Science at Risk from a History of Science Perspective
Fred Gibbs

As librarians, curators, and archivists think more about archiving online science content for future use, they are challenged to strike a practical balance between the wealth of savable data on one hand, and the work required to make it into a meaningful and accessible collections on the other. After all, content needs to be not only gathered and stored, but also made useful and visible, a process that takes substantial human work, even if heavy automation can aid in the process. This challenge is often framed in terms of properly identifying what to collect, or perhaps as a challenge in filtering the great mass of content from which one must carefully select.

Needless to say, selection processes remain important. Even if one believes that storage space is cheap, and simple file formats are likely to be available many decades from now (as many already have been), content needs not only to be collected and stored, but also to be made visible, a process that takes substantial human work, even if the process can be heavily automated. The work of collecting, organizing, as well as making visible and available is simply impossible given the magnitude of digital material and increasingly limited resources to conduct these complex processes.

This essay argues, from the point of view of a historian of science (and to some extent of a digital historian), that librarians, curators, and archivists must address the difficult value question of what content to save with three important but often neglected considerations in mind: the varied audience for science content (e.g. scientists versus historians), the importance of collecting science content that departs form what might be considered good or mainstream science, and the changing nature of archival use.

Varied audiences

Science at Risk workshop participants agreed that it is helpful to think of three stages of archival life: creation, near-term, and long-term. This tripartite scheme nicely encompasses the varied challenges of 1) collecting from diverse sources that employ diverse technologies; 2) making such content immediately available for immediate research needs, and 3) preserving it for posterity and future reference.

In addition to this scheme, we also must consider the different audiences that will benefit at those various stages. In the near term, other scientists and perhaps policy makers will likely be the primary audience--and thus dictate near-term strategies both in terms of what to collect and how it should be made visible and available. In terms of the long term, however, historians--especially historians of science--will benefit most. Collection development should be made with both audiences in mind. While there is substantial overlap in the kinds of materials that each group will be interested in, there are significant differences that must factor into collection strategies.

The disciplinarily diverse audience and presenters at the Science at Risk workshop showed how many participants are actively creating and curating online science content according to their varied needs and interests. Workshop presenters associated with science blogging or citizen science projects, for example, demonstrated their distinct interest in preserving discussions about current science issues, whether from professional scientists or science enthusiasts--with their content ranging widely across natural philosophical discussions, methodological questions, historical essays, or arguments about what species of bird appears in a particular photo. Open notebook enthusiasts demonstrated their interest in preserving a narrow but deep view of science in action. There is no doubt that all of these constitute sources worth saving. Such sources will be of use to scientists (or civic scientists) struggling with similar problems; parts will be useful for historians who want deeper insight into the messy processes of science that do not emerge from official and polished publications.

Yet for these generators of online science content--as seemed true for many participants at the workshop--the emphasis of what was at risk leaned heavily toward what the creators and managers of these resources, as well as those tasked with archiving such sources considered to be "good" science. There is no question that, when considering the near term use of scientists or future historical uses to learn about mainstream science, archives of content from publications like science blogs and open notebooks will prove to be fantastic and largely unprecedented resources.

Longer-term archival materials, however, are useful to a rather different audience that does not share the same agenda as many creators of online science content. From a historian's perspective, it would be deeply problematic for future research if content selectors focused on preserving a narrow--and to some extent arbitrary--selection of content that a particular set of insiders thought was "good." Of course it is true that historians' ability to understand and interpret the past will continue to be mediated by the stewards of our cultural artifacts: librarians, curators and archivists who, laboring under various practical constraints, must often save what is or will be of obvious value. This value is often determined by the context in which it is collected. Science content, then, is likely to be collected because it reflects upon the activities of a recognized scientific community and is said to constitute "good" science.

Yet some of the most fascinating work from historians, philosophers, and sociologists of science examines how societies (at various levels) demarcate science from non-science or how various communities embrace (or not) various explanations or theories. Such research often attempts to establish the ways in which historical actors determine the boundaries of science, or to examine how historians have chosen to portray them. Being able to determine the boundaries of science, regardless of their epistemological origin, are entirely crucial to the success of these historical efforts. As a result, thinking about such future historical use should encourage different kinds of selection processes from those that have been previously employed. Archivists and curators must select the broadest possible spectrum of science content that represents a wide range of attitudes and understandings about science, even when they contradict what would be generally considered good science. In other words, we must prioritize breadth over depth even when limitations on the content collection process do not allow a more cohesive or thorough cataloging effort. It will be helpful to broaden the filters, even if the catch from a wider net cannot be fully processed or cataloged per the usual rigor. As will be discussed below, historians are gaining greater facility with processing such mountains of data and in fact need less parsing done for them.

For example, we must actively preserve materials that can easily be labeled as pseudo-science, including creationist blogs, anti-climate change blogs, and generally science-skeptic blogs, regardless of their religious or political motivation. For the historian of science, the historical record that outlines ideas and attitudes about creationism, phrenology, and alchemy have been just as important as those that outline evolution, psychology, and chemistry. Similarly, science bloggers (and sites that aggregate such content) often publish invectives against what they consider pseudo-science or bad science. If collected together, they provide an unusually complete discourse around science in the popular realm. To attempt to separate "real" science and knowledge claims from the complex interactions of politics and science is to ignore or deny the vast historical analyses that reveal the social and cultural constructions of science and judgments about it.

One facet of the historical record that historians of science never seem get enough of is the "popular" attitudes, views, and understandings about science. In terms of targeting specific content, these might include blog posts and user comments about--and especially in response to--scientific or science policy articles that run in online newspapers, or other web periodicals with web forums of some sort. For example, the violent storms that swept through the Washington, D.C. area in the summer of 2012 were the subject of numerous newspaper articles that prompted user comments that

mentioned climate change as a possible explanation for the rare storm system. Many comments (perhaps in a coordinated effort) explicitly challenged any connection between global warming and severe weather or the scientific status of man-made climate change. This is a wonderful and new (historically speaking) venue for getting at a variety of attitudes about science, including the kinds of arguments people do or do not not make in the course of such debates about the viability or applicability of certain scientific theories. And it perfectly exemplifies the so-called gray literature--writing that does not fall into traditional archival categories--than can be easily neglected, especially by scientists and others interested in promoting "real" science, which can unfairly minimize the voices of those who do not agree with it.

Especially if the mainstream science blogging sites or other official publications turn their back on what they deem as "bad" science, the cultural heritage community must redouble its efforts to capture this rhetoric. This would, for example, allow future historians to see how effective such rhetoric was at important political moments, how it has changed, or how it correlates with other data, like demographic or election data. It can provide a fascinating window onto a much broader scientific discourse that lies outside the typical venues of official science publications.

Apart from the discourse itself, one of the potential values of science content captured from online sources will be to help historians to understand the wide diffusion, perhaps even the popularization, of scientific knowledge. To study (at least effectively) larger social phenomena such as diffusion, though, requires careful and relatively precise metadata about the content, such as when and where a particular post or comment came from--information that is sometimes not visible on the webpage where the content resides. Historians will hope for as much metadata as possible, and their analysis will be as rich as the metadata is complete. As websites may balk at collecting and/or sharing data about posts, archivists are seriously limited when working only in content-ingestion mode. Rather, librarians, archivists, and curators must work with content providers to capture as much metadata about the posts as possible (even if not publically visible, such as IP addresses that reveal geographic data) in a way that is sympathetic to privacy concerns without being a slave to them.

When trying to understand the diffusion of scientific knowledge, not only is content essential, but also some sense of its influence. One obvious example would be to capture the viewing or download statistics for various publications, or perhaps how often (and when) it was posted to Facebook or retweeted. But the many kinds of statistics that one might find associated with a particular online publication (and thus might want to preserve) do not necessarily overly complicate the archival process. It is important to remember collecting can be done in ways that preserve metrics without thinking too much about exactly what needs to be preserved. Websites, services, and publishers often display this kind of information on webpages that contain the original content.

At the same time, it is also important to think about the ways in which diffusion might be measured in ways that are not already explicitly quantified and displayed on pages. Participants at the workshop repeatedly lauded the value of alt-metrics in measuring the value of scientific work or its uptake in the community. But once publishers start to foreground alt-metrics for whatever purpose--as they already have done--then they are not really "alt" anymore, and thus they lose some of their value that they had when they were truly outside mainstream measures. Truly "alt" metrics are not, by definition, clearly visible. The implications for archiving--as with content--is to save as much metadata as possible--not just what is obvious value now, whether considered mainstream or "alt." Of course it is difficult if not impossible to anticipate what future alt-metrics might be, and truly alt-metrics will comes from historians discovering new patterns and trends from whatever combinations of data are available to them. And this is yet another argument for casting as wide an archival net as possible for not only content but metadata as well. Future researchers might, for example, use various text mining methods to understand influence of a particular blog or article and correlate it to other historical events--but this depends on having as much data and metadata as possible, not only

what is prejudged to be of sufficient scientific quality or to have an established value for measuring diffusion. Certainly, such determinations will yield different kinds of historical analyses in the future.

Lastly, it is worth pointing out that content that might normally be deemed outside mainstream science are crucial not only for historical research, but also for contemporary policy research as well, a potential use that several workshop participants emphasized. Policy decisions are based as much on rhetoric as "real" science, and policy research will be more effective if a wider range of arguments and contexts can be preserved.

Upstream intervention

Some participants wondered if librarians, archivists, and curators now face a paradigm shift with respect to traditional archival practices. The notion of a sea change is certainly a useful heuristic to make the question more approachable, but it is one that foregrounds the difference between potential processes and perhaps distorts the nature of the challenges in archiving web content. It recalls (I can hardly resist a history of science example here) the sixteenth-century choice between heliocentric and geocentric systems, which is often taken as an exemplar of a paradigm shift. But historical research has shown that this wasn't really a choice dictated by mounting evidence, or necessarily a choice at all. Many natural philosophers embraced both models, using whichever one best fit a particular purpose. When rethinking archival practices we must bring finer nuance to the question of what is changing and what is not.

The basic premise of the archivist--to collect, label, organize, and preserve--is not fundamentally different now than it has been. However, some crucial aspects of archiving now demand fundamentally new approaches and processes. For instance, preserving web science from rapidly-changing online sources has precipitated considerable scrambling on the part of archivists to respond to changes in website design, dynamic content, fleeting video formats and proprietary players, and so on. Such a process is wholly unsustainable. It simply cannot keep up with current rates of production--to say nothing of the additional technology migration issues each day. In other words, the technology to ingest content will never keep up with technology (and its nuanced variations) to produce it.

One possible response is to narrow selectivity even further. This is problematic, however, because 1) identifying things like "good" science blogs is unfairly judgmental; and 2) it automatically filters out those blogs that have not reached a threshold of notoriety or publicity. From a historian's point of view, what's unusually intriguing about blogs as historical sources is that they can be from *anybody*. Considering the broad range of online science content that will be relevant to future historians, as well as the range of publishing platforms that host such content, new collection strategies are required.

To mitigate some of these new collection challenges, curators and archivists must become more active in upstream intervention—in making arrangements to automatically collect content from some sites, or possibly encouraging sites or even individuals to apply to have site content preserved. Some of this content will never be worth preserving, some will be of obvious value; other content might not be worth collecting initially, but will become something of greater interest over time.

Lower level goals toward upstream intervention might include, for example, producing Wordpress plugins (or something similar for other platforms) that allows users to configure their blogs to be more easily archived. They might also include encouraging online newspaper or magazines to insert tiny bits of code that make the job of archival crawlers easier. Such development efforts could be complemented by tutorials and other educational and outreach efforts that provide clear and concise instruction not only about the technology itself, but also how bloggers or other sites with potentially

useful content can understand the challenges of preservation and the value of their own content for science policy, historical study, and so on--likely an attractive possibility for those who consider themselves marginalized by mainstream publication practices.

At a higher level, curators and archivists must maintain active relationships and communication channels with partners (blog aggregators, for example) who collect content worth saving. For larger sites like *Scientific American* or major newspapers which host content like user comments that might not normally be archived, there may be an easy way to collaborate with those sites to allow such content to be easily archived with little effort on the part of an outside archiving agent who, given unfairly tight budget constraints, will always be hard pressed to keep up with constantly changing technologies used on various sites that impede preservation efforts. These techniques of course cannot capture everything, but it allows the archivist and overarching collection agencies to focus on the greyest matter, so to speak, that resists such automation.

Access and future methodologies

Upstream intervention may make it easier to collect content, but that does little to lessen the substantial archival work of proper labeling and sorting for future visibility. Traditional historical research has been both circumscribed and facilitated by archival practice in which the researcher depends on archivists to properly catalog and retrieve relevant materials for a particular research question. In many respects, these limitations still and will always exist, and any limiting effect is easily overstated. Still, for better or for worse, historical research has traditionally utilized one model for accessing archival materials: the historian goes to the archive and works with the librarians and archivists, who bring relevant materials to the researcher.

New methodologies and expectations of access must shape current archival practices because historians will be using the library in fundamentally different ways. Of course they will want access to physical books, articles, and manuscript papers. But they will also expect to be able to download large swaths of data that they can subject to various kinds of analysis. Providing data in this way might sound like an additional layer of complexity that adds to an already overburdened archival staff--and to be fair, it does require different kinds of virtual interfaces to libraries and archives than are now common. But the expectation of large data acquisition can also be seen as a tremendous freedom in the sense that historians are beginning to use tools and processes that don't require archivists and librarians to catalog everything as carefully as they have in the past.

In terms of future use and visibility, it may become less important for archives to provide access points mediated through careful curatorial cataloging. In other words, visibility through full-text searching will become far more important than precise classification or cataloging. This has direct implications for collection practices. It allows collection efforts to expand the collection net, so to speak, to gather more material than they normally could. It will allow libraries and archives to direct resources from cataloging to making items visible through their content rather than classification. Obviously not all items lend themselves to full-text searches, but many do, of course, and lend themselves to new kinds of historical analyses that are becoming popular in the digital humanities community.

Given the way that new searching and analysis might work, the work of the archive must change as well. One important new service that libraries must provide, for example, will be facilitating data exchange. Given a variety of cross-sections of science content that a historian might gather, historical questions about correlation and causality have new possibilities, but only if archival materials are visible through very high level searches and API queries.

New relationships

So far I've emphasized broad content selection, steps to minimize the resources required for collecting it, and suggested a new emphasis for how this data will be useful for future researchers. In the last section, rather than focus more on specific content sources (mostly because I want to deemphasize the value of pre-selection), I want to outline what I see as some of the most important strategic initiatives for improving the historical utility of online science content. In short, it is to facilitate new kinds of relationships that can help make preserving web science content a manageable enterprise. These grow out of the workshop conversations, but they maintain my bias as a historian of science.

Relationships between historians and cultural stewards

The scholarly community must transcend the typical disciplinary divides between historians and archivists. In particular, historians of science are well positioned to make insightful recommendations about the kinds of science content that will be useful for future historical research. We can hardly rely on a few SMEs (Subject Matters Experts) to know of all possibilities across such a broad range of science disciplines and sub fields. There is simply too much to know. Even with the most vigilant efforts toward objectivity, the gravitational pull of mainstream science and higher-profile spaces of discussion remains strong.

Historians of science are uniquely positioned to know and think about the alternative venues. Those engaged in science content preservation might reach out to a wide audience of historians and sociologists of science and technology to discover what kinds of sources they now use and what they hope their students will use in the future. They will be especially helpful for understanding how current historical research questions and answers would be different if certain kinds of materials would have been saved. Those who consider themselves digital historians are worth consulting as well, to understand growing importance of data, new techniques for exploring it, and future expectations of access.

Closer partnerships with other collection efforts

Cultural heritage institutions must facilitate and actively maintain more clearly articulated relationships and missions between various foci of institutions with special collections. This kind of "divide and conquer" strategy allows for a more sustainable way of integrating various archiving practices so that these sources can be recombined in the future. This can also help offload and outsource some of the immediate science content preservation to more local production sites, freeing larger repositories to focus on the truly gray literature that cannot be easily slid into any other preservation domains.

The Workshop participants' many and varied vocational interests (scientists, publishers, archivists, historians, etc) clearly demonstrated quite varied perspectives, concerns, and levels of interest in archival work. This dramatically increases the amount of material that needs to be collected, the ways it should be collected, and the uses to which it can be put. It also means the necessity of more collaboration with other repositories and publishing platforms. Considering the variety of possible technological solutions is nothing if not dizzying. Perhaps as a result of the variegated interests in technologies and strategies that generate science content, and the uses to which content might be put, there is little agreement about best practices for archiving it. Yet because no single institution is likely to craft the definitive standards and best practices for archiving science content, it remains crucial to create and maintain a relatively stable topography of collection efforts.

Visible Leadership

Even if no single repository will ever be the first place that comes to mind when considering best practices for archiving online science content, archives that feel like they have sound practices in place should be more vocal in terms of their recommendations for best practices. As with many web

technologies, if not technology in general, standards and best practices do not need to be fully worked out and agreed upon before their implementation. Standards generally emerge from practice and community consensus over time. But visible leadership--even if conducted jointly--is paramount. It prevents, for example, smaller repositories or collection efforts to avoid reinventing the wheel, or making unnecessary deviations from established, successful practice. A combination of top-down and bottom-up (or perhaps explicit and implicit) directives will drive consolidation of collection strategies and ways to facilitate the process.

Without both high-level and low-level action, collection efforts will continually be at the mercy of fragmented, incomplete, and abandoned localized archival efforts, adding yet an additional layer of complexity to the archival process. It is not as important to provide "correct" answers as to help bridge the gap between content generators and preservers with experience and advice directed toward, and differentiated for, various publishing platforms, institutional repositories, and individuals.