Subject: **Response to question of free access to digital data**
Date: **November 16, 2011 2:51:36 PM EST**
I am attaching my response to the question of free access to digital data from the perspective of a historian and an editor. With thanks,

Holly C. Shulman
--
Holly C. Shulman
Editor, Dolley Madison Digital Edition
Founding Director, Documents Compass
Research Professor, Department of History
University of Virginia

The Cost of Free
Holly Cowan Shulman
26 October 2011

Digital publication is neither free nor cheap. Despite enthusiasm for scholarly publication at no cost to the end user, free publication will not solve the dilemmas that have upended the ecosystem of scholarly communications since the invention of the World Wide Web. As a community with many different stakeholders, ranging from scholars to universities, libraries to publishers, and funders to programmers, we need to think more clearly about the future of scholarly electronic publication.

The World Wide Web burst into public awareness as an open space for freely available communications in 1993 with the introduction of Mosaic, the precursor of Netscape. By the end of the 1990s that space had been further transformed by increasingly effective search engines, especially Google in 1998. As we all know, the impact was, and remains, a virtual tsunami on the world of communications. The WWW has already changed the environment for newspapers, magazines, music, political organization, and personal communications. It threatens cable and broadcasting. It is in the process of upending publishing and scholarly communications as one piece of this global shift.

The impact of this revolution has played out differently in various arenas. The cost of journals – especially those in Science, Technology, and Medicine (STM) – has skyrocketed while the budgets of libraries and institutions of higher education have decreased. In STM fields the importance of early publication is important, and publication online is a solution to the time factor. In some areas of the humanities online publication has posited a solution for the difficulties of finding traditional scholarly outlets. As the wages of most Americans over the past generation, including most academics, has flattened, the prospect of reducing the cost barriers of purchasing books, subscribing to journals and buying newspapers has been seductive, especially as new reading devices such as Kindle and iPad appear and their prices drop. The role of agents is threatened as new organizations are formed for authors who seek to self-publish their books. We are now at the point of vertical integration of the publishing industry as Amazon expands beyond a virtual bookstore to become a publishing house. As a recent report published by the British Research Information Network (RIN) said: "At a time of financial stringency for universities, research funders and publishers, it is important that all the stakeholders in the scholarly communications system work together to find the most cost-effective ways of fulfilling their joint goal of increasing access to the outputs of research."

The United States government has weighed into this debate. Research funded by public monies, they argue – especially STM – should be freely available to the American public. STM journals have responded by charging their authors to publish, a practice now known as "author-side payments." Already 40% of biomedical journals work this way. There is also the policy of online publication after a time barrier, which most of us know through JSTOR and MUSE as two examples of aggregators that sell their product rather than give it away. Some members of the digital humanities community passionately believe that online scholarly communications should be freely available to all as a matter of policy in order to open academic discourse to the widest audience possible.

Libraries are integral to this new mix. Where once they provided shelves, cataloging, reference, and an occasional rebinding services, they now must subscribe to new online publications and host their own catalogs and other works. In some cases libraries embrace their university publishers or have become close working colleagues of the university publishing house. There is a movement for open repositories where any scholar in that institution is expected (if not required) to post their work. Libraries have also hosted a world of experimentation such as the Institute for Advanced Technology in the Humanities and the Scholars' Lab at the University of Virginia. Adding to the seduction of library publication are copyright rules that differ for educational versus commercial use. If a humanities scholar wants to publish images, for example, she or he may find it far simpler to stay within presently defined academic boundaries and publish their work through their library.

Within this shifting ecosystem of scholarly communications, lies the small world of reference works, and within that world the smaller arena of documentary editions, and even within that the very small field of documentary editions in history. This is the world of the "the papers of….": Thomas Jefferson, Albert Einstein, Elizabeth Cady Stanton, Margaret Sanger, George Washington, the Freedmen and Southern Society Project, Thomas Edison, James Madison, Eleanor Roosevelt, Naval Documents of the America Revolution, Frederick Douglass, Abraham Lincoln, Cordell Hull, Andrew Jackson, Martin Luther King, and dozens more including my own Dolley Madison Digital Edition. While some have private funding, most operate through grants received from the federal government, specifically the National Endowment for the Humanities and the National Historical Publications and Records Commission. These funders are now considering the pros and cons of demanding not only electronic publication, but also free access to those electronic publications. The implicit question is: what are the costs of free? Do we need to unbind ourselves from the world of publishing in order to nurture creativity and open dialogue – or are these illusory goals that will destroy the armature of editions by replacing cost of purchase with cost of production, hosting, and maintaining?

*

It is my belief, to begin with, that a documentary edition should be neither a website nor an electronic book. Information on a website moves through a series of links that allows the reader to go from page to page. Its basic language is HTML. It is better suited to an electronic exhibition than an electronic archive. An electronic book transforms the printed page into displayable and searchable pixels that visualize a book for the consumer. It adheres closely to the traditions of print, plus perhaps a slider on the bottom, a page marker, a percentage number to locate the reader in the book, and various ways of highlighting the text and taking notes.

As a reference work and research tool, the documentary edition is best suited to the format of a digital archive, or a storage repository. After all, the edition may include thousands, if not millions of documents. Unlike a novel or an essay or a scholarly monograph, there is no argument being made, no narrative, no real spine at all on which to hang the bones of the edition. It is a collection

that is searchable through a process of identification and constraint. Each piece of information is like a pebble on a beach, and when you do a search you pull up from the beach a certain number of those pebbles. When you're done, those pebbles drop back to the beach. The scholar or student can read it from end to end, or browse through it, or search for what they want through different search tools. This representation, however, requires that whoever is publishing the work tag it in XML, most likely using the encoding guidelines of the Text Encoding Initiative (TEI), in a manner conformant to a whole series of programming demands. It also means that its display be reasonably intuitive so that the reader can navigate the text. In sum, it means that a well-done digital documentary edition should not be composed of PDFs taken from a printed book, or a series of pages simply coded in HTML, and posted on line. The internal structure of a digital edition is important. It may be altered; it may be expanded; but it cannot simply be thrown up on the web as can a blog, for example.

The central argument, however, revolves around what are the advantages and disadvantages of digital versus print. After all, a print edition is a wonderful object, replete with nicely formed pages, all kinds of fore matter such as preface and introduction, bibliographical information and a table of contents. It has an index that allows the contents to be searched. Why bother to move over to an electronic format? Are the reasons cost or scholarly benefit? And do these align on one side or the other of this argument over free?

There has long been an assumption that digital is cheap, in fact so cheap it can be given away for free. The hypothesis is in part built upon the way the medium grew in the 1990s when newspapers, for example, started publishing for free electronically as well as for a price in print. They rue the day. Today newspapers are establishing cost firewalls to protect their bottom line, while they fight off competition from bloggers and radio and television online products. Who wants to pay for The New York Times if you can get it for free? The print world of daily news and magazines is demanding subscription fees while also adding new content to give additional value to the online subscriber reader. The Times is a leading example of value added electronic materials rendering a superb electronic publication – at a cost. We all know that the print world is struggling to balance production costs with revenue. Amazon.com set a radically low fee to entice readers to purchase e-books so they could sell the Kindle, and as observed above are now trying to further cut their own costs through vertical integration of production by starting their own publishing organization. Meanwhile, the rest of the book-publishing world struggles to keep up. The American Association of University Presses recently issued a report tackling the problem of communications transformation entitled Sustaining Scholarly Publishing in which they lay out their own issues and explore viable solutions.

The cost of creating quality electronic scholarly publications in the field of documentary editions is, in fact, not cheap. Currently, funding grants carry a project through to the point where they are assumed to hand it over to someone who will then take care of it and somehow get it online. For starters, this model portends the elimination of a number of important editorial interventions that publishing houses undertake. They copyedit. They provide peer review. They help develop a product. They market. All of this has long been true in print. In the new world of digital, presses must also check data for consistency, extend code for new features, create new data interfaces, and so on.
They must have someone who can deal with the challenging world not only of markup languages but also of the specifications provided by the TEI. Without these standards there can be no consistency. Without this regularity there can be no interoperability or cost scaling or industrial output; electronic publication would instead remain a craft in which each producer created their own unique widget that would not play with their neighbor's or peer's widget.

Every edition needs to be "served," or hosted on a server. A good one is costly; even purchasing a license to use a piece of it is expensive. Moreover open source software needs customization. Most editors simply cannot open up a box entitled Drupal, an open source content management system, and make it work for their project without external help. There are start up costs and there are platform costs and there are new content costs. And there are the continuing costs of maintaining a product in a world of constantly shifting technology in which what works today will not necessarily function tomorrow, of sustaining the machine and the human resources to maintain that content.

Imagine that you are the editor of a small project, and you have decided – eagerly or reluctantly – to be born-digital. That said -- to whom would you go? Who would you find as a programmer, an information systems person, or a designer? You would need a new way to control and track your work and your deadlines. Where would you find an appropriate, and free, content management system? What would your overall costs be? Whereas in print you would have simply handed over the manuscript to your publisher, who would have counted on recouping all costs through the sale of books, now you can no longer do so. You may fare well at a large university, but suppose you are an editor residing at a small institution. Or perhaps you are a retired professor and want to edit an edition as a retirement project. To whom do you turn for advice; what questions do you even ask? And if you have conquered the challenges of DTDs, XML, and TEI for encoding ordinary text, then what do you do when you later want to add different kinds of records such as inventories or ledgers? Two years on, can you find the same team who first set you up? Without a publisher, who guarantees that the work is not only quality scholarship, but even legitimate? Large-scale, well-funded projects can hire developers. Small projects usually run on a shoestring. And yet to date there has been NO talk about adding subvention costs to grant funding. The bottom line is that digital publication is neither free nor cheap. Three months of a programmer's time might cost $25,000. The cost of a production-level license for a first-rate XML publishing environment like MarkLogic Server runs upwards of $30,000 at current prices.. And every time you wanted to add a new installment you would need editorial and technical work that would easily cost $5,000 to $10,000. And that is assuming there is a structure somewhere out there that can do for you what you want.

Electronic publication has important advantages: but only if done well. To begin with there is the compelling allure of cross-searchability, interoperability, and aggregation. The execution of this vision is, to date, tied to server and tagging issues. Thus Rotunda, the electronic imprint of the University of Virginia Press, can publish compilations such as Founders Online and make these editions cross searchable. There is a huge scholarly reference advantage here, but had Hamilton, Madison, Washington, and Jefferson all been published in discrete models this integration would not yet be possible. The Dolley Madison Digital Edition is now beginning to include lists of 300 names or so that will require a new kind of tagging; Rotunda is there, ready and willing and able to help accomplish this goal. Single volumes in a series such as the Papers of George Washington are now tied together and the researcher can do one simple search that carries her or him through all 60 volumes. Documents Compass, a unit in the Virginia Foundation for the Humanities, has been working on collecting all of the names and biographical references in all the founding editions and creating both a biographical dictionary and a prosopography. Funded by the Andrew W. Mellon Foundation, this will be a Rotunda publication.

The model of author-side payment is not a solution for documentary editions. Nor may individually built editions have the resources or technical infrastructure to create quality publications. I know of no digital humanities tools that confront these basic issues, so while it may be useful to have citizen transcribers using a tool known scripto (scripto.org), it does nothing to solve the basic

problems of publication, while it is unclear that it resolves issues of costs in order to professionally establish the texts at hand. While technology is upending the world as we once knew it, we should approach the problems of electronic editions more carefully and beware of solutions that promise free – at great cost.

---

[1] There were other web solutions in 1993 besides Mosaic but it is this author's memory that Mosaic stole the show. By 1994 a number of search engines started up including Yahoo! and Lycos, which the following year were joined by AltaVista, Magellan, and others. Google was launched in 1998. The combination of browser and search engine transformed communications.

[1] STM fields tend to be rapidly developing areas of research in ways not true in the Humanities and Social Sciences, and even before the WWW depended upon prepublication of journal articles.

[1] Julie Bosman, "New Service of Authors Seeking to Self-Publish E-Books," *The New York Times,* 2 October 2011, http://www.nytimes.com/2011/10/03/business/media/perseus-creates-new-service-for-authors-seeking-to-self-publish.html

[1] *Heading for the open road: costs and benefits of transitions in scholarly communications* http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/heading-open-road-costs-and-benefits-transitions-s

[1] *Heading for the open road.*

[1] There is the Coalition of Open Access Policy Institutions (COAPI) in the United States. See Jennifer Howard, "Universities Join Together to Support Open-Access Policies," *The Chronicle of Higher Education,* 27 October 2011, http://chronicle.com/blogs/wiredcampus/universities-join-together-to-support-open-access-policies/32632

[1] *Sustaining Scholarly Publishing: New Business Models for University Presses* (AAUP, March, 2011).

[1] Matthew Gibson to Humanist Discussion Group (Willard.mccarty@mccarty.org.uk), 26 October 2011.