1650-1850: Ideas, Aesthetics, and Inquiries in the Early Modern Era

Volume 18 Article 15

2011

ON THE ROAD WITH DIGITAL HUMANITIES

David Hill Radcliffe

RESEARCH METHODS UPDATE

ON THE ROAD WITH DIGITAL HUMANITIES

David Hill Radcliffe

"Are we there yet?" "Ten more minutes." "It was ten minutes ten minutes ago." How long until funding comes, practitioners are trained, and the enterprise of digital humanities begins to realize its full possibilities? Ten more minutes.

Scholars are often wary of digital publication, rightly so: those who work with centuries-old materials are bound to be skeptical of a medium whose long-term survival depends on perpetual care from parties yet to be identified. Meanwhile, institutional commitments are deferred and Wikipedia and Google Books rock the foundations of scholarly publication. We live in curious times.

I here offer some thoughts on where we've been, where we are, and where things might go supposing scholars might take a more proactive role in building digital infrastructure. Despite two decades of experience in digital publication I cannot claim any great authority in the matter; partial understanding is all anyone can offer given the complexity and flux in digital publishing. My focus will be on editing books, one corner of a large field.

✳ I. Where We've Been ✳

We might compare the advent of digital books to developments in sound recording a century ago. In the early days progress was hampered by rival technologies and platform-specific products: recordings were made on cylinders or discs, in vertical- or lateral-cut processes. Your Edison product was not playable on your Victor product. After a couple of decades of creative destruction, sharing of patents, and adoption of standards, things shook out in such a way that by 1925 the technology became commercially viable to the extent that millions of phonograph records were being made and sold.

In roughly the same space of time an equivalent process has unfolded with respect to digital books with a similar surge in production. In 1925 the adoption of the microphone dramatically altered the landscape, enabling comparatively high fidelity recordings and radio broadcasts. The invention of the internet browser had analogous effects, making digital publication cheap, ubiquitous, and standardized. Within a remarkably short period of time (five years, was it?), the Web, like radio before it, became the dominant medium for the dissemination of information, music, and popular culture.

Yet digital documents, like sound recordings of old, leave much to be desired and face a similar risk of obsolescence and loss. It is comforting to note that cylinders, wire-recordings, and acetate discs remain playable today. Who could have imagined that so many historical broadcasts would have survived, albeit in uncouth forms? Like a shoebox full of eight-tracks, currently un-viewable digital documents from the 1980s and '90s seem likely to become objects of interest over time. Do not toss those floppy discs.

Where are digital books on the technological arc? We seem to be stuck in 1925, lacking equivalents to what electrical recording brought to sound recording: high fidelity, long-play, stereo. Digital books, though abundant, are remorselessly low-fi: Google Books, Project Muse, ECCO and the like afford us page images searchable by means of uncorrected and often unreliable OCR text with minimal metadata and few hyperlinks. They are "machine-readable" only in primitive ways.

✳ II. Where We Are ✳

To explain what machine-readable means, let us pass from the retrospective to the present tense. The Internet browser, like a radio receiver, works by translating electrical impulses into human-readable forms, either page images or web pages rendered by HTML tags into paragraphs, italics, images, links, buttons, menus. The markup is chiefly concerned with visual rendering as opposed to the machine-readable information that renders a digital book the equivalent of "high fidelity": with better markup, a digital document becomes something you can compute with as well as look at.

Markup is where humanities scholarship intersects with computer technology: oddly enough, human intervention is required to render a document properly machine-readable. Machines can be trained to clumsily guess that the string of characters "Miller" refers to a person and not a profession, but are seldom able to infer that the string refers to Cambridge Massachusetts as opposed to Cambridge England, or Andrew Miller as opposed to William Miller, especially when OCR has misrendered the string as "Millen" or when the original document has only "M*ll*r." OCR (optical character recognition) is particularly weak on names—"John Galt," almost always rendered as "John Gait," becomes invisible in Google Books.

Once names have been identified and marked, a machine-readable document can do much with that information: a person can be indexed, noted, linked, glossed; they can be linked to a genealogy chart, a bibliography, or a statistical diagram. But most of the clever things that computers can do require editorial intervention by knowledgeable scholars working with standards and conventions. There are specialized markup languages for things like musical and mathematical notation but the most common employed in humanities computing is TEI (for Textual Encoding Initiative).

TEI editors can encode documents in ways that describe their formal, linguistic, rhetorical, or bibliographical structures, providing information that can then be used for various kinds of computational analysis and display. But perhaps the most useful kind of markup is simply that used to disambiguate names of persons, places, and titles for purposes of finding needles in the digital haystack.

Some examples will illustrate what this looks like. Where a TEI personal name element is given an attribute, it can be used to discrimi-

nate the poet Alexander Pope from the actor Alexander Pope, thus: <persName key="AlPope1744">Alexander Pope</persName> versus <persName key="AlPope1835">Alexander Pope</persName>. Markup can get around alternate spellings: <persName key="JoDryde1700"> John Driden</persName> and even no spellings: <persName key="JoDryde1700">******</persName> (The machine-readable text in brackets is not displayed.)

Markup languages like TEI have been around for years and have undergone considerable testing and refinement; this is mature technology whose assets and liabilities are well known. Among the liabilities of TEI is the inherent difficulty of reconciling standardization with flexibility; TEI lends itself well to sophisticated project-specific solutions at the expense of cross-project data-exchange. If and when we have a common framework for exchanging scholarly data, TEI documents should be adaptable to it. Ten more minutes.

The problems besetting digital humanities seldom have much to do with technology per se, which has by and large overcome the compatibility, migration, and interoperability issues that plagued the field in the 1990s. There *are* technical means for exchanging data across variously encoded projects, though their implementation requires human intervention and cooperation. Like so much else this waits upon the economics of scholarly publishing, digital preservation, and the academic rewards system—things that evolve much more slowly than does technology.

The chief objection to textual markup is that it is a laborious business and hence time-consuming and costly compared to the automated process that can be used to create digital books out of page images and OCR. It is, after all, a form of editing and as such requires human judgment as well as human labor. Perhaps not so much labor as is commonly thought, since even a digital book with light markup and metadata can take advantage of the powerful economies computation enables.

A further word then about textual markup. At bottom, it is a matter of dividing strings of characters into nested containers: volume, chapter, paragraph, stanza, line—and within these text blocks, into additional containers for in-line items like names, notes, dates, and titles. The containers are called "elements" and often have descriptive "attributes" such as a locating number for a paragraph, or an identifying

key for a person, place, or title, or a standardized, or computer-readable form for a date.

As crafts go, markup is not particularly complicated. While TEI workshops typically run for a week, it actually only takes two or three hours to learn how to use a text editor to block out a simple document. One can be doing useful work in a week's time. Since documents like letters, plays, novels, and histories tend to have similar forms encoding work soon becomes fairly routine—though with variations and conundrums enough to keep things interesting.

TEI is descriptive markup only; to render the document on the screen with fonts and margins requires a style-sheet involving a different kind of coding and a different degree of craft. This is not particularly difficult to learn either, but to translate a machine-readable document into a pleasing, human-readable screen image requires practice, experimentation, and a degree of taste. Crafting style sheets resembles pot-making: unexpected results happen frequently. But it is not difficult to apply or adapt a ready-made style sheet to a structurally similar letter, play, novel, etc. and one can become reasonably proficient in a month or two.

To do actual computing with textual markup—to convert tagged names to links, generate indexes, compile tables, import information from databases or other documents, to send queries and return structured responses—requires a programming language. The learning curve here is steeper: in my experience a couple of months to get results and a year or two to get proficient. It is at this level that humanities computing moves from being a documentary to a creative enterprise.

The time frame for getting up and running as a digital humanist is thus something between two weeks and two years. How long does it take to mark up a document or to build a digital project in TEI/ XML?

This varies with the size and nature of the documents, the amount of annotation, and the density of the markup, but it seems useful to throw out some numbers. In my experience: a variorum edition of Scott's *Lay of the Last Minstrel:* a summer project; a collection of seventy-five Civil War letters: four months. In the case of my major project, *Lord Byron and His Times* (LordByron.org), perhaps twenty years (this is in its third year at the time of writing). LBT is a collection of pamphlets, letters, articles, memoirs, and biographies related to Byronic controversies. The tagging is comparatively light, but with more than common attention to visual formatting since the documents are presented as web pages instead of page images.

Starting with uncorrected OCR, a 400-page octavo with a substantial amount of structure (inset poems and letters, notes and running heads) takes about two weeks to encode: two days to block the text into numbered chapters, paragraphs, and stanzas, the remainder of the time to correct the OCR, add visual formatting, and to supply names and titles with identifying keys. A substantial essay (say a *Quarterly Review* article) might take a long day to mark up. This includes time spent entering names and titles into separate data files.

Lord Byron and His Times is not a critical edition but an archive of documents mapping social relationships among people Byron wrote about, people who wrote about Byron, and the social network of "friends of a friend" within which the drama of Byron's career was played out and recorded for posterity. The editorial labor does not go into establishing a critical text (most documents were printed only once) but into establishing relationships between documents and people.

That is the sort of thing computers are good at: at the time of writing we have some 250 documents, chiefly biographical, linked to data files containing names of 8,000 people and 3,000 titles. The links are made using keys like those described above. The keys and data files generate pop-up notes as well as making the archive cross-searchable. In this project markup is used to bring to light relationships among persons and titles whose names are often suppressed, mangled, or left implicit—in cases where a name was unknown to the original author it can be supplied in the markup.

Markup transforms a digital book into a database in which chapters, stanzas, and paragraphs become so many records. When keys are used for names and titles, a digital book can be parsed like a relational database; for example, where ancillary files contain information about education, it becomes possible to pull up documents or references in documents for persons who were Byron's contemporaries at Harrow, or to books published by those persons, and to filter those results in a variety of ways: references to *Childe Harold* in writings by Harrow contemporaries.

✳ III. The Semantic Web ✳

In *Lord Byron and His Times* markup is used to identify persons and titles whose names may or may not appear in the text, and by means of

ancillary files to compute with those names in various ways. But what if, instead of just using LBT data files to do this work, we could link its documents and data files to other resources on the Web for purposes of exchanging data?

When Tim Berners-Lee and his colleagues established the World Wide Web they created HTML as a markup language for documents, but also began work on a second language called RDF (Resource Description Framework) for expressing information about objects. These objects might be web documents or authors, but could be anything—persons, things, events, relationships—that could be uniquely identified using a URI.

This schema, the basis for what is referred to as the "Semantic Web," has been slow to take off: simple in concept, it has proven a challenge to implement. RDF takes the form of statements about URI-specified objects referred to as triplets, in effect machine-readable syllogisms used for simple reasoning. Here are some examples of triplets, including two that use the LBT namespace as a unique identifier: "Lord Byron wrote Childe Harold," "Childe Harold was published by John Murray," "LordByron.org/LdByron is Lord Byron," "LordByron.org/LdByron.Harold is Childe Harold," "LordByron.org/JoMurra. 1843 is Murray, John (1778-1843), publisher," and so forth.

Imagine a repository containing millions of such RDF statements that can be parsed for relationships among persons, titles, events, and web objects, among them information uploaded from the Byron archive: it becomes possible to parse the Web at large in ways analogous to those described above for key relationships within LBT. A search for books on the web originally published by John Murray will turn up the document in LBT, but also information about those documents and their authors from other archives and data repositories around the globe.

An LBT person-record—say for an obscure contemporary at Harrow—might use an RDF query statement to find and display a list of the person's progenitors pulled from genealogy sites, a list of their publications from library catalogues, and lists of biographical and critical works about them pulled from bibliographies. Data exchange being a two-way street, one can also imagine a genealogy record elsewhere on the web pulling up passages relating to this Harrow student that appear in Byron correspondence published in LBT.

The much-to-be-desired Semantic Web will make available for cross-querying all manner of information of interest to humanists: structured-data lists of school graduates, baptismal records, pension lists, publisher's catalogues, exhibition catalogues, dictionaries, business directories, subscriber lists, census records. Getting such information published would be a huge task, but then because the number of potential beneficiaries is also huge it might just happen. The Semantic Web was conceived as working in a piecemeal, distributed way with little centralized planning—like the World Wide Web itself.

How long will it take to make the Semantic Web a reality? There are RDF-driven sites already up; among them the NINES index of nineteenth-century material is the most familiar to digital humanists. The technology exists, though it is cranky and difficult to implement. The chief bottleneck is again the labor required to get printed and manuscript information into machine-readable form. Ten more minutes and we'll be there.

✻ IV. Whither We Tend ✻

In the 1930s, music listeners were treated to a series of incremental refinements as producers learned about microphone placement and groove width. Behind the scenes experiments were underway with long-playing records and stereo recording, developments that would be two decades coming. Just so, we see a variety of incremental improvements to web interfaces and transmission rates even as much larger and transformative developments are underway.

Advances in sound recording were the work of multinational corporations with access to laboratories, patent-lawyers, factories, retail chains, and broadcast facilities. By contrast, advances in digital humanities are contingent upon open-source software, open-access documents, public-commercial partnerships, and distributed computing. These postindustrial developments have had the effect of making producers of consumers, putting the means of production (as it were) within reach of comparatively impoverished scholars and academic institutions.

All agree that academic publishing will be in a very different situation ten minutes hence. If humanities scholars elect to engage deeply with the information revolution (which is by no means certain) the technology will be there—is here already. The difficulty lies in the

fact that "high-fidelity," information-rich documents do not wait upon the invention of a better class of machine, they wait upon a different class of scholarship and a new economic model.

There is much time-consuming work to be done that would be prohibitively expensive if hired out but that could be done better, and much more cheaply, if incorporated into the general scholarly research program and so accomplished by redirecting existing resources. It could happen; scholars might well find in editorial work and information architecture the same kinds of professional satisfaction they take in critical work—I have. So far from rendering old-fashioned scholarship obsolete, information technology underscores its importance and amplifies its traditional philological strengths.

It is more difficult to imagine how the scholarly publishing business might adapt to a situation in which documents and data-files are made available open-access on the Web—as they need to be to reap the potential benefits of information technology. To make that happen, we might see a return to a pay-to-publish as opposed to the pay-to-read scenario, as with subscription-printing of old. Where profits used to be made by selling books, in future they are might be made by selling reading equipment and software designed to parse masses of distributed, open-access data-files. Selling access to documents, particularly when those documents are in the public domain or when they are mere OCR with page images, does not seem like a sustainable business model over the long term.

The next stages of digital evolution promise to be particularly bumpy as various interested parties—publishers, funding agencies, libraries, universities, and scholars—learn to operate in a digital environment in which authorship, ownership, and authority divagate from the norms of print. But let us lift our eyes to the horizon where our imagined destination beckons. Cue the *Wizard of Oz!* I cannot help but think of the coming Bibliotopolis as a return to the future as imagined in the 1930s, that great era of philology and urban life.

Implicit in the metaphor of information-architecture is the notion of an Emerald City, the construction of which will be a slow, multi-generational proposition. Its infrastructure will be developed piecemeal, partly planned and partly not, with document-buildings linked and supported by the digital equivalent of transportation and utility grids. Much of what we build today is destined to be pulled down and

replaced, but the better things will be preserved and developed because good work is destined to last.

In this imagined city books will figure as information portals, their pages so many windows onto the Web. The browser used to parse a scholarly book will use machine-readable code to identify appropriate contexts—historical, biographical, textual, linguistic, aesthetic—and selectively gather pertinent information into an apparatus assembled on the fly. No doubt this will be filtered through knowledge of a particular reader's contexts and preferences, and the apparatus will enable readers to contribute their own mites to the evolving Bibliotopolis.

The idea of a city of books suggests the continuity-amid-change that typifies both print and digital publications. We are learning to fashion digital texts that are modular—capable of being integrated with other documents or archives—and migratory—capable of being recast into new digital genres and formats. Over time, what one writes or encodes is likely to be consumed or repurposed in unanticipated ways. But what of that? Books and buildings and scholarship have always been handled so. It just takes place more quickly now…or does it?