 works from the Documenting the American South project

The EEBO and ECCO texts come from collections that will grow to 25,000 and 10,000 volumes respectively. All these texts will pass into the public domain after 2015. So will another 6,000 texts chosen from the Evans archive of American imprints and encoded in a similar fashion.

All the texts in the current Monk corpus have their origin in digitization projects at Michigan, Virginia, and Indiana that have a strong family likeness and follow the Guidelines for TEI Text Encoding in Libraries (http://www.diglib.org/standards/tei.htm), which were developed largely by a group of librarians at those institutions. The texts were converted to a TEI format that follows the new P5 standard. They were then tokenized, lemmatized, and morphosyntactically tagged.

2.5.1 Shared baseline encoding in the Monk Project

The conversion to a shared TEI P5 format was the work of Brian Pytlik Zillig and Stephen Ramsay at the University of Nebraska. This format is called TEI-Analytics. Its purpose is to create a 'machine-actionable' text so that users can instruct a machine to perform various analytical results .

Although the MONK texts originated in very similar shops, their conversion to a common format turned out to be a non-trivial task. While different projects made sensible decisions about how to do this or that they paid little attention to the needs of users who wanted to mix texts from different collections. The problems involved in such mixing are trivial if the uses of the digital text stay limited to looking up words and reading bits of text. But problems mount quickly if your goal shifts from 'extending' access (more people doing the same thing with more texts) to 'enhancing' access (doing more with the same texts). For this you need a higher degree of interoperability among texts. The soft hyphens mentioned above are the best example of a little thing that can make a big difference if it is handled in the same way, or at least in compatible ways, by different projects.

The conversion of different text archives into a common TEI P5 interchange format is similar in spirit to the *Kernkodierung* or 'baseline encoding' of the German Textgrid project (http://www.textgrid.de/). Textgrid aims at creating a distributed environment in which scholars can produce digital editions. Each of these editions uses markup to realize its particular goals, but the markup can be reduced to a baseline encoding that makes the texts in Textgrid interoperable. This is a particularly good example of reconciling the different perspectives of intratextual and intertextual analysis. There is much to be said for an environment in which different projects pursue their special needs on a high plateau of shared baseline encoding. The higher that plateau the higher and more granular the potential for intertextual analysis. In practice, the implementation of this principle means agreeing to do a lot of little things in the same way.

2.5.2 Linguistic annotation in the Monk Project

Linguistic annotation in the Monk Project was done with MorphAdorner, an NLP toolkit developed by Phil Burns at Northwestern University. MorphAdorner works with a tag set that can describe morphosyntactic phenomena from late Middle English (Chaucer) to the present. In addition to providing a part-of-speech tag for every word token, it also maps the spelling or surface form of each token to a standard spelling and to a lemma. Lemmatization in MorphAdorner takes a 'lumping' rather than 'splitting' approach and is similar to the 'hyperlemma' used by TextGrid. The modern form of a broadly defined lemma bundles diachronic and dialectal variance. Thus the form 'sote' in the first line of the *Canterbury Tales* (more often spelled 'swote' in Chaucer) is lemmatized as 'sweet' so that a search for 'sweet' will retrieve this dialectal variant.

2.6 Adding more texts to EDDAC

The operations that have been performed on 1,800 TCP texts can be readily extended to any or all of the 40,000 texts envisaged for EEBO, ECCO, and Evans. Thus one can claim that for texts prior to 1800, a version of EDDAC already exists or can be easily created. While the texts are not yet in the public domain they will pass into it within a decade, and in the interim they are available to the large community of scholars at the major research universities in the English speaking world.

For texts from 1800 on, if the texts do not already exist in a reliable TEI format, the best choice is to work with texts created by OCR, whether Google Books, the Open Content Alliance, or similar sources. Optical character recognition has made much progress over the past few years, and it is superior in some ways to double keyboarding because the optical transcription automatically retains the layout of the page block and makes it much easier to align the digital text with its facsimile image. For scholarly purposes this ability to return to the page image serves as an important security blanket. On the other hand, texts created with optical character recognition still require a lot of orthographic clean-up to serve as good enough 'diplomatic' editions of their source.

The layout of a printed page is full of implicit 'metadata' that readers tacitly process. There is now good software that transforms this layout into a kind of 'whitespace XML' from which you can derive a TEI-format through a combination of algorithmic processing and manual editing. Current experiments at the University of Illinois and Northwestern University suggest that you can create 'good enough' digital editions in reasonable time with editorial assistance from readers who are literate, have an interest in the book, and are willing to pick up modest technical skills of digital editing. Many undergraduate English majors meet those criteria.

The German TextGrid project is built around the idea of a platform that supports distributed editing and the sharing of results in a common corpus. Some version of such a design could support the creation of hundreds or thousands of 'good enough' editions that are designed from the ground up to promote digital intertextuality.

In extending EDDAC beyond 1800, there are good reasons for focusing first on '1001 novels' as a project that can stand on its own but can also be part of a larger enterprise. Substantial portions of '1001 novels' do exist:

1. fiction before 1800 is adequately covered through the TCP texts
2. American fiction between 1851 and 1875 is exhaustively covered in the Wright project
3. the public domain sections of the Virginia Early American Fiction project provide adequate coverage for fiction from the first half of the nineteenth century (some crucial texts, however, are not in the public domain)
4. The Library of Southern Literature provides good coverage of its field

What is missing is British fiction from 1800 to 1923 and American fiction from 1875 to 1923. Coverage of those area with 500 texts would go a very long way towards creating a quite robust module of fiction in EDDAC. And if EDDAC never proceeds beyond that initial module, a digital annotated corpus of 1001 (or more) public domain novels in English will be a useful resource for many scholars.

Fiction has some other advantages. It is the genre most widely read by readers at very different levels of sophistication. And from the perspective of data curation, novels are relatively easy texts to handle. Many of the tasks involved in creating good enough digital editions of novels lie within the range of amateur readers. Thus fiction is the perfect guinea pig for distributed and collaborative data curation.

3 Textkeeping or Distributed Collaborative Data Curation

Over the past two decades thousands of texts have been encoded by volunteers for Project Gutenberg. The Distributed Proofreaders Foundation has very effectively channeled the desires of many individuals who care about orthographic accuracy. The disregard of Project Gutenberg for provenance issues and accurate bibliographical description rules out most of the texts as candidates for good enough editions in EDDAC. But the project is a remarkable testimony to the cumulative power of the work of many hands. Can the energies and passions of scholars be harnessed to a similar enterprise so that, as in the case of life scientists and their gene banks, textual data can to some extent be curated by the scholars and critics who have the greatest long-term interest in having data of sufficient quality?

Textual data curation takes at least three different forms:

1. the creation of new digital editions
2. the correction of errors in existing editions
3. the adding of additional layers of encoding or annotation to existing digital texts

With regard to these three different forms of activity, it is necessary to rethink the opposition of mechanical and manual routines. I exaggerate only a little if I say that textual projects tend to be located at the two extremes of a range. There is the boutique project in which scholars lavish unlimited attention to the details of a text important to them, and there is the institutional project, typically housed in libraries, where you shudder at the thought of manually intervening in a text, rely on automated workflows as much as possible, and are willing to live with a level of textual error that no self-respecting teacher would tolerate in a basic composition class.

3.1.1 Correcting orthographic and similar errors

If you look at human editorial activity -- proofreading is a simple example -- there are three stages to the task:

1. finding the passage that needs attention
2. deciding what needs to be done
3. recording what you have decided to do

Of a minute's editorial labor, five seconds might be given over to the actual exercise of human judgment. 55 seconds are spent on getting there and reporting on what you have done. Can you build systems in which you drive down the time cost of human editorial labor and employ human judgment more effectively and also more consistently?

The answer is 'yes', although it is not easy and involves considerable 'up front' costs. Let me give an example. The 15,000 EEBO texts transcribed so far are a remarkable achievement. They are, however, full of errors. There are several million words where the transcriber could not identify one or more letters. There are countless examples of words that are wrongly joined or split. Sometimes paragraphs or whole pages are missing because they were missing in the microfilm on which the transcription is based.

Passages in some foreign languages, e.g. Greek or Hebrew, were not transcribed to begin and appear as marked lacunae. The EEBO texts include millions of untagged French or Latin words.

These are things that can and should be fixed, and they are best fixed by people who use the texts and care enough about them. If the texts are not used in the first place, there is no virtue in fixing them. If they are fixed as they are used, users collectively decide priorities as they go along.

If the texts are linguistically annotated, as they are in the MONK Project, every word is a 'word object' with a known address to which various kinds of new annotation can be attached without overwriting the text itself. When in reading such a text I come across an incomplete word, I can fix it in a few seconds. If I care enough about a text, I might look for all its lacunae and fix them.

What I can do with this text someone else can do with another. Data curation can be the work of many hands at many times in many places. There are two fundamental requirements for this to happen, and both of them are well within reach of current technology. First, you need a stable framework of Internet accessible data that makes it really easy for users to contribute in a casual or 'snacking' mode. Users should not have to take out the china and set the table. Secondly, corrections or additions by users should never overwrite the source text but should be submissions that are subject to editorial review (which could be an automatic procedure).

The community of potential contributors to such an enterprise is large, diverse, and global. It begins in the high schools. There are tens of thousands of high school students in the world who could do a little textkeeping here or there and whose collective results would be very large. At the other end of the demographic spectrum there are the 'little grey cells' of millions of educated retired people who can be recruited to the task of doing something useful for a book or author they care about. In the middle, there is a world of teachers and scholars who can perhaps be coaxed into contributing something, however busy they claim to be. My colleagues, not excluding myself, are all a little like Chaucer's Sergeant of the Law:

> Nowher so bisy a man as he ther nas,
> And yet he semed bisier than he was.

If you think of the tasks of textkeeping from the perspective of the volunteers who do it, you want to create a framework in which the volunteers can also do things for themselves while doing things that are helpful to others. It may therefore be productive to think of the software environment as a general framework for annotation. The correction of an orthographic error, a missing word, or the like is easily modeled as an annotation. Think of an annotation as a bundle of key-value pairs including a userID , a time stamp, the wordID that is the target of an annotation, the annotation type, which might be 'correctSpelling' or 'addNote', and finally the suggested correction itself.

Such a framework for annotation is more than what is needed for the specific tasks of error correction. But it may well be more effective in recruiting volunteers because it embeds their textkeeping in other forms of interacting with the text. These other forms have their own value for many scholarly, pedagogical, and recreational purposes.

A proper framework for digital data curation is both more controlled and more spontaneous than 'manual' editorial work. It is more spontaneous because it allows for casual

work along the way. It is more controlled because the forms of user intervention are more specified. Above all, the system is much better than any human at keeping consistent records of who did what, when, and where.

3.1.2 Creating digital editions in a collaborative fashion

It is a more complex task to create a properly structured digital edition of, say, *Bleak House* from the digital facsimile and OCR text of the first edition of 1853. The task is not as easily broken down into 'atomic' acts that can be done in any order, as is the case with proofreading. It does not rely on skills that educated readers possess qua educated readers. Instead it requires some knowledge of mark-up language, and it has to be done as a single project. But current experiments at Northwestern and UIUC suggest that with a proper framework and good documentation you can turn scholars with little technical skills into good enough digital editors of texts that they care about sufficiently to spend a few days of their life on. It may be harder than Googling but it is a lot easier than learning how to play the violin. Moreover, final proofreading, which will remain by far the most time-consuming task of creating a good enough digital edition from OCR texts, can be farmed out in the manner of Distributed Proofreaders. Alternately, you can think of proofreading as a chore that needs to be done, whether you live in a print or a digital world.

3.1.3 Correcting morphosyntactic errors

Automatically applied linguistic annotation has an error rate of ~3 percent. Whether such errors are ever worth fixing is a nice question. Generally speaking, the tolerance of users for morphosyntactic errors will be much higher than for orthographic errors. Orthographic errors are always visible, while morphosyntactic will typically be hidden even from users who take advantage of such tagging. From an analytical perspective, linguistic annotation is the basis for many quantitative operations. An error rate of 3% is unlikely to affect many results.

POS tagging errors are distributed very unevenly across texts and cluster in typical errors, such as not distinguishing correctly between the past tense and past participle of a verb, which usually happens in 15-20% of cases. The flipside of such clustering is that you can target errors if you care enough about them. If a text is modeled as a sequene of 'word objects' with metadata, you can create a tabular or vertical representation of the text in which every data row consists of the word, its unique ID, its reversed spelling, its POS tag, and forty characters before and after. Such a display gives you a storable concordance with enough context to support the correction of at least 99% of all errors.

This 'KWIC view' of a text has been an important part of the back office operations in the MONK project. It is an obvious way of presenting EDDAC data to users in a variety of contexts, and it can in principle serve as a framework for targeted correction of morphosyntactic or orthographic data.

3.1.4 Adding new levels of metadata: identifying spoken language

Error is endless. In a corpus of any size there will always be a need for textkeeping that consists of the humble tasks of getting it right. But there are ways of adding value to EDDAC that go beyond correcting mistakes, and they do not have to wait until the last error has been corrected.

As an example, I discuss the opportunities for identifying spoken language. The spoken language of the past is largely a mystery to us. We have no direct records of spoken language from before the age of Thomas Edison. Many of the records from the early period of the grammophone or radio represent formal ways of speaking that may be closer to writing. Extensive documentation of the way people actually talk has been with us for less than a century--roughly since the 1930's.

What we know about the speech of earlier ages is thus largely an extrapolation from its written representations. From comparing the dialogue of movie scripts with the transcripts of what people actually say we know that the differences are very large. Still, the written representations of spoken language are better than nothing, and they are all we have. There are many research scenarios for which it is helpful to distinguish between spoken and narrated text, whether or not the 'written spoken' is used to form hypotheses about the 'real spoken'. The distinction between speech and narration is an important part of much fiction. In most novels before 1900 the distinction is clearly marked by typographical indicators. In fact, the distinction between spoken and narrated language is probably the only typographical distinction that readers expect to find in a conventional novel.

Through a combination of automatic routines and manual review and correction it is possible to tag spoken language with <said> tags. From some experiments, I conclude that for a novel of ordinary complexity, this can be done in less than two hours per novel. In a second step, it is also possible -- though more time-consuming -- to identify speakers, as in a play, or to classify them by sex or social status. The utility of that procedure was demonstrated by John Burrows in his study of the different speech habits of Jane Austen's characters (Burrows 1987). But even without this additional granularity, the coarse binary division into speech and narrative is useful for many purposes. Reasonably comprehensive and accurate encoding of spoken language in a Book of English creates at least a diachronic record of what writers thought speech sounded like. That is in itself a useful thing.

4 EDDAC, Digital Intertextuality, and the Role of the Library

EDDAC is about 'enhancing' rather than 'extending' access, about doing more with the books you already have rather than adding more books. ('E-humanities', a term more popular on the Continent than in America, plays with its initial vowel, which originates in 'electronic' but moves from 'extending' to 'enhancing'.) 'More' involves activities that go beyond reading or simple cross-collection searches for character strings with or without secondary constraints, such as a search for 'love' near 'death' in texts with dates between 1589 and 1612.

Here are some search scenarios that illustrates this 'more':

1. You choose a set of texts and look for other texts that are 'like it' in terms of lexical or syntactic habits.
2. You select a group of texts, e.g. several hundred sermons between 1500 and 1800 and see whether the distribution of lexical or syntactic phenomena divides them into groups that are useful for subsequent analysis.
3. You take a syntactic pattern like "the king's daughter", gather instances across a collection of texts and visualize the results by creating images in

which the owners are 'nuclei' that are defined by the 'rays' of their possessions.

4. You extract names of people and places from a group of texts and look for patterns in their distribution by genre, region, or date.
5. You define a sub-corpus --e.g. novels by George Eliot-- and ask what words are disproportionately common or rare in that subcorpus when compared with some other corpus -- e.g. novels written during her life span.
6. You take a word or a concept defined by a 'basket' of words and track its frequency over time in different text categories defined by the intersection of genre, and sex or origin of author.
7. You take a word in different texts and explore the ways in which its use is inflected by the company it keeps.
8. You look for phrases of varying length that are shared between one work and another and use them as point of departure for allusive relationships -- intertextuality in a very traditional sense.

These are all search scenarios that are currently supported by programs like Monk, Philologic, WordHoard, or visualization projects like Many Eyes. They depend on familiar techniques in statistics, corpus linguistics, and bio-informatics, with names like supervised/unsupervised classification, log likelihood statistics, collocation analysis, named entity extraction, or sequence analysis.

While all these search scenarios are available somewhere, it is not the case that they are available in a single environment where they can be used by literary scholars with average technical skills on a wide variety of texts, including texts they might want to add to an already existing archive. Who should build such an environment, maintain it, and provide guidance to literary scholars and other humanists whose relationship with digital technology is as yet insecure?

4.1 Libraries as the natural institutional home for EDDAC as a cultural genome

The most obvious and in some ways quite traditional institutional framework for EDDAC is a university library or a group of libraries acting in a consortial manner. The CIC Libraries come to mind, because they have a strong tradition of consortial activity with a strong focus on digital text archives, albeit of an 'extending' rather than 'enhancing' kind. A librarian might at this point object that the enhancement envisaged in EDDAC are really the reader's responsibility and that the Library's responsibilities have been fully met by making digital texts available. That is a serious argument, but it can be countered by drawing attention to the peculiar role that primary texts play in humanities scholarship.

Scientists encounter the primary objects of their attention in their laboratories, which nowadays contain much more complex and expensive tools than the Bunsen burner, the paradigmatic tool of the 19th century chemist. The scientist's library holds the 'secondary literature' or just 'literature' about their field. Primary data in the sciences may be held in laboratories, but increasingly they are held in library-like environments, partly to share the cost, but mainly because shared repositories greatly increase the circulation and analysis of data. GenBank, already mentioned in this essay, is a prime example.

Let us return to EDDAC as a 'cultural genome', an 'annotated collection' if not of all, then of many important 'publicly available' texts, where 'annotation' refers not to critical commentary but the standardized identification of words or 'text molecules' such that the annotated texts become 'machine actionable' and allow scholars to gather and organize textual data for analysis and integration at a higher level. This is another step in the 'allographic journey' of texts -- a migration from scrolls. codices, and printed books into a digital world that supports all the affordances of these previous technologies of the word but adds new forms of contextualization and analysis.

What is the appropriate institutional framework for such a cultural genome? The offices of individual literary scholars will not support a network of collaborative exploration of shared data beyond small communities. Neither will Departments of English, separately or together: their administrative, financial, and technical infrastructure is simply not suited to such tasks.

The best answer to the question is 'the library', and this answer derives fairly directly from the traditional role that libraries have played as keepers and mediators of the primary data of Literary Studies. The answer becomes more obvious once you free yourself from the idea that digitized books are somehow more technological than printed books: the written word has always already been technologized. If you go into a Rare Book Library, you would not be surprised to see an old printing press that was in its day a high-tech tool. Now it would sit there for decoration rather than use, but it is a reminder that the written word has always had a high place in the technical pecking order of its day. The habits and practices of the Rare Book Library in fact set useful precedents for the work required for EDDAC. Rare Book libraries are about highly curated data. Their achievements have rested on close cooperation between scholars and librarians and on the conviction that in any large library there will always be special data that require and justify high levels of data curation.

There are different ways of being special. In the rare book library of my university, the most valuable and jealously guarded treasure is not a Gutenberg Bible or First Folio, but the autograph of a Beatles lyric. Rarity or fragility are often the causes for taken special care of items in a collection. The high level of data curation in a genome project, on the other hand, is not justified by the fact that the data are rare but by the fact that their elaborate curation supports inquiries that would not otherwise be possible. The same is true of EDDAC.

Digital data curation involves not only metadata that describe an object at the item level -- e.g. a manuscript--but derivative data structures that may be many times the size of the original object. You can think of a morphosyntactically tagged text as a derivative data structure. An even clearer example is sequence alignment, which depends on the prior existence of an annotated corpus. In this technique, common in Bio-informatics and plagiarism detection, you ignore the 100-200 most common function words, which account for at least half of the words in a text, map the surface forms of the remaining content words to their lemmata and look for repeated lemma strings or n-grams of variable length. When a new text is added to the collection, an initial algorithm checks for matching n-grams and keeps track of them. The resultant derivative data structure of repeated n-grams weaves a web of intertextual echoes made up of literal and fuzzy string matches. Mark Olsen and his collaborators have used this technique to model the relationship of Diderot's *Encylopédie* to its sources. Sequence alignment is a conceptually simple but

computationally intensive procedure. If used across a large data set like EDDAC, it is a powerful way of leveraging the analytical potential of the underlying textual data.

Can you distinguish clearly between content and tools? Their fluid boundaries are further dissolved by digital technology. Curated data typically add layers of information created by tools, and these layers may be extracted, aggregated, and used to create additional data structure. If you follow this way of thinking, the distinction between tools and content eventually disappears: 'content' is layers of value added with the help of tools. It then becomes a pragmatic decision whether to deposit the output of a particular tool as a new layer of content or whether to produce it on demand.

Consider the act of cataloguing a book. What value is added by the activities that end with sticking a label with a call number on a book? Does the 'content' of the book change? If you engage in the thought experiment of 'uncataloguing' a million-book collection, removing the call numbers and scattering the books at random across the floors of the stacks, you would say that the 'contextualization' that a catalogue affords to a book has a very substantial impact on the reader's assessment of its content. Through the act of cataloguing the original layers of content of a book are surrounded with a new layer of output. That layer takes the physical form of the surrounding books. The catalogue is a very powerful tool and shapes the knowledge space within which books circulate.

Librarians have a deep and honourable reluctance to come between readers and their books. They see their task as pointing readers towards the books and then getting out of the way. Freedom of inquiry is greatly helped by this self-effacing ethos. But the truth is that the librarian's work always comes between the readers and their books. A library is an 'instrumentarium' in which a hierarchy of tools adds value to the single book, which is itself a tool to begin with.

Enhanced access to primary digital texts may be understood as a way of extending the findings aids of the catalogue to move below the item level into the digital object itself. The techniques may be new, but the questions are not. "Could you help me find passages in *Hamlet* that echo earlier plays or are echoed in later plays" is a proper question addressed to a reference librarian, although she would be hard put to answer it. It is a kind of question that is readily answered by EDDAC with an appropriate analytical instrumentarium and user-friendly interface.

The primary texts of Anglophone literatures make up a relatively small percentage of all the books in their library catalogue ranges, and the percentage drops further if you 'deduplicate' virtually identical items. With regard to these primary texts and for their scholarly users the library is both library and laboratory. The image of a Renaissance scholar working at a book wheel points to the close relation of reader, tool, and the 'books', which themselves are tools.

As these primary texts migrate into the digital sphere, we need to think of them as existing in a digital laboratory in which scholars can take advantage of their digital affordances. If you do not care for the scientific metaphor think of the digital laboratory as a kind of kitchen, perhaps a witch's kitchen. Even in a predigital world, scholarly reading and writing are a form of cooking in which texts are sliced or diced, kneaded and rolled, boiled, steamed, baked, or roasted. In a digital kitchen it is not enough to have the repositories of pantry, refrigerator, and freezer. And you need a few more tools than a microwave oven.

5 Works Cited

Burrows, J. F. *Computation into Criticism : A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press, 1987

Darnton, Robert. "Google and the Future of Books " *New York Review of Books*, 12 February 2009

Firth, J. R., and F. R. Palmer. *Selected Papers of J. R. Firth, 1952-59*, Indiana University Studies in the History and Theory of Linguistics. Bloomington,: Indiana University Press, 1968.

Goodman, Nelson. *Languages of Art; an Approach to a Theory of Symbols*. 2d ed. Indianapolis, Ind.: Hackett, 1976.

Homer. *Homeri Ilias*. Edited by Martin L. West. Bibliotheca Scriptorum Graecorum Et Romanorum Teubneriana. Stutgardiae: B.G. Teubner, 1998.

Lüdeling, Anke, Thorwald Poschenrieder and Lukas Faulstich. "Deutschdiachrondigital -- Ein Diachrones Korpus des Deutschen." *Jahrbuch für Komputerphilologie* (2004): 119-36.

Ong, Walter J. *Orality and Literacy : The Technologizing of the Word*, New Accents. London ; New York: Methuen, 1982.