Lessons Learned in Born-Digital Preservation

Miguel Beltran

As more government documents are created in digital mediums, it is increasingly important that agencies could preserve and make them available to the public. This article discusses one group of government documents related to the war in Afghanistan and the landscape that would potentially preserve them. Based on the current conditions, there is a possibility that these documents and those of a similar nature may be overlooked and lost to future generations.

In 2019, a series of articles published by *The Washington Post* provided an outlook of the war in Afghanistan mostly unknown to the public entitled "At War with the Truth." Citing government documents, they [the documents] reveal, that despite the oversight of three presidential administrations, billions of dollars spent, and thousands of lives lost, the government failed to tell the truth about the conflict through its first eighteen years.[1] Drawn primarily from the Lessons Learned Reports produced by the Special Inspector General for Afghanistan Reconstruction (S.I.G.A.R) and various other government documents, a story unfolds of inconsistent strategy amplified by intentionally misinforming the public about the war's progress.[2] These documents were largely unclassified until *The Washington Post* sought to obtain them through a Freedom of Information Act request prompting the government to then restrict some documents.[3] A move that was overturned following a nearly three-year legal battle.[4]

These documents are now available online for anyone to peruse.[5] Hosted by a .mil web address, the Lessons Learned Reports in their PDF format may stand the test of time and be preserved for researchers and the public alike. Reading them empowers the people to insist that their government representatives do not repeat the mistakes of the past. But will they be preserved? If they are preserved, how will the documents be discovered in the future? How will they be authenticated as legitimate government documents? These important questions determine whether lessons are learned from these reports. The current transition from the production of tangible government publications to primarily born-digital content brings new challenges in addition to technologies and formats. Clear strategies and widespread collaboration are necessary to preserve government information on these mediums.

Without discussing the Government Publishing Office (GPO), no conversation about preserving government information is complete. Its mission is to "publish trusted information for the Federal Government to the American people."[6] In addition to the printing and distribution of tangible government publications, the GPO produces and distributes government information for all three branches while providing permanent access to this information via the Federal Depository Library Program (FDLP) and the website govinfo.gov.[7] Historically, the FDLP was the workhorse of government preservation efforts. Documents in various physical formats were distributed to nationwide participating libraries, which then stored them for public access.[8] The increase in born-digital government information is shifting preservation strategies in government organizations. The GPO defines preservation as "initiatives, programs, and processes designed to maintain useful access to information assets, serving the information needs of both present and future generations."[9]

Note that this definition describes preservation as an ongoing process requiring the collaboration of many people and programs regardless of the medium in which the information asset exists. If documents and websites like the Lessons Learned Reports and the S.I.G.A.R website are to be preserved for both present and future generations, legal mandates by the federal government paired with sustainable funding are necessary. The GPO is the main government organization responsible for preservation and is currently a leading example to the world on how to do it.

Another major resource the GPO uses to fulfill its missions is the website govinfo.gov. This website is an online repository

that provides "free public access to official publications from all three branches of the federal government."[10] A one-stop shop for all major publications ranging from bills and statutes to judicial and regulatory information, govinfo.gov is the one website anyone interested in government information should know. As stated earlier, govinfo.gov is more than just a website. It is an online repository utilizing metadata-powered search engines, content management, and digital preservation compliant with ISO 16363.[11] This standard certifies its holders as a trustworthy digital repository. "As of July 2020, GPO is currently the only organization in the world to hold ISO 16363:2012 certification."[12] Govinfo.gov may be the best example of how to preserve online materials in the world.

As we all know, not everything on the internet is true. How can the GPO reassure citizens that the content on govinfo.gov is authentic and true? "Because many of the official publications GPO provides online are in PDF format, GPO uses digital signature technology to provide evidence of authenticity and integrity and safeguard against unauthorized changes to these files."[13] This feature enables the use of digital versions of government documents for legal purposes. As an extension of the content already uploaded onto govinfo.gov, the GPO has invited members of the FDLP to help grow the digital national collection as digital preservation stewards, digital access partners, and digital content contributors.[14] These commitments, if fully realized, enact a plan to digitize the entire FDLP collection through the use of PURLS and ingesting content onto govinfo.gov. It is worth noting that many items are also discoverable online via the Catalog of US Government Publications (CGP).[15] This website is also maintained by the GPO and acts as an index for federal publications as well as a finding tool for historical and current publications. Direct links to documents are sometimes available.

We can assume that most of these digitization efforts will produce PDF images of documents and will therefore have the option to be authenticated when downloaded from govinfo.gov. However, not all government information is a document that can easily be transferred into a PDF format. Some government information is posted on websites without official publication in the form of a document. Since 2011, the GPO has partnered with the Internet Archive with the goal to "provide permanent public access to Federal Agency Web content, the Federal Depository Library Program harvests selected U.S. Government Web sites in their entirety."[16] The program is called the FDLP Web Archive. The key limitation of this program is that only selected government websites will be included for preservation. More importantly, information on harvested websites that is not published as a PDF currently cannot be authenticated.

Through the GPO, the government has taken commendable steps to ensure that born-digital government information and documents are preserved for future generations. There are, however, holes that some materials can slip through. The earlier example of the Lessons Learned Reports is one of these. While the reports themselves are, in fact, PDFs, which would allow them to be authenticated if they were housed on govinfo.gov. They, unfortunately, reside on a government website that is not harvested by the FDLP Web Archive. Now that US forces have withdrawn from Afghanistan, it stands to reason that S.I.G.A.R will be decommissioned if it has not been already. What will happen to the website if the program and its lead official no longer exist in the coming years? Will the documents created because of S.I.G.A.R be made accessible through other means? How will people come to discover those documents if they are unaware of their existence? Something that is accessible without discoverability is nearly unusable. The threat of these and other important government documents disappearing from public access increases as more of them are born digital.

While everything produced by the federal government cannot be captured at this time, the efforts of the GPO to preserve born-digital government information are commendable. Programs improve over time if adequate funding is provided and can expand appropriately. An example of one such program is the National Digital Information Infrastructure and Preservation Program (NDIIPP), formerly led by the Library of Congress. Although the NDIIPP is no longer an active program, "its success is evident in the diverse and mature digital preservation community that is now thriving in the United States."[17] This program began by focusing on three areas: building a network of partners; developing a technical infrastructure of tools and services; and capturing, preserving, and making available significant digital content.18 All three of these focal points can be observed in the GPO's programs, initiatives, and technologies.

There must be widespread interagency collaboration to have the best results in preserving born-digital government documents and information. The current environment for dissemination of government publications flows through the GPO. "Federal agencies are required by statutory mandate to provide Federal publications to the Federal Depository Library Program (FDLP) and Cataloging & Indexing Program (44 U.S.C. §§ 1710, 1902-1903)."[19] There is an inherent limitation in the definition of government publications that excludes some types of born-digital government information. "'Government publication' . . . means informational matter which is published as an individual document at Government expense, or as required by law."[20] PDF documents fit neatly into this definition hence the

emphasis undertaken by the GPO to authenticate them. However, by definition, websites, audio recordings, video, and all other digital mediums are not required to be preserved. "Congress should establish a collaborative interagency process, and designate a lead agency or interagency organization, to develop and implement a government-wide strategy for managing the lifecycle of digital government information."[21] This may require expanding Title 44 of the United States Code or creating additional legislation to include new technologies, such as those used to produce born-digital content, or both.

This brief exploration into the preservation of born-digital government documents and information is just the tip of the iceberg regarding the future of preservation. We march towards a time when tangible mediums are rarely created, and most government information is born-digital. In this new environment, it may become increasingly difficult for the GPO to fund all of its preservation programs.

> Only about 12 percent of GPO's funding is appropriated directly to the Agency to cover the cost of congressional work, the Federal Depository Library Program, and supporting distribution programs. The rest of GPO's revenue comes from reimbursements by customer agencies for work performed or sales of publications to the public.[22]

It was, after all, a cessation of funds that ended the NDIIPP. Creating laws that mandate preserving born-digital government information and determining responsible agencies to oversee the process is the only way to ensure their transmission to future generations.

"These publications document the fundamental rights of the public, the actions of Federal officials in all three branches of our government, and the characteristics of our national experience."[23] It appears that the Lessons Learned Reports are, in fact, government publications and should have been submitted to the GPO for dissemination. They are not, however, easily discoverable on govinfo.gov or in the CGP. As the website which houses them ages and maintenance decreases, it is possible that these documents and the lessons they contain will be lost to public access and discoverability: the title of this group of documents is ironic. The medium they have been published in and the strategies for preserving them may indeed demonstrate a lack of lessons learned. While the Internet Archive may harvest these webpages apart from the FDLP Web Archive, that would be haphazard preservation. There is no guarantee that the Internet Archive will capture the PDF documents. In fact, there is no guarantee that the Internet Archive will survive at all. In the hundreds of years this democracy has existed, there have been many attempts to find and preserve the documents produced by our government. Our collective responsibility is to ensure that they survive despite changing technology. Failure to do so can, as in the case of Afghanistan, cost lives. Surely, we all can agree that it is something worth preserving.

Miguel Beltran (miguel.beltran.jr@gmail.com) is a Master of Library and Information Science graduate from the University of Illinois at Urban-Champaign School of Information Sciences. This paper was written for IS 594—Government Information, Spring 2023, Professor Dominique Hallett.

Notes

1. Craig Whitlock, "At War With Truth," *Washington Post*, December 9, 2019. https://tinyurl.com/bds9ebv7.
2. Whitlock, "At War With Truth."
3. Whitlock, "At War With Truth."
4. Whitlock, "At War With Truth."
5. "Lessons Learned Reports," Special Inspector General for Afghanistan Reconstruction, https://tinyurl.com/t6cz3saw.
6. "Mission, Vision, and Values," Government Publishing Office, https://www.gpo.gov/who-we-are/our-agency/mission-vision-and-values.
7. "Mission, Vision, and Values."
8. "FDLP Basics," Federal Depository Library Program, https://fdlp.gov/basics.
9. "Preservation at GPO," Federal Depository Library Program, https://fdlp.gov/preservation/preservation-at-gpo.
10. "About Us," GovInfo, https://www.govinfo.gov/about.
11. "About Us."
12. "Trusted Digital Repository ISO 16363:2012 Audit and Certification," Federal Depository Library Program, https://fdlp.gov/preservation/trusted-digital-repository-ISO-16363-2012.
13. "Authentication," GovInfo, https://www.govinfo.gov/about/authentication.
14. "The National Collection of U.S. Government Public Information," Federal Depository Library Program, https://fdlp.gov/about-the-fdlp/the-national-collection.
15. "The National Collection of U.S. Government Public Information."
16. "Federal Depository Library Program Web Archive," Archive It, https://archive-it.org/home/FDLPwebarchive.

17. "Digital Preservation," Library of Congress, Digital Preservation, https://www.digitalpreservation.gov/.
18. "Program Background," Library of Congress, Digital Preservation, https://www.digitalpreservation.gov/about/background.html.
19. "Dissemination Program," Government Publishing Office, https://www.gpo.gov/how-to-work-with-us/agency/services-for-agencies/dissemination-program.
20. 44 U.S.C. § 1901 (2021), https://www.govinfo.gov/content/pkg/USCODE-2008-title44/html/USCODE-2008-title44.htm.
21. National Academy of Public Information, *Rebooting the Government Printing Office: Keeping America Informed in the Digital Age*, January 2013, p. 3, https://tinyurl.com/yp6bkud5.
22. US Government Publishing Office, "America Informed: Strategic Plan 2023–2027," https://www.gpo.gov/docs/default-source/mission-vision-and-goals-pdfs/gpo_strategicplan_fy23-27.pdf.
23. "Preservation at GPO."