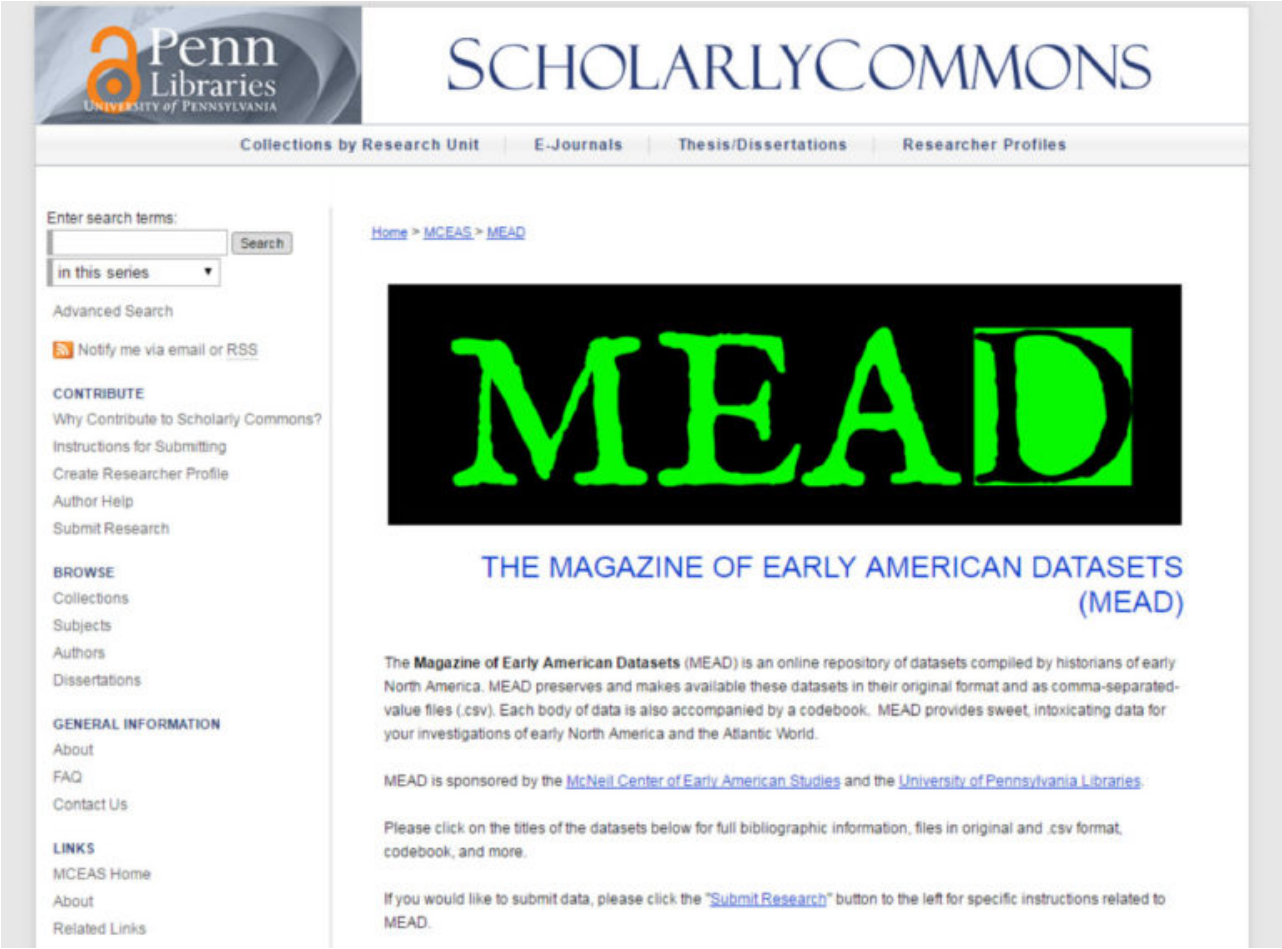
Constructing the Magazine of Early American Datasets (MEAD): An Invitation to Share and Use Data about Early America



The Problem of Disappearing Data

Data. Before postmodernism, or environmental history, or the cultural turn, or the geographic turn, and even before Brent Spiner's character on *Star Trek: The Next Generation*, historians began to gather and analyze quantitative evidence to understand the past. As computers became common during the 1970s and 1980s, scholars responded by painstakingly compiling and analyzing datasets, using that evidence to propose powerful new historical interpretations. Today, much of that information (as well as the data compiled since) is in danger of disappearing. As a profession, we are experiencing a generational shift, and much of the data created several decades ago has already been lost. More will disappear in the coming years if not preserved soon.

Disappearing data is not unique to early American historians, although some of

the problems may be related to the academic culture of the humanities. In social science fields like economics, political science, psychology, and sociology, multiple factors have resulted in a scholarly culture that preserves a good deal of its data. Many projects attract funding from the federal government, often from the National Science Foundation, which requires the resulting data to be made available to the public. Researchers in the social sciences often work in teams, and so are used to collaboration. Their disciplines frequently value new research by junior scholars and graduate students using previously generated data. Most journals that publish data-driven articles in the social sciences also require the posting of datasets as a condition of publication. Furthermore, with the Inter-university Consortium for Political and Social Research (ICPSR) and other repositories, practitioners in these disciplines have access to considerable infrastructure to preserve their data.

By contrast, as humanists, historians tend to think of research as a more solitary enterprise. We treat our compiled data much as we do our notes: as raw materials of little value to anyone but ourselves. Because we think of data preservation as a personal rather than an institutional issue, our datasets are vulnerable to all the vagaries caused by quickly changing technologies. Physical media likewise degrade and become obsolete, making data recorded in those formats hard to decipher even if one can find the machinery to read them. Today, for example, most university Information Technology departments no longer possess apparatuses to read floppy discs or Zip drives.

Software packages have also changed dramatically during the past few decades and will continue to do so. Data in some formats favored by past generations of historians, like Statistical Analysis System (SAS) or Statistical Package for the Social Sciences (SPSS), can still be transferred to newer formats, but more arcane ones can be almost impossible to decipher without a great deal of highly technical work. As Jeff Rothenberg, a computer scientist, wryly observed two decades ago, “It is only slightly facetious to say that digital information lasts forever—or five years, whichever comes first.” The problem has grown increasingly severe ever since.

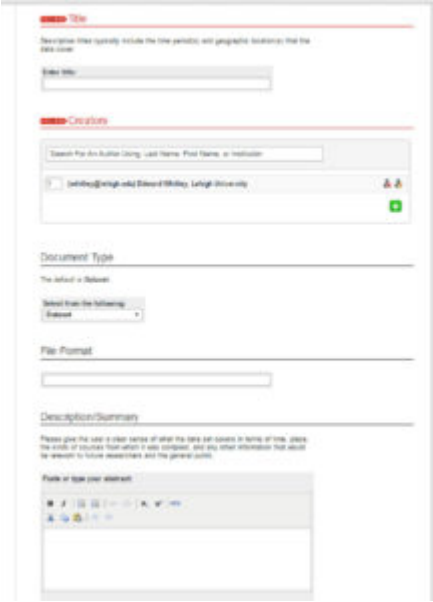
Preserving Early American Data in MEAD: The Magazine of Early American Datasets

1. The homepage of MEAD: The Magazine of Early American Datasets. Courtesy of the University of Pennsylvania Libraries.

In an effort to combat the problem of disappearing data, we have developed MEAD: The Magazine of Early American Datasets, which has been designed to preserve and share, openly and freely, statistical data about early America. We appeal to all early American historians to take the time both to preserve and to share their statistical evidence with present and future scholars. It will not only be a legacy to the profession, but will also encourage historians to share their evidence more openly and provide a foundation on which scholars can continue to build.

We created MEAD with the generous support of the McNeil Center for Early American Studies, the University of Pennsylvania Libraries, and Bepress (the open-access electronic publishing platform behind the Digital Commons software that many universities and research centers rely on for their institutional repositories). A handful of scholars—including Gary B. Nash, Robert E. Wright, and Thomas J. Humphrey, and the authors of this essay—have already loaded datasets to MEAD. Even with the minimal datasets it currently hosts, MEAD is already of interest to many people around the world: more than 1,500 people from the United States, Russia, India, and China have accessed and downloaded these datasets.

To make the initiative successful, we need more datasets. It is in this spirit that we make the following appeals for scholars of early America to share their statistical evidence. We would love to have your datasets, your huddled 1’s and 0’s (and other numbers and letters) yearning to be freely used by other scholars. Once loaded onto the Website, the files will be enduring. Each dataset in MEAD has its own unique, permanent URL, which means that users can link to them with the assurance that they will never die or require updating. In addition, the files are available to any scholar, teacher, genealogist, or layperson who can access the Internet.



2. Uploading data to MEAD. Courtesy of the University of Pennsylvania Libraries.

The process of submitting a dataset is (relatively) painless. Scholars may mount datasets themselves. Contributors can go to the site, click on the "submit research" link, and get started. Once data suppliers open a free account (no information required beyond name and email address, which will never be shared), they are directed to a page with fields for bibliographic information, both about the original source material and about the dataset generated from it. Users upload the dataset and any additional explanatory material, such as a codebook or relevant information about how the dataset was generated. The directors of MEAD take it from there. Alternatively, if they prefer, contributors can e-mail us the files and the associated information, and we will fill out the forms and upload the files. To download files is even easier. After setting up an account, simply click on the dataset and follow the directions to download. The entire process of submitting a file takes less than fifteen minutes, and we approve submissions within a few days to make certain they appear on the site.

MEAD and the University of Pennsylvania Libraries' ScholarlyCommons

MEAD is hosted by the Penn Libraries' ScholarlyCommons, the institutional repository of the University of Pennsylvania. The mission of the ScholarlyCommons is, in part, to support open access to innovative research and new publishing models; visibility of Penn scholarship, and sustainability of Penn's academic endeavors

Because MEAD's purpose coincides with the goals of the ScholarlyCommons, the two are natural partners. All relevant submissions are accepted regardless of the creators' association with the University of Pennsylvania because MEAD furthers the mission of the McNeil Center.

ScholarlyCommons uses Bepress' Digital Commons platform, which was developed by

scholars at the University of California, Berkeley, as an alternative to the for-profit dominated academic journal publishing industry. Hundreds of institutions use Digital Commons to host publicly accessible scholarly material.

There are numerous advantages to using a repository such as ScholarlyCommons. One is search engine optimization. Scholars employ the Internet to connect with scholarly works and raw data, which creates a need for machine-readable materials, quality metadata, and the accessibility of the materials to ensure that researchers are able to discover the most useful resources available. Everything in ScholarlyCommons is fully indexed in search engines. Each dataset can be tagged with associated disciplines that, in addition, place them in Digital Commons' subject-based repositories, thereby further increasing the visibility and discoverability of the datasets.

Another advantage is persistent URLs. Each dataset in MEAD has its own unique, permanent URL, which means that users can link to them with the assurance that they will never die or require updating. The metadata created for the series is designed to provide context for the datasets—when and how the information was gathered, coverage, publications based on the datasets, related datasets, and the like—as well as to help users find similar materials.

There is an increasing need not only to make information openly accessible but also to convey how that information can be reused. Submissions to the MEAD series are required to have a Creative Commons Attribution License. This license allows users to share and adapt the content for any purpose while maintaining the integrity and maximizing the accessibility of the original data. All users must provide a link to the license, indicate if changes were made, and avoid any suggestion that the licensor endorses the use made of the data. Furthermore, users may not apply legal terms or technologies that restrict access to the original information.

Current and Future Issues

One major struggle we have encountered, as we anticipated, is convincing historians to submit their datasets. It has not been for lack of trying. We have been advertising the project publicly, and this essay belongs to that effort. We have privately contacted dozens of scholars, each of whom spent many mind-numbing hours sweating over tax lists, census registers, merchants' ledgers, women's journals, men's diaries, plantation records, or newspaper advertisements for runaway slaves. We appreciate their efforts to enter thousands of pieces of information into computer files and then analyze the evidence to help deepen our understanding of early America. We would love to have these scholars make one more pass through their data and to load it online to MEAD.

On the one hand, scholars whom we have approached are enthusiastic to preserve and share their records. Few of them want their hard work to disappear. On the

other hand, only a handful of historians have taken up the offer. One reason, as suggested earlier, may be rooted in how we treat our datasets—as notes for ourselves alone, rather than as a community resource. As a result, most historians' datasets are messy: they might have inconsistent spellings or capitalization, extraneous columns or rows, or other anomalies that the person who compiled it understands and accepts. Some of those reflect the source material; others are artifacts of the process of compiling or data entry. Regardless, the scholar who created the datasets assumes that she is the only one who will use the data, so she did not take the time to make it cleaner for someone less familiar with the sources or the process. Revisiting and cleaning datasets does require time and energy. Most conversations or e-mail exchanges with potential contributors result in the perfectly understandable response that their data requires some more work to be useful to others. Moreover, historians would rather spend time on their next project rather than on cleaning up data for their old one. It is all quite understandable, but it does mean that the data will continue to disappear and our profession and understanding of history may be the worse for it.

We will keep reminding scholars of the value of preserving their datasets. In the meantime, if you have created any files full of numbers, names, dates, prices, or places, or even jokes, then please consider submitting them to MEAD.

This article originally appeared in issue 16.3 (Summer, 2016).

Billy G. Smith, bgs@montana.edu, professor, Montana State University
Nicholas Okrent, Research and Instructional Services, University of Pennsylvania
Andrew M. Schocket, aschock@bgsu.edu, professor, Bowling Green State University
Sarah Wipperman, Repository Services, University of Pennsylvania

We encourage *Common-place* readers to add to our list of digital resources for the study of early American history and culture by contributing to the Zotero group Suggestion Box for Common-place Web Library or by emailing suggestions to Web Library editor Edward Whitley (whitley@lehigh.edu).