

Building infrastructure in the digital age: case study of the Isis Bibliography of History of Science, 2002-2018

Stephen P. Weldon¹

Abstract

The IsisCB Explore went online in 2015 as a foundational digital resource for historians of science. Built on the History of Science Society's 100-year-old Isis Bibliography of the History of Science, this service is meant to lay the groundwork for a digital infrastructure to support historical work in the relatively new digital environment where so much modern scholarship now takes place. In order to create this resource, the director of the project, Stephen Weldon, has learned how to shape traditional historical methods, practices, and resources to fit the new digital paradigm. Computer and networking technologies have been built out of the needs and practices of technologists, natural scientists, and business innovators, all of whom employ it in very specific ways, quite different from the practices of humanistic scholarship, and history in particular. As a result, the digital environment is not especially friendly to historical work or products. It has consequently taken a great deal of effort to understand and refactor historical data so that they function well within a digital knowledge ecology, a "knowledge infrastructure," as Christine Borgman refers to it. This paper describes the difficulties (epistemological, cultural, and economic) that make the creation of tools like the IsisCB Explore service challenging for historians and suggests some ways forward.

Keywords

Digital Humanities; Bibliographies; Databases; Digitization; Open Access; History of Science Society; Isis Bibliography of the History of Science

¹ Editor of the Isis Bibliography of the History of Science and Associate Professor of History of Science, University of Oklahoma, USA. ✉ stephenpweldon@gmail.com. The present paper derives from a presentation at symposium Doing History of Science in a Digital, Global, Networked Community: Tools and Services Linking Scholars and Scholarship, 25th International Congress of History of Science and Technology, Rio de Janeiro 22-29 July 2017.

How scholars practice their craft in the twenty-first century is radically different from how they did so in the centuries before. Nowhere is this more apparent than in the tools that they use to do research, which have been radically transformed in the wake of the digital revolution. The statement on “Digital Scholarship” on the History of Science Society website says: “Digital tools and platforms are clearly shifting the ways in which we do history: from research to presentation and publication. We now have people curating digital exhibits with multimedia, building web interfaces for traditional scholarly resources, doing computational analytics that reveal patterns in the historical data, using popular social media platforms to foster community and scholarship around the globe, teaching gamified classes, and more.”² For those who build resources for research, this is a challenging time and it bears close scrutiny and study. In the pages that follow, I explain the impact the digital transformations had on my work as the official Bibliographer of the History of Science Society from 2002 until the present. During that time, I worked on two very different kinds of research tools for historians of science, one to prepare the annual Current Bibliography by collecting citations of all recent work in the discipline, the other to enable worldwide access to a new and sophisticated tool developed from the bibliographical references collected from 1913 onward. The two tools can be interestingly compared because both were based on a single dataset, the bibliographical records of secondary material in the discipline. Just how those two tools came about with very different purposes is the focus of this paper. Through this story of the building of two “bibliographical” systems, one can begin to understand more about the forces at work (technological, social, and conceptual) that have driven the rapid development of digital scholarship in the directions that it has taken. It is meant to be a first stab at a more comprehensive history of research tools and their place in the formation of disciplinary knowledge communities.

Part I: The first decade at the University of Oklahoma, 2002-2012

Organizing the editorial office

I am the fourth permanent bibliographer to assume editorship of the *Isis Bibliography of the History of Science*. The first was George Sarton, the Belgian scholar who established the bibliography in 1913 as part of his effort to establish a field of scholarship around history of science, and who experimented with its form over its first forty years. After Sarton’s retirement, the task was relegated to a committee for about a decade, at which point two bibliographers were hired for different tasks: Magda Whitrow, a bibliographer working at Imperial College, London, for nearly twenty years, spearheaded the publication of a cumulative edition of the previous half-century’s semi-annual lists. And John Neu, a University of Wisconsin—Madison librarian took over the editing of all new annual volumes. My work on the *Isis Bibliography* began in 2002, when I was hired to assume the new permanent position after the retirement of Neu. The History of Science Society had arranged to move the operation to the University of Oklahoma, where there existed one of the oldest history of science departments in the United States and a valuable and prized collection of rare books in the field. The conjunction of the History of Science Department

² See <http://digital.hssonline.org/digital-scholarship/>.

and the History of Science Collections along with a supportive University administration made the school an attractive place to relocate the bibliography project.

In 2002, the Bibliography was almost 90 years old and my task was quite clear: I was to ensure that every year I would publish about 4000 new classified citations and over 1000 book reviews in order to maintain the reference tool as it had developed up to then. In addition to publishing an annual printed volume, I was required to upload the citations to the History of Science, Technology, and Medicine database (HSTM), an online reference tool that incorporated the Isis CB and three other bibliographies in the discipline. At the time, HSTM was hosted by a non-profit consortium, the Research Libraries Group.³

I came to this job with three experiences that shaped the way I worked. First, I was trained in history of science at the University of Wisconsin. As practitioner in the discipline, I approached bibliography with less theoretical understanding of bibliographic management than either Neu or Whitrow, who were students in the area of library and information studies. My approach to the bibliography, as a result, was oriented toward the specific intellectual and social matters of the discipline of history of science. Neu had worked at Wisconsin since the 1960s, and was a pivotal member of the science community in Madison. As a librarian, not a faculty member, Neu was a bridge between graduate students and faculty in ways that were unique.⁴ I regarded Neu as a friend and important resource, though I never worked for him as an assistant on the bibliography. After getting my degree, I worked for Margaret Rossiter at Cornell University as managing editor for the journals *Isis* and *Osiris* and was appointed to manage the publication of the Society's 75th anniversary issue, *Catching Up with the Vision*. Throughout my time in these editorial capacities, the social ties in the discipline were most striking to me, as I came to see the Society as a scholarly community held together by intellectual interests. The fruition of this line of thinking I've written about earlier in 2013 where I explore the social meaning of the *Isis Bibliography* project as I have come to understand it.⁵

Second, my experience in both traditional and electronic publishing impressed upon me the significance of designing a rigorous publication workflow. I learned at *Isis* and *Osiris* how to get publications "out the door" and into people's hands. Also, at about the same time, I was hired to manage a cutting-edge electronic journal called *G-Cubed*, a scientific publication produced by the American Geophysical Union.⁶ At the time, *G-Cubed* was one of

³ Over the years, HSTM moved to different library database hosting services: in 2006, RLG was acquired by OCLC; and in 2012, HSTM was acquired by EBSCO, which is the current host.

⁴ He would always, for example, hold the table and buy the first pitcher on Friday afternoons, for the weekly happy hour gathering, a time when we all as a group all socialized together.

⁵ Stephen P. Weldon, "Bibliography Is Social: Organizing Knowledge in the Isis Bibliography from Sarton to the Early Twenty-First Century," *Isis* 104, no. 3 (September 1, 2013): 540–50, <https://doi.org/10.1086/673273>. Reading this draft, I realize that it tends to have a similar narrative tone to the articles in the special commemorative issue of *Isis* that I edited some years ago. My intention is to have this article serve, in part, the same purpose, since I would like to ensure that there is a historical record of the development of this project, recognizing that the social and personal interactions will have a bearing on the understanding of how this project developed for anyone approaching it from a social-historical point of view. See Margaret W. Rossiter, ed., *Catching up with the Vision: Essays on the Occasion of the 75th Anniversary of the Founding of the History of Science Society* ([Chicago]: University of Chicago Press, 1999).

⁶ See <https://agupubs.onlinelibrary.wiley.com/journal/15252027>.

very few entirely electronic journals, having no hardcopy distribution at all. It was instructive to see the ease with which the young internet was transforming the nature of information dissemination around the globe and exciting to play a role in that transformation.

Third, I came to this work with experience in coding and a historical understanding of the role of computers in society. During the height of the PC revolution of the late 70s and early 80s, I worked as a programmer for a public school district in Texas as part of a small research team.⁷ There I gained a deep understanding of algorithmic thinking by doing extensive coding for the school administration. That experience, along with my undergraduate studies in STS (Science, Technology, and Society) at both Cornell University and MIT, as well as a brief stint as a docent at the Computer Museum in Boston, provided me with an orientation toward the nature of modern computing—the logic and practice of coding—and its social functions—in particular, the intersection between computers and society. In a recent article, I have tried to spell out the ramifications of this intersection for historians. In essence, I argued that the ways in which historians understood and employed facts and evidence was quite different from the way that those same things (usually termed “data”) were used in statistical and computational analysis, and thus, for historians to begin to use the tools of computer science, both they and tool developers would need to rethink their assumptions and practices.⁸

Editing the *Isis Bibliography* requires special tools and a rigorous structure. My first several years as bibliographer were devoted to simply creating the tools and structure needed to collect, edit, and print more than 5,000 citations per year. This required a robust database, which I had to construct from scratch. (When I took over the project, it had been two years since Neu had published a volume, and he was now retired. Extenuating circumstances required that the Society run a second search for a bibliographer and hire an interim replacement.) As a result, the old database that Neu had used—a system that had been homegrown at Wisconsin in the 1970s—was long outdated and, indeed, was no longer accessible to me.⁹

The challenge when I was hired was to create and build an entire citation management system, while at the same time continuing to collect citations so as to keep up with the publication schedule. The printed volume remained central to my mission. Collecting the entries, proofreading them for errors, and classifying them, was challenging enough, but on top of that I had to find a way to organize and format a print-ready, digital file to send to the University of Chicago Press to be printed and mailed to members with the December issue of *Isis*.

⁷ See <http://www.ronniejhastings.com/2012/08/26/high-school-student-researchers/>.

⁸ Stephen P. Weldon, “Historians and Their Data,” in *Crossing Oceans: Exchange of Products, Instruments, Procedures and Ideas in the History of Chemistry and Related Sciences*, ed. Ana Maria Alfonso-Goldfarb and Márcia H. M. Ferraz (São Paulo, Brazil: Coleção CLE, 2016), 299–322.

⁹ Neu wrote a short history of the IsisCB, presented at the 2013 meeting of the History of Science Society, which I reference here. Neu, “The Isis Bibliography, 1966-2000,” presented at “Session T1: The Isis Bibliography and the Profession: Past, Present, and Future” on Thursday, November 28, 2013, at the History of Science Society Annual Meeting in Boston, Massachusetts.

The department appointed a new graduate student, Sylwester Ratowt, to be one of my two assistants when I arrived. Ratowt had obtained his B.S. degree in astrophysics and was now starting a dual degree in library science and history of science. With his keen, inventive mind and off-beat sense of humor, we hit it off well and made a good project team. He immediately understood the nature of the work at hand and played a central role in the development of the Bibliography both conceptually and technologically over the following decade and a half.

Developing a curation system for an old-style bibliography

The system we developed used the popular FileMaker Pro database application, which we supplemented with a set of scripts written in Perl. The resulting tool allowed us to create our own tiny distributed network in which local machines were synchronized with each other by sending data to a server directory that stored each machine's newly modified records. Like so many small development teams, we injected humor into the programs we wrote, and Ratowt was proficient in this. At one point, he had to design a way to transfer data back and forth in chunks between machines, and his solution resulted in giving these files animal names. Elephants, tigers, giraffes, and other exotic beasts were roaming the cyberspace from one laptop to the next, and for a year or two, every time we synchronized we would see a parade of animal names marching across the screen. Later, Ratowt created a more efficient system that got rid of the animal names but was overseen by the Egyptian god Horus. In the end, only a single "moose" remained, and it became a critical button used to proofread each citation for accuracy.

The resulting home-built tool was ideal. It was flexible and easily modifiable; it allowed the graduate assistants (GAs) to work anywhere, something that was still somewhat novel in 2002. You could be in the library, at home, or even travelling out of town. One year, we completed the bibliography during a semester when I was living in Japan, and another year, one of the GAs spent some months at the Chemical Heritage Foundation. The design we developed also made system resilient and protected it against data loss. There was never a danger that the master copy would be destroyed because the entire database was replicated in each of our machines as well as on the server. The system also contained precise tracking tools that allowed us to mark records as they progressed from one stage to the next: Entered, Proofed, "Moosed," Authorized, Printed, and Published online.

One especially time-consuming task was producing a print-ready text for publication of the annual volume to be packaged and mailed with the December issue of *Isis*. Ratowt spent many hours working on code that would create a LaTeX file (a standard format used especially by scientists and mathematicians for creating printed output) from which we could produce a PDF that the Press would accept. The challenge arose in part because the formatting of citations in the Chicago Manual of Style form I had adopted required manipulating data elements in ways that the data organization was not well suited for. Author and editor names, for example, were terribly frustrating because the last and first names were printed and arranged differently depending upon the context. We used only first initials in the book review section, for example, but full names elsewhere. The first

author in a multi-authored collection was printed last name first, but all others were printed in first-last order. Asian name order was, unlike Western name order, always last-first. Organizations that were authors as well as some medieval, Arabic, and ancient names broke the rules in other ways. In a nutshell, names extremely tricky to work with. In addition, we had to develop our own means to mark and print special diacritic marks, italics, and similar objects. And the ordering of citations and their spacing on the page presented other problems. In all, the creation of a print publication required attention to details that an ordinary database would not have needed.

The PDF we produced was the culmination of a year's work and was considered the authoritative version of the data. Proofreading this in page-proofs before it went to press allowed us to catch errors and helped me rethink the classification of some entries. This produced a product that had a unique connection to both the digital and print worlds. We spent a lot of time and energy worrying about typesetting, not data modeling, and sometimes the data was entered and even structured in fields so that it would *look* right when printed even though it made no sense in the database. This was even more so in Neu's data where the fields were strictly for typesetting purposes.¹⁰

The system grew and developed as Ratowt and I continued to tinker with it over the years. We streamlined the layouts, and the interface became a complex sophisticated tool, with hundreds of scripts activated by buttons and triggers of all sorts. The goal was always to make the work faster and more accurate. (See Figure 1 for an example of the interface as it looked in early 2015.) The system enabled me to manage about 220,000 citation records by the spring of 2015. Finally, we established a type of version control for all citations; every time a record was changed, the previous version was stored as a small separate file on the server and later archived.¹¹

After the day-to-day curation was under control, I turned my attention to the digital files that John Neu had left behind. These data were in a folder on the desktop computer I was given when I arrived in Oklahoma in 2002. Although the database itself was no longer accessible, the citation data that Neu had collected for 25 years, between 1974 and 1999, was readable in a standardized albeit anomalous format.

From early on, I accepted that one of my tasks as bibliographer was preservation, and the Neu data were in a precarious state. Although the Neu digital records had been uploaded to become part of the HSTM database, the host changed twice in only a few years, which merely reinforced my view that these data files were HSS property, and we should never rely on an unaffiliated data provider to preserve them. Preservation was important both from a historical-archivist perspective as well as from the Society's economic point of view. That data was valuable, and I knew that I would eventually need those data to fulfill my long-term goals, which required creating a single dataset containing all of the citations (going back all the way to 1913 if possible) in a single, well-described, standardized, digital format.

¹⁰ I want to thank Sylwester Ratowt for reminding me of the importance of producing the print volume and its unique role in this project.

¹¹ All of that data exist still in the archives, so one could theoretically piece together the entire set of transformations for all records over time.

This meant that I needed to convert the archived data into a format that could be ingested into the FileMaker system. So, over about five years, I directed Ratowt, along with several current and former graduate students, to work on the data. This expanded the database by about 120,000 records, and in the process, I rediscovered most of the book review citations that Neu had collected and printed in the annual editions. These had never made it into the HSTM database, but they were all in these files except for several years that were mysteriously missing and will have to be replaced at some point. In recent years, the University of Oklahoma Libraries has developed their digital infrastructure, and they have been able to ensure the preservation of this data in their SHAREOK repository.¹² All of IsisCB records, new and old, will be maintained in that repository under the current agreement between the University of Oklahoma and the History of Science Society.

Initial goals for the Isis Bibliography, 2002-2012

In terms of data curation, then, this first period of my tenure was devoted to the development of a curation and publication system for the *Isis Bibliography* that would replace the one that became defunct when Neu retired. My initial goals were modest insofar as I simply needed to build an efficient working tool that would accomplish the tasks of maintaining a resource that was vested in my care by the History of Science Society, a group that was devoting a significant portion of their own resources into its preservation. I brought to bear my historical sensibilities, my understanding of the value of this resource to the community, and my skills in publishing and working with data and computers.

If there was one other goal that I felt passionate about from the moment I began, it was to expand access to the resource. I had learned from my time under Rossiter in the *Isis* office that institutions can sometimes be quite blind to the struggles of their members, and Rossiter was exceptional in her desire to stick up for those whom she considered to be underdogs. When I started working as bibliographer, I wanted to find a way to utilize the burgeoning internet technology to foster greater access.¹³ As an insider with some power over how this tool could be employed, I felt I had the responsibility to make it as open as possible. I intuitively understood that there was no reason that access to this resource had to cost a lot, since it simply relied on an internet connection to an electronic file.

The problem that I faced in 2002 when I started was that the bibliography was not easily or inexpensively accessed outside of North America. The printed form went to libraries and scholars with annual subscriptions to the journal *Isis*, but overseas mailing costs made print subscription quite expensive, which disadvantaged scholars in less developed countries. The HSTM database was likewise a library subscription service especially costly for many libraries overseas. Although the History of Science Society had made the site free for members, other historians did not have this advantage. My desire to maintain control of the data (especially Neu's old data) for HSS was based in large part on

¹² See <https://shareok.org/>.

¹³ It was 2002 and Google, for example, was just four years old; Gmail wouldn't appear for another two years.

these idealistic economic motivations to spread the tools of scholarship more widely around the world.

The flexibility of the electronic resource was not lost on us. Ratowt at one point spearheaded the design of a system that would allow us to capture collections of records that could be organized and sent via email to subscribers. The idea was that we could encourage people to interact with the bibliography more directly by providing them with monthly updates of the newest citations in the bibliography—long before they would appear in print in the annual volume. In return, we hoped that people getting these subscription emails would be more inclined to tell us about new scholarship that we might have a hard time finding. Ratowt's plan, which we never had the time or manpower to enact, would have allowed anyone to subscribe to the service and give us a list of their interests so that the emails could be specialized to each person. The software would easily have allowed this kind of system, but it was going to require more programming and more manpower to operate than we had, so the system never materialized. Nonetheless, it was through brainstorming ideas like this with Ratowt that I gradually came to understand the projects as potentially interactive in a way it had never quite been.

Even though I didn't think that the bibliography needed to be very expensive, I did not initially consider the possibility of making a service that was entirely open access. That notion only came to me over the course of the 2013 Norman conference (which I will discuss later) when I finally realized that the open access model of information distribution made certain functionality possible that even minimal pay-for-use models would impede. The argument for maintaining some restrictions on the data was that free access would erode the Society's membership benefits and, thereby, its bottom line. Would some people give up their membership, if they knew that they could get the bibliography for free? Surprisingly to me, I found that when I proposed the complete open access model to the Society, many members supported it; indeed, though I was not party to the council discussions on the matter, I heard no objections. Indeed, even the University of Chicago Press, which managed our membership and had a lot of experience with journal subscriptions of other societies, found the prospect of an open access version of the bibliography to be unobjectionable. In the Society, Jane Maienschein, in particular, who was HSS president in 2008 and 2009, gave strong support to this notion, as did Bernie Lightman, the Society Editor who succeeded Margaret Rossiter. No doubt the Press's experience contributed to their thinking on this. It was at this same time that Lightman and the Press arranged to make all of the special Focus sections in *Isis* as open access—not just the *Isis Bibliography*. As it stands now, the PDFs of the annual print bibliographies are available to anyone to download for free. Moreover, with the opening of the IsisCB Explore service and the companion IsisCB Cumulative, open access extended both to a database search service and to all of the society's older data. The Explore database goes back to 1974 and the Cumulative HTML pages cover the period from 1913 to 1975. It is these two projects, Explore and Cumulative, that I will focus on in the rest of the article.

Part II: Rethinking the bibliography project: toward IsisCB 2.0

Understanding scholarship in the digital age

Out of the day-to-day manipulation of the bibliographic data, I gradually developed new sensibilities, and it became clear to me that there were new things the data (which we already collected) would allow us to do if it were somehow made more accessible. It was only after I had begun talking with younger historians, digital archivists, and colleagues doing computational history that I realized the great potential of the bibliographical database. To accomplish any of these things, the data would have to be restructured, and this would take much more effort than anything I'd done before.¹⁴

Although I had, from the start, understood the management of citations as primarily manipulation of data, it was only when I started to focus on the structure of the information itself that I began to reconceptualize the resource as more than an electronic version of a paper bibliography. The data that exist in each citation were comparatively rich in information about the field of history of science itself. After all, the list of citations contained all of the field's authors, all of the publishers, journals, and schools (via dissertation citations), and a classified and structured list of the field's subject matter. With the right tools, it would be possible to break apart those citations and recombine the data so that they would highlight the intertwined network of social, institutional, and intellectual relationships that forms the discipline. In that way it could create a new kind of research tool. This would require extensive rebuilding and manipulation of the data.

The most important element of the change involved establishing a rigorous means of authority control for all of the people, institutions, and concepts that were in any way related to the published scholarship. In other words, I had to find a way of dealing with authors, publishers, journals, and subjects in the same way that I treated the citations themselves, as entities in their own right. The bibliography in its classical form is simply an ordered list of citations. The new conception of the data was as a network of entities of

¹⁴ There is a growing list of works about the nature of doing digital work in the humanities and how one needs to start thinking about data in different ways. Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out Classification and Its Consequences*, Inside Technology; Variation: Inside Technology. (Cambridge, Mass.: MIT Press, 2000); Roy Rosenzweig, *Clio Wired: The Future of the Past in the Digital Age* (Columbia University Press, 2011); Ann Blair, *Too Much to Know: Managing Scholarly Information Before the Modern Age* (New Haven [Conn.]: Yale University Press, 2010); Christine L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (Cambridge, Mass.: MIT Press, 2007); Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities" 3, no. 4 (2009), <http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html%20/000077.html>; Gavan J McCarthy and Joanne Evans, "Principles for Archival Information Services in the Public Domain," *Archives and Manuscripts* 40, no. 1 (March 1, 2012): 54–67, <https://doi.org/10.1080/01576895.2012.670872>; Gavan McCarthy, Ailie Smith, and Michael Jones, "Looking Beyond the Archive: Utilizing Encoded Archival Context in a Broader Societal Context," *Journal of Archival Organization* 12, no. 1–2 (April 3, 2014): 143–64, <https://doi.org/10.1080/15332748.2015.1001206>; Daniel Pitti et al., "Social Networks and Archival Context: From Project to Cooperative Archival Program," *Journal of Archival Organization* 12, no. 1–2 (April 3, 2014): 77–97, <https://doi.org/10.1080/15332748.2015.999544>; Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscopic*, Reprint edition (London: Imperial College Press, 2015); Vladimir Geroimenko, "The Concept and Architecture of the Semantic Web," in *Visualizing the Semantic Web*, ed. Vladimir Geroimenko DSc MSc and Chaomei Chen BSc MSc (Springer London, 2006), 3–18, http://link.springer.com.ezproxy.lib.ou.edu/chapter/10.1007/1-84628-290-X_1.

different kinds all related to each other in different ways. All of these entities—people, places, concepts, and the scholarly works themselves—formed a web of relationships. Authors who wrote articles and books were just as much a part of the relationship network as the cited works that they produced. This meant that the bibliography was not just a bibliography any longer, it was a database that could provide information about all aspects of the discipline. You can use the database to learn more about authors, publishers, journals, and schools, just as easily as you can discover new bibliographic citations.

There is nothing new about this way of looking at bibliography. Experienced historians do this all the time when they look at lists of citations, comparing them with what they have come to know about the field, where people work, what they study, how to evaluate journals and publishers. Indeed, I know from talking with other historians that one of the primary uses of the annual printed bibliography is scanning it to get a mental map of the recent year's citations in their area of scholarship. By simply reading through the bibliography, they are mentally updating what they know of the people and institutions they are interested in. But these historians do all this internally and often unconsciously, and they rely on their training and their participation in the community over the course of their career to do so. A reconfigured bibliography could make some of these learned relationships explicit, so you didn't need to have years of experience in the field to understand and evaluate material in a more sophisticated way. Or at least, it would make many of these tasks easier for people new to the field. By looking at a bibliographical entry in a well curated network of authorities, one could see how the author of the work was integrated into the intellectual and social map of the discipline. One could explore the various subject areas related to the work. Given the right tools to browse and visualize the data, it would be possible to evaluate the people and institutions and subjects in a given area simply by looking at the many relationships.

The reconceptualized networked data came about through my work with colleagues in related projects that the bibliography brought me in touch with. The most important of these was the World History of Science Online. WHSO, as it came to be called, was an effort to establish a more open and accessible research tool available to the international community. I had come to be involved with this group early on in my tenure with the *Isis Bibliography* and the process of working on it forced me to confront several issues that were crucial to my understanding of the evolving information ecology of the 21st century. Without the conversations and contacts that I made working on this project, I have a hard time imagining how the IsisCB Explore would have emerged at all, let alone in any form similar to its current one.

The WHSO project had grown out of conversations among bibliographers who had been meeting regularly at different venues over the previous decade or so. In the original discussions, WHSO was to be an open access international version of HSTM, with the difference being that the focus would be on documentation of sources and archives, not on current publications. The major institutional form where this work took place was within the International Union for the History and Philosophy of Science / Division of History of Science.¹⁵ In particular, it was the Division's Commission on Bibliography and

¹⁵ It is now abbreviated IUHPST/DHST with the inclusion of Technology.

Documentation, composed of historians, archivists, and librarians, that oversaw work on the WHSO. The project was established in 2003, and the Secretary General of the Division at the time, Juan José Saldaña, secured a large grant from the International Council for Science to fund it.¹⁶

One of my closest associates in this group was Gavan McCarthy, director of the eScholarship Research Centre (ESRC) at the University of Melbourne. Over the years, after the initial funding had dried up and the project had languished for a few years, the two of us picked it up and grappled with the challenge of building a robust WHSO search service.¹⁷ By the early 2010s, we began working on ways to create WHSO as a directory of online resources in the field. It would be hosted on the data infrastructure system that McCarthy had developed at ESRC and rely on some graduate students at the University of Oklahoma to collect data using a variation of my Filemaker Pro system.¹⁸ McCarthy's deep insights into the nature of information in the digital age came from his long work with information management at ESRC and his interaction with other pioneers in the digital archival community. Much of my understanding of database design and the significance of the authority as a data form comes from discussions with him.¹⁹

The two Oklahoma graduate students whom I hired to help get this project started, Margaret Gaida and Amy Rodgers, also played a critical role in my thinking. As we developed our plan of action, our brainstorming sessions turned us to ideas about the very nature of scholarship in the digital age. The insights that Gaida and Rodgers brought to the discussion revolved around the dynamic nature of interaction that was taking place in digital communities across the globe. Scholarship was being transformed by social media and other similar interactive resources. They pressed me to think about reframing the bibliography for a new generation of scholars. As we talked during our periodic meetings on WHSO, a new more interactive paradigm involving the IsisCB became more clear in my mind.

Thus, WHSO became a critical turning point for three reasons. It introduced me to the international history of science community and the somewhat different framing of information and access to sources than what I was used to in my conversations with members of the primarily North American members of HSS. Second, it helped me confront the nature of database design in ways that were unfamiliar to me. Looking at McCarthy's work and understanding the networked infrastructure with its reliance on authority data

¹⁶ Stephen P. Weldon, "The World History of Science Online: A Project of the Commission on Bibliography and Documentation," *Acta Baltica Historiae et Philosophiae Scientiarum* 1, no. 2, no. 2 (2013): 97–101. See also the minutes to the Paris meeting of 2003, <http://www.dhst-whso.org/about/history-of-whso/paris-minutes/>.

¹⁷ McCarthy had worked with the Australian historian of science Rod Home, a counterpart in many respects to John Neu insofar as Home had edited a bibliography of history of science in Australia.

¹⁸ See <http://www.dhst-whso.org/2012/04/gathering-data-for-whso/>. The ESRC employs what they call the Online Heritage Resource Manager (OHRM), which is "a contextual information management system capable of integrating data from a wide range of sources" (<https://esrc.unimelb.edu.au/about/tools-and-technology/ohrm>). Primarily designed to host archival information records, the OHRM system is the host of most of their databases are instances of an OHRM that stores the data in flexible and extensible ways.

¹⁹ We met at the 2005 International Congress in Berlin and began working together from that time on. After the 2009 Congress in Budapest, we applied unsuccessfully for a couple of grants that would help solidify the project and its international collaborators.

shifted my perspective in critical ways. Finally, the project simply brought me into contact with new people. McCarthy, Gaida, and Rodger shown light on the dynamic ways in which scholarship was being done, and which I had simply missed in my narrower focus on collecting and classifying citations for a print-based bibliography.

By this time, I had a very different vision of what could be done with the bibliography and an awareness that there was some urgency about this. Things had not stood still during the previous ten years that I was working on the bibliography. Indeed, Ratowt had graduated by this time and was on the verge of stepping aside because he also recognized the somewhat clunky nature of the existing bibliography project. A survey of HSS members conducted by Gaida, Rodgers, and me further cemented the idea that it was time to move in new directions. Unless the project was refashioned in some form, the *Isis Bibliography* would gradually cease to be relevant in the new world of digital scholarship.

I approached the Alfred P. Sloan Foundation to see if they would fund a major project of this sort, and I received encouragement and advice from Joshua Greenberg, one of the Sloan grant officers responsible for digital project funding. Greenberg had received his doctoral degree in Science and Technology Studies from Cornell University, having completed a dissertation on the history of the videocassette recorder. His work at George Mason University's Center for History and New Media on such open source projects as Zotero and Omeka as well as his later position at the New York Public Library dealing with digital scholarship allowed him to quickly see how the *Isis Bibliography* data could play a role in a new digital infrastructure for the discipline.²⁰ The Sloan Foundation provided funding, first, for an initial development conference in 2013, and later, for actual development of the IsisCB Explore system between 2014 and 2016.

The development conference that took place in Norman in April of 2013 included scholars from the US and abroad involved in digital projects related to history of science.²¹ This included board members of the Commission on Bibliography and Documentation, members of the *Isis Bibliography* Advisory Board, many of whom had been with me from 2002, and members of another important group that I was just becoming familiar with, the Digital HPS Consortium. This Consortium was a growing community of scholars who had been meeting informally for over a decade to discuss their own digital work. Though based in the United States, it also included several European scholars working on online projects.²² One of my colleagues at Oklahoma, Kerry Magruder, who had just recently become curator of the History of Science Collections, was part of this group, and Magruder was especially

²⁰ See his biography at <https://sloan.org/about/staff/joshua-m-greenberg>.

²¹ Participants included the following: Jennifer Alexander (University of Minnesota); Colin Allen (Indiana Philosophy Ontology Project, Indiana University); Emanuela Appetiti (National Museum of Natural History, Smithsonian Institution); Marcia H. M. Ferraz (Center Simão Mathias of History of Science, Pontifical Catholic University); Daniel Goldstein (University Library, University of California, Davis); Manfred Laubichler (Arizona State University); Henry Lowood (Stanford University Libraries); Jane Maienschein (Arizona State University); Gavan McCarthy (eScholarship Research Centre, University of Melbourne); Birute Railiene (Wroblewski Library of the Lithuanian Academy of Sciences); Sylwester Ratowt (Independent Scholar); Sage Ross (Wikimedia Foundation); Sean Takats (Roy Rosenzweig Center for History and the New Media, George Mason University); Alex Wellerstein (Center for History of Physics, American Institute of Physics).

²² See <http://digitalhps.org/>.

helpful in ensuring the library's support of this Sloan conference on the rethinking of the IsisCB.

This gave rise to a second proposal to Sloan that launched the development of what came to be called the IsisCB Explore and the IsisCB Cumulative projects.²³ We were able in two years to build a complete replacement for the Filemaker Pro data management system that Ratowt and I had developed. The new service not only managed the data, but it did so through a server that supported both a curation and public interface. This meant that the line between the curation of records and their distribution was gradually breaking down. Records that are approved are immediately visible on the public site. People no longer needed to wait a year to find the newest citations. Ratowt's earlier ideas about bringing citations to the researchers had finally come to fruition, albeit in a somewhat different form.

In addition, my goal of ensuring that HSS really did own the data was fulfilled as well. Although it is not in a database form at present, all of the volumes of the Isis Cumulative Bibliographies going back to 1913 were photographed and digitized and partially marked up in TEI. All of those texts are now available as standard web pages at the *IsisCB Cumulative* website.²⁴

Building a social-semantic network out of citation data

The technical framework behind the new IsisCB database arose out of an intersection of several different institutional and personal connections, and I ended up with a coherent framework for dealing with data in an entirely new way from what I was thinking when I began as bibliographer. When I started, it was a technical problem that required building a system that would recreate something that already existed; after a decade, I embarked on a fundamental rethinking of what a bibliography was.

The authority tables were the key to the transformation of a bibliographic database into a relational network based on both social and semantic data in the discipline. By turning the focus of the project away from citations to the entities that make up the citations (e.g., the authorities), a range of new possibilities emerges. Authority control at this highly specific level makes it possible to understand the underlying structure of the scholarship through patterns of publication, relationships among authors and ideas, and institutional roles and behaviors. Authority control also makes it possible to remove the isolation of the Isis Bibliography by creating links to outside resources and enabling other resources to link to

²³ Though the project was primarily run by me at the University of Oklahoma, the planning and coordination took place most intensely between me, McCarthy, and Ratowt in the early part of the project. The governing board included the three of us as well as Gaida and a philosopher of science from Indiana, Colin Allen, who does extensive work on computational history, including topic modeling.

²⁴ The IsisCB Explore service was created by the consultant firm A Place Called Up, run by two graduates of the Arizona State University history and philosophy of science program, Julia Damerow and Erick Peirson. IsisCB Cumulative was created through the work of the University of Oklahoma Digitization Lab, which photographed about 5000 pages of the cumulative bibliographies, a double-key transcription service, Apex CoVantage, that converted the imaged text to extremely accurate text files, and extensive scripting by Conal Tuohy in Brisbane, Australia.

the Isis data. The enabling of this kind of interlinking among different databases across the web creates what is called “linked open data.” Databases that are closed and whose records are not linked this way are known as siloed and considered to be much more limited in their ultimate usefulness as the internet’s data infrastructure grows and becomes more integrated.

To implement this vision, much needed to be done. When I set up the Filemaker database in 2002, I had created a relatively flat file. There were some relational components in the database, but they were trivial and non-essential. They provided no inherent semantic or functional relationships. Citation records were mostly composed of text fields: there was a title field, an author field, a publisher field, and so forth. The author field, for example, simply contained one or more names separated by semicolons. (See Figure 1.) There was a separate field for editors, and another one for publishers. Names were entered as published and there was never any attempt to establish a definitive list of names in a separate authority file. Subjects and classifiers were slightly different insofar as I had a thesaurus that kept track of controlled vocabulary, but they, too, were added to the citation record in more or less the same way as author names, as terms separated by a delimiter in a single text field. The only consistent relational structure was the one I used to link citation records to each other: articles were linked to the journals in which they appeared, reviews to books that they reviewed, and chapters to books where they were published. The database, in other words, was designed primarily to ensure that citations would be carefully tracked and managed so that I could accomplish my main objective, to publish an annual print bibliography.

The new database that runs the Explore system is relational at its core. It breaks up the citations so that their components are all stored in separate tables. The two primary tables are citations and authorities, but each of those contains relatively little information—mostly a title or a name and possibly some descriptive information.

It is the supplementary tables that do all the work. They contain the critical data elements. The attribute table stores dates, place data, and similar kinds of discrete information. The external link table contains the URLs and similar information about outside resources like VIAF and OCLC. The relationship tables are the workhorses of the database. They create the crucial ties that define such things as which authorities are linked to which citations and in what way. Without the relationship records there would be no way of knowing which authority was the author of which citation. There are three of these relationship tables: the citation-to-citation table, the authority-to-citation table, and the authority-to-authority table. What this means is that the complete citation as an object does not exist as a single entity in the database like it did in the Filemaker system; instead, the citation record is essentially hollow, missing any indication of the author, subject, date, and publisher. The complete citation must be extracted from the various tables. No single table or record can, by itself, reconstruct the citation.

Book ID 1210257

Author: Nielsen, Kristian H.; Harbsmeier, Michael; Ries, Christopher J.

Title: Scientists and Scholars in the Field: Studies in the History of Fieldwork and Expeditions

Place of Publ.: Aarhus: Aarhus University Press Year 2012

Physical Det.: 476 pp.; ill.; maps; index; bibl. ISBN 9788771240146

Description: Collection of seventeen essays concerning the history of fieldwork and expeditions in scientific research. Contents:

Subjects: Field work [4669] // Scientific expeditions [4663] // Travel; exploration [454] // Collectors and collecting [245] // Observation [16320] // Scientists [22227] // Research methods [15215]

Category: History of science, general works

Abstr. / Descript.: {Metadata}{AbstractBegin} For centuries the practice of undertaking fieldwork and

Lang: English

Source: hardcovv:

Entered 2012.08.02 Jared N Moose Proofread 2012.08.02 Jared N SPW 2012.09.18 stephen

Fig. 1. Book record input layout from the IsisCB database interface as of early spring 2015 (created in FilemakerPro). Note that most of the fields are simply text fields. The list of editors is a name list separated by semicolons. The subjects, though controlled, are simply listed by name with their ID in brackets. In this book, the relational component is shown in the yellow field to the right, where the titles of linked chapters are listed. At the bottom of the record you see the tracking buttons, including the “Moose” button that allows the user to proofread the citation as it will appear in the printed bibliography.

Comparing Figure 2 with Figure 1, you can see how different the data appear in the curation interface in the two systems. Although much of the same data are shown in both, the data are displayed more discretely in the Explore database than in the Filemaker database. This is even more clear when you look at the tabs that show data from the relationship tables. Figure 3 shows the authority tab in the citation record. There you can see what authorities are connected to the citation and what their role is, author, editor, subject, publisher, and so forth. Figures 4a and 4b show an individual record from one of the relationship tables; in this case, you can see the authority-to-citation table for an editor of the book. Figure 4b shows the dropdown list of possible relationships. This structure makes it possible to define and track any type of authority relationship that is embodied in a bibliographical citation. It is extensible, so if new types of relationships arise for such things as digital records, for example, we can define them and add them.

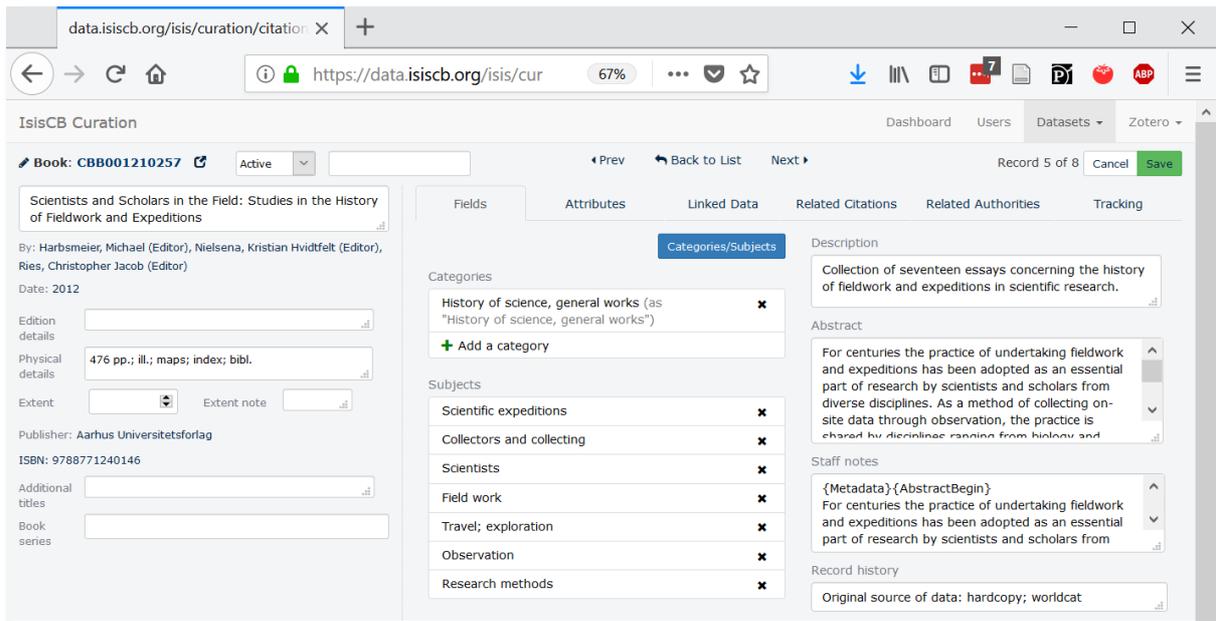


Fig. 2. Book record (same record as in Fig. 1) as seen in the curation interface from the IsisCB Explore system as of early 2018. Note the differences from the data in the Filemaker database. First, the entire system exists within a browser window, meaning that it is accessible without regard to specialized software. Second, there are very few simple text fields outside of the title, abstract, physical details, and some notes. Third, all authors, contributors, publisher, and subjects (see Fig. 3) are individually linked to this record. Their data is found in a separate authority table.

For recording citations, the new system as a whole is much more flexible. It can in principle support an unlimited number of different types of data. It can record information about entirely new types of records: digital material can now be easily included, as well as conference presentations. We can even add such things as notices of awards and honors, such as the winners of annual prizes for excellent scholarship that are offered by many academic societies. In all, the greater flexibility of this system anticipates a wide range of new uses as we enter into a period of rapid change in the nature of scholarly creative work.

The Explore system was meant to do more than simply show a bibliography, however. It was designed to provide a much richer picture of the discipline in general. The refactoring of the data enables users to shift their point of view and see the data from different angles, so to speak. Because we are now collecting data about the authority entities—that is, individual authors, publishers, and topics—we can use Explore to tell us interesting things about them. When you start looking at the network of relationships around any given authority record, you find a rich store of information that comes from the citation data. In Figures 5a and 5b the advantages of this system become apparent. Being able to see the range of an author's works on a single page is the kind of perspective that Explore can provide, which was inaccessible in the Filemaker system. It is here that we can begin to see how the Explore records can function in new ways to help users see the data in ways that experienced researchers see them. In Figure 5b in particular, the display shows a timeline, a visual analog to the work that an experienced historian might do in his or her head with the years of knowledge that they've internalized.

Book: CBB001210257 Active

Scientists and Scholars in the Field: Studies in the History of Fieldwork and Expeditions

By: Harbsmeier, Michael (Editor), Nielsena, Kristian Hvidtfelt (Editor), Ries, Christopher Jacob (Editor)

Date: 2012

Physical details: 476 pp.; ill.; maps; index; bibl.

Publisher: Aarhus Universitetsforlag

ISBN: 9788771240146

Fields Attributes Linked Data Related Citations **Related Authorities** Tracking

Filter by relation type

Editor	Has Personal Responsibility For	Harbsmeier, Michael (as "Harbsmeier, Michael")	✕
Publisher	Is Institutional Host Of	Aarhus Universitetsforlag	✕
Subject	Provides Subject Content About	Scientific expeditions	✕
Subject	Provides Subject Content About	Collectors and collecting	✕
Subject	Provides Subject Content About	Scientists	✕
Category	Provides Subject Content About	History of science, general works (as "History of science, general works")	✕
Subject	Provides Subject Content About	Field work	✕
Subject	Provides Subject Content About	Travel; exploration	✕
Subject	Provides Subject Content About	Observation	✕
Subject	Provides Subject Content About	Research methods	✕
Editor	Has Personal Responsibility For	Nielsena, Kristian Hvidtfelt (as "Nielsen, Kristian H.")	✕
Editor	Has Personal Responsibility For	Ries, Christopher Jacob (as "Ries, Christopher J.")	✕

+ Create new relation

Fig. 3. The authorities tab in the citation record from Fig. 2. The list shows all of the authority records that are linked to this citation record. The blue and green tags explain the how entity, named in bold, is related to the citation. The lighter grey text in parentheses allows us to display the citation as published, if an author, for instance, uses a different version of their name in different publications.

data.isisb.org/isis/curation/citation

https://data.isisb.org/isis/cur 67%

IsisCB Curation Dashboard Users Datasets Zotero

ACR000086036 | ACR record Active Back to Citation Cancel Save Delete

Name for display in citation: Harbsmeier, Michael

Type controlled: Editor

Data display order: 20.002

Confidence measure: Certain/very likely

Staff notes: Editor

Citation record: CBB001210257 Book
Scientists and Scholars in the Field: Studies in the History of Fieldwork and Expeditions
Harbsmeier, Michael (Editor), Nielsena, Kristian Hvidtfelt (Editor), Ries, Christopher Jacob (Editor)

Authority record: CBA000042661 Person
Harbsmeier, Michael

Harbsmeier, Michael

Fig. 4a. One of the authority-to-citation relationship records in the curation interface.

Name for display in citation: Harbsmeier, Michael

Type controlled: Editor (selected)

Data display order: Author

Confidence measure: Editor

Staff notes: Editor

Other roles in dropdown: Advisor, Contributor, Translator, Subject, Category, Publisher, School, Periodical, Committee Member

Fig. 4b. A close up of the dropdown list for type of relationship in the authority-to-citation record above. Most of these designations are created automatically during the ingesting of authority records, but they can be individually created or changed by hand, as you see here.

data.isiscb.org/isis/curation/author/...
 IsisCB Curation
 CBA000042661 | Harbsmeier, Michael
 Person: Harbsmeier, Michael
 Authority Details: Type: Person, Record status value: Active
 Fields | Attributes | Linked Data | Related Citations | Related Authorities | Tracking
 Filter by relation type: All
 Total number of ACRelations: 11
 List of citations with roles: Narrating the Arctic: A Cultural History of Nordic Scientific Practices (as "Harbsmeier, Michael"), Scientists and Scholars in the Field: Studies in the History of Fieldwork and Expeditions (as "Harbsmeier, Michael"), Studying Scientists and Scholars in the Field: An Introduction (as "Harbsmeier, Michael"), Fieldwork avant la lettre: Practicing Instructions in the Eighteenth Century (as "Harbsmeier, Michael"), Das Buch als magisches und als Repräsentationsobjekt (as "Harbsmeier, Michael"), Towards a prehistory of ethnography: Early modern German travel writing as traditions of knowledge (as "Harbsmeier, Michael"), Nabuh's Method, Early Scientific Expeditions and Local Encounters—New Perspectives on Carson Nabuh and "The Arabian Journey": Proceedings of a Symposium on the Occasion of the 350th Anniversary of the Royal Danish Expedition to Arctic-Eurasia.

Fig. 5a.: The curator's view of an authority record for an editor of the work in Fig. 2. The related citations tab allows one to see all of the roles that this person has had in the 11 different citations that he has been involved with.

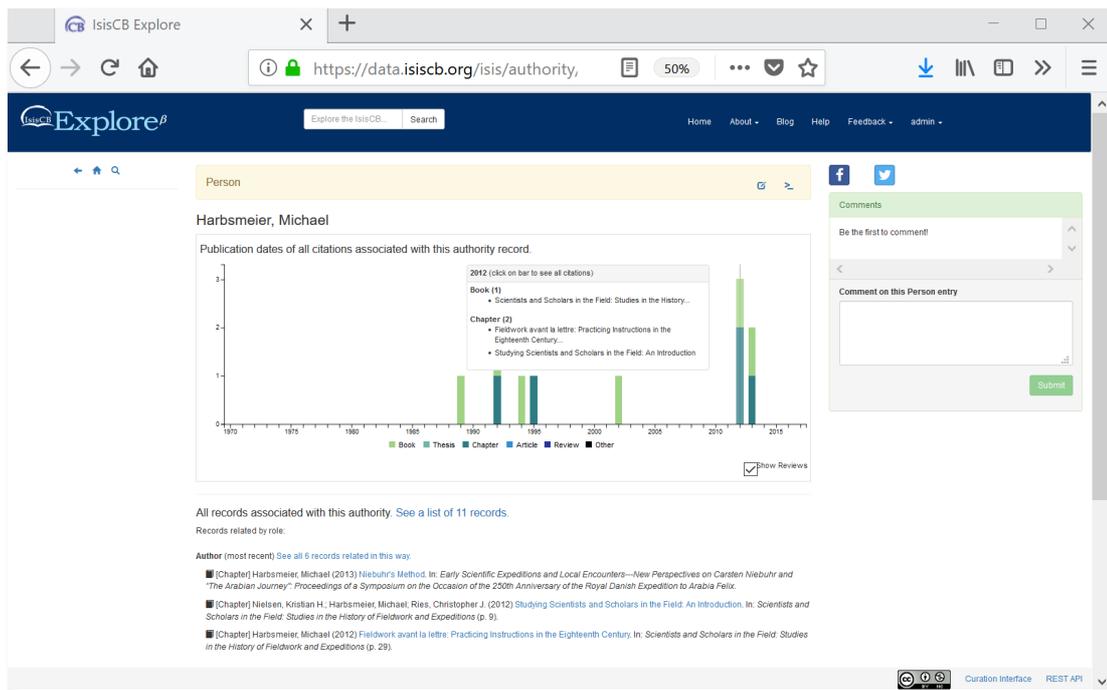


Fig. 5b. The public view of the same authority record as in Fig. 5a. This interface shows the advantage of the new system, which can provide views of information to users that provide new ways of understanding, such as this timeline that indicates the author's publication record, and you can easily get a sense of the scope and nature of this author. (The citation list in the box in the middle of the timeline is generated when the user hovers over one of the time bars.)

The transformation comes with some serious challenges, however. My task has now changed from simply managing citation records to managing authorities. This is more than simply an increased amount of data to manage; it requires new procedures and techniques that are not fully resolved even in the information science community. Authorities turn out to be notoriously tricky to manage. Tracking name similarities and differences requires time and information that is sometimes simply unavailable: consider how you might determine if J. Smith is the same as Jane Smith, or whether two instances of Jane Smiths refer to the same person, or whether Jan Smith is simply a misspelling of Jane Smith, or whether K. J. Smith-Wilson is a variant of Jane Smith. These problems are not especially well solved anywhere.

Digital scholarship has transformed the discipline of history of science in a number of ways, some of which, I have sought to illustrate in this personal history of the History of Science Society's Isis Bibliography project. The story of how people and institutions and technological drift instigated changes in the nature of the tool development reveal details about the nature of the transformation of the project that are sometimes obscure.

This period turned out to be extraordinarily challenging for me, and I suspect this is true for most people managing scholarly resources in the early twenty-first century. The fact that I ended up building two very different kinds of resources in less than fifteen years is telling. Whereast the preceding period was remarkably stable in terms of the tools used and the products produced, the rapid shifts in the technological, economic, and social infrastructures of scholarship in the digital age have been destabilizing. All of this led to my dramatic rethinking of the nature of bibliographic and historical data, and it happened in an

effort to ensure that the institution of the Isis Bibliography, which had been instrumental to the development of the discipline from its outset, would remain a viable and useful resource in the future. That was the intention, and whether this reconstructed resource will provide the payoff that I and others have envisioned remains to be seen. I am hopeful that it will.