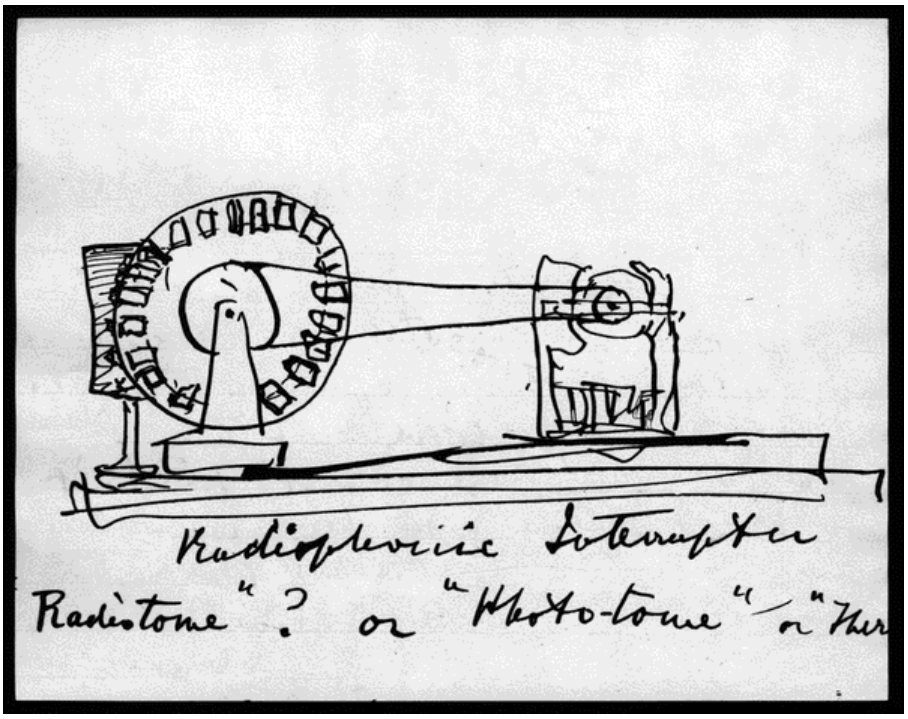
SCIENCE @RISK



Alexander Graham Bell's sketch of a radiophonic interruptor, May 27, 1893. box 205, "Subject File: Drawings by Alexander Graham Bell, 1881-1911." *Alexander Graham Bell Family Papers, Manuscript Division, Library of Congress.*

November, 2012

Toward a National Strategy for Preserving Online Science

A report of the National Digital Information Infrastructure and Preservation Program focused on identifying valuable and at-risk science content on the open web. Topics include, science blogging, open notebook science, citizen science and ideas for approaches to ensuring long-term access to this content.

Contents

Science @ Risk

TOWARD A NATIONAL STRATEGY FOR PRESERVING ONLINE SCIENCE

Fifty years from now, what currently accessible web content will be invaluable for understanding science in our era? What kinds of uses do you imagine this science content serving? Where are the natural curatorial homes for this online content and how can we work together to collect, preserve, and provide access to science on the web? These were the three principal questions up for discussion at *Science at Risk: Toward a National Strategy for Preserving Online Science*, a recent National Digital Information Infrastructure and Preservation digital content summit.

The Blue Ribbon Task Force on Sustainable Digital Preservation and Access recommended that "leading stewardship organizations should convene stakeholders and experts to address the selection and preservation needs of collectively produced web content."[1] Thanks to generous support from the Alfred P. Sloan Foundation, The Library of Congress was able to invite a small but diverse set of science bloggers, representatives from citizen science projects, and individuals working on innovative online science publications to talk about and share their work with archivists, librarians, curators, and historians from a diverse array of cultural heritage organizations to work through and explore these questions.

This report summarizes the discussions and findings from the meeting, suggests a number of calls to action for stewardship organizations, and includes two perspective papers and a brief case study from different participants to represent the view of creators and future users of online science.[2] The first perspective essay comes from Fred Gibbs, Assistant Professor of History at George Mason University and Director of Digital Scholarship at the Center for History and New Media. Gibbs provides a perspective on the diversity of web content that historians of science are likely to be interested in and why. The second essay—from Bora Zivkovic, Blogs Editor at Scientific American, visiting Scholar at NYU School of Journalism, and organizer of the ScienceOnline conference—provides the perspective of a content creator on the development of science blogging. This is followed by a case study of the U.S. National Library of Medicine History of Medicine Division's Health and Medicine Blogs collection pilot. This collection exemplifies how cultural heritage organizations' existing collecting goals can translate into a targeted web archive collection development strategy. The report closes with an appendix briefly listing examples of similar ideas for web archive collections that cultural heritage organizations could create based on the priorities identified by meeting participants.

---

[1] Blue Ribbon Task Force on Sustainable Digital Preservation. (2010). *Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information*, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf, p. 68

[2] This report was compiled by Library of Congress staff with Abby Smith Rumsey.

WHY FOCUS ON ONLINE SCIENCE?

For the purposes of the meeting, participants defined online science as the products or results of scientific activities—as well as the community of discourse among scientists, policymakers, funders, citizens, and future scholars, historians, and scientists—shared on the web. To capture online science would be to capture the informal and largely non-peer-reviewed network of blogs, projects, forums, and innovative publications that connect scientists, science journalists, and the interested larger public.

The digital versions of traditional peer reviewed journals are also of high interest and importance. However, because the ownership and distribution structure are known and relationships are in place with libraries and other stewardship organizations, this group of content is not as immediately at risk. Setting digital journals aside, the primary focus of the meeting was on science discourse outside of the traditional publishing model whether analog or online.

To date, issues around the preservation of scientific data have been the primary focus of born digital preservation efforts. The Blue Ribbon Task Force on Sustainable Digital Preservation made noteworthy suggestions for the preservation of research data.[3] Projects like DataONE and the Data Conservancy have made significant headway to ensuring ongoing access to research data.[4] In this space, NDIIPP has been particularly involved in supporting work related to the preservation of social science data through the Data Preservation Alliance for the Social Sciences (Data-PASS).[5] The four-year project developed and maintains a collaborative infrastructure for preservation and access to social science data. In contrast to concerted efforts for scientific data, the preservation of new modes of scientific discourse occurring on the web has yet to be substantially addressed.

STAGES OF ARCHIVING

Archiving science on the web presents many challenges: the distributed nature of web-based content; the fact that in blog comments, forums, and citizen engagement projects authorship/ownership is often unclear; and affiliation between online science projects and organizations are loose and changing. Three stages in the life cycle of stewardship and archiving were identified to help frame discussion about types of action that could be taken to preserve scientific discourse online.

1. **Self-archiving.** This involves taking steps to keep data in order and saved during the process of creation and use.
2. **Near-term archiving.** These are steps done by hosting sites or repositories, personal Web sites, publishers' sites, or those that share products and data to keep content preserved while organizational affiliations are in place.

[3] *Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information*
[4] For information on DataONE see https://www.dataone.org/ for information on the Data Conservancy see http://dataconservancy.org/about
[5] For information on Data-PASS see http://www.icpsr.umich.edu/icpsrweb/DATAPASS/

3. **Long-term archiving.** It is the responsibility of libraries, archives and museums to provide stewardship that endures over hardware and software upgrades, organizational changes, and generations.

WHY IS ONLINE SCIENTIFIC DISCOURSE AT RISK?

Scholarly discourse and interaction among scientists and the public is rapidly changing. The ephemeral nature of this online discussion leaves it at substantial risk of being lost. Science blogging has become a major mode of scientific discourse. The last ten years have seen significant growth in large science-focused blogging communities and platforms. In this space, sites like ScienceBlogs, PLOSBlogs, and Scientific American's Blog Network are playing an important role in science communication and may be prime targets for partnerships with digital preservation organizations and other stakeholders. At the same time, many scientists are running their own individual blogs, either through generic blogging platforms like Wordpress.com and Google's Blogger service, or through their own content management systems. These individual blogs present more complicated issues for selection and preservation.

A range of other novel online modes of publication have emerged, and are continuing to emerge, which require attention. Various projects for sharing pre-prints of articles, like SSRN, RePEc, and ArXive.org, are already developing new preservation approaches.[6] However, new models of publications, like the video Journal of Visualized Experiments (JoVE), and science podcasts present non-textual information. These digital objects present particular risks for loss because they are not published through traditional library acquisition channels.

Citizen Science initiatives are engaging members of the public to participate in data collection and interpretation. Much of the work of citizen science is evident in the collected data and reported in scholarly literature. However, a considerable amount of important work occurs in online forums and discussion spaces. That information will likely be an important set of source material for understanding the role that these systems have played in the history of science. For example, much of the work involved in the discovery of a new kind of galaxy in the Galaxy Zoo project resulted from discussions in the project's web forums[7].

Much of the content that participants in the preserving online science summit thought most valuable are also most at risk of loss because they do not clearly fall into the existing collecting practices of libraries, museums and archives. Discussion forums and a range of rather ephemeral websites offer considerable value as historical records. As noted in Bora Zivkovic's essay on science blogging, an outage on a popular science blogging network last year underscores just how easy it would be for a single point of failure to result in the loss of content documenting changes in

[6] For example, see ArXive.org's digital preservation plans with Cornel University's Library http://arxiv.org/help/support/whitepaper

[7] Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S. P., Bennert, N., Urry, C. M., Lintott, C., et al. (2009). Galaxy Zoo Green Peas: discovery of a class of compact extremely star ‐ forming galaxies. *Monthly Notices of the Royal Astronomical Society*, *39 9*(3), 1191-1205. doi:10.1111/j.1365-2966.2009.15383.x

science communication, and a diverse collection of responses and reactions to scientific research.

WHY IS ONLINE SCIENTIFIC DISCOURSE VALUABLE?

Below are three kinds of value the participants identified in this content. These are not meant to be exhaustive, but instead as a starting point for explaining why this web content is important.

The Record of Scientific Knowledge, Discovery, and Innovation:

Much of the history of science, technology, medicine, and mathematics is built from primary records of scientific publication and unpublished materials of scientists. Traditionally, material has been preserved through a combination of collecting the personal papers of scientists and their published work in books and journal articles. With the emergence of practices like open notebook science, science blogging, and science discussion forums a considerable amount of this content is being produced and presented on the web. If we do not act to collect this contemporary material, we may end up with more complete records of scientists' unpublished notes and personal communication from previous eras than we do from our own.

Related, the emergence of citizen science projects has resulted in some discoveries and advances in science happening on the open web. For example, the discovery of the green pea galaxies occurred entirely on the discussion forums that accompany the Galaxy Zoo website. The forums, where these kinds of discussions occur, document the process and contributions of individuals in scientific discoveries.

Changes in Scientific and Scholarly Communication:

Aside from documenting the record of science and discovery, the new media of blogs, websites, and forums are themselves documentation of significant changes occurring in scholarly communication. Much as work on the history of the book documents an array of changes in culture, the history of online communication media are themselves of considerable value in understanding science and scholarship in contemporary society. In this respect, these sites are going to be of interest as valuable primary sources in the history of technology, communications, and media.

Public Understanding and Perception of Science and Science Policy:

Conversations and reactions to science from members of the general public represent one of the most exciting prospects for historians of the future to understand science in our times. In particular, various controversies around topics like evolution, vaccines, and climate change have stirred up an enormous amount of online discussion. Records of these discussions will be invaluable for historians and policy analysts for understanding and exploring public reactions and perspectives on science. Furthermore, various pop-cultural developments that touch on science topics (for example, videogames like Spore) are similarly likely to generate substantive online discussion and offer potentially unique perspectives on science in our times.

USE AND REUSABILITY IN COLLECTION DEVELOPMENT

Because the purpose of preservation is reuse, participants urged that data be as well documented and standardized as possible. What those terms mean depends very much upon the data and potential uses. Raw data, for example, should be in standard formats to ease processing for pattern recognition, mining, simulation, longitudinal studies, and so forth.

Participants also suggested there be some measure of collecting samples of records of online scientific discourse *just in case*, specifically, gathering data at scale and keeping in relatively low levels of curation to reduce costs required for cataloging and description. This is recommended for data that seem relevant but may have no short-term demand. For example, embracing an all-hands-on-deck approach to documenting significant events, such as tsunamis, earthquakes, and hurricanes could include lots of data in an archive for later analysis. It would be impossible to predict exactly what future researchers will want access to. The Blue Ribbon Task Force on Sustainable Digital Preservation and Access recommended the capture of such data at a very low level of curation so that they may be discovered and processed in the future if deemed desirable. [8]

Simultaneously, there was consensus around the need to collect small, highly-curated topical collections of web content focused on ensuring long-term access to small representative sets of material in which scientists and historians of science see long-term value. The idea here would be to ensure high levels of quality assurance for collected content and a strong curatorial role in organizing and arranging collections as a point of entry into the much broader swath of content.

CALLS TO ACTION

As a result of the discussion at the summit, and the following essays, we suggest four calls to action for cultural heritage organizations.

Call for Engaging, Assisting, and Supporting Content Creators:

The scientists and science communicators who participated in the summit were eager to learn more about how they could help to manage and steward their content. Eventually, the personal documents of scientists often make up special collections at libraries and archives. There is considerable value in the cultural heritage community creating guidance materials for managing personal digital information. Specifically, reaching out to scientists and science communicators to help them better steward their own content can help creators self archive. The Library of Congress provides personal archiving guidance to the general public that can be customized and redistributed to a specific audience. [9]

Call for Developing Relationships with Online Science Communities:

[8] *Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information*, p 68
[9] htpp://digitalpreservation.gov/personalarchiving/

The organizations or communities that host or contribute to online science projects or discourse must care for their assets in the near term. Cultural heritage institutions have the mission and expertise to serve as long-term stewards. Relationships at the institutional level can be built to give guidance on preservation practices during the life of a project and advise on future curatorial homes for data when organizational affiliations change.

Call for Targeted Web Archive Collections:

To meet the challenge of stewarding this content, we suggest cultural heritage organizations begin to develop focused web archive collections related to their particular institutional goals and needs. For example, a focused special collection on open notebook science, or a collection focused on controversies around vaccines, or the web presence of its scientists and science centers. Cultural heritage organizations are uniquely positioned to, based on their own particular focuses, identify and collect around particular themes and topics that can collectively serve as part of a distributed national and international online science collection. The case study of U.S. National Library of Medicine's Health and Medicine Blogs collection provided in this report can serve as an exemplar. Also included are examples of a series of different kinds of special collections we could see different cultural heritage organizations developing as an appendix.

Call for Outreach to Historians and Other Researchers:

Stewardship organizations must establish a user community which values the content they are preserving. There is not yet substantive interest from historians of science and other researchers in online scientific discourse. While researchers and scholars of literature and the arts have been engaged in helping develop practices around the collection and preservation of born digital artwork and literature, there has not been a similar reaction in the history of science community. Archivists, librarians, and curators ought to reach out to historians of science and make them aware of the born-digital primary resources that can be collected. Simply put, without intervention, much of this online discourse is likely to disappear before historians of science take an interest. Engaging professional organizations and associations for these researchers will be a critical component in developing sound collection approaches and policies.

The Historical Value of Ephemeral Discussion of Science

FRED GIBBS, ASSISTANT PROFESSOR OF HISTORY AT GEORGE MASON UNIVERSITY AND DIRECTOR OF DIGITAL SCHOLARSHIP AT THE CENTER FOR HISTORY AND NEW MEDIA

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 collection 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. 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 AUDIANCES

Science at Risk summit 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 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 attending the Science at Risk summit showed how many participants are actively creating and curating online science content according to their varied needs and interests. Summit 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 are 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 summit—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— 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 online forums of some sort. For example, the violent storms that swept through the Washington, DC. area in the summer of 2012 were the subject of numerous newspaper articles that prompted user comments mentioning 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 grey 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 web page where the content resides. Historians will hope for as much contextual metadata and *paradata* as possible, and their analysis will be as rich as that 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 is needed. 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 web pages 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 summit 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 of 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 come 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 online discussions and presentations of science 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 summit 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 that arise 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*. New collection strategies are required, 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.

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 newspapers 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. These efforts should 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. Regarding 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 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 allocate resources from cataloging to making items visible directly through their content, rather than classification. Obviously not all items lend themselves to full-text searches, but many do, 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 summit 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 subject matter experts (SMEs) 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 grey literature that cannot be easily slid into any other preservation domains.

The Science at Risk 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

collections from 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.

Ten Years of Science Blogs: A Definition, and a History

BORA ZIVKOVIC, BLOGS EDITOR AT SCIENTIFIC AMERICAN, VISITING SCHOLAR AT NYU SCHOOL OF JOURNALISM AND ORGANIZER OF SCIENCEONLINE.

What makes a science blog? Who were the first science bloggers? When did they start? How many science blogs are there? How does one differentiate between science blogs and pseudo-science, non-science and nonsense blogs? The goal of this article is to try to delineate what is, and what isn't a science blog, what are the overlaps between the Venn diagram of science blogging and some other circles, and what out of all that material should be archived and preserved forever under the heading of *Science Blogging*.

DEFINING A SCIENCE BLOG

Defining a science blog—or for that matter, just defining a blog—is difficult. After all, a blog is just a piece of software that can be used in many different ways.

What is considered a science blog varies, and has changed over the years. Usually it should satisfy one or more of these criteria:

- blog written by a scientist,
- blog written by a professional science writer/journalist,
- blog that predominantly covers science topics,
- blog used in a science classroom as a teaching tool,
- blog used for more-or-less official news and press releases by scientific societies, institutes, centers, universities, publishers, companies and other organizations. '

But is a blog written by a scientist that never covers science really a science blog? Is a blog by a PhD in dentistry who spews climate denialism in every post a science blog?

What is considered a science blog also changes with the advances in technology. There is now a fine-grained division of blogging into macro-, meso- and microblogging. Initially, this distinction was made by technology. Macroblogging happened on platforms like WordPress or Blogger, mesoblogging on sites like Posterous or Tumblr, and microblogging on social media like Twitter and Facebook. But technology moves, and now it is possible to do all three *sizes* (or is it *speeds*?) on any of those platforms—and some people do.

Is a one-liner posted on a blog the same as a one-liner posted on Twitter? Some posts on Facebook and Google Plus are longer and more thorough than some others that use the more traditional blogging platforms like WordPress, Blogger or Drupal. Yet Google Plus is very new and Facebook, until recently, had quite a short word-limit. Many people used blogging software to do very brief updates back when that was the only game in town. Today, quick updates, links etc. are done mainly on social media and many bloggers use the traditional blogging software only for longer, more thorough, one could even say more *professional* writing.

Finally, blogging is not just about text. There is photoblogging, videoblogging, podcasting, etc. And for each of these specialized types of blogging, one can potentially use a traditional blog software, or instead choose to do it on social networks, or on specialized sites, e.g., Flickr, Picassa, Instagram, Pinterest, Tumblr, YouTube, DeviantArt, etc. Does all of that count?

THE BEGINNINGS OF SCIENCE BLOGGING

Pin-pointing the exact date when the first science blog started is a fool's errand. Blogs did not spring out of nowhere overnight. The first bloggers were software developers who experimented with existing software, then made some new software, fiddling around until they gradually hit on the format that we think of as a *blog* today. The evolution was gradual in the world of blogging. It was also gradual in the more specific world of science blogging.

The earliest science bloggers were those who started out doing something else online—updating their websites frequently, or participating in Usenet groups—then moving their stuff to blogging software once it became available in the late 1990s and early 2000s.

As much of the early online activity focused on countering antiscience claims, e.g., the groups battling against Creationism on Usenet, it is not surprising that many of the early science bloggers came out of this fora and were hardly distinguishable in form, topics, and style from political bloggers. They brought a degree of Usenet style into their blogs as well: combative and critical of various antiscience forces in the society. And certainly, their online activity had real-world consequences and successes. For example the Dover trial for which a decade of resources accumulated by the bloggers and their community, in some cases presented at the trial itself by those same bloggers, helped defeat a Creationism bill in a resounding manner that, in effect, makes all future efforts to introduce such bills relatively easy to defeat.

Phil Plait, Chad Orzel, Razib Khan, Derek Lowe, David Appell, Sean Carroll, P.Z. Myers (whose blog started as a classroom teaching tool), Tim Lambert, John Wilkins, Chris Mooney, and Carl Zimmer were some of those early science bloggers. Panda's Thumb blog and Larry Moran's Sandwalk are for all practical purposes direct descendants of the old Usenet groups. Real Climate has, I believe, similar origins. Among early adopters of blogging software, rare are the exceptions of people who instantly started using it entirely for non-political (and non-policy) purposes, just to comment on cool science, or life in the lab, etc., e.g.,

Jacqueline Floyd, Eva Amsen, Jennifer Ouellette, Zen Faulkes, and Grrrlscientist.

In those early days, we pretty much all knew, read, linked, blogrolled, and responded to each other, despite a wide range of interests, backgrounds, topics, etc. As the blogosphere grew, the nodes appeared in it, concentrating people with shared interests. Those nodes then grew into their own blogospheres. Medical blogosphere, skeptical blogosphere, atheist blogosphere, and nature (mostly birding) blogosphere used to be all part of the early science blogosphere, but as it all grew, these circles became separate with only a few connecting nodes. Those connecting nodes tend to be veteran, popular bloggers with large readerships, as well as bloggers on science blogging networks (e.g., at Scientific American, Discover, PLOS, or Wired) which tend to want to have representatives from many areas, e.g., medical bloggers mixed in with paleontology bloggers mixed in with space bloggers, etc.

SOME KEY MOMENTS IN THE EVOLUTION OF SCIENCE BLOGGING

I will now try to identify some of the events and developments in the history of science blogging that, in my opinion, were especially important in the direction science blogging evolved: the changes in styles, the growth in size, and the rise in respectability.

SCIENCE BLOG CARNIVALS

What is a blog carnival?

It is a crowd-sourced online magazine, occurring at a regular interval (e.g., weekly, monthly), usually rotating hosting blogs for each edition. Bloggers submit their best posts from a particular period or on a particular topic to the next editions' host who accepts (or rejects) the entries, and edits a blog post that contains nicely arranged and introduced links to all the entered posts. Thus, it is a well-defined, well-archived, regular, rotating linkfest. Usually all the included bloggers link back to the carnival from their blogs (as well as other online sites, e.g., social networks) thus bringing attention and traffic to the host, as well as to all the bloggers whose work is included in that edition.

The very first such "rotating blog magazine" was started in 2005 under the name "Carnival of Vanities" (from which the phenomenon got its name) and the concept quickly spread like wildfire.

One of the very first carnivals was started by P.Z. Myers. This was Tangled Bank (unfortunately, the archive appears to be gone). This weekly rotating linkfest helped science bloggers discover each other, promote themselves and each other, encourage new people to start blogging, and start building a community. Several spin-offs showed up later, e.g., Grand Rounds (medicine), Skeptics' Circle (countering pseudoscience), I and the Bird (birds), Circus of the Spineless (invertebrates), Berry Go Round (plants), Change of Shift (nursing), Friday Ark (animals, mostly photos), Encephalon (neuroscience), The Accretionary Wedge

(earth science), Carnival of the Blue (marine science), The Giant's Shoulders (history of science), Festival of the Trees, Carnival of Mathematics, Carnival of Space, and a few dozen others. Some of those are still around, but most have closed after a good multi-year run.

With the more recent development of social media, the carnivals are not seen as important for community building as they once were. First came the feed readers, and feed aggregators (especially FriendFeed) that made it easier for one to track and filter blog posts and other content by topic or some other criteria. The primary function of the carnivals—to build community—could easily be done in these new spaces. Then Twitter came along, though it took some time for people to figure out how to use it, to invent various Twitter norms (e.g., RT, hashtags, @reply), and to build apps that make Twitter more useful.

A little bit later, Facebook bought FriendFeed and imported all of its good functionalities (e.g., "Like" button, "Share" button, "Friend of Friend", "Pages", video embed, toggling between "Top stories" and "Most recent" on the homepage feed, etc.), lifted the word-limit on status updates, made importing other feeds easy, and made long-form blogging easy as well. Finally, about a year ago, Google Plus was launched—essentially FriendFeed on steroids, linked more and more intimately to all the other Google stuff, from Gmail to Google Docs to YouTube to Picassa. Give them another year, and G+ will become what FriendFeed would have been if it was not sold and continued to be developed.

All of those platforms make community-building easier than traditional carnivals. It is easier to do. It is easier for newbies to join in and get noticed. It is easier for one to individualize a degree of engagement with that community. But the easier the community-building gets, the harder it is to perform the second key role of carnivals—as archives. Each edition of a carnival is a magazine, a snapshot of the moment, and a repository of pieces that both their authors (by submitting) and hosts (by accepting) thought were good and important. And when a carnival dies, and the archives' host subscription expires, all those historically important links are gone!

In place of carnivals, what people tend to like these days are linkfests done by individuals who serve as trusted filters. I started doing it myself a couple of months ago, picking perhaps a third of the links I tweet over a period of a week and organizing those links in a single blog post. In the very first installment of my Scienceblogging Weekly, I wrote:

> *Ed Yong's weekly linkfests and monthly Top 10 choices he'd pay for are must-bookmark resources.*
>
> *Some other bloggers are occasional or regular sources of links I pay attention to, e.g., John Dupuis on academia, publishing, libraries and books, Chad Orzel on academia and science—especially physics, Mike the Mad Biologist on science and politics, and the crew at the Knight Science Journalism Tracker for the media coverage of science. And at the NASW site, Tabitha Powledge has a must-read 'On science blogs this week' summary every Friday.*

These one-editor carnivals seem to be the fashion of today. But old-style carnivals were, in my opinion, better both at community building and as historical archives.

RESEARCH BLOGGING

Second important moment was the start of a new blog, Cognitive Daily, written by Dave and Greta Munger. They pioneered the form of blogging that was later dubbed *researchblogging*—discussing a particular scientific paper (which is referenced and linked at the bottom), usually in a way that lay audiences can understand.

At the time, science blogging was developing its own norms, as there is no such thing as *word limit* online (blog posts tend to be much longer than traditional news articles, not cutting out any relevant context out of the post), bloggers instinctively understand the value of links (which forces them to research much more thoroughly than the usual daily news article), blogs tend to have a more chatty and personal style, yet most science bloggers are either experts in their fields (thus no need to interview other experts just to get the quotes) or have acquired expertise by covering a topic for decades (e.g,. Carl Zimmer on evolution), thus can speak with authority.

Even today, but especially in the early days, bloggers usually did not care to cover brand new papers the moment the embargo lifts. In the early days, coverage of papers was quite rare. Apart from debunking pseudoscience, much of early blogging was more educational than journalistic—covering decades of research on a topic, or explaining the basics. If they covered a paper, bloggers were just as likely to cover an old, historical paper as a new one.

But when Dave and Greta started their blog, others took note. With the researchblogging style, not only can the blogger report on a paper, but there is also a way to embed videos, polls, animations, etc, to make the readers engage much more actively—which their readers did. In many a post they did a sort of quick-and-dirty replication of studies online, with readers as volunteer subjects.

This format of blogging rapidly took off—many bloggers started emulating it (especially new bloggers)—probably vastly outnumbering the anti-pseudoscience bloggers today. Formation of the ResearchBlogging.org site, with its icon, code and aggregator, also made this type of blogging attractive to newcomers. Probably the best example is Ed Yong, who instantly took to the format, blogging about at least one paper per day, often covering nifty papers that the rest of the media missed. And Ed covered new papers. The moment embargo lifted. This was obviously journalism even to the most traditional eyes. This was something that other journalists, or people hoping to get into journalism, could also do. So they did; in droves.

BLOG NETWORKS

Third important moment in the history of science blogging was the start of science blogging networks. The first one was Nature Publishing Group's Nature Network. It was essentially an accident—the site was supposed to do something else, but ended inviting people to write blogs instead. Unfortunately, due to technical architecture, it is not well connected to the rest of the world (for example: posts, if they show up on Google Blogsearch at all, show up with several days of delay). One had to remember to go there instead of having the links thrown in one's face wherever one may be online. Also, the initial strategy of the network was to ask researchers to blog, but very few of them took to the format very well—most of their blogs had one post and then died. Those few who did start blogging well, found themselves isolated, not knowing who is reading them, or even how many did. After a decade, the network has undergone some changes, the bloggers have rotated in and out with some excellent writers there now, and it appears to be more visible now than it used to be when it first started due to it's move to a new domain—Scilogs.com.

The second network (launched in January 2006), Seed Media Group's Scienceblogs.com was what really made a difference. Here was a media organization vouching for the quality of bloggers they hired to write on their site. And they picked bloggers who already had large readership and traffic, as well as clout online, the likes of P.Z. Myers, Orac, Grrrlscientist, Tara Smith, the Mungers, Revere, David Kroll, Tim Lambert, Ed Brayton, Razib, etc. This gave the network's bloggers respectability, and the rest of the mainstream media got into a habit of checking Scienceblogs.com as their source of science news online.

A couple of other networks started relatively early in the history (Scientificblogging.org which was later renamed Science2.0, Discover, Discovery News, Psychology Today, Smithsonian, . . .), but mainly dwelled in the shadow of Scienceblogs.com until the infamous #Pepsigate affair (when the addition of a blog written by Pepsi PR people resulted in a fast mass exodus of a large number of bloggers).

OPEN LABORATORY

The fourth important moment was the first edition of the Open Laboratory, an annual crowdsourced anthology of the best writing on science blogs. After five years of getting published at Lulu.com, the sixth edition was published in September 2012 by FSG, imprint of Scientific American at MacMillan. Here was, as early as January 2007, a collection of some amazing blog writing about science, in traditional book format, built by the community itself. It really helped the community define itself. Gaining an entry into the anthology became a big deal. The Open Laboratory was a project designed to go together with the first ScienceOnline conference, and although the publication date is now completely different from the date of the meeting, the books are still a project of the ScienceOnline organization. The conference itself added to the feeling and spirit of the community in a way that gatherings of techie, skeptical, atheist or political

bloggers could never accomplish.

For many people, seeing words printed on paper still carries a certain dose of respectability. After all, the real estate of paper is expensive. A book is a result of a large investment of time, money and effort—either bottom-up, by the author (sometimes perceived as a result of a big ego), or top-down, with an editor choosing what material is worth the investment.

Open Laboratory turned that on its head. Authors submit what they think is their best work, trusting that a jury of peers will fairly assess them, choose the best pieces, perhaps improve them a little bit (more this year than in previous years), and that the entire community will help promote the final product. Inclusion of a blog post in #openlab is not just a result of the whim of an editor, but a result of two or three rounds of judging by multiple people all of whom are also science bloggers and writers. This mutual trust matters.

AWARDS

Early on there were Koufaxes, later Webbies, and all sorts of other blogging awards. Some of those had awards for science blogging. But if the managers of the award allow bloggers who only pretend to be scientists and use seemingly-scientific language to push pseudoscience (e.g., global warming) into the Science section of the awards, then real science bloggers react with disdain, and ignore that particular award in the future. When the award is set up essentially as a popularity contest, and when such antiscience bloggers, due to hordes of followers, win such contests, then there is no real reputation linked to that victory. Thus there is no need for science bloggers to expend their energies or in any way promote such awards.

Fortunately, over the last few years, a reputable award for science blogging emerged (the fifth important moment in the evolution of science blogging), the 3 Quarks Daily Award, with three rounds—one with reader voting, one with jury voting, and final judgment by the prominent judge who declares the final winners out of ten or so finalists. The winners get money, and proudly sport the 3QD buttons on the sidebars of their blogs.

THE AFTERMATH OF #PEPSIGATE

The sixth important moment was #Pepsigate, when Scienceblogs.com broke up and about a quarter of the bloggers left. The time was ripe for it—there were too many science bloggers around, yet only blogs at Scienceblogs.com got any traffic or respect. That was an unstable situation. So many good bloggers were out there, writing wonderfully, but were essentially invisible under the shadow of “The Borg.”

In the wake of #Pepsigate, existing networks (e.g., Discover, Nature Network) redesigned their sites and brought in some of the bloggers fleeing Scienceblogs.com. New networks sprung up almost instantly to lure in more of these

blogging veterans. There were new networks started by organizations like Wired, The Guardian, PLOS, NatGeo, AGU, ACH, as well as self-organized science blogging collectives like Scientopia, Field Of Science, Science3point0, and Lab Spaces. The last one to launch was the Scientific American network, which celebrated its first anniversary in July 2012.

Being on one of these networks became a stamp of approval for the bloggers, and we quickly built the Scienceblogging.org site (which is about to undergo a thorough rebuild and redesign, and also a project of ScienceOnline organization) to help people find all of the networks, collectives, and key group blogs all in one place. While the inclusion there is not as stringent a process as it is on ScienceSeeker.org, this site is also a proxy for quality in some ways, as most of the blogs appearing there wear the imprimatur of traditional organizations, be it the media, publishers, or scientific societies, or the warranty by their colleagues who invited them to join their collectives. This site has, to many in the mainstream media as well as bloggers and readers, replaced scienceblogs.com as the homepage where they start their day.

AGGREGATORS

I have already mentioned above that an important moment in the history of science blogging was the start, by Dave Munger, of the website ResearchBlogging.org which aggregates blog posts from science blogs, but only if the posts contain the code indicating that the post is covering a paper. The code also renders the citation correctly in the post itself. As the site has editors who decide which applicants can be accepted (or rejected), this became an unofficial stamp of approval, the first method of distinguishing who is and who is not a science blogger.

A couple of years later, when PLOS started accepting bloggers onto their press list, being a member of ResearchBlogging.org was the criterion used for acceptance to the press list (I know this as I was the one doing the approval at the time as their blog/online manager). A little later, PLOS introduced its Alt-metrics on all of their papers. One of those metrics counts the number of blog posts written about the paper. Going through Google Blogsearch and Technorati brings in all sorts of spamblogs, or people who use blogging software to post copies of press releases, instead of genuine science bloggers. Thus PLOS used ResearchBlogging.org as a filter on their papers.

As ResearchBlogging.org is owned by Seed Media Group, now controlled by NatGeo, and as there seems to be no technical support, financial support, or development of the site any more, people who are using it are advised to switch instead to the successor site, ScienceSeeker.org—another project of the ScienceOnline organization, which is a much better site that serves the same purpose but also does much more, has some funding (and is asking for more) and is in constant development. Dave Munger is, again, one of the key people involved in the development of this site. At ScienceSeeker.org, one can filter by discipline, or only show posts that have the ResearchBlogging.org code in them, or only show posts that ScienceSeeker editors have flagged as especially good. Both

ResearchBlogging.org and ScienceSeeker.org now count (as far as I know) around 1200 blogs on their listings (with much, but not complete, overlap). More blogs need to be added for the site to become a more comprehensive collection, but blogs that are on there are a pretty good snapshot of the core of the scientific blogosphere today.

SIZE OF THE SCIENCE BLOGOSPHERE

It is relatively easy to count science blogs in *smaller* languages, e.g., German, Italian, French, Spanish, or Portuguese, with several dozen each at most. It is much more difficult to count science blogs written in English, Russian, Chinese, or Japanese—those most likely count in multiples of thousands. But it is impossible to make a good estimate as it depends on one's definition.

Searching Google or Technorati brings up many blogs with a "science" tag that have nothing to do with science—or worse (spam blogs, antiscience blogs, etc). Researchblogging.org and ScienceSeeker.org are still too small to be useful for counting the total size of the blogosphere.

How does one count blogs that have not been updated in six months—on hiatus or dead? How does one count multiple blogs by the same person, perhaps not even updated simultaneously but successive editions of the blog (e.g., as the person moves from one network to another)? One blog or many? Does one count classroom blogs, at least those that are not set on *private*? How about institutional news blogs? Are they *real blogs* or just an easy software to use to push press releases? And do press releases count? We can fight over this forever, I guess, so I'd rather concede that blogs are uncountable and leave it at that.

RISING POWER AND RESPECT

I have written recently, in an article for a Croatian newspaper, about the history of science blogging and the problem of delineation of who is in and who is out. In that article I also mentioned some events that added to the respect of science blogs, e.g., Tripoli 6 affair, George Deutch affair, the PRISM affair, and #arseniclife affair, though there have been many other cases in which science bloggers uncovered wrongdoing, or forced media to pay attention to something, or forced action on something important. Some of those cases involved clearing the record within science, others had effect on broader society or policy.

Each one of these cases strengthened the respect for science bloggers. In some cases they did a much better job reporting than the mainstream media did. In others, they tenaciously persisted on a story until they finally forced the mass media to pick up the story and broadcast it to bigger audiences that, in turn, could effect a change (e.g., by calling their representatives in Washington). In many ways, science bloggers shocked the old system and built a new system in its place.

Increased reputation also came from cases in which bloggers solved scientific problems online, in public, for everyone to see. The most famous case is, of course, the Polymath Project, in which Tim Gowers and his readers solved an old

mathematical problem in the long comment section of his blog post. The details of the project, as well as why it was so important for open science, were wonderfully detailed in Michael Nielsen's book *Reinventing Discovery: The New Era of Networked Science*.

The best such example to date is the #arseniclife affair because it did two things simultaneously. First, the scientists with relevant expertise took to their blogs to critique, criticize, and debunk the infamous paper about the uptake of arsenic instead of phosphorus by the DNA of a strange bacterium living in a Californian lake. That is not so new—bloggers criticize studies all the time, with expertise and diligence and thoroughness. But importantly, a second thing also happened—the attempt at replication of the experiment was live-blogged by Rosie Redfield, describing in painstaking detail day-to-day lab work, getting technical feedback from the commenters, resulting in the Science paper demonstrating that the experiment could not be replicated. This was a powerful demonstration of the process of Open Notebook Science as one of the things that scientists these days can do with their blogging software.

PROFESSIONALIZATION OF SCIENCE BLOGGERS

You may have noticed recently the so-called Jonah Lehrer affair. In the aftermath, Seth Mnookin used his blog to further explore, in three long blog posts, the professionalization of blogs, and the blurring of the lines between blogging and mainstream journalism.

One of the most interesting reactions by some of the Scienceblogs.com bloggers during #Pepsigate was "we are not journalists, I am not the media." But they were. If your blog is indexed by Google News, hosted by a media company, you are the media. New media perhaps, but it's still media. More personal, more conversational, but still media.

The issue with Jonah Lehrer was something people called *self-plagiarism,* i.e., re-using one's own old words in a new article (true plagiarism was uncovered a little later). This is the clash between old media ("our content is exclusive!") and new media ("my blog is my writing lab where I develop my ideas over time"). Judging from all the discussions, journalists, bloggers and readers are all over the place regarding this issue. Is it OK to reuse one's old words if one is not paid? Is it OK if one is transparent (perhaps using links to old posts, or quotes—I am all for it and do it myself a lot)? Is it OK on a blog but not in an article (and how does a reader know what is what)? Is it OK to reuse one's own tweet or Facebook update (because it is not always thought of as blogging, an attitude which I find silly), but not OK to reuse words that occurred on a WordPress platform? What is the real difference here?

Obviously, the times are in flux. Some science bloggers would rather not be considered media, and not asked to write the way journalists write. Some prefer to use their blogs as writing labs—often repeating and reiterating ideas and words and sometimes entire passages in new contexts, with a new angle or twist—gradually adding and changing their own thinking over the years, introducing new

readers to old ideas (after all, who digs through the years of archives?), with no intention of ever turning that material into commercial fare, e.g., a magazine article or a book.

If your beat is debunking anti-vaccination misinformation, how many ways can you do that if you post every day? And getting a couple of hundred dollars per month for editor-free posting on someone else's site is not really *professional writing* in a traditional sense. Writing under the banner of a well-known media organization, while it confers respectability by virtue of being chosen to be there, does not automatically mean that blogging is the same as reporting news or writing professional op-eds. There is much more freedom guaranteed. More editorial control would require much more money in exchange.

On the other hand, some science bloggers see their blogs as potential marketing tools for themselves as writers. Their blogs are a different kind of a *writing lab*—a place to write more fine-tuned kinds of pieces, more *journalistic*, in hope of being seen and then getting gigs and jobs in the media. They tend to cover new papers, rather than write broader educational pieces. They try to proofread and polish their posts better. And why not? Nothing wrong with that. Just like there is nothing wrong with NOT wanting to do that either. Many scientist-bloggers really have no journalistic ambitions. Others do. Each has different goals, thus different writing styles and forms, slightly different ethics, and a different understanding of what their blogs are all about.

During one of the debates about professionalization of science bloggers, I heard a sentiment that bloggers with no journalistic ambitions should not confuse everyone by being on networks hosted by media organizations. As an editor of one of those networks, I beg to differ. I want all kinds of bloggers, all styles and formats, because I want to diversify our offering, I want to have something for every kind of reader—from kids to postdocs, from teachers to researchers and more. I want to blur the line between old and new media, make it so new, more web-native forms of stories become a norm, not just the old tired inverted pyramid.

The world of media is rapidly changing and, in many ways, returning to the many-to-many communication that we are used to, the 20th century broadcast model being the only weird exception in history. Mixing and matching various styles of communication in one place, especially a highly visible place, is a good thing for science as each piece will be interesting to a different subset of the potential audience, which will keep readers coming back for more, looking around, and learning how to appreciate other styles as well.

I want cool science to be everywhere in the media ecosystem—from movies and television, to theater and music, to newspapers and magazines, to books and blogs and tweets. I want the science communicators to practice the new journalistic workflow which assumes, almost by definition, that a lot one says will be repeated over and over again in various places in various contexts. Self-plagiarism does not make sense as a concept in this model. Self-plagiarism IS the new model—that is how good ideas get pushed (as opposed to pulled) to as many audiences, in as many places, over as many years as possible.

On one hand, bloggers need to adjust. Moving from independent blogs to Scientific American put a lot of our bloggers into a phase of self-reflection. They sometimes try to write perfect posts (and sometimes need encouragement to just throw things up on their blogs even if they are not entirely perfect). But blog posts are not supposed to be, with occasional exceptions, polished, self-contained pieces. A blog post is usually one of many in that person's series of posts on the same topic, reflecting personal learning and growth over the years. Or a post on something new to the person, a way to organize one's own thoughts about a very new topic. That post is also a part of an ongoing conversation the blogger has with regular readers and commenters. That post is also part of a broader online (and sometimes also offline) conversation.

A blog post is just a very long tweet in a series of other very long tweets, usually, but an occasional polished diamond is certainly welcome as well. It is a writing lab, after all, so occasionally a perfect article may appear. But focusing on that goal is misguided—a blog is a place to think in public. And if the media host understands that, then there is no question or problem of *self-plagiarism*.

On the other hand, readers also need to adjust. When they arrive at a media site, they should learn not to expect a self-contained inverted pyramid every time. Blogs have been around for fifteen years, they are not so novel any more. It's easy to see if a site is a blog, if it reads like a blog, and one should know what one should expect on a blog. I think that most complaints in the comments are really trolling—people who dislike what scientific research concluded complain about typos, or format, or length, or tone, in order to divert the discussion that makes them personally uncomfortable. Our bloggers have full moderation powers to deal with such comments in any way they see fit.

CONSIDERATIONS ON PRESERVING SCIENCE BLOGS

After #Pepsigate, many bloggers feel the freedom to move from one network to another, or on and off networks, with considerable ease and speed. What happens to the archives? A couple of months ago, someone at National Geographic flipped the wrong switch and years of archives from almost a 100 science blogs were gone. Completely gone, even blocked from viewing at Wayback Machine and Google Cache. It took a dozen of tweets to get the attention of some of their bloggers, who contacted the relevant person who flipped the switch back on Monday morning, making all those historically very important archives accessible again. See how easy it is to erase history? Perhaps with Radio2+River2, if it is universally used, this would not be a problem. Wait and see.

For a huge archive to be useful to users—and that's what such an archive is for—it has to be organized in a meaningful way. Should it be by topic or by person? By narrow area, or by a whole discipline (human genome or entire genetics)? Or by technological platform (tweets to the left, datasets to the right, blog posts straight ahead)? Or separate independent blogs from network and institutional blogs? If all of the stuff all of the science bloggers in the world have ever posted on all of their blogs is to be archived and preserved, how should that material be organized?

Chronologically, minute by minute? Or in chunks akin to blog carnivals? Or sorted by topic? Should papers be connected to blog posts that discuss those papers? Should #arseniclife be its own *unit*?

Another problem is privacy. Facebook has many privacy settings. Tweets, and some blogs, occasionally switch from private to public to private—what is a repository to do with stuff that is uncertain if it is private or public at any given time? Should the archiving be opt-in? In that case, how does one ensure that most of the people opt in so the repository is of decent completeness?

Also, many blog posts are reactions to other sites. A blog post may debunk a claim from a creationist, or anti-vax or GW-denialist blog, linking to it and quoting from it. If science blogs are preserved, but antiscience blogs are not, there will be link rot right there, preserving reactions without the context of the reactions. So perhaps all those antiscience and pseudoscience blogs should also be preserved—they may be bad science, but they are an important aspect of today's society and will be interesting to future historians. In which case, how does one label them? They are clearly not science blogs (although some of them pretend to be), so they should not be just thrown into the same bag. Which is why this delineation between *real* science blogs and other stuff has to be made.

And how will this decision be made and by whom? Should something like ScienceSeeker be used as an edited, peer-reviewed collection of respected science bloggers? If so, how does one get more bloggers to know about this and apply to it?

Developing a "Health and Medicine Blogs" Collection at the U.S. National Library of Medicine

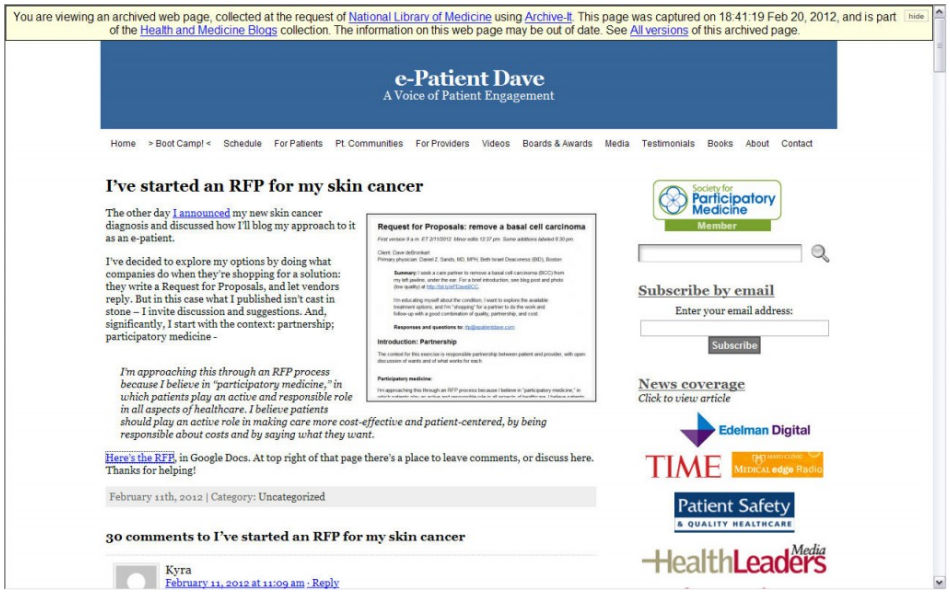
CHRISTIE MOFFATT, ARCHIVIST IN THE U.S. NATIONAL LIBRARY OF MEDICINE HISTORY OF MEDICINE DIVISION AND PROGRAM MANAGER OF NLM'S DIGITAL MANUSCRIPTS PROGRAM, AND JENNIFER MARILL, CHIEF OF THE TECHNICAL SERVICES DIVISION FOR THE NATIONAL LIBRARY OF MEDICINE

The United States National Library of Medicine (NLM) has a mandate to collect, preserve and make accessible the scholarly biomedical literature as well as resources that illustrate a diversity of philosophical and cultural perspectives not found in the technical literature. New forms of publication on the web, such as blogs authored by doctors and patients, illuminate health care thought and practice in the 21st century. In June 2011, the NLM Web Collecting and Archiving Working Group engaged in a pilot project to understand better the processes and challenges of collecting born-digital web content to expand the Library's collecting strategy for digital formats.[10]

The NLM working group gained a practical understanding of web archiving workflows and began the Health and Medicine Blogs collection, presenting the perspectives of physicians, nurses, hospital administrators and other individuals in healthcare fields. So far, the authors of these blogs are physicians, nurses, hospital administrators, and other health professionals in different stages of their careers. NLM also collected blogs of patients who are chronicling their experiences with conditions such as cancer, diabetes, and multiple sclerosis.

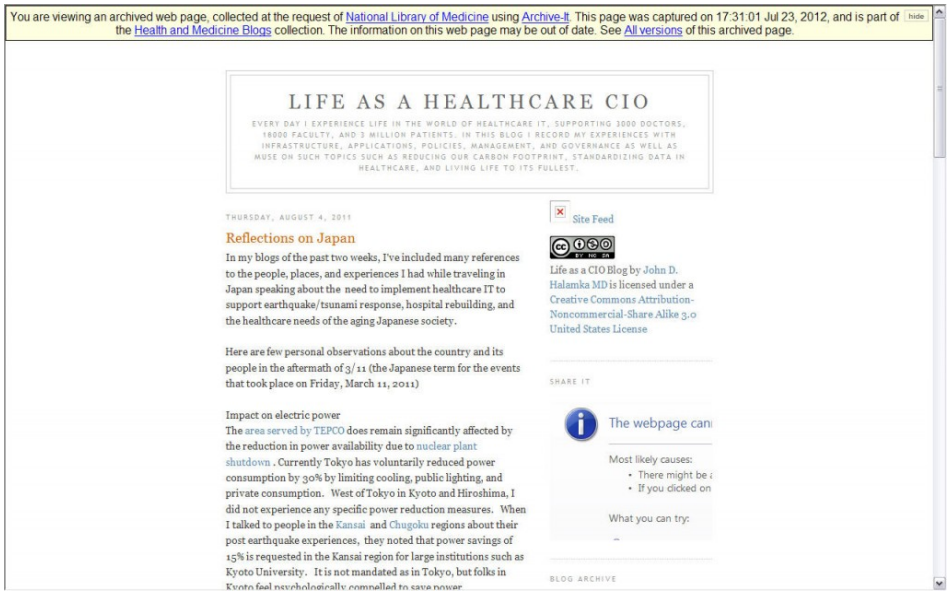
[10] http://www.archive-it.org/organizations/350?show=Collections

E-PATIENT DAVE



"E-patient Dave" is the blog of Dave deBronkart, a cancer patient and blogger who has become a noted activist for healthcare transformation through participatory medicine and personal health data rights.[11] Mr. deBronkart writes in this post as a newly diagnosed skin cancer patient who is taking action to make his treatment most cost-effective.

LIFE AS A HEALTH CARE CIO

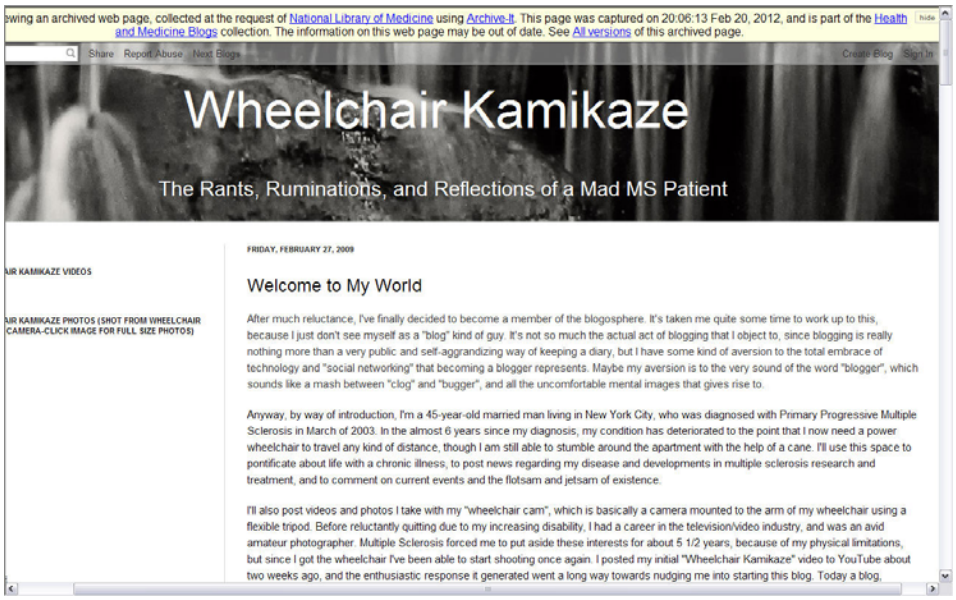


"Life as a Health Care CIO" is the blog of Dr. John Halamka, Chief Information Officer of Beth Israel Deaconess Medical Center in Boston, Massachusetts.[12] He is

[11] http://epatientdave.com/
[12] http://geekdoctor.blogspot.com/

also Professor at Harvard Medical School and a practicing Emergency Physician. In this captured blog post, Dr. Halamka reflects on his work in Japan on the implementation of health care IT to support earthquake/tsunami response.

WHEELCHAIR KAMIKAZE



"Wheelchair Kamikaze" is written by an individual named Marc with multiple sclerosis (MS). He drives his wheelchair at full speed and takes videos and still photos using a camera that he has attached to his chair.[13] He posts these images on his blog and writes about his experience living with MS. This is a screenshot of his first blog post on February 27, 2009.

During the pilot, NLM crawled selected blogs monthly over the course of a year using the Internet Archive's Archive-It service. NLM staff conducted monthly quality control reviews of the archived pages and made adjustments to the crawling instructions to better capture the look, feel, and functionality of the content. Throughout this effort the working group explored issues of selection, quality control, metadata, copyright, and the workflows needed to develop web-based collections.

Through a learn-by-doing approach, the group found that it was able to capture selected blogs fairly well despite known limitations to web archiving. One significant challenge that the group faced included dealing with frequent links to outside, *out-of-scope* sources, raising questions about what it means to completely capture a blog, and the extent to which linked content should be preserved. Other challenges included capturing content protected by passwords or blocked by robots.txt files, and some types of video files. NLM learned that capturing web content remains a moving target, and that with each new post, and certainly with overall structural changes to a blog, problems can quickly arise. The group's

[13] http://www.wheelchairkamikaze.com/

experience confirmed the value of early and thoughtful attention to scope, crawling frequency, and crawling duration, as well as the importance of thorough quality review. Test crawls were very helpful for identifying and addressing problems in advance.

Some of the biggest challenges were non-technical and included determining collection scope (this is a *big picture* question—which blogs should be captured?). Other issues were permissions (weighing the fact that these blogs can be quite personal) and monitoring when blogs end, change focus, or move to a new URL. NLM staff learned the importance of both curatorial and technical expertise and the need to keep up with new tools to get a better handle on working with this content. Perhaps most significantly, NLM gained first-hand appreciation for the importance of acting now, despite the imperfect methods of collection.

The Working Group has recommended that NLM expand traditional collecting capabilities to include born-digital web information and to participate in collaborative efforts to capture at-risk web content. As NLM moves forward, other areas of interest include:

- Capturing web-only grey literature, especially content from small *at-risk* organizational websites that do not already have affiliations with repositories or that lack the resources to archive their web content themselves
- Developing thematic collections, such as the intersection of medicine and art on the web
- Event-based collecting (for example, both official and non-official responses to epidemics, or in the aftermath of a disaster)
- Web content that complements traditional manuscript collecting (laboratory websites, online laboratory notebooks)

Eleven Brief Ideas for Web Archives of Online Science Discourse

100 Science Blogs: A highly selective collection intended to broadly sample the diversity of science blogs in terms of science topics, background of bloggers, and approach to presentation and format.

21st Century Public Science Controversies: A selective collection of blogs, forums, and subsections of sites like Reddit that focus on discussion of evolution and vaccines.

Citizen Science on the Web: A collection focused on preserving a range of significant citizen science projects. This would best focus on preserving both the look and feel of the online citizen science tools themselves, and the associated discussion forums and other online spaces in which users are communicating with each other.

Mathematics Discussions Online Collection: A collection of discussion forums in which professional mathematicians discuss issues and trouble shoot problems. This collection might also include discussion of particular mathematical software, like Mathmatica.

Open Notebook Science Collection: A collection of the wikis, blogs, and other kinds of content management systems that different scientists are using to practice Open Notebook science.

Popular Science on the Web Collection: A collection focused on collecting the work of amateurs, hobbyists, and science enthusiasts outside the world of professional science.

Professional Programming Discussion Forums Collection: A collection of discussion forums where programmers discuss technical solutions.

Science in Pop Culture Online: This would focus on targeting points of contact between science and popular culture. For example, TV-show forums for science fiction shows, and video game forums for games like Spore that focus on science topics,.

Science on the Web from Your Institution: Collect across a broad range of different kind of content produced by an institution's faculty and staff. This might

include everything from science research center's web pages, to staff and faculty blogs, to department websites and any faculty projects on the web.

Teaching Science on the Web: A collection focused on sampling the diversity of different kinds of sites and content that is being created to teach science on the web.

U.S. Science Policy on the Web Collection: A collection focused on collecting various kinds of online community sites where members of the public and experts are presenting and responding to science policy in the United States.