H-Diplo FORUM 2021-2

H-Diplo Forum on Scholars and Digital Archives: Living the Dream?

6 October 2021 | https://hdiplo.org/to/Forum-2021-2
Commissioning Editor and Chair: Richard H. Immerman
Editor: Diane Labrosse
Production Editor: George Fujii

Contents

Introduction by Richard H. Immerman, Williams College 2
Essay by Matthew Connelly, Columbia University 6
Essay by Kaeten Mistry, University of East Anglia 13
Essay by Christopher J. Prom, University of Illinois at Urbana-Champaign Library 20
Essay by Joseph C. Wicentowski, the Office of the Historian 26

Introduction by Richard H. Immerman, Williams College

For the vast majority if not the entirety of readers of this forum, the appeal of digital archives is almost irresistible. Generations of historians and other researchers, particularly those with decades-long careers, have spent countless days and weeks applying for funds, planning schedules, booking flights and accommodations, and then travelling in order to spend days or months poring through folders and boxes. Make no mistake. These journeys are often rewarding and exciting, and I've always found the sight of a cartload of boxes to be energizing. I even enjoy their smell. Yet these treks are also very expensive, time consuming, and difficult to balance with teaching schedules, family vacations, and other commitments and opportunities that are integral to our lives. Then came the COVID-19 pandemic and the compulsory closure of archival facilities throughout the United States and the world. As I am writing the introduction to this forum, some but not all are reopening, in many cases slowly and encumbered with many restrictions. At many, access by researchers is severely restricted. Many historians are all but paralyzed. Graduate students cannot complete theses and dissertations; junior faculty cannot finish books even as the tenure-clock keeps running; grant recipients have time and money but no place to go. The scholarly enterprise is crippled.

The intrinsic appeal of digital archives is that they can provide a solution to such challenges. Rather than having to secure the time and money necessary to travel, or waiting until a closed facility opens and even extends its hours, we can do our research from home. I vividly recall a conversation I had with my good friend and wonderful historian, the late Nancy Tucker, back in the 1980s. Nancy was focusing on the U.S. presidential library system, a small piece in a much larger archival mosaic. She was bemoaning the logistical nightmare created by the need to travel from Independence, Missouri to Abilene, Kansas, to Waltham (soon Boston), Massachusetts to Austin, Texas to conduct research for her current project (this in addition of course to Washington, D.C., and other domestic and international sites). She suggested that the Society for Historians of US Foreign Relations and other historical organizations collectively propose a consolidation of the presidential library system. Nancy knew her proposal was a non-starter. Nevertheless, it reflected a widespread sentiment. There must be a better way for underpaid, resource-starved, and time-constrained historians to access their essential documents; they should not have to book flights, rent cars, and/or take buses in order to get to and shuttle between West Branch Iowa, Hyde Park, New York, and Independence, Missouri.

One can only imagine how Nancy would have reacted when she learned that President Barack Obama decided that he would not deposit his "papers" in a library, in Chicago or anywhere. They would all be digital, held in a cloud, and managed by the National Archives and Records Administration out of Washington. And the evidence points toward Obama's decision becoming a model for all successive presidents (as with everything else, Donald Trump is likely to remain a flame-throwing wild card). The archival component of any future presidential library will be digital.

Much the same may hold true for America's National Archive. A June 28, 2019, memorandum jointly issued by Russell Vought, at the time the Acting Director of the Office of Management and Budget, and Archivist of the United States David Ferriero directs agencies to digitize all their remaining paper records by December 2022.[1] The National Archives and Records Administration (NARA) will no longer accept paper records after that date, at which time it will manage all its permanent records electronically. Many agencies will receive waivers. Nevertheless, albeit perhaps delayed due to the pandemic, all are busily turning their records into digital archives.

There are drawbacks to these projects to facilitate access (and conserve space) that go beyond the time and cost of digitizing paper records and making discoverable those and the growing volume of records that are 'born digital' (a key concern of

[1] Russell T. Vought and David Ferriero, Memorandum for Heads of Executive Departments and Agencies (M-19-21), "Transition to Electronic Records," June 19, 2021, https://www.archives.gov/files/records-mgmt/policy/m-19-21-transition-to-federal-records.pdf.

these essays).[2] Traveling hither and yon domestically and internationally is expensive and often difficult. Nevertheless, where else could one find other researchers, at times an entire community of researchers, with similar interests and questions? Informal conversations in cafeterias and boarding houses frequently proved invaluable. And then there are the archivists, many of whom have unparalleled knowledge of and insight into the collections, both those that the researcher has identified as crucial to the project at hand and related ones that otherwise would remain unfamiliar to him or her. In the digital world, researchers can both access documents and outline their interests and communicate their questions to archivists. But that's not the same as a face-to-face conversation, nor can an email or a telephone call replace the discoveries found by accident when combing through folders housed in one large archival box of documents. Further, the jury remains out as to whether an archivist at a digital archive will over time develop the same expertise in the holdings as one at a brick-and mortar archive. It's too early to tell.

There's another unknown: without a "natural" home for their papers, where will private individuals "deposit" their papers in the digital age. We all know of discoveries of valuable archives in basements and archives. Will those historical actors whose papers do not fall under federal guidelines keep their papers on private servers? Delete some or maybe even all of them? How will historians know if they are retained, and if so, whether they have been curated? What incentive will public actors have to leave their papers to a library or a historical society, let alone a federal depository?

The following essays do not and cannot address each of these questions. But they do identify, explain, and evaluate the many opportunities and challenges that digital archives present. And they do so from diverse perspectives and expertise. Matthew Connelly is a pioneer in the digital archive universe. His introduction to archival holdings began while he was a graduate student and evolved from his profound interest in information technology. Rather than pursue an archival track, he experimented with and tested these developing technologies—hardware, software, and a variety of services and databanks in the course of his study and research. In doing so, he created his own "personal archive," while recognizing that this collection at best resembled a digital library.

The more expert he became, the more sensitive Connelly became to the inherent problems, most notably those of locating or discovering documents within the collections—finding needles in the haystack. Key-word searches yielded not only limited results but also often inaccurate and misleading ones. He explains in his essay how this appreciation of this phenomenon led him to examine, using the State Department Central Files (RG 59) as his source base, how a voluminous digital archive might actually promote secrecy as opposed to discoverability and transparency. The explosion of born-digital documents has produced a bonanza of riches for historians but also is marked by serious and in some cases intractable impediments to harvesting them. Connelly does not minimize the potential benefits to historians of the riches, but he underscores the impediments. He also proposes ways, using himself as an example, that historians can mitigate these impediments, by developing more sophisticated understandings of if not extensive expertise in data science and by collaborating with data scientists and archivists. Most important, historians must increase their awareness that the digital turn holds as much potential to turn into our nightmare as it does our dream.

Kaeten Mistry's essay provides us with a glimpse into that nightmare. Mistry is among the few historians whose scholarship has required deep immersion in the archives of the Central Intelligence Agency. That archive is virtually entirely digital. Indeed, it may well be the largest digital archive in the world (some paper CIA records are on deposit at NARA II in College Park, MD, but most of these remain classified or otherwise off limits to non-agency personnel). For that reason, digitalization should have been a boon for a historian like Mistry, who is based in England. He should in theory have gained much more ready access to it. He explains that initially the archive was in fact accessible solely by consulting the CIA Research Search Tool (CREST) that could be found only on set number of computer terminals housed at National Archives II in College Park, MD.

[2] Why Do Historians Still Have To Go To Archives?" *Contingent Magazine*, March 25, 2019, https://contingentmagazine.org/2019/03/25/mailbag-march-25-2019/.

That problem (not just the need to travel to College Park but also the limited number of terminals) suggests to Mistry a theme that runs through his essay: that the CIA created a digital archive, and only a digital archive, to deter if not to prevent scholars from discovering its contents. That situation changed less than half-dozen years ago, when the CIA uploaded all its declassified records to an Electronic Reading Room that its website hosts and made CREST available from any home computer. Yet because the archive manifests so many of the problems pinpointed by our essayists, it remains, to use Mistry's words, "while not entirely useless… not useful in any meaningful way." Probably the most glaring problem is the search engine, which is based on key words. But a key-word search will invariably turn up hundreds if not thousands of documents that lack context, chronological order, or even logic. The essential metadata is absent. The researcher can never evaluate how complete or comprehensive the search's yield is on any topic. The multiple redactions in the documents exacerbate the problems, but they are not the fundamental cause. Mistry judges the inutility of the CIA's digital archive purposeful, arguing that it is rooted in the agency's commitment to secrecy. While the Agency proclaimed the creating of its digital archive as progress toward CIA transparency, the practical consequences for researchers have been the opposite.

Christopher Prom's expertise and experiences contrast sharply with Mistry's. While still a Ph.D. candidate, he began working as an assistant archivist. That turned into his career, and his career trajectory paralleled that of the digital archives. Indeed, among a handful of scholars steeped in the craft of both history and archival management (for this reason the American Historical Association appointed him to its inaugural NARA Review Committee), Prom, as he writes, became a charter member of "group of like-minded professionals who are working in the corners of our digital economy to address an issue that most people assume has been solved, but hasn't: how to preserve a digital record of society." The goal is to build "discoverable, usable, and sustainable digital archives."

Prom's essay identifies what achieving that goal requires, the progress being made, and the obstacles and challenges archivists confront. A digital archive, Prom's contribution reveals, is far more than a collection of digitized documents, whether they are scans or were born digital. Unless expertly created and carefully managed, the digital archive runs the risk of obscuring contextual information (metadata) that is essential to understanding a record and identifying its relationship to other records. It can also conceal gaps in the record. In other word, it can turn into a CIA-like archive.

Joseph Wicentowski's essay demonstrates unambiguously how different the outcome can be when the objective of an archivist, or an archive's creator, is to serve the researcher as fully as possible. Connelly and Mistry refer positively to the State Department's *Foreign Relations of the United States* series in their contributions, as would almost any reader of this forum. Wicentowski, who works in the State Department's Office of the Historian (OH) as the digital history advisor, is most responsible for the series' digitization. Because of his expertise and attention to detail, in fact, and the dedication of his OH colleagues, the digitized *FRUS* volumes that are now available on the office's website (history.state.gov) not only meet Prom's definition of a digital archive, in contrast to a collection of digitized documents, but they contrast sharply with that of the CIA.

With commendable specificity Wicentowski explains not only how to use the *FRUS* archive, but also the methods and technology that went into making it such a valuable resource for all historians (not just those who concentrate on U.S. foreign relations). His essay illustrates the vast potential of digital archives, but warns that reaching that potential requires vast amounts of time, effort, and know-how. Many H-Diplo readers and subscribers have used the digital editions of *FRUS*. An added value of Wicentowski's essay is that it provides a manual on how best to exploit the full spectrum of the archive's capabilities.

Because plain text format will not support images and graphs, with the exception of this introduction, H-Diplo opted to publish this forum exclusively as a PDF, accessible via https://issforum.org/to/Forum-2021-2. You will learn a great deal by doing so.

Participants:

A former president of the Society for Historians of American Foreign Relations, **Richard H. Immerman** is ending more than a decade-long tenure as chair of SHAFR's Historical Documentation committee. He is also ending an equally long tenure as chair of the State Department's Advisory Committee on Historical Diplomatic Documentation (the HAC). He continues as chair of the American Historical Association's recently established NARA Review Committee. Retired from Temple University, Immerman will serve as Williams College's Stanley Kaplan Distinguished Visiting Professor in American Foreign Policy in 2021.

Matthew Connelly is a professor of international and global history at Columbia. He is co-director of the Institute for Social and Economic Research and Policy, and principal investigator of History Lab, an NSF and NEH-funded project to apply data science to the problem of preserving the public record and accelerating its release. His publications include *A Diplomatic Revolution: Algeria's Fight for Independence and the Origins of the Post-Cold War Era* (Oxford University Press, 2002), and *Fatal Misconception: The Struggle to Control World Population* (Harvard University Press, 2008). His current book project, to be published by Random House, is titled "The Declassification Engine." Matt has written research articles in *Nature Human Behaviour,* the *Annals of Applied Statistics,* Com*parative Studies in Society and History, The International Journal of Middle East Studies, The American Historical Review, The Review française d'histoire d'Outre-mer,* and *Past & Present.* He has also provided commentary on international affairs for *The Atlantic Monthly, The New York Times, The Washington Post,* and *Le Monde,* and has hosted documentaries for BBC Radio.

Kaeten Mistry is Associate Professor of American History at the University of East Anglia. Among his publications are *The United States, Italy and the Origins of Cold War: Waging Political Warfare* (Cambridge: Cambridge University Press, 2014), with Hannah Gurman, *Whistleblowing Nation: The History of National Security Disclosures and the Cult of State Secrecy* (New York: Columbia University Press, 2020), and articles in *Diplomatic History*, *Journal of American History*, and the *Washington Post*. His current project explores the culture of modern secrecy.

Christopher J. Prom is Associate Dean for Digital Strategies in the Library at the University of Illinois at Urbana-Champaign and holds a Ph.D. in History from the University of Illinois. He is a Fellow of the Society of American Archivists and previously served as its Publications Editor and Chair of the Publications Board. He is currently directing two grant projects: Email Archives: Building Capacity and Community and Email Archiving in PDF: From Initial Specification to Community of Practice.

Joseph C. Wicentowski is the Digital History Advisor at the Office of the Historian, U.S. Department of State. He completed his Ph.D. in History at Harvard University in 2007 and joined the Office of the Historian the same year. He co-authored *XQuery for Humanists* (Texas A&M University Press, 2020).

Essay by Matthew Connelly, Columbia University

What I Learned from Thirty Years at the Bleeding Edge of Historical Research

I have a love-hate relationship with digital archives, one that began in graduate school three decades ago. I love how the information revolution has given us tremendous new opportunities to make discoveries. But I hate the fact that most historians still think of the digital turn as just providing an easier way to ferret out archival finds, and have hardly begun to grapple with the downside risks. When I reflect on the technological changes we have already seen just in the last thirty years, I cannot help but think about what all this means for students starting out today, and what kind of work they will be doing thirty years from now. Will it be the kind of bright future one can easily imagine from reading Chris Prom and Joe Wicentowski, in which dedicated professionals steadily surmount technological challenges to make historical records ever more accessible? Or will it be a plunge into a "digital dark age," with more and more of these records all but lost in "data dumps," like the CIA collection well described by Kaeten Mistry?

Whether you love it or hate it, changes in information technology are already transforming the nature of historical research. This is especially true for those who choose to study any history since the 1970s, when electronic records started to replace paper files. But the past may still be our best guide to the future, and everyone will have to make their own choices about how to respond to these trends. So I will offer some personal reflections on how I have navigated these technological transformations, and what kinds of choices I think all of us will face going forward.

Starting with my first year at Yale, in 1991, I was an early adopter of digital technology. I had learned to use database software in my first job after college, and used the last dollars I saved to buy a Macintosh desktop. It had 1 MB of memory, and a 9-inch display screen. Unprompted, I would preach about the many advantages of being able to sift, sort, and Boolean search my notes, and looked on with enormous condescension as my classmates compiled their research in legal pads, or created Microsoft Word documents that were hundreds of pages long. How could they possibly find anything when it came time to write their dissertations?

Then one day my hard drive failed, and my backups turned out to be corrupted. If the redoubtable technicians at Tekserve hadn't recovered my data, I was fully prepared to abandon my dissertation and find a new career. It was an early lesson in how easy it is to lose information stored electronically without even realizing it is gone.

I also used bibliographic software, and felt triumphant the day I finally figured out how to download entries directly from Sterling Library's online catalog into my own Endnote database. Professor Paul Buskovich was passing by, and I explained this breakthrough. He asked me why I trusted the bibliographic entries were accurate without actually looking at the physical book or journal article. I took him for a Luddite, but it was a very good question, one that we ignore at our peril: electronic records do not come to us *deus ex machina*, and the technology does not make human errors go away -- it can just make them harder to notice.

One day, my dissertation advisor, Gaddis Smith, told me that he had met with people to discuss the possibility of digitizing the *Foreign Relations of the United* States – an early stage in the long process Wicentowski describes. It seemed like an odd choice, since *FRUS* volumes were already easy to access at any federal depository library and are fully indexed. At the time it did not occur to me that there could be more to digitizing archives than just making primary sources more accessible through the World Wide Web. It took years before I began to see how, for instance, digitization would make it possible to analyze how many records were being released each year, what periods and topics were emphasized or ignored, and even look at the people and places that were disproportionately likely to be redacted from the official record.

To me -- and I think this is still true for most historians -- digital archives seemed like a big deal mainly because we could more easily access and search them. The first one I ever encountered was Lexis-Nexis, which included contemporary newspaper, magazine, and wire service stories. I well remember how I went to Sterling and sat down in front of the

designated terminal, and my first keyword search blew my mind. I had entered the name of my most famous professor, Paul Kennedy, and it returned dozens of articles about his globe-trotting and media battles. I wasted hours doing that, when I might have been working on my dissertation or reading an actual book. Technology might make it possible to do path-breaking work, but I realized early on that it can also be completely distracting.

When I finally finished my dissertation, and revised it for my first book, I had thousands of photocopied pages of documents. When, after I started my history of the population control movement, it became possible to take digital images of documents, I collected thousands of them from dozens of different archives. I could now link to the image of the document from my database, so that all my notes and all my documents could be kept in one place. This too seemed like a real advance, and I well remember debating archivists about why they would not allow researchers to bring cameras into reading rooms. When I first heard how optical character recognition (OCR) might one day make it possible to search for words and phrases in the thousands of fuzzy jpegs I had started to collect, I thought this would be a quantum leap: Soon, I thought, each one of us would have our own fully-searchable digital archive we could haul around inside our laptops.

But of course, this was not really an 'archive.' As Chris Prom points out, archivists know that an archive is more than just a collection of documents, where everyone gets to be their own collector. A real archive results from the preservation of the documentary record of the originating institution, with the records arranged in the same way they were originally created and stored. They know that historical understanding often depends on being able to explore the context and the connections between these different records. The kind of 'archives' that I and others were creating, with cherry-picked documents, were very nearly the opposite: we were extracting individual documents, and sometimes just pages from these documents, and then rearranging them according to our own research agendas. The ability to search the contents to find 'keywords' would only compound the problem: Now we were just cherry-picking words and phrases.

At least when we created these 'archives' for ourselves, we had some sense of the original order and arrangement, and what individual pages represented in terms of the larger whole. Not so with the kinds of digital archives that commercial publishers began to produce, whether the Digital National Security Archive, or the Declassified Documents Reference System. How many of us have searched these and other collections with keywords, and found a juicy quote or two, without ever asking how they were put together, what has been left out, and how the search engine actually works (and how it might not work)? An even more basic problem is that many historians are left out: these and other collections are only accessible to researchers at institutions that can afford to pay for them.[3]

The vast majority of digitized document collections are made searchable through OCR, and the quality of the resulting text is highly variable. This is common knowledge among digital historians, since we confront massive quantities of garbled text when we try to do a textual analysis. But I still find many other historians who are blissfully unaware. They plug in their keywords, and come to all kinds of conclusions -- sometimes based on what they *do not find* -- not realizing that the words they are looking for might be illegible. The bigger problem is that they do not necessarily know which words are 'key' -- in controlled experiments, people who rely on keyword searching miss four out of five relevant records.[4] And whereas in the past they might have read or at least scanned thousands of words while doing traditional archival research, poring through files and folders while picking out which pages to capture and take home, many of us now take as many document photos as we can, reading as little as we can as quickly as we can.

Some of these problems can be addressed with better training. At the least, every introductory methods course should include readings on how every historian is already a digital historian, whether they realize it or not, and how to avoid the

[3] Heidi J.S. Tworek, "Digital History and Global Publics," in *Global Politics: Their Power and Their Limits, 1870-1990,* ed. Valeska Huber and Jürgen Osterhammel (New York: Oxford University Press, 2020), 332-334.

[4] For the classic study, see David C. Blair and Melvin E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM* 28:3 (1985): 289-299.

most elementary mistakes.[5] In the meantime, we will likely see a lot of public shaming, which will make it risky for historians to use digital tools without quite understanding them. Other problems might diminish with improved technology, and more transparency about the tools we are already using. OCR is already much better than it was ten years ago. But I fear Wicentowski and the Office of the Historian may be exceptional in how carefully they have stewarded *FRUS*. Commercial publishers have little incentive to redigitize and reprocess existing collections, and rarely explain their methodological choices.

But what happens when the archives -- the original documentary records and the systems used to store them -- are 'born digital'? This presents a whole host of novel problems. I only started to become aware of how both archiving and historical research would change with electronic record systems when I learned about the fate of the State Department's Central Foreign Policy Files. I was starting to do research on the history of secrecy, and was struck by how many of the records since 1973 have vanished. To be sure, archivists and records managers have long destroyed large parts of the original documentary record -- it was only at this point that I learned how large: an estimated 97-98%.[6] But what was new was the sheer volume of this collection, and the fact that the records were no longer arranged in files and folders. Archivists therefore used a crude form of sampling to decide what was worth preserving and what could be deleted. They ended up destroying all the records related to scientific exchanges, cultural diplomacy, international sport, and much else besides. Moreover, the content of many thousands of records that *were* deemed historically significant was lost because, at some point along the way, the data had become corrupted. In place of the original messages of State Department cables, there were error messages. When what remained was made available through the National Archives website, there was no finding aid, only an FAQ.

From conversations I've had over the years with archivists, I've come to understand that more and more 'born-digital' records will be made available much like these post-73 Central Foreign Policy Files, which make old archival principles like arrangement and description seem irrelevant. For the foreseeable future, they will be 'data dumps,' with no means of access other than filtered searching. Certainly, this is how the CIA prefers to make their records available. As Kaeten Mistry shows us, there are some eleven million pages, with no finding aid, or even an FAQ.[7] And the collections to come are even bigger. The Coalition Provisional Authority classified email network reportedly included nearly one million files, or 222 gigabytes, and the CENTCOM email consists of more than four million files, or about 1.8 terabytes. The State Department produces as estimated two billion email each year.

Back in 2003, Roy Rosenzweig wrote an article about what this might mean for future researchers with the title "Scarcity or Abundance?"[8] I think we now have an answer to that question: It's both. We will have an abundance of records to explore contemporary history, far more records than historians ever had to cope with previously, if only because we no longer have the kind of file structures and finding aids to help us isolate the subset of records that might actually be relevant to our own research interests. Instead, we will confront a huge volume of *potentially* relevant records, and will have to come up with new ways to sift and sort through it all.

---

[5] See, for example, Tim Hitchcock, "Confronting the Digital. Or How Academic History Writing Lost the Plot," *Cultural and Social History* 10:1 (2013): 9-23, and Lara Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *The American Historical Review* 121:2 (2016): 377–402.

[6] Wendy Ginsberg, "Common Questions About Federal Records and Related Agency Requirements," February 2015, Congressional Research Service, www.crs.gov.

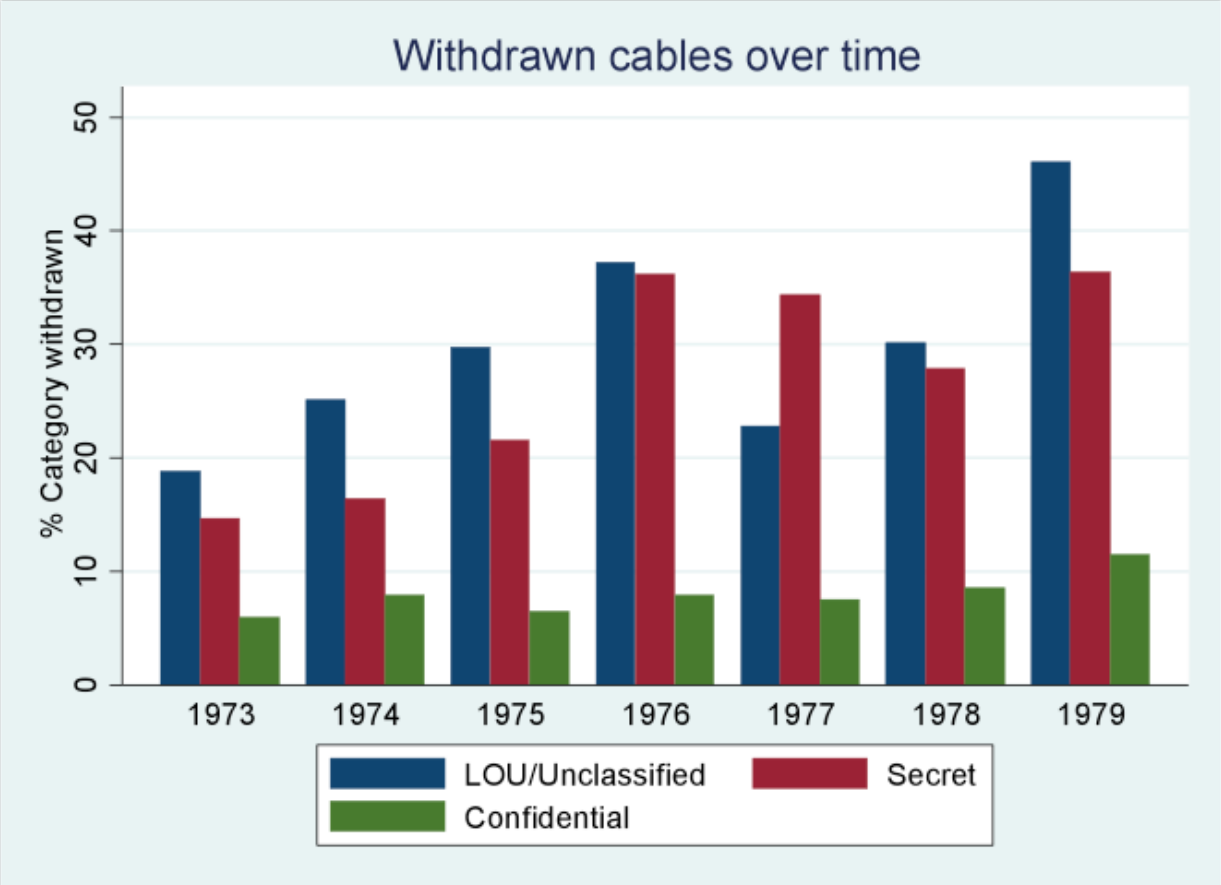
[7] See the contribution by Kaeten Mistry to this forum.

[8] Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *The American Historical Review* 108:3 (June 2003): 735-762.

But at the same time, what we are looking for could be scarce or non-existent, because we will likely see relatively less and less of the original documentary record. Parts of it will have been lost along the way in migrations between different hardware platforms and software packages. And the labor-intensive system for reviewing records -- which is still largely based on page-by-page inspection -- cannot possibly cope with the quantitative and qualitative transformation of the task. As difficult as it was looking through paper files for sensitive information, now reviewers have to do that one electronic record at a time, staring for hours into their screens, often with no sense of the context and connections between records. Moreover, they have to observe increasingly stringent requirements not just for withholding national security information, but also for protecting personal information.

The results can already be seen with the same Central Foreign Policy Files: Relatively more and more records are being withheld rather than released. Virtually no top-secret cables have been released at all. But this is true even of records that were never classified to begin with, likely because of the presence of personal information. And whereas for much of the last decade another year's worth of records were being released annually, there has been no new release for almost five years.

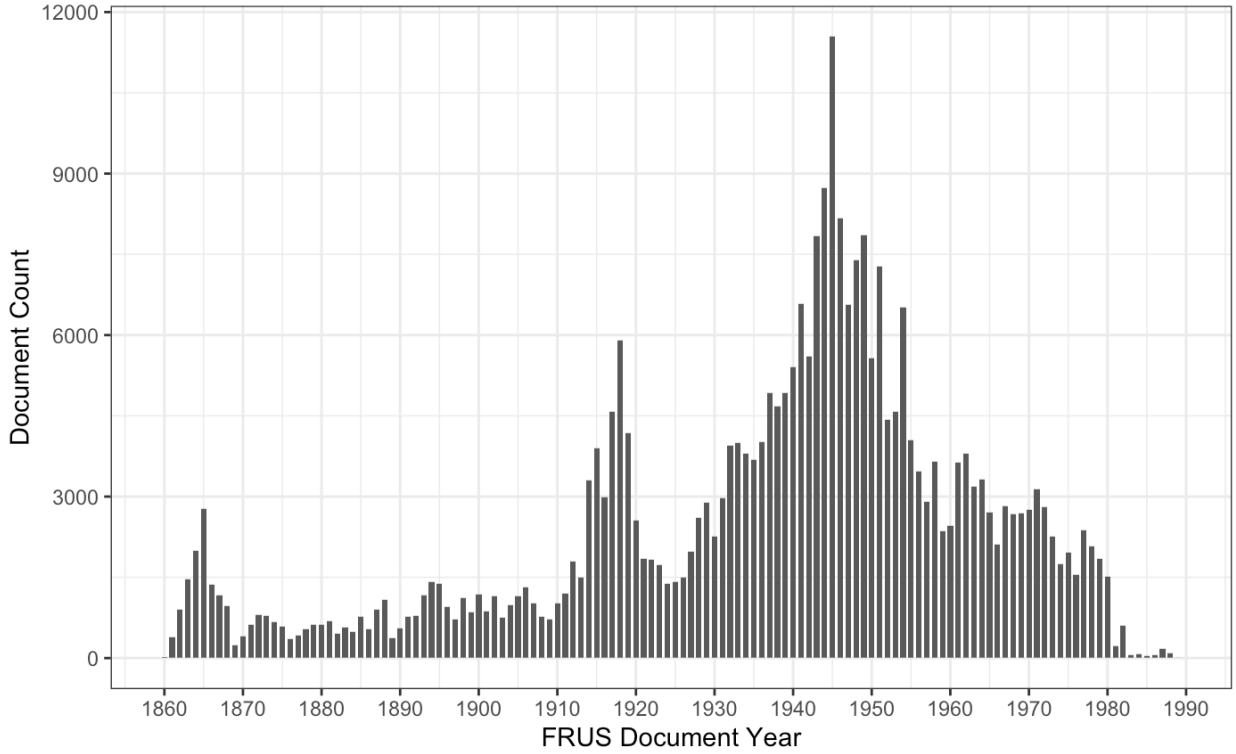
FIGURE 1: THE NATIONAL ARCHIVES HAS REVIEWED SOME 3.2 MILLION RECORDS FROM THE STATE DEPARTMENT CENTRAL FOREIGN POLICY FILES, MAINLY DIPLOMATIC CABLES. BUT WITH EACH ANNUAL INCREMENT A LARGER AND LARGER PERCENTAGE OF THESE RECORDS IS BEING WITHDRAWN RATHER THAN RELEASED.



As for the *Foreign Relations of the United States,* despite the heroic efforts of State Department historians, it is taking steadily longer to see fewer and fewer records. There are far more records available to study the 1950s compared to the 1970s, even

though the number of countries, the size of the national security apparatus, and the different kinds of diplomacy all increased dramatically in the intervening years.

FIGURE 2: THE ANNUAL VOLUME OF RECORDS RELEASED THROUGH THE FOREIGN RELATIONS OF THE UNITED STATES IS HIGHLY VARIABLE, BUT THE OVERALL TREND IS CLEAR: RESEARCHERS WHO WORK ON ANY PERIOD SINCE THE EARLY COLD WAR HAVE FEWER AND FEWER DOCUMENTS, DESPITE THE GROWING SCOPE AND INTENSITY OF U.S. ACTIVITY ABROAD



The transition from paper to digital archives has been happening at a time of severe budget constraints at the National Archives and Records Administration (NARA). While decisions to make large collections of records 'temporary,' i.e., destined for destruction, seemed particularly dubious when they involved the Trump administration and Immigration and Customs Enforcement, cost is clearly a factor. Senior NARA officials have made it abundantly clear that they do not have the staff or funding to provide services that generations of historians have taken for granted. The National Archives no longer favors the creation of new presidential libraries -- starting with Obama's, all new libraries will be digital -- and soon it will not even accept paper files. But NARA does not even have the technical capacity to safeguard the electronic records it already holds. For instance, it has been reported that NARA staff cannot access the email of the latter years of the Ronald Reagan and the entire George H.W. Bush administration. Ironically, these emails were the subject of much litigation and a landmark ruling requiring the preservation of electronic records.[9] Now they may remain unavailable to researchers because the National Archives cannot manage even the first generation of email records, much less the 300 million that it received from the Obama White House.

[9] David A. Wallace, "Preserving the U.S. Government's White House Electronic Mail: Archival Challenges and Policy Implications 2–7 (1998)," (unpublished manuscript) http://www.ercim.eu/publication/ws-proceedings/DELOS6/wallace.pdf.

Everyone -- especially graduate students, or anyone training them -- needs to be aware of what all this means in terms of the conditions for doing historical research going forward. As soon as I saw what was coming, I realized I would need to work in a new way. I took some time to learn how to code -- badly -- if only to better understand the nature of this work, and how I could work more effectively with true professionals. I began to redesign my courses to try to turn them into laboratories, in which I could run experiments working alongside my students. But I also started to spend a lot of time applying for grants, so I could assemble a team, History Lab, and join the effort to develop new tools and techniques commensurate with the challenges we face.

Over the last seven years, we have been working to make collections like the Central Foreign Policy Files and the CIA CREST collection more useful to researchers. Instead of trying keyword searches across multiple websites, they can find eight different collections on a single website, with user-friendly visualization tools to identify the most important people, places, locations, and topics. Researchers can also export the data through an Application Programming Interface to build their own tools and conduct their own analyses.

My experience in recent years working with data scientists, developers, and engineers will likely not be typical, at least as long as professional norms discourage collaborative, multi-disciplinary work. Instead of just managing my own database on my own computer, I work with a team in aggregating and analyzing whole collections, storing the data in the cloud, and using a high-performance computing cluster to run complex experiments. It is still "bleeding edge," because all this has taken a lot of time to organize, and there have been many false leads and blind alleys. But the experience has made me truly excited about the enormous opportunities that lay before us. If the past is any indication, the tools will likely become easier to use, and many more people will be able to pick them up. There is still a lot of low-hanging fruit, such that even someone working alone with limited coding skills can carry out some really interesting work.

One thing is clear: For an increasing number of periods and topics, we cannot work in the same way historians once did, and should not pretend otherwise. Every time we cite an online source as if we found it in a library and read it on paper, we are denying reality. Instead, we could be embracing the opportunity to analyze entire archives, answer old questions with new rigor, and devise entirely new lines of inquiry. For me, one of the most intriguing is to systematically identify patterns and anomalies in official secrecy and archival destruction. It is more important than ever to know what we do not know, and we can now establish some ground truth by data-mining large collections of digitized or born-digital text.

But all this depends not just on historians keeping abreast of new developments in data science, but also working closely with librarians and archivists. I have been attending their conferences and workshops, and joined a task force on email archiving. I have also become more active in professional organizations of historians, like SHAFR and the AHA, to raise awareness of what we are facing, and try to tackle some of the larger, structural factors that have compounded our problems.

In recent years, as the academic jobs crisis has worsened, many have all but given up on these professional organizations and attack them on social media. But if the AHA did not exist, we would need to invent it. In fact, for anyone who hopes to be doing historical research twenty or thirty years from now, organizations like the AHA are the only ones able to effectively represent us when key decisions are being made, whether about restoring funding for archives, reforming antiquated declassification procedures, or a host of other more obscure but critical issues, like deciding how email providers support archiving. No individual researcher can hope to be active and knowledgeable about all these issues, and we cannot make progress on any of them if we do not act collectively. The AHA and other organizations could be doing more, of course, on archives no less than advocacy for contingent labor. Amazingly, the AHA did not even have a committee on archives until 2019. But it will not be able to do anything going forward unless younger historians -- who have the most at stake -- do not stop attacking it and instead join efforts to make it stronger.

When I started out thirty years ago, I had the great luxury of being able to only focus on my research, and could pick and choose whatever technique or technology seemed best. I simply assumed that the archives would always be there, documents would continue to be declassified, and archivists would be standing by and ready to answer my questions. But I've come to realize that, as a profession, we take these things for granted at our peril. Archives, after all, are not just about history. The

whole point of building an archive is to create something that is 'future proof' to better safeguard the historical record. But the future keeps changing, and it turns out that a digital archive -- like my old FileMaker database -- may already be more fragile and perishable than paper files in a brick-and-mortar building. If my experience is any indication, the next thirty years may see even more dramatic changes in the nature of archiving, and the nature of historical research. It is the most profound challenge facing our profession, and we can only surmount it if we do so together.

Essay by Kaeten Mistry, University of East Anglia

The Hollow Archive: The CIA's Digitized Declassified Records

Born a digital archive in the twenty-first century, the declassified files of the Central Intelligence Agency (CIA) are a curious, perplexing, and frustrating collection. This database is effectively the only available CIA archive (paper records are off limits), yet since its creation in 2000 until very recently, the only way to access the digital database was a distinctly analogue process. Consulting the CIA Records Search Tool (CREST) involved trekking to the National Archives facility at College Park, Maryland, negotiating the airport-grade security checks, descending to the cavernous basement to store personal belongings, before ascending to the reading rooms.

While this routine will be familiar to many researchers who have worked at NARA II, delving into CREST meant heading to the fluorescent-lit fourth floor reading room where a bank of four computers were hardwired to a large black database. Conspicuous CCTV cameras pointed at the terminals and a sign warned anyone contemplating mischief that all information about the visit was automatically logged by the Agency. Getting a computer terminal was on a first come, first-served basis. Files could be printed but not downloaded or saved. On my last visit only two computers and one printer were working, the toner and paper supply running low. The NARA archivist informed the five of us jostling for position that someone from Langley periodically came to check the terminals and restock the printers. However, when the next maintenance visit would take place was impossible to say.

Figure 3: The CREST database (right) and the four computers that could access it at NARA II, College Park, MD, before the collection was placed online in 2017.

Today researchers no longer face the rigmarole of securing a prized spot at a functioning computer on the fourth floor in College Park. Anybody with an internet connection can use CREST from anywhere in the world. In 2017 the entire collection was placed online on the CIA's Electronic Reading Room, where CREST material is augmented by further records released through Freedom of Information Act (FOIA) requests and various Agency release programs. Over 13

million pages can be consulted and downloaded from the comfort of one's home or place of work.[10] The declassified CIA records are arguably the most digitized archive of the entire U.S. government. Reading the primary sources of one of the most secretive arms of the Executive branch is as routine as scanning the newspaper or ordering takeaway online. Everyone appears to win. Files are accessible to all, our carbon footprint has been dramatically slashed, and Langley makes savings on the cost of ink and paper.

Yet the newfound ease of access cannot detract from fundamental problems with the content and form of the CIA archive. Simply put, the collection is piecemeal, lacks context, and huge questions surround the scope and utility of the documents that have been made available in the public arena. Furthermore, the archive was originally created in the context of broader Agency struggles *against* declassification and preoccupation with image management. While not entirely useless, the collection is not useful in any meaningful way. Indeed, its existence speaks to the perils of digital archives and superficial transparency.

*****

CREST is the de facto archive of CIA primary resources. It is not an electronic copy of a paper archive, nor a supplement to textual records. A small amount of paper files selected by the Agency are deposited at NARA II, primarily the writings of the CIA's in-house History Staff and documents involving miscellaneous early Cold War themes (some of these special topic collections have been digitized and placed online by NARA staff).[11] The bulk of the textual records remain with the CIA and/or are classified. In theory, all declassified documents more than 25 years old are on the database. In reality, the declassification program lags far behind schedule, with the bulk of the available material dating back to the early decades of the Cold War.

A fundamental problem is that the search engine identifies documents without any context. A keyword search can pull up hundreds, sometimes thousands, of files with no sense of order or place. Results appear randomly on a page, listed neither by date nor relevance, with the researcher's chances of stumbling across a relevant file uncertain. On occasion the same keyword search pulls up a different set of documents. A conventional google search has more logic and perspective. The absence of metadata is not a quirk of the system but a feature of the archive itself. There are no finding aides, guides, or overviews as to what is and isn't held. The holdings are not organized in any discernible fashion. Not only is it impossible to tell what may precede and follow a document (thematically or chronologically) in a hypothetical box or folder, but also, the place of an item in the archive as a whole is unfathomable. Since the researcher is thus unable to see the forest for the trees, the fragmentary nature of the material makes it difficult to assess the relevance of documents and ascertain what material would sit adjacent, thereby complicating the process of filing FOIA requests.

The only attempt to organize material is an Agency-curated selection of "Special Collections," a categorization that bears only a remote resemblance to a folder or even box. The eclectic range of topics include CIA Animal Partners, former Directors of Central Intelligence (DCI), the Missile Gap, and Doctor Zhivago.[12] A deeper dive into a collection released in 2019 – the "Argentina Declassification Project - The 'Dirty War' (1976-83)" – highlights a further issue: records are replete

[10] CIA Freedom of Information Act Electronic Reading Room, https://www.cia.gov/readingroom/.

[11] For an overview of CIA textural records deposited at NARA see https://www.archives.gov/research/guide-fed-records/groups/263.html. Digitization of some of the special topic collections are available at https://www.archives.gov/research/intelligence/cia.

[12] CIA Special Collections Archive, https://www.cia.gov/readingroom/special-collections-archive.

with redactions.[13] Although redactions are a common feature of national security information, their scale can render declassification next to impractical. One 122-page report from the collection is, with the exception of just over two pages, entirely redacted (including details about the sender, recipient, date). A footnote clarifies that the declassification process focused "on Argentina only information" with everything else beyond the scope of review, "whether or not it has been previously released."[14]

Another feature of the archive is an abundance of press clippings. This is noteworthy on a couple of levels. Firstly, it reveals an underappreciated aspect of CIA culture. The Agency was acutely attentive to its public image, systematically collating clippings and news reports about itself. While scholarship has begun to consider CIA public relations,[15] the obsession of this most secretive arm of the U.S. government with respect to its public image remains under-examined. Secondly, it highlights the nature of the modern secrecy regime whereby anything that enters the system is bound by the rules of classification. Press clippings, speeches, transcripts, academic articles, popular magazines, memoirs and biographies, and other items already in the public realm can be stamped secret on account of having crossed the desk of an intelligence official. Over 25 years can pass before that item is declassified. Running information that is widely available in the public sphere through the secrecy regime is not merely absurd but a waste of everyone's time.

Beyond press clippings and redacted reports, the collection is made up of mundane office memos and assorted titbits. While some add colorful details to what is already known, the majority of documents are worthless. As one academic review put it, "the documents made available through CREST are at best uninteresting."[16] All of which begs questions about the purpose of an archive that does little to help reconstruct the past and whose contribution to public understanding is scant. Historians of U.S. foreign relations have found little of value; a cursory glance at seminal works on major U.S. conflicts in which the CIA was involved, including Vietnam, show CREST material does not feature amid vast source bases. In my own work on U.S. political warfare in Italy, I did not draw on the archive when examining the inaugural CIA covert campaign of the Cold War. Even recent histories of the Agency rarely cite more than a handful of items from the principal archive of the subject of inquiry.[17]

*****

[13] See the recent H-Diplo Forum, "The Argentina Declassification Project: A Model of "Declassification Diplomacy" to Advance Human Rights—and History" for a larger discussion of the Argentina Declassification Project: https://issforum.org/forums/Forum-2021-1.pdf.

[14] "Resistance Movement Outside Chile," Document Number (FOIA)/ESDN (CREST) 03304956 (Approved for Release 2018/10/01), CIA Electronic Reading Room, https://www.cia.gov/readingroom/docs/THE%20RESISTANCE%20MOVEMENT%20O%5B15515140%5D.pdf; Note on Collection page, https://www.cia.gov/readingroom/collection/argentina-declassification-project-dirty-war-1976-83.

[15] Simon Willmetts, "The Burgeoning Fissures of Dissent: Allen Dulles and the Selling of the CIA in the Aftermath of the Bay of Pigs," *History* 100:340 (April 2015): 167-188; Willmetts, "The CIA and the Invention of Tradition," *Journal of Intelligence History* 14:2 (2015): 112-128; David S. McCarthy, *Selling the CIA: Public Relations and the Culture of Secrecy* (Lawrence: University Press of Kansas, 2018).

[16] David M. Barrett and Raymond Wasko, "Sampling CIA's New Document Retrieval System," *Intelligence and National Security* 20:2 (2005): 332-340, here 332.

[17] See for instance, Fredrik Logevall, *Embers of War: The Fall of an Empire and the Making of America's Vietnam* (New York: Random House, 2012); Kaeten Mistry, *The United States, Italy and the Origins of Cold War: Waging Political Warfare* (Cambridge: Cambridge University Press, 2014): 127-152, 176-199; Richard H. Immerman, *The Hidden Hand: A Brief History of the CIA* (Malden: John Wiley, 2014).

The problem of content is tied to the creation and subsequent evolution of the digital archive itself. The origins of CREST lay in a post-Cold War era where debates about secrecy and the role of intelligence agencies in a world without the Soviet Union brought calls for reform and greater transparency. Although the 1991 proposal by Senator Daniel Patrick Moynihan to disband the CIA did not take hold, there was increasing public and political pressure to release historical records.[18] The CIA sought to get ahead of the curve through an initiative of greater "openness." Robert Gates, Director of the CIA under George H.W. Bush, outlined three areas this would manifest: greater communication with the press and public; more work with the academic community, including increased activity by the Center for the Study of Intelligence (CSI) in organizing conferences and publishing documentary collections on topics like the Cuban Missile Crisis and early Cold War intelligence; and increased declassification of historical documents.[19]

"Openness," however, proved to be fleeting and hollow. The CIA considered releasing a trove of documents in the 1990s, including on well-known historic CIA covert operations, even staging an event at Independence Hall in Philadelphia to celebrate the initiative. Agency promises to release records led to only one study, the history of Operation PBSUCCESS, the 1954 covert intervention in Guatemala written by historian Nick Cullather. But because it feared the implications for an upcoming Guatemalan election, publication was vetoed. Cullather's study was eventually published by a university press but other histories never materialized, again on the grounds that it would compromise contemporary U.S. relations with the countries in question.[20] Reflecting on his time on the CIA's History Review Panel (HRP), a group of outside scholars brought in to advise the Agency on releasing records during the first half of the decade, George Herring concluded it was a "brilliant public relations snow job."[21]

A 1995 Clinton administration Executive Order forced the declassification issue. EO 12958 established a system for classifying and safeguarding national security information that included an unprecedented effort to automatically declassify records over 25 years old that were of "permanent historical value." Although there were exemptions for federal agencies to protect sources, methods, and material related to national security, the expectation was for a publicly available "Governmentwide database of information that has been declassified."[22] To assist the process of identifying historically valuable records, the HRP was revamped with a new cohort of non-governmental historians and political scientists meeting twice a year and providing recommendations directly to the DCI.

CREST launched in 2000, although momentum for declassification was soon lost in the post-9/11 era. In fact, it was reversed when the George W. Bush administration amended the Clinton EO to allow more information to be classified and kept so for longer, stymying requirements around declassification and removing the request for a database. In 2009 the

[18] Daniel Patrick Moynihan, "Do We Still Need the C.I.A.?" *New York Times*, May 19, 1991: E17. Moynihan led a bipartisan study on the problem of government secrecy, which published its final report in 1997. The appendix – an institutional history of the secrecy system – was expanded and published as a book: *Report of the Commission on Protecting and Reducing Government Secrecy* (Washington, D.C.: Government Printing Office, 1997); Moynihan, *Secrecy: The American Experience* (New Haven: Yale University Press, 1998).

[19] Robert M. Gates, "CIA and Openness," Remarks at Oklahoma Press Association (February 21, 1995), https://fas.org/irp/eprint/gates1992.html.

[20] Nick Cullather, *Secret History: The CIA's Classified Account of Its Operations in Guatemala, 1952-1954* (Stanford: Stanford University Press, 1999); Mistry, "Approaches to Understanding the Inaugural CIA Covert Operation in Italy: Exploding Useful Myths," *Intelligence and National Security* 26:2-3 (2011): 246-250. See also McCarthy, *Selling the CIA*, 77-97.

[21] George Herring, "My Years with the CIA," Speech at American Historical Association annual conference (January 1997), https://fas.org/sgp/eprint/herring.html.

[22] Executive Order 12958, "Classified National Security Information," April 17, 1995, https://www.govinfo.gov/content/pkg/WCPD-1995-04-24/pdf/WCPD-1995-04-24-Pg634.pdf.

Obama administration produced a new executive order that sought to rekindle the flagging declassification agenda.[23] In the background, CREST continued to operate on its own track, with fewer declassified documents being added to the database.

The struggle for meaningful declassification is of course a larger issue. But the critical point is that the CIA favored standalone initiatives like CREST and CSI publications over collaboration on programs like the compilation of *Foreign Relations of the United States* (FRUS) volumes and reviewing documents for presidential libraries and NARA.[24] This in turn has caused friction with the HRP advisory group, as well as the State Department's Office of the Historian and its advisory group (which, unlike the HRP, is statutorily mandated). One of the first reports of the revamped HRP in 1996 noted concerns that CSI outputs reflected "a 'scattershot' approach" that "do[es] not substitute for the opening of archives that would allow historians not affiliated with the Agency to work directly with its records."[25] Over two decades later, the latest HRP continued to underscore the importance of collaborating on *FRUS* volumes and working with presidential libraries, emphasizing how they represented "great value to scholars and the public."[26] In 2019, the HRP made its final public statement, announcing that the Trump administration had effectively shut down meetings and that membership of the group was "being restructured." With one exception, all panel members, including long time chair Robert Jervis, were removed.[27] The Biden administration has resurrected and reconstituted HRP as the Historical Advisory Panel (HAP), with meetings slated to resume in 2021. While some of the chaos in recent years can be attributed to the Trump administration's modus operandi, problems with the government's system for declassifying historical documents is deep-rooted.[28]

The travails of the HRP point to a wider problem with the CIA's declassification program. Simply put, initiatives are strewn with exemptions and are reliant on Agency collaboration. The dictates of protecting 'national security' allow vast swathes of documents to be partially or fully redacted, if never released at all. The lack of CIA engagement on declassification efforts that are widely recognized to be of great utility (*FRUS*, the NARA system) has become a familiar lament across administrations of contrasting political persuasion by different advisory panels. In the meantime, the CIA continues to trumpet its commitment to transparency by curating collections on topics largely defined by itself (unilaterally

[23] Executive Order 13292, "Further Amendment to Executive Order 12958, as Amended, Classified National Security Information," March 25, 2003, https://www.govinfo.gov/content/pkg/WCPD-2003-03-31/pdf/WCPD-2003-03-31-Pg359.pdf (for an annotated version showing all deletions see https://fas.org/sgp/bush/eo13292inout.html); Executive Order 13526, "Classified National Security Information," December 29, 2009, https://www.archives.gov/isoo/policy-documents/cnsi-eo.html.

[24] The Foreign Relations statute of 1992 mandates that volumes in the *FRUS* series are "thorough, accurate, and reliable," and for that purpose a process exists to acknowledge covert actions. It is cumbersome though and essentially provides the CIA with veto authority.

[25] John Lewis Gaddis to DCI John Deutch, "Report: CIA Historical Review Panel Meeting, February 5, 1996," March 6, 1996, https://fas.org/sgp/advisory/ciarev.html.

[26] Robert Jervis (Chair), "Public Statement from the CIA's Historical Review Panel," July 24, 2017, https://networks.h-net.org/node/28443/discussions/188287/public-statement-cia%E2%80%99s-historical-review-panel.

[27] Robert Jervis (Chair), "Announcement from the CIA's Historical Review Panel (HRP)," January 14, 2019, https://networks.h-net.org/node/28443/discussions/3569932/announcement-cia%E2%80%99s-historical-review-panel-hrp. The interim task of monitoring CIA cooperation over *FRUS* volumes fell to the Advisory Committee on Historical Diplomatic Documentation (HAC) to the Department of State. For minutes of HAC meetings see, https://history.state.gov/about/hac/meeting-notes.

[28] William Burr, "Trapped in the Archives," *Foreign Affairs*, November 29, 2019, https://www.foreignaffairs.com/articles/2019-11-29/trapped-archives.

deciding what documents it should release in conjunction) and drip-feeding material onto the FOIA Electronic Reading Room.

Yet even the decision to publish the CREST records online underscores the inherent resistance to making declassified records more easily accessible. In 2014, MuckRock, a non-profit group focused on government transparency and sharing publicly available information, sued the CIA for failing to respect FOIA processes and impeding public access to documents they it had declared were no longer secret. The Agency offered a series of remarkable rationales for why the database could only be accessed via four computers in College Park, although the excuses brook no argument. Facing a legal defeat, in late 2016 CIA announced that "given the high public interest in the CREST database" it had "recently surged resources" to place the entire collection online.[29] The newfound altruism and concern for the public interest was ostensibly unrelated to the lawsuit.

*****

Bemoaning the CIA's perfunctory approach to declassification and stonewalling is a logical reaction. Even by the standards of an intelligence agency defined by secrecy, CIA resistance to mandated transparency is remarkable. Yet the reality is that historical and popular understanding of the Agency is not particularly helped by initiatives like CREST. Scholarship has shown it is far from a necessary primary resource for writing histories that include the CIA.

Other research centers, archives, transparency initiatives, and organizations are more useful for working with declassified documents and CIA records than the Agency's much touted digital archive. The most effective forums for sources and analysis comes from the likes of the National Security Archive, the Federation of American Scientists program on Government Secrecy, and MuckRock, among others. History Lab has been working to process the CREST collection as part of a larger database of declassified records, which will soon be available online, that makes it more useful to researchers.[30] The national security reporters at newspapers and news sites, new and old, are among the most knowledgeable individuals on contemporary matters. The State Department has a range of programs and the *FRUS* volumes, for all their issues, continue to represent an essential resource. The NARA system handles significant national security material, with the chances of releasing further information arguably greater via Mandatory Declassification Review requests at presidential libraries.[31] Research in non-state repositories and international archives can help examine the issue from the perspective of those allied or opposed (or both) to the CIA. Whatever one's views on national security whistleblowers, their revelations have placed important information in the public realm that triggered important discussions and had political

[29] Bruce Falconer, "Inside the CIA's (Sort of) Secret Document Stash," *Mother Jones*, April 3, 2009, https://www.motherjones.com/politics/2009/04/cias-open-secrets/; Michael Morisy, "How We Sued the CIA and (Mostly) Won," *MuckRock*, December 14, 2016, https://www.muckrock.com/news/archives/2016/dec/14/lawsuit-cia-crest/; Kel McClanahan, "Our Three-Year Saga to Release 13 Million Pages of CIA Secrets," *MuckRock*, January 19, 2017, https://www.muckrock.com/news/archives/2017/jan/19/three-year-saga-behind-CIA-release/.

[30] See the National Security Archive, including its Virtual Reading Room, https://nsarchive.gwu.edu/; FAS project on Government Secrecy, including the *Secrecy News* blog, https://fas.org/issues/government-secrecy/; MuckRock FOIA requests, https://www.muckrock.com/foi/; and History Lab's FOIArchive, http://history-lab.org/search, as well as Matt Connelly's essay in this forum.

[31] Traditional state archives (NARA, State Department records, presidential libraries, Library of Congress) have detailed finding aids online and varied initiatives to facilitate research, including processes for requesting material remotely. Helpful resources are also listed by research centers such as the Miller Center (Virginia), Cold War Studies and International History (Santa Barbara), IDEAS (LSE), Cold War International History Project (Wilson Center). In addition to their reporting, many respected national security journalists – for instance Jane Mayer (*New Yorker*), Charlie Savage (*New York Times*), Dana Priest (*Washington* Post), Mark Mazzetti (*New York Times*), Spencer Ackerman (*Wired*, *Guardian*, *Daily Beast*), Jeremy Scahill (*Nation*, *Democracy Now!*, *Intercept*), Barton Gellman (*Washington Post*) – have produced important books on contemporary intelligence, surveillance, and secrecy.

ramifications.[32] In short, there are other avenues to find primary documents. The CIA's digitized records not only fall short as an archive but have proven less useful than other repositories, traditional and contemporary, physical and online. CREST demonstrates that a computer terminal or keyword search does not substitute for conventional leg work.

As the other contributors to the roundtable argue, there are numerous possibilities and pitfalls to digital archives. This is a moment, accelerated by the COVID-19 pandemic, where the future of archival research can be imaginatively reimagined. There are numerous paths ahead. What is certain is that having a digital archive that represents an instrumental tool for an organization that has a lukewarm approach to declassification and that releases records online without context is to be avoided. It is superficial, disingenuous, and arguably pointless.

[32] Mistry, "Approaches to Understanding the Inaugural CIA Covert Operation," 251, 261-65; Kaeten Mistry and Hannah Gurman, eds., *Whistleblowing Nation: The History of National Security Disclosures and the Cult of State Secrecy* (New York: Columbia University Press, 2020).

Essay by Christopher J. Prom, University of Illinois at Urbana-Champaign Library

What Are Digital Archives?

What are digital archives? What do we gain (and lose) when archives move from analog to digital media?

These questions can be addressed in many ways. After all, digital archives, like their paper-based cousins, are produced by many different people and organizations. They come in many flavors, and they can be used for many purposes. And in the archives and library community, many people seek to preserve them and to make them discoverable. Ultimately, digital archives exist to be used, whether a researcher is seeking personal enrichment, protecting his or her rights, or conducting historical analysis.

I am one of the people who is working to shape society's digital legacy. While the main focus of my work is at my home institution, I'm part of a digital archives community that spans the globe. We're a group of like-minded professionals who are working in the corners of our digital economy to address an issue that most people assume has been solved, but hasn't: how to preserve a digital record of society. (After all, isn't digital storage free or next to free?)

We're making significant headway in building discoverable, usable, and sustainable digital archives, even while we recognize and juggle a continuing responsibility to steward access to paper and other analog collections. A key part of our jobs is defining and implementing best practices and standards to ensure that our institutions identify and preserve authentic digital records (either trustworthy replicas of paper records or copies of 'born-digital' materials that never existed in analog form). We seek to do this by creating descriptive records and contextual information (metadata) that makes both types of records understandable and usable—the type of work described so well in Joseph Wicentowski's essay on the origins of the *FRUS* digital edition.[33]

Like some in the community, I entered through the backdoor: Trained as a historian, I worked for many years as an archivist. (My path was a bit atypical, since most archivists now hold a master's degree in library, Information, or archival science. Subject area training, often in history, is also common.) I now oversee digital operations for a major research institution, the University of Illinois at Urbana Champaign Library system. This twenty-year evolution provided me an opportunity to learn about digital archives and even to make a few contributions of my own. By sharing my experiences, I hope that H-Diplo readers and other members of the H-Net Community will likewise learn a little bit more about the records my community is seeking to preserve and how to best interact with them.

I first encountered archives (pre-digital) in 1994 and was promptly confused. It would be easy to chalk that up to the fact I was an inexperienced Ph.D. candidate in history. Like most trainees in my generation, I had read the work of Michel

[33] The Society of American Archivist's Dictionary of Archives Terminology defines metadata as "a characterization or description documenting the identification, management, nature, use, or location of information resources (data)." While users of archives will be most familiar with metadata that describes an information resource, there are other metadata types, including administrative, preservation, and structural metadata. Taken as a whole, metadata is often defined as 'data about data,' but I prefer to think of it as 'data about a record." It which helps ensures not only that the record itself, as an output of some human or machine activity is preserved, but also that those using it are aware of its provenance, chain of custody, internal structure. Ideally, metadata also documents a record's connections to other records, record creators, or record-creating activities, or to those whose lives or activities are documented by the records. Metadata" and "Record," in *Dictionary of Archives Terminology* (Society of American Archivists), accessed July 30, 2021, https://dictionary.archivists.org/entry/metadata.html and https://dictionary.archivists.org/entry/record.html.

Foucault, but I knew little to nothing about how archives are formed, much less how to find and interpret them.[34] It was on-the-job training. For three weeks, I toddled around UK country record offices wearing my multi-colored jacket and trainers, giving myself away as an American graduate student even before archivists puzzled over my Wisconsin accent. I eventually came to figure out that the archivists differentiated between two types of archives: the records of organizations, archives proper (or 'the archive,' in the UK), and personal papers or family collections (manuscripts). Why had someone not let me know that beforehand?

If they had, I might have realized that there was a good reason why, while thumbing through dozens of binders at the Historical Manuscripts Commission office in Chancery Lane, why I still hadn't found what I was looking for (in the words of the U2 earworm that was looping through my head at the time). I had quite unintentionally stumbled into a room filled with finding aids: the inventories, registers, and lists that archivists use to make available descriptive information about the holdings of materials relating to a particular person, family, or organization.

I didn't realize it at the time, but such lists can describe materials at many different levels of specificity, from short paragraphs describing a record series, down to and including individual folders or even items.[35] Archival arrangement is probably easier to show than it is to explain, but it proceeds from some core assumptions: namely that records of one creating person or organization should be kept together (*respect des fonds*) and that those records should be preserved in the manner of their creation and use (original) order. Based on these principles, archivists typically organize materials in a hierarchical fashion with various "levels" of description such as the repository level, record group level, series level, box level, file level, and item level being the most prominent. Mercifully, the metadata in those HMC binders has now been incorporated into other indices and online search tools, even as the core principles described above are applied equally to print and digital materials.[36]

If the HMC granted me an inauspicious introduction to archives, it also helped me understand that there is a big difference between *indices* to archives and the archives themselves, a point that takes on more salience in 2021's digital-first world. When I began working as a Visiting Assistant Archivist (my PH.D. was delayed so that I could put bread on the table), one of my first jobs was to convert a paper-based index system into an online digital database. Many people working in the archives community spent the better part of the early 2000s doing something similar, developing some data structure standards along the way: most notably Encoded Archival Description and Encoded Archival Context.[37] These efforts, and

[34] Michel Foucault, *The Order of Things: An Archaeology of the Human Sciences*, 1st American ed., World of Man (New York: Pantheon Books, 1971). Many years later, I remediated the gap in practical knowledge by reading Richard Cox and James O'Toole's book *Understanding Archives* (Chicago: Society of American Archivists, 2006).

[35] Dennis Meissner, *Arranging and Describing Archives and Manuscripts* (Chicago: Society of American Archivists, 2019), 164.

[36] The National Archives (UK). n.d. "Historical Manuscripts Commission," accessed June 25, 2021, http://www.nationalarchives.gov.uk/archives-sector/our-archives-sector-role/historical-manuscripts-commission/.

[37] "Development of the Encoded Archival Description DTD (EAD Official Site, Library of Congress)." n.d., accessed June 25, 2021. https://www.loc.gov/ead/eaddev.html.

; Daniel V. Pitti, "Creator Description: Encoded Archival Context," *Cataloging & Classification Quarterly* 38:3-4 (October 25, 2004): 201–26, https://doi.org/10.1300/J104v38n03_16.

many others, resulted in the several kinds of digital catalogs: Local,[38] regional,[39] or national.[40] In addition, aggregator services for digital objects can be quite helpful in identifying source materials.[41] The catalog resources that people in my profession build are often historian's first line of attack (or source of confusion) in finding potential source materials, and potentially, in using some digital materials.

But as with the paper binders, it didn't take me or the rest of the profession long to discover the limited utility of the online indices, which mimicked their paper predecessors in one key respect: they were long, nested lists of folders. They were less digital archives and more a shadow of one, both demarcating the contours and concealing the details within in their hazy outlines (aka 'finding aids').[42] It's not surprising that, since the mid-90s, the archival and especially the Library community began investing heavily in projects that digitize paper or other analog collections so that those lists could be linked to something, a bit like the proverbial pot of gold at the end of a rainbow, since there is so much to digitize.[43]

Initially, much of the focus of this work was branded as that of building "Digital Libraries," defined in 1998 as "organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities."[44] It's still a workable definition, with a welcome focus on the organization and institutions that sit behind digital objects. As an archivist, I would supplement it only lightly, by adding that digital archives are a type of digital library, namely the organized, non-current records of a person or organization, preserved in a way that preserves and makes apparent their provenance and context of creation. That last bit is important.

Some of the early projects to digitize paper-based archives or photographs stumbled a bit out of the gate—a fact noted at the time as the possibly apocryphal "on-a-horse" problem. The story goes that a library technician transcribed captions of a Teddy Roosevelt image as "On a horse (1)," "On a horse (2)," etc., but neglected to include the key detail: the name of the rider. Whether or not this story is true, it illustrates a fundamental problem. Digital library and archives systems, particularly those that emphasize item-level descriptions, can easily obscure critical contextual information that can help archivists and scholars understand records or identify related materials, such as similar documents connected to the same records creating person or body. In my own work, we tried to remediate this problem by integrating two complex systems (a website application that turned the pages of digitized correspondence, with the webpages that represented our encoded

[38] Many institutions provide access to their finding aids and descriptive records via database applications like ArchivesSpace (https://archivesspace.org) or Access to Memory (https://www.accesstomemory.org/en/).

[39] "Online Archive of California – Help," accessed July 12, 2021. https://oac.cdlib.org/help/.

[40] For Example: OCLC, "ArchiveGrid," April 23, 2021. https://www.oclc.org/research/areas/research-collections/archivegrid.html. The Archives Hub, "About the Archives Hub Service," accessed July 12, 2021. https://archiveshub.jisc.ac.uk/about/, "The National Archives Catalog," accessed July 30, 2021, https://catalog.archives.gov/.

[41] Digital Public Library of America, https://dp.la/.

[42] Society of American Archivists, "Dictionary of Archives Terminology: Finding Aid," n.d., accessed June 25, 2021, https://dictionary.archivists.org/entry/finding-aid.html.

[43] Many of these projects have been documented in the interviews conducted by the Digital Pioneers Project: "Digital Pioneers," accessed July 12, 2021, http://digitalpioneers.library.du.edu/.

[44] Donald Waters, "What Are Digital Libraries?" *CLIR Issues*, July 13, 1998, https://www.clir.org/1998/07/clir-issues-number-4/.

finding aids), work that I discussed in my first presentation and publication as an archivist and assistant professor on my Library's tenure track.[45]

What was I learning as I sought to adapt the principles of paper based archival management to digital files? The lesson was simple: That the specific ways that digital archives systems are designed, as well as the practices staff members use when creating descriptive metadata, have an outsized impact on the ability of people to discover the existence of records relevant to a research need. Some systems support a 'flat' metadata model, in which a name, subject, or location is simply a text field on that particular record. Others enable authority control and networking of records to other records, laying base social connections or hidden relationships both to the records creators and between the people and organizations who create records.[46] Many if not most of the current systems provide many clues that help scholars understand the relationships between specific digital files, other files in the collection, and other associated materials.

In addition, the archives community has developed and supports a range of standards to ensure that digital copies of analog materials retain the fidelity of the originals and that such records are created with appropriate metadata for preservation, discovery, and access.[47] The most prominent and widely implemented set of such standards are those that emerged from the Federal Agencies Digital Guidelines Initiative (FADGI).[48] Recently, for instance, the National Archives promulgated draft standards that government agencies would be required to follow when scanning textual documents to be submitted under the mandate that the vast majority of permanent federal records be deposited to the National Archives in digital form. This mandate implanted a policy established by OMB/NARA Memorandum, Transition to Electronic Records (M-19-21).[49]

Directives and directions like this may seem worrisome to scholars and members of the public, and there are important questions to address. Will government agencies prioritize digitization of records that have no immediate business use? Will the National Archives and Records Administration be funded at a level sufficient to provide an infrastructure to care for these materials? How will records be reviewed for declassification? While these questions deserve answers, it is worth noting that the best practices and baseline expectations for digitization and digital access have been widely accepted in the archives community, as well as by equipment manufacturers and vendors, illustrating that the field is increasingly achieving maturity. As Matt Connelly notes in his essay, it benefits all historians to learn a bit about how the tools they are using have been constructed, and in many cases digitization will open up untold access possibilities.

---

[45] Christopher J Prom, "E-content on a Shoestring: Using the EAD Cookbook and Ebind XML to Deliver the James Scotty Reston Papers." Computers in Libraries 2001 Proceedings (Medford, NJ Information Today): 192-198.

[46] R. Larson, D. Pitti, and A. Turner, "SNAC: The Social Networks and Archival Context Project - Towards an Archival Authority Cooperative," *IEEE/ACM Joint Conference on Digital Libraries, Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference On*, September 1, 2014, 427-428, https://doi.org/10.1109/JCDL.2014.6970208. The SNAC project website is available at https://snaccooperative.org/.

[47] Donald Waters and John Garrett, "Preserving Digital Information, Report of the Task Force on Archiving of Digital Information" (CLIR: 1996) https://www.clir.org/pubs/reports/pub63/, is a touchstone report. More recent works provide an overview of the entire range of planning and preservation decisions that each institution needs to address in curating access to digital materials. See Trevor Owens, *The Theory and Craft of Digital Preservation* (Baltimore: Johns Hopkins University Press, 2018) and Aaron D. Purcell, *Digital Library Programs for Libraries and Archives: Developing, Managing, and Sustaining Unique Digital Collections* (Chicago: ALA Neal-Schuman, 2016)

[48] "Federal Agencies Digital Guidelines Initiative." n.d., accessed July 1, 2021, http://www.digitizationguidelines.gov/.

[49] U.S. National Archives, Federal Register, "Federal Records Management: Digitizing Permanent Records and Reviewing Records Schedules," December 1, 2020, https://www.federalregister.gov/documents/2020/12/01/2020-26239/federal-records-management-digitizing-permanent-records-and-reviewing-records-schedules.

As a recent book about audiovisual archives noted, digitization is often the only preservation option, at least the only one that provides meaningful access. If we are truly interested in ensuring the long-term longevity of materials that depend on inherently unstable media (such as magnetic tape) and obsolete playback equipment, they must be converted to forms that make them easy to copy, preserve, and use.[50] Standards such as those promulgated by the FADGI project ensure that these materials will remain discoverable, accessible, and usable. Regardless of format, single physical copies remain at risk of catastrophic loss, as was tragically illustrated by the destruction of the Cape Town University's special collections library in spring 2021.[51]

Scanned paper documents and photographs, as well as digitized audio and video materials compose one type of 'digital archive.' Archivists also seek to identify and preserve a variety of 'born digital' materials, such as emails, digital photographs, documents, websites, social media, and myriad other record types. As with digitized records, the digital archives community has developed a range of standards to help ensure that these formats can be identified and preserved, such as the Digital Preservation Coalition's Technical Watch Reports.[52] And many of these 'born digital' records are now beginning to enter the holdings of established institutions, or partners like the Internet Archive, which offers a web archiving service. My current work focuses on a collaborative project to define standards for the preservation of email, which is one small example of what Trevor Owens has called *The Theory and Craft of Digital Preservation.*[53] There are many, many others involved in other aspects of this work, too many to name here.

When attempting to locate and use digital archives, it is important to keep in mind that, whatever system you are using, you'll be navigating between descriptions, digital objects (either digitized or in a native digital format), and back again. And not everything can or will be found online, due both to copyright limitations and the fact that many resources have not been digitized, or may have been digitized by private corporations, which then sell access to libraries. A good jumping off point for any seeking to find digital archives will be the local catalogs that nearly every archival repository now places online, as well as some of the online resources I've cited above. And don't forget to talk to an archivist: When the database isn't helping, the archivist may know where or how to look. After all, the metadata and systems they create can never be fully complete or adequate, given both the very real resource limitations, as well as the inherent challenges of fully documenting a record or its context.[54]

Unfortunately, not every digital resource follows practices such as these. As Kaeten Mistry describes in his essay on CREST files, the lack of such standards can mask other, deeper problems with such collections, such as an insufficient statutory or regulatory basis to facilitate access to public records, even long after need for classification has passed. A combination of factors like these make many digital archives little more than a scrap heap of administrative trivia.

---

[50] Anthony Cocciolo, *Moving Image and Sound Collections for Archivists* (Chicago: The Society of American Archivists, 2017).

[51] Christina Goldbaum and Kimon de Greef, "Wildfire Deals Hard Blow to South Africa's Archives," *The New York Times*, April 19, 2021, sec. World, https://www.nytimes.com/2021/04/19/world/africa/cape-town-table-mountain-fire.html.

[52] Digital Preservation Coalition, "DPC Technology Watch Publications," various dates, https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports.

[53] EA-PDF Working Group, "A Specification for Using PDF to Package and Represent Email" (Board of Trustees of the University of Illinois, January 2021), https://www.ideals.illinois.edu/handle/2142/109251; Trevor Owens, *The Theory and Craft of Digital Preservation* (Baltimore: Johns Hopkins University Press, 2018).

[54] See Donald C. Force and Randy Smith, "Context Lost: Digital Surrogates, Their Physical Counterparts, and the Metadata That Is Keeping Them Apart," *The American Archivist* 84:1 (June 24, 2021): 91–118, https://doi.org/10.17723/0360-9081-84.1.91, for a detailed description of some practical challenges, as well as suggestions for improvement.

At the end of the day, archivists seek to preserve a record of society that is both useful and used. Many individuals benefit from the work that academic, government, and nonprofit archives pursue. At its core, that work seeks to identify, preserve, and provide access to records that can be used to understand and interpret the immense economic, social, cultural, political, and diplomatic stressors and events that are shaping life in the twenty-first century. The task is enormous and unending, and technologies, methods, and expertise are still in their adolescence. Yet they will surely mature. What is more, we can take some comfort in an indisputable fact: Only a small portion of the records from prior ages have escaped destruction. The surviving record, however, is becoming more and more accessible, through digital methods and technologies. And digital methods and technologies hold the potential of preserving and making accessible future records to an extent prior generations of historians and archivists could not even imagine—provided that institutions have the will, the means, and the support to use them.

Essay by Joseph C. Wicentowski, the Office of the Historian

Tour de FRUS: The origins and evolution of the Foreign Relations of the United States digital edition[55]

In 2018 the Office of the Historian completed its decade-long project to digitize all 535 printed volumes in the *Foreign Relations of the United States* (*FRUS*) series, providing the public free and unparalleled full-text access to the official documentary record of U.S. foreign relations through the Office's public website, history.state.gov. Today, the site houses over 310,000 documents from nearly 550 volumes, covering 1861 to 1989. Progress has already begun on digitizing the final segment of the series: the thirteen microfiche publications released in the 1980s and 1990s, which are indisputably the least accessible portion of the entire 160-year-old series. With the Office having reached this advanced stage in the *FRUS* digitization project, this is an opportune time to review how researchers can take full advantage of the fruits of this initiative. This article offers a brief account of the creation of the Office's website and the modern *FRUS* digital edition, details the scope of the holdings, explores the capabilities of the website, and suggests additional possibilities for advanced research with the corpus using its open data. Along the way, it acknowledges some limitations of the current offerings and identifies opportunities for improvement.

Creating history.state.gov and the Modern FRUS Digital Edition

The Office of the Historian launched its public website, history.state.gov, in March 2009. Since then it has become an enduring online presence for the *FRUS* digital edition and the Office's many publications and datasets on the history of U.S. foreign relations and the institutional history of the U.S. Department of State. In establishing a site independent from the Department's main website, state.gov, the Office's goal was to overcome the fragmentation of its online publications and to create a stable, dedicated, high quality, scholarly resource for accessing the Office's publications.

The Office's earlier online publishing initiatives, which reached back to the Department of State Foreign Affairs Network (DOSFAN) collection[56] hosted by the University of Illinois at Chicago (1990-97) and various incarnations of state.gov (1998-2009), were responsible for the online release of 88 volumes in the *FRUS* series (covering largely the Kennedy through Ford administrations), including the series' first electronic-only volumes, and a growing collection of online adaptations of printed publications.[57]

By the late-aughts, it had become apparent that the Office's online publications needed a new home. *FRUS* was fragmented in location and in form. Depending on the vintage of a digital release, a volume might appear on DOSFAN or one of two sections of state.gov, and it might appear in plain text, in chapter-sized HTML pages, in groups of ten documents, in chapter-sized PDFs, or as monolithic PDFs. Like many websites of the time, the *FRUS* pages struggled to present essential features of scholarly editing, such as footnotes. These developments in digital publishing mirrored and amplified the

[55] I would like to thank Joshua Botts, Mandy Chalou, Renée Goings, and Kathleen B. Rasmussen for their comments on earlier drafts of this paper. I delivered a public presentation of the same title at a virtual meeting of the Advisory Committee on Historical Diplomatic Documentation (HAC) on August 30, 2021. In due course, a video recording of the presentation, which included detailed illustrations of the resources described in this essay, will be posted to the Office of the Historian's collection of HAC Meeting Notes, at https://history.state.gov/about/hac. For inquiries, please contact the Office of the Historian at history@state.gov.

[56] "Transfer of DOSFAN Site to GPO," September 24, 2018, Federal Depository Library Program, https://www.fdlp.gov/news-and-events/3794-transfer-of-dosfan-site-to-gpo.

[57] See U.S. Department of State, Archive, "Office of the Historian" (website), 1 January 1997-20 January 2001, https://1997-2001.state.gov/about_state/history/index.html and U.S. Department of State, Archive, "Bureau of Public Affairs: Office of the Historian" (website), 20 January 2001-20 January 2009, https://2001-2009.state.gov/r/pa/ho/index.htm.

fragmentation of the series between print and microfiche in the decade preceding the arrival of the World Wide Web. Researchers now had to look in even more places—and use markedly different tools in each place—to exploit the entire *FRUS* corpus. There was no single, full-text searchable online source that held every volume and document in the *FRUS* series.

The University of Wisconsin Digital Collections Center's (UWDCC) online edition of *FRUS*,[58] which was built from 2003-2008, demonstrated to historians in the Office and throughout the field the power of a unified portal into the series. The University had selected *FRUS* for UWDCC's first experiment in mass digitization, and—without assistance from the Department of State—assembled a collection of 375 printed *FRUS* volumes from the inaugural 1861 volume through the 1958-1960 subseries, stopping at the 1961-1963 subseries where the Department's online offerings began.[59] UWDCC unbound the volumes and scanned their contents to yield high resolution images, and prepared an interface for searching and browsing the volumes. That interface allows visitors to list all volumes in the collection, browse a volume's table of contents, navigate to a chapter, and view the scanned image of any page. (For a roundtable on using the UWDCC's edition of *FRUS* in teaching, see the April 2011 issue of *Passport*.[60]) If the UWDCC edition of *FRUS* had any weakness, it was the quality of the digital text that powered the site's search function, since optical character recognition (OCR) technology could only achieve a certain level of accuracy. As a result of inevitable OCR typos, a user's search for a word or phrase might not return a complete set of results. The UWDCC site, however, proved the utility of an improved digital interface to the series.

In 2007-2008 the Office of the Historian conducted a review of its online publications and set out the following goals for an improved website: (1) To bring the 88 *FRUS* volumes released online to date on its various sites under a single roof, with a capable search engine.(2) To prioritize presenting visitors with the text of the documents, rather than displaying scanned images of pages from the printed volumes, even if this meant slower progress or more costly conversion. Federal accessibility laws require scanned images to be accompanied with transcriptions for use by visitors who rely on screen reader technology to access content. As the official documentary record of U.S. foreign relations, the online *FRUS* series could not be rife with OCR errors. A clean text would benefit all users. (3) To improve navigation within volumes by displaying internal cross-references as hyperlinks, showing footnote text both inline and at the foot of a page, and placing reference aids from a volume's glossaries right beside the text of a document instead of on a separate page. (4) To build a stable foundation for the future. A modern digital format for *FRUS* could be extended, as resources allowed, with new layers of analysis and annotation. If designed correctly, the *FRUS* data would not need to be continuously overhauled as the Office's website underwent inevitable redesigns or server migrations. Persistent, readable URLs would facilitate long-lasting citations. (5) To adopt open, standards-based solutions and avoid proprietary ones—a vital strategy in an uncertain budgetary environment—and to identify means of adding earlier volumes from the series to the new website.

These goals led the Office to adopt a new master digital format for *FRUS* and all of its other article- and book-length content based on the Text Encoding Initiative (TEI).[61] Why TEI and not some other format? As the de facto standard for digital humanities projects, TEI offered a sophisticated, mature, and stable model for preparing texts of any kind for digital

[58] See University of Wisconsin-Madison Libraries, "Foreign Relations of the United States," various dates, https://digital.library.wisc.edu/1711.dl/G5OAT7XT7HRHX84.

[59] Beth Harper and Melissa McLimans, "Preserving the Past and Increasing Access through Digitization," April 4, 2012, Federal Depository Library Program, https://www.fdlp.gov/all-newsletters/community-insights/1281-preservingdigitization.

[60] Vicki Tobias, Richard Hume Werking, Brian Clancy, Robert M. Morrison, and Nicole Phelps, "Using Digitized Documents in the Teaching of The University of Wisconsin's *Foreign Relations of the United States* Series," *Passport,* April 2011, pp. 14-21, https://shafr.org/sites/default/files/Using-Digitized-Documents.pdf.

[61] See Text Encoding Initiative (TEI) Consortium, https://tei-c.org/.

application. The TEI Guidelines describe how to annotate and enrich text with structural and semantic information that can be used for analysis and transformation.[62] TEI's standard vocabulary contains 99% of what a documentary edition of modern documents like *FRUS* needs: divisions for each section of a volume, headings for document titles, datelines, dates, place names, person names, signature blocks, footnotes, cross-references, page breaks, tables, and paragraphs. (The Guidelines' discussion in Chapter 3 of the standard TEI vocabulary for handling names and dates, etc., is illustrative.)[63] For the remaining 1% of the Office's requirements for *FRUS*, TEI is deeply customizable and provides a facility for extending the vocabulary and annotation types to meet a project's specific needs.[64] TEI annotations are expressed in XML, the World Wide Web Consortium's open, non-proprietary, plain-text based standard for encoding documents and data.[65] The Office was attracted to TEI's plain text-based format because it did not require any special software to open or edit. The Office would not have to fear that a company going out of business or changing business models could imperil the Office's core investment in reformatting these publications. The TEI community, through its active mailing list[66] and annual meetings,[67] was welcoming and generous with the Office's inquiries.

Because TEI focuses far more on the structure and content of text than its typography or presentation, the Office's decision to adopt TEI facilitated kinds of scholarly analysis that were not possible in print.[68] TEI turned each *FRUS* volume from bound paper into a database that could be queried and interrogated with continually improving digital tools and not simply displayed for readers online. Adopting a format used widely in the humanities could even allow researchers in a variety of fields to incorporate *FRUS* documents into their databases.

However, the choice of format was only the first of a series of practical questions. How would the Office reformat its existing publications into TEI? What software was capable of making a large TEI corpus browsable and searchable on the web? If the Office could meet its initial goals, could it marshal resources to expand the project to include a larger set of volumes, or even the entire series?

Thanks to insights and contributions from two university partners, the Office was able to answer these questions. First, in 2008, the University of Virginia Press's digital imprint, Rotunda, and the Documents Compass project at the Virginia Foundation for the Humanities shared invaluable insights into the process of digitizing documentary editions, based on

[62] See TEI Consortium, "P5 Guidelines," n.d., https://tei-c.org/guidelines/p5/.

[63] See TEI Consortium, "P5: Guidelines for Electronic Text Encoding and Interchange, Version 4.3.0. Last updated on 31st August 201, revision b472b1ff," [subsequently TEI P5] "3 Elements Available in All TEI Documents," "3.6 Names, Numbers, Dates, Abbreviations, and Addresses," https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CONA.

[64] See TEI P5, "23 Using the TEI," "23.3 Customization," https://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html#MD.

[65] See the World Wide Web Consortium (W3C), "Extensible Markup Language (XML)," 11 October 2016, https://www.w3.org/XML/.

[66] See TEI P5, "Support," "Advice and queries: TEI-L mailing list," https://tei-c.org/support/#tei-l.

[67] See TEI, "Members Meetings," n.d., https://members.tei-c.org/Events/meetings/.

[68] Joseph C. Wicentowski, "history.state.gov: A case study of Digital Humanities in Government," *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1:3 (2011), https://knowledge.uchicago.edu/record/457?ln=en.

their expertise from having launched the American Founding Era Collection.[69] Their experience gave the Office confidence in this model for digitizing and publishing documentary editions online. The Office identified a vendor who could scan and prepare TEI-encoded texts at high quality (minimizing transcription errors), and learned about native XML databases, such as eXist-db—a free and open-source software package that could power a TEI-based website and search engine.[70] The Office adopted eXist-db, which powers history.state.gov to this day. The eXist-db community grew to include many TEI-based projects from around the world, and in recent years these efforts have coalesced around the free, open source TEI Publisher project to build reusable components that individual researchers and organizations small and large can now use to publish their TEI collections more efficiently and effectively.[71]

Second, in 2009 the Office proposed a partnership with the University of Wisconsin Digital Collections Center—the creator of the aforementioned *FRUS* portal—to exchange digitized *FRUS* assets: UWDCC shared its collection of scanned images of the 1861-1960 volumes, and in return, the Office enriched UWDCC's images into clean, high quality, reviewed TEI files and provided these to UWDCC, along with the post-1960 volumes it scanned according to UWDCC's specifications.[72] This partnership allowed the Office to save the cost of scanning hundreds of thousands of pages of printed books (again!) and to direct these resources into preparing a highly accurate digital text edition of every volume.

Thanks to these exchanges and the availability of mature open standards and free open-source software, the essential resources provided by the Department were sufficient to launch the website with the envisioned features and complete the digitization of the printed portion of the *FRUS* series by 2018. The Office has also actively exchanged insights and expertise with peer offices in foreign ministries abroad that produce documentary editions, resulting in a set of best practices for digital documentary editions presented at the 2019 International Conference of Editors of Diplomatic Documents.[73]

Reading FRUS on history.state.gov

The feedback the Office receives from visitors to the website indicates the site is generally intuitive and easy to use. Even so, the site houses more resources than many students and scholars of international and diplomatic history may realize, and a deeper understanding of the considerations underlying its organization and features (both basic and advanced) can help visitors take full advantage of the resources published on the site. Seasoned visitors to the website may opt to jump ahead to the next section on searching *FRUS*, but the details here in this section are certain to earn readers an advanced badge in history.state.gov spelunking.

[69] See University of Virginia Press, "Rotunda," https://rotunda.upress.virginia.edu/, 2004; Virginia Humanities, https://virginiahumanities.org/, 2021; and University of Virginia Press, "American Founding Era Collection," https://www.upress.virginia.edu/founding-era, 2021.

[70] See eXist Solutions, "eXist-db," https://exist-db.org/, 2018.

[71] See tei Publisher, https://teipublisher.com, n.d. The Office's website uses the TEI Publisher libraries to present all of its article- and book-length publications. In addition, version 7.1 of this software, which was released in August 2021, added a digital annotation facility, allowing users to apply annotations (e.g., people, places, and organizations) linked to an extensible set of open authority records to their TEI-encoded documents from a convenient web browser-based interface. The Office recently begun using this tool as part of a workflow for preparing back-of-book indexes.

[72] "General Guidelines for Digitization," University of Wisconsin Digital Collections Center, 2006, https://minds.wisconsin.edu/handle/1793/6731.

[73] "Best practices for digital diplomatic documentary editions," https://diplomatic-documents.org/best-practices/digital-editions/ and "15th International Conference of Editors of Diplomatic Documents," https://diplomatic-documents.org/berlin-2019/.

The history.state.gov landing page contains announcements about recent publications and activities, as well as a site-wide menu and search field. The *FRUS* series is located under the Historical Documents section;[74] this section also houses Status of the Series[75] (a continually updated listing of volumes that are currently under research or in production), as well as text of the Office's 2015 monograph on the history of the *FRUS* series—an invaluable resource for understanding the series in its institutional and political context.[76] The Department History section houses Principal Officers and Chiefs of Mission (a database of ambassadors and senior leaders in the Department), Travels of the President and Secretary of State (a database of their official travel abroad), Visits of Foreign Leaders (a database of official visits of foreign leaders and heads of state to the U.S.), and the Office's newest publication: Administrative Timeline of the Department of State.[77] The Countries section houses essays on the history of U.S. relations with every country in the world, focusing on dates of recognition and key changes in bilateral relationships,[78] and the About section houses the minutes from every meeting of the Advisory Committee on Historical Diplomatic Documentation (HAC) since 1996; additional minutes and HAC-related memoranda can be found in the *FRUS* history section.[79]

Moving into the Historical Documents landing page, visitors will find a listing of presidential administrations. Selecting an administration leads to a listing of the *FRUS* volumes that cover the dates of that administration. This presentation aids researchers as they in navigate the chronological breadth of the series, but grouping *FRUS* volumes by presidential administration is actually somewhat anachronistic. The series has only been organized into presidential sub-series since the Johnson administration. Prior to Johnson, *FRUS* was grouped into annual, triennial, or special topical sub-series. Thus, volumes from these periods may contain documents from more than one presidential administration. To ensure that readers see an accurate listing of volumes, the website uses each volume's earliest and latest documents' dates to determine which administration(s) it should be listed under. For example, the Lincoln administration listing includes an 1894 *FRUS* volume because it contains three documents from 1864.[80] Listings for administrations that the Office is actively working on show the titles of volumes that have not yet been published; for example, see the Reagan, Bush, and Clinton administration listings. And, as the Kennedy administration page shows, microfiche supplements that have not yet been digitized are also listed. To aid in administration-specific research, a link appears at the top of each administration listing that initiates a keyword search for documents from the dates of the administration.

Selecting a volume takes the reader to the volume's landing page, which presents the names of editors responsible for compiling the volume, its year of publication, and a table of contents. The right sidebar of this page contains a "Search inside this volume" field for performing keyword searches on the volume itself, links to download eBook (and, when available, PDF) editions of the publication, and a listing of subjects covered in the volume, each of which is linked to the

[74] See U.S. Department of State, Office of the Historian [henceforth DOS-OH], "Historical Documents," https://history.state.gov/historicaldocuments, n.d.

[75] See DOS-OH, "Status of the Foreign Relations of the United States Series," https://history.state.gov/historicaldocuments/status-of-the-series, n.d.

[76] William B. McAllister, Joshua Botts, Peter Cozzens, and Aaron W. Marrs, *Toward "Thorough, Accurate, and Reliable": A History of the* Foreign Relations of the United States *Series*, DOS-OH, 2015, https://history.state.gov/historicaldocuments/frus-history.

[77] See DOS-OH, "Department History," https://history.state.gov/departmenthistory, n.d.

[78] See DOS-OH, "Countries," https://history.state.gov/countries, n.d.

[79] See DOS-OH, "About Us," https://history.state.gov/about, n.d., and "Documents," https://history.state.gov/historicaldocuments/frus-history/documents, n.d.

[80] See DOS-OH, "Abraham Lincoln Administration (1861–1865), https://history.state.gov/historicaldocuments/lincoln, n.d.

Office's taxonomy of subjects in the history of U.S. foreign relations.[81] Selecting any of these subjects leads to listings of other volumes that have also been associated with the same topic. (For a list of all eBooks available for download, see the eBooks page.[82])

Selecting a chapter or subchapter brings the reader to a listing of documents contained within. The document list shows each document's heading, dateline, and source note. (For example, see the document list for Chapter 1 in the Reagan Soviet volume covering March 1985-October 1986.[83]) These document lists can be a quick way to scan the content of a chapter.

Selecting a document displays the text of the entire document on one page, including its footnotes. Readers can display the text of any footnote by hovering or tapping on its reference number. Readers can also jump to the section at the bottom of the page containing all footnotes by clicking on any footnote reference link. From the footnotes section, clicking on the "arrow" icon at the end of any footnote jumps back to the place in the text where the footnote appeared. Footnotes often contain cross references to other documents in a volume or to other volumes in the series; the website presents these cross references as links for quick navigation. (See, for example, the 9 cross references in the footnotes to Document 27 of the Carter Soviet Union volume.[84])

For volumes whose front matter contains a "List of Persons" and/or "List of Terms and Abbreviations," the document's right sidebar also contains a listing of persons and terms that appear in the document. Hovering over any of these entries reveals a description of the item. (See, for example, Document 12 in China, 1973-1976.[85]) The right sidebar may also, in some cases, offer a downloadable PDF containing the scanned images of the original archival document. These are only available in microfiche supplements and certain electronic-only volumes from the Nixon-Ford series. (See, for example Document 7 from Volume E-7, Documents on South Asia.[86])

To navigate forward and backward through the documents and divisions in a volume, use the "<" and ">" icons that float in the middle of the left and right edges of the page. This may be more convenient than navigating back to the landing page or document list just to advance to the next document.

The design of the website emphasizes document-based views but also provides affordances for browsing volumes by printed page. When viewing a document, any page breaks appear as right-aligned links, such as "[Page 320]". Selecting this link leads to a view of a screen resolution view of a scanned image of the appropriate printed page. To navigate forward and backward through the page images, use the "<" or ">" icons that float in the middle of the left and right edges of the page.

---

[81] See DOS-OH, "Tags," https://history.state.gov/tags, n.d.

[82] See DOS-OH, "Ebooks," https://history.state.gov/historicaldocuments/ebooks, n.d.

[83] See DOS-OH, *Foreign Relations of The United States (FRUS), 1981–1988, Volume V, Soviet Union, March 1985–October 1986*, "March 1985–July 1985, 'Now we have to begin everything anew': Gorbachev's Debut," https://history.state.gov/historicaldocuments/frus1981-88v05/ch1.

[84] See DOS-OH, *FRUS, 1977–1980, Volume VI, Soviet Union*, "27. Memorandum of Conversation," Washington, April 12, 1977, 4-4:40 p.m., https://history.state.gov/historicaldocuments/frus1977-80v06/d27.

[85] See DOS-OH, *FRUS, 1969-1976, Volume XVIII, China, 1973-1976*, "12. Memorandum of Conversation," Beijing, February 17-18, 1973, 11:30 p.m.-1:20 a.m., https://history.state.gov/historicaldocuments/frus1969-76v18/d12.

[86] See DOS-OH, *FRUS, 1969-1976, Volume E-7, Documents on South Asia, 1969-1972*, "1. Memorandum Prepared by the National Security Council Staff for President Nixon," Washington, undated, https://history.state.gov/historicaldocuments/frus1969-76ve07/d1.

Recently digitized microfiche supplements even present two parallel runs of page links: one run leads to the image of the scanned microfiche, and the other run leads to the image of the typeset edition. (See, for example, Document 2 in the Cuban Missile Crisis supplement.[87])

Eagle-eyed readers will notice that the URL of every *FRUS* resource follows a regular pattern. By understanding this pattern, readers will be able to deduce the URL for another volume, document, or page. For example, examine the URL:

https://history.state.gov/historicaldocuments/frus1947v07/d329

The first slug (or section of a URL), "historicaldocuments," identifies the Historical Documents section of the site where all *FRUS* resources reside. The next slug, "frus1947v07" is the identifier for *Foreign Relations*, 1947, Volume VII. The final slug, "d329," is the identifier for document 329 in that volume. Changing the end of the URL to "d330" and hitting the return key will cause the website to advance to document 330.[88] Changing the end of the volume ID from "v07" to "v08" will yield *Foreign Relations,* 1947, Volume VIII.[89] Page numbers take the form "pg_1," so changing the end of the previous URL to "pg_386" will advance directly to page 386 of volume VIII.[90] Visitors can always navigate to a resource through the website's interface, but understanding the structure of its URLs can save time.

The four views described above—the table of contents, document list, document-centric text view, and page-centric image view of *FRUS* volumes—are simply different transformations of the same TEI source data. The eBook edition of *FRUS* publications is a fifth transformation. This flexibility is one of the key benefits of TEI's media-neutral representation of text and its use of XML technologies widely supported by open-source software. This inherent flexibility enables the transformation of a source TEI document into many formats. It also enables visitors to perform sophisticated searches across the entire *FRUS* series.

Searching FRUS on history.state.gov

The search engine on history.state.gov is a powerful tool for locating documents for research.[91] The best way to learn how to exploit the capabilities of the search engine is to read the Search Tips page, which provides a thorough walk through of the search engine's options and features, and then experiment with your own queries.[92] The website's search engine provides options that go beyond the typical "Google-style" keyword search; visitors can perform phrase, Boolean, wildcard, and proximity searches and can use filters to limit the results to specific criteria. The Search Tips article provides live examples

[87] See DOS-OH, *FRUS, 1961-1963, American Republics; Cuba 1961-1962; Cuban Missile Crisis and Aftermath, Volumes X/XI/XII, Microfiche Supplement*, "2. Memorandum from Schlesinger to Kennedy, March 3," March 3, 1961, "Subject: The Crisis in Bolivia," https://history.state.gov/historicaldocuments/frus1961-63v10-12mSupp/d2.

[88] See DOS-OH, *FRUS, 1947, The Far East: China, Volume VII*, 893.00/3-147: Telegram, "The Minister-Counselor of Embassy in China (Butterworth) to the Secretary of State," Nanking, March 1, 1947—11 p.m. [Received March 1—11:50 a.m.], https://history.state.gov/historicaldocuments/frus1947v07/d330.

[89] See DOS-OH, *FRUS, 1947, The American Republics, Volume VIII,* 124.241/6-1347: Telegram, "The Secreatry of State to the Embassy in Bolivia," Washington, June 23, 1947—7 p.m., https://history.state.gov/historicaldocuments/frus1947v08/d330.

[90] See DOS-OH, *Foreign Relations of the United States, 1947, The American Republics, Volume VIII*, p. 386 [scanned image], https://history.state.gov/historicaldocuments/frus1947v08/pg_386.

[91] See DOS-OH, "Search," https://history.state.gov/search, n.d.

[92] See DOS-OH, "Search Tips," https://history.state.gov/search/tips, n.d.

for each of these options showing how they can be used. Rather than repeat the information in that resource, this section briefly introduces what the Office means by describing the *FRUS* digital edition as a "full-text searchable" resource and clarifies what logic is used when visitors perform a search. In an age when many services' search algorithms operate as opaque "black boxes," visitors to history.state.gov deserve full transparency about the behavior of its search engine.

In the context of the Office's website, "full-text searchable" means that the complete text of every digitized *FRUS* volume—from cover to cover—has been ingested into the search engine, and this complete text is what is searched when visitors perform a "keyword search." When the website ingests a *FRUS* document (which happens when an Office historian uploads a new or changed volume), the search engine compiles a "search index," which it uses to quickly locate documents when visitors perform searches. This index contains (1) every word in the document (chopping the text into individual words wherever a space or punctuation character appears, ignoring case), (2) a count of how many times each word appears in the document, and (3) a count of how many words total appear in the document. This information about each document is added to the site-wide search index. When a visitor performs a keyword search, the search engine simply looks at this index of words and returns the documents that contain them all. No other factors contribute to the behavior of the search engine when carrying out a keyword search.

The keywords in a visitor's search are treated literally. No automatic expansion or "stemming" is applied. For example, a keyword search for "modern" will return documents that contain this form of the word, but it will not return documents that *only contain other forms* of the word, such as "modernization." To construct a keyword search that finds two or more forms of a word or multiple synonymous terms, place the all-caps "OR" between the terms. For some variants, using an asterisk ("*") can help to match all words beginning with a common prefix. For example, "modern*" would match instances of the base prefix, "modern," as well as "modernity" and "modernization." The Search Tips page covers these and other powerful options and provides live examples of each. All to say that unless visitors use such options, no automatic options are applied. This design ensures that the user is squarely in the driver's seat when performing searches.

There is one significant limitation in the current operation of the search engine. It is the result of the fact that, as noted above, the search engine is case-insensitive and ignores punctuation. For example, when a visitor performs a search for the acronym "GOA" (Government of Australia), the search engine will include results about "Goa," India. And a phrase search for "S/S" (Executive Secretariat) will return results containing "U.S.S.R." This is because the search engine both completely ignores punctuation and treats it as a word boundary; "S/S" is split into two words ("s" and "s"), and "U.S.S.R." is split into four. The Office is tracking these problems and plans to add case- and punctuation-sensitive search capabilities to the website.[93] This functionality is high on the Office's list of development priorities. In the meantime, visitors searching for acronyms containing punctuation may encounter false positives. The best approach to limit false positives is to wrap the term in double-quotes forces the search engine to treat the letters within as a phrase.

Visitors may be concerned not just about the algorithms search engines use for finding documents, but also about how search engines rank and order results. The Office's website offers two methods for controlling how search results are sorted. These options are straightforward and configurable by visitors. After the search engine finds all of the documents that match the keyword(s), it offers two methods for sorting the results: "relevance" and "dates." Visitors can toggle between the two options via the "Sort by" dropdown menu at the top-right corner of the search results.

The first sorting option, "relevance," is useful for quickly accessing those documents that best match the keyword(s) supplied by the user. To determine which documents should appear first, the search engine considers how many times the keyword(s) appears in the document, as well as the length of the documents. Thus, if a user searches for the term "ideology," the search engine would first return a short document that mentions the word 10 times, and second a much longer document that mentions it 11 times. Even though the second document contains one additional mention, the shorter

[93] See Joe Wicentowski, "Assemble list of quirks with our full text search engine (esp. punctuation) #255," hsg-shell (history.state.gov redesign) Github repository, 26 October 2016, https://github.com/HistoryAtState/hsg-shell/issues/255.

document is more likely to focus on "ideology," whereas the longer document only covers it in passing. Thus, the search engine considers this document to be more "relevant" and ranks it higher in the listing of results.

The second sorting method, "dates," simply sorts documents from oldest to newest (or the reverse), based on the date of the document. Most *FRUS* documents are explicitly dated, but some documents are undated or imprecisely dated. To ensure every document could be searched for and sorted correctly by date, the Office undertook a major project to review every document and identify a Gregorian calendar date or date range for it. The project identified dates from many calendar systems and entailed extensive research to resolve missing or imprecise dates. Given the range of methods used to date documents, the Office developed a taxonomy for indicating which of the 25+ methods was used.[94] The search engine does not currently display the method, but this information is captured in the master TEI files for each volume to aid in future queries. (For more details on obtaining these files, see "Beyond history.state.gov" below.) The search engine does reveal the TEI-encoded date in each document's search results, along with the document's title, recorded date and location, and a "keyword in context" display of a handful selected instances of the keywords and surrounding text.

Keyword search is a powerful tool for locating documents, but with over 310,000 documents in the corpus, this method alone is likely to yield an overwhelming number of results. To help researchers further refine their searches and craft more specific queries, the search engine offers date and volume filters. To employ these filters, select "Historical Documents" as the scope of the search in the "Sections" filter in the left sidebar, and immediately the two new filters will appear beneath it. The date filter allows visitors to provide start and/or end dates for their search. The date filter works even without keywords, returning all documents from the date range. Dates and date ranges can be expressed with any degree of specificity from the year down to the minute. A date range search for 1600-1860 returns the 467 documents that predate the series.[95] Fine-grained date ranges are also possible: a visitor could search for documents between President Richard Nixon's announcement of his resignation on August 8, 1974, at 9 pm and the delivery of his resignation letter to Secretary of State Henry Kissinger the following morning at 11:35 am.[96] The search engine assumes the U.S. Eastern time zone. The volume filter allows visitors to include or exclude certain volumes, providing a useful proxy for topic or chronology.

To provide visitors with even more powerful tools to narrow searches, the Office is investigating adding filters for the rich metadata found in each *FRUS* document's heading and source note, such as sender and recipient, document type, provenance (i.e., source repository), original classification, and people, places, and organizations mentioned. The largest obstacle to expanding the selection of filters is the fact that the contemporary annotation practices for capturing this information varied across and within epochs in the series. Human readers can cope with such variation, but in its current form, this information is not sufficiently regularized to be 'machine readable.' Just as the date filter and sorting options required a major effort to identify dates in documents and make them uniform and machine readable, every additional 'facet' or 'dimension' of metadata that the website could expose will require considerable investment of effort and resources. The Office continues to investigate possibilities for enhancements in these areas.

As the links in the preceding paragraphs' footnotes indicate, search results from the Office's website can be shared by copying the search page's URL after crafting a search. All search parameters (i.e., keywords and filters) are captured in the URL. If one visitor emails a search URL to another user, the recipient will see identical results. The only reason that search

[94] See Amanda Ross with Joe Wicentowski and Virginia Kinniburgh, frus.odd schema [XML code], frus (Foreign Relations of the United States - TEI XML source files), Github repository, 28 June 2020, https://github.com/HistoryAtState/frus/blob/master/schema/frus.odd#L1515-L1655.

[95] See DOS-OH, "Search," https://history.state.gov/search?q=&within=documents&start-date=1600&end-date=1860&sort-by=date-asc, n.d.

[96] See DOS-OH, "Search," https://history.state.gov/search?q=&within=documents&start-date=1974-08-09&end-date=1974-08-09&end-time=11:35&sort-by=date-asc, n.d.

results for a saved URL might change would be the addition of new sources into the website. So, while search results are not permanent, the queries can be saved and submitted again.

Beyond history.state.gov

The Office of the Historian recognizes that researchers may conceive of types of analysis that the history.state.gov website does not facilitate. To allow researchers to perform more advanced analysis, the Office maintains complete, up-to-date repositories of the website's source data via its "HistoryAtState" organization on GitHub.[97] GitHub is a free website that many researchers and developers who are working on humanities, government data, and open-source software projects use to publish their data and/or source code.

The *FRUS* source files can be found in the "frus" repository within the HistoryAtState organization on GitHub.[98] Every new *FRUS* volume published to history.state.gov is simultaneously uploaded to GitHub, and every edit is logged and timestamped, with descriptive comments and views showing the precise changes. Visitors can follow the daily work of Office historians publishing and maintaining these publications by looking at each repository's list of Commits, or changes; for example, see commits to the *FRUS* repository.[99] By establishing a GitHub account, readers can even report problems (e.g., typos, bugs) and suggest fixes directly to Office historians, who will evaluate them. The same applies to every publication, dataset, and piece of code found on the history.state.gov website.

In fact, researchers can download and install a complete, live copy of the history.state.gov website on their personal computers, following the same directions Office historians use.[100] Doing so could offer a practical way to perform research when working without a stable internet connection. Since all of the Office's publications are encoded in TEI and XML, any TEI- or XML-aware technology can ingest the source materials as is or transform them into other formats. Free, open-source applications like the aforementioned eXist-db and TEI Publisher allow users to load their own documents and documents from other sources for analysis and publication. Programming languages for exploiting TEI and XML sources, such as XPath and XQuery, are readily accessible to historians and researchers in other humanities fields who may not have a background in computer science.[101] In fact, every page and function of the history.state.gov website is written in XQuery. For scholars who have advanced text mining skills—or collaborate with those who do—the Office's sources are natural targets for the application of natural language processing, text modeling, and other computational analysis techniques.

The Office hopes that posting *FRUS* and all of its publications' source data will enable and attract scholars to perform new kinds of research with these materials. An example of the new kind of scholarship enabled by access to the *FRUS* TEI is Eun Seo Jo's research. Her 2020 dissertation, "*Foreign Relations of the United States* Series, 1860-1980: A Study in New Archival History," is a critical investigation of the applicability of computational linguistics, natural language processing, machine

[97] See DOS-OH, Office of the Historian, U.S. Department of State Github, https://github.com/HistoryAtState.

[98] See DOS-OH, Github, frus Repository, https://github.com/HistoryAtState/frus.

[99] See DOS-OH, Github, frus Repository, "Master Commits," https://github.com/HistoryAtState/frus/commits/master.

[100] See Joe Wicentowski, "Set up a history.state.gov Development Environment," hsg-project Github repository, 23 April 2021, https://github.com/HistoryAtState/hsg-project/wiki/setup.

[101] For example, see *The Programming Historian* (ISSN 2397-2068) at https://programminghistorian.org, Elisa Beshero-Bondar's course and workshop materials at https://newtfire.org/courses/, and Clifford B. Anderson and Joseph C. Wicentowski, *XQuery for Humanists* (College Station: Texas A&M University Press, 2020), https://xquery.forhumanists.org/.

learning, and artificial intelligence to archival sources for the purposes of historical research.[102] Jo argues that the selectivity inherent in textual archives requires a new theoretical framework, called "new archival history," that recognizes that archives embody changing historical processes. The availability of the *FRUS* series as an open, full-text collection made it a suitable locus for her investigations, which include topics as disparate as the impact of National Security Advisor Walt Rostow's notion of modernization theory in foreign policy and changes in the language of diplomacy during the Cold War.[103]

Conclusion

Thanks to invaluable university partnerships, the Office of the Historian has established a modern foundation for the *FRUS* digital edition on the basis of open standards and open-source software, with a complete collection of the printed volumes now available in a full-text, searchable format and as downloadable eBooks on history.state.gov and as open government data via its HistoryAtState repositories on GitHub. At the projected pace of digitizing one microfiche supplement per year, the Office's goal to incorporate these least accessible of all volumes into the *FRUS* digital collection could be completed within a decade or so. The Office continues to evaluate and improve the utility and quality of all of the Office's publications for students, scholars, and the general public. Readers' feedback is welcomed at history@state.gov.

[102] Eun Seo Jo, "Foreign relations of the United States series, 1860-1980: a study in new archival history," Ph.D. Thesis, Department of History, Stanford University, 2020, https://purl.stanford.edu/km610md5945.

[103] This example is used for illustrative purposes only. The Office of the Historian does not endorse the research or conclusions of private scholars.