Liladhar R. Pendse

Collaborating to create the At-Risk Afghanistan Web Archive (ARAWA)

A project at the University of California-Berkeley Library

Twenty years later, as the U.S. "Afghanistan Project" concluded in 2021 with the U.S. troop withdrawal and civilian evacuation, it became clear that despite our good intentions to develop and foster a democratic state in Afghanistan, it was only a partial success.[1,2,3] Despite the corruption within the ranks of certain Afghan officials and the frustrating outcomes for the planned projects, a semblance of functioning civil society had emerged across several urban centers in Afghanistan. The levels of corruption, ineptitude, and missed chances have been documented in the series of reports with the title "What We Need to Learn: Lessons from Twenty Years of Afghanistan Reconstruction," prepared by the Special Inspector General for Afghanistan Reconstruction.[4] The central government in Kabul had a web presence through several departmental websites and those of the regional, provincial governments. Besides governmental websites, several educational institutions of higher learning, artists, social activists, and nongovernmental organizations (NGOs) described themselves, expressed their opinions, reported policy decisions, and communicated other information through their web presence on official websites and social media sites.

The rapid takeover of Afghanistan by the Taliban and their arrival on August 15, 2021, at the Presidential Palace in Kabul symbolized how Afghanistan will be governed. It also meant that the Taliban would appoint new ministers and implement new policies that will replace the existing governmental websites. Also, the change indicated the way civil society would function. The activists, NGOs, and others with websites and social media presence, might be forced to take these websites down or delete them. The "disappearance" of these websites implied leaving lacunae in the reconstruction of an evolving society in Afghanistan. The rich substrate of differing opinions these websites represented was at risk of being lost forever. Thus, Liladhar R. Pendse decided to act in a timely fashion and began crawling some of the obvious websites that he thought were bound to change. However, the project could not have been successful if it were not for collaboration from faculty members and students who are Afghanistan specialists and are fully versed in the cultures and languages of Afghanistan.

Liladhar R. Pendse is librarian for East European, Eurasian, and Latin American Studies, University of California-Berkeley, email: lpendse@library.berkeley.edu

Background on web archiving

What is an archive, printing, and technologies has been examined by philosophers Jacques Derrida and Walter Benjamin in their respective works, "Archive Fever: A Freudian Impression" and "The Work of Art in the Age of Its Technological Reproducibility."[5] While the theoretical works of these individual authors and philosophers can be perceived as fundamental to our understanding of the concept of what an archive is, the web archive strives to reproduce and preserve a "web document," which can be lost if the site goes offline or gets modified or erased. Similar excursion about the archival nature of electronic records, whether these are texts or tweets, have been investigated by Luciana Duranti in her works such as *Preservation of the Integrity of Electronic Records,*[6] "Archives in a Eigital Society = Les Archives Dans Une Société Numérique,"[7] and *Trusting Records in the Cloud.*[8]

Web archiving was initially examined in great depth by Julien Masanès in his 2006 work "Web Archiving."[9] Niels Brügger and Ralph Schroeder have provided a methodological framework in their 2017 book *The Web as History: Using Web Archives to Understand the Past and the Present.*[10] While digital archives can be construed as the repositories for historical exercises and preservation of "memories," the question of why ephemera should be preserved is open to interpretation, mainly due to the complicated nature of social media-based proclamation, exclamations, or utterances.

On the one hand, these exclamations or utterances can serve as a meaningful way to understand social issues and problems. On the other hand, one can question the authenticity and integrity of such proclamations. Lastly, the questions about these utterances being weaponized have been examined by P. W. Singer and Emerson T. Brooking in *LikeWar: The Weaponization of Social Media.*[11] The recent article, "Inevitable Weaponization of App Data is Here: A Substack Publication Used Location Data from Grindr to Out a Priest Without Their Consent," by Joseph Cox on the popular investigative journalism site Vice News, allows us to postulate that "these utterances" can have evidentiary values and can lead to long-lasting consequences just as the archives of analog documents.[12] Recently, the use of web archiving techniques to preserve cultural and endangered websites for research purposes has been examined by several authors.[13, 14, 15]

Methodology and issues

At the University of California (UC)-Berkeley, out project had enthusiastic supporters in the library administration. First and foremost, our Senior Associate University Librarian Elizabeth (Beth) Dupuis, Associate University Librarian for Scholarly Resources Jo Anne Newyear-Ramirez, and Associate University Librarian for Digital Initiatives and Information Technology Salwa Ismail were keen on helping out with necessary administrative processes, as they understood the sense of urgency. The At-Risk Afghanistan Web Archive (ARAWA) tries to preserve the recent past for future researchers and scholars.[16]

As the librarian for the East European and Central Asian Collections, I collect electronic and print materials from the region and provide research assistance to a diverse student body, faculty, and visiting scholars. Afghanistan sits at the crossroads of Central Asia, South Asia, and the Middle East. It is a multiethnic country that is intimately connected to the histories of both Central Asia and South Asia. At UC-Berkeley, the curator for South Asian Studies is responsible for the collection development for Afghanistan. Also, it was important from the onset to invite faculty members who were familiar with the culture and history of Afghanistan. Faculty members whose research focuses on Afghanistan can help in the processes of evaluating the contents of one site over the

other. Furthermore, they can help curators in their choices of websites that are bound to be either taken down or changed significantly.

When the first website was crawled, I invited Professor Mariam Ghani of Bennington College, Professor Shah Mahmoud Hanifi of James Madison University, and Sherine Ebadi, a doctoral student at UC-Berkeley's Geography Department to participate. In addition, I reached out to engage Adnan Malik, our curator and cataloguer for South Asian Studies. These individuals were the initial core group of advisors to the project.

Besides these collaborators, Professor Robert Crews of Stanford University and Professors Shah Wali Ahmadi and Sanjyot Mehendale of UC-Berkeley agreed to serve as the faculty mentors to the project. Mejgan Massoumi of Stanford also decided to collaborate with the project. These individuals were essential to the project's success. They contributed with recommendations about the Afghanistan-based websites in Dari and Pashto languages that were either bound to change or cease to exist. Both Ebadi and Massoumi and other faculty members served as "insider" experts for information that needed to be preserved from the earliest stages of the project.

The project that we undertook at the UC-Berkeley library was based on a simple proposition: the need to preserve the reporting of the new government and associated transitions and the subsequent changes to existing governmental websites, including the ephemeralities of websites of the traditional press in Kabul, which was not only urgent but a responsible course of action. I avoided implicit biases by inviting participants from different academic institutions to contribute to the project. The following fundamental questions confronted us in selecting the websites for harvesting to represent all sides of the unequal equation defined by the opposition members, governmental entities, and the media websites that were the markers of Afghanistan's civil society:

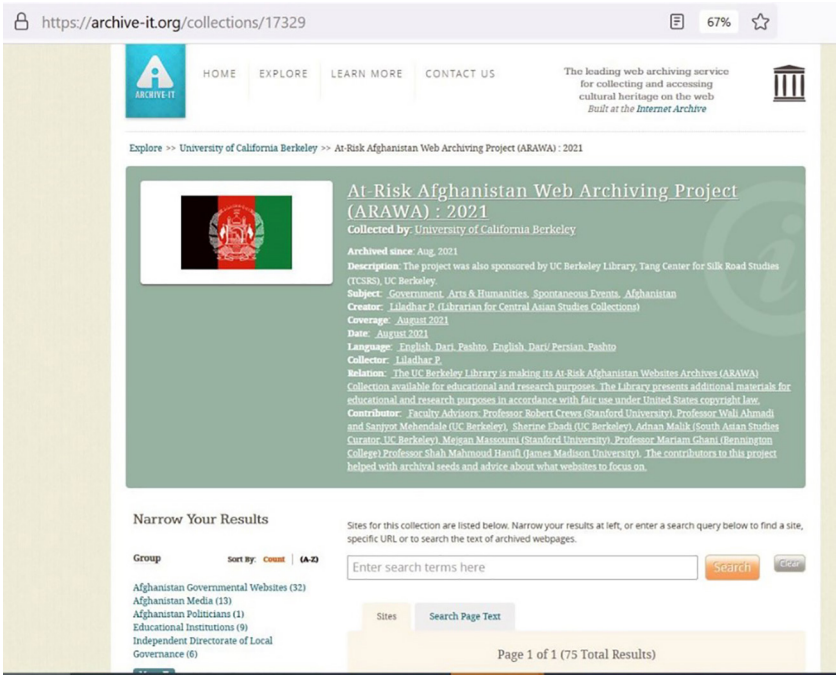

Figure 1: The landing page of the ARAWA Project.

1. What were some of the websites that required immediate archiving due to possible "taking down?"
2. What were some of the criteria that would help the curators to propose the websites for archiving?
3. Should scope and limitations of the project include depth of crawling and frequency of crawling a particular website that were fixed for the total duration of the pilot project?
4. What were some of the social media sites that were deemed necessary for crawling? We recognized that one could not crawl all of the websites manually, and a frequency for crawling was determined by automating the process.
5. What were some of the ways we would seek permission for crawling from the creators of these websites?

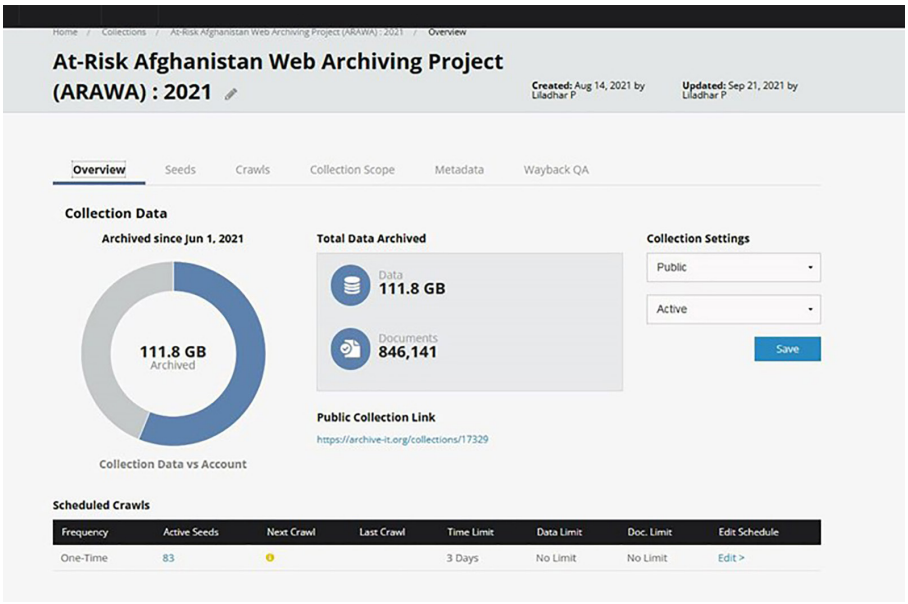

Figure 2: The backend of the landing page of the ARAWA Project.

Consistent with archival practice, we decided to preserve only select websites, given the limited duration of this pilot project. The anticipated time of the pilot project was a maximum of one-and-a-half months. The purpose, as mentioned earlier was to preserve at-risk websites. We decided to use the Internet Archive's web-archiving platform to preserve these websites (see Figure 1). The archive's backend page looks like the one that is shown in Figure 2.

We archived a total of 83 websites. However, we did not make all of the archived websites available to the public, as some of these belonged to social activists and female social figures. Given the sensitivities of their social media utterances, the project curators and collaborators agreed that it would be best to restrict access to some of the sites, as perhaps they would be a target of interest of the newly established government. These websites were grouped into several broad subject categories: Afghanistan Governmental Websites, Afghanistan Media, Educational Institutions, Islamic Emirate of Afghanistan, and Twitter. Currently, the ARAWA project contains more than 846,000 individual documents. These include policy documents, official journals, forms, and audio and video files. For example, Figure 3 shows an archived copy of the official Ministry of the Interior journal Pulis (Police) that was published before the Taliban takeover. Figure 4 is an example of a governmental website that was archived.

Figure 3: An issue of the Ministry of the Interior's Magazine: Pulis (Police).

Lessons learned

The project itself had several challenges. The first challenge was to archive the websites at risk of immediate erasure due to the Taliban's takeover of Afghanistan's central government. The second was to assemble a team of enthusiastic collaborators with cultural, linguistic,

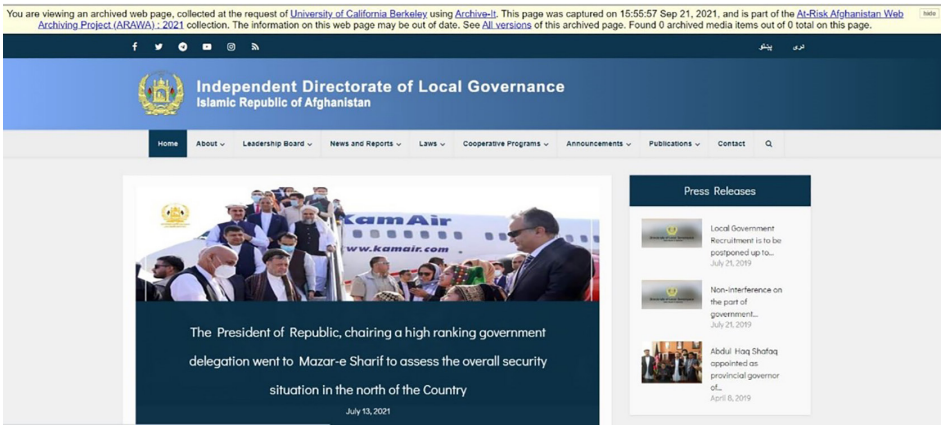

Figure 4: An archived webpage of Independent Directorate of Local Governance.

and area expertise. To create the group, one had to be realistic about the possibilities and the time constraints placed on faculty and student collaborators. Lastly, one had to consider the administrative support required to pull off such a time-sensitive project. The library's administration, particularly Elizabeth (Beth) Dupuis, senior associate university librarian, provided the needed approval on August 15, 2021, at short notice.

It was decided to sunset this pilot project at the end of September 2021 for several reasons. First, the Taliban were firmly in charge of Afghanistan after 20 years of successful insurgency fighting against the occupying forces. Second, the faculty mentors and the participants in the project felt comfortable with the fact that we were able to selectively preserve the websites from potential erasure in a constructive and timely manner.

Assembling the team of willing faculty and scholar participants took some time, but the core group was assembled very quickly. The timely response was an essential factor to make this project viable. In the final analysis, librarians or curators do not function in a vacuum. It takes a whole village of faculty, students, and colleagues to make projects like ARAWA a success.

Notes

1. Agreement for Bringing Peace to Afghanistan Between the Islamic Emirate of Afghanistan Which is Not Recognized by the United States as a State and is Known as the Taliban and the United States of America. February 29, 2020, which corresponds to Rajab 5, 1441 on the Hijri Lunar Calendar and Hoot 10, 1398 on the Hijri Solar Calendar, U.S. Department of State, https://www.state.gov/wp-content/uploads/2020/02/Agreement-For-Bringing-Peace-to-Afghanistan-02.29.20.pdf (accessed November 8, 2021).
2. Carter Malkasian, "The Karzai Regime," in *The American War in Afghanistan: A History* (New York: Oxford University Press, 2021), 80-102.
3. Kaamil Ahmed, Kevin Rawlinson, Caroline Davies, and Helen Sullivan, "Chaos at Hamid Karzai International Airport in Kabul—As It Happened," *The Guardian* (International Edition), https://www.theguardian.com/world/live/2021/aug/16/afghanistan-taliban-kabul-evacuation-live-news-updates (accessed August 16, 2021).
4. John F. Sopko, "What We Need to Learn: Lessons from Twenty Years of Afghanistan Reconstruction" (Arlington, VA: Special Inspector General for Afghanistan Reconstruction, 2021), https://www.sigar.mil/pdf/lessonslearned/SIGAR-21-46-LL-Executive-Summary.pdf (accessed October 27, 2021).
5. Jacques Derrida and Eric Prenowitz, "Archive Fever: A Freudian Impression," *Diacritics* 25, no. 2 (1995): 9–63, https://doi.org/10.2307/465144. Peter Osborne and Matthew

Charles, "Walter Benjamin," in *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), ed. Edward N. Zalta (Stanford, CA: Metaphysics Research Lab, Stanford University, 2021), https://plato.stanford.edu/archives/fall2021/entries/benjamin/ (accessed November 8, 2021). Walter Benjamin, "The Work of Art in the Age of its Technological Reproducibility: Second Version," translated by Edmund Jephcott and Harry Zohn, in *The Work of Art in the Age of its Technological Reproducibility, and Other Writings on Media,* eds. Michael W. Jennings, Brigid Doherty, and Thomas Y. Levin (Cambridge, MA: Belknap Press of Harvard University Press, 2008: 19-55).

6. Luciana Duranti, Terry Eastwood, and Heather MacNeil, *Preservation of the Integrity of Electronic Records* (Dordrecht; London: Springer, 2002).

7. Luciana Duranti and Corinne Rogers, "Archives in a digital society = Les archives dans une société numérique" (2015), https://www.deslibris.ca/ID/244920 (accessed November 8, 2021).

8. Luciana Duranti and Corinne Rogers, *Trusting Records in the Cloud* (London: Facet Publishing, 2019), https://doi.org/10.29085/9781783304042 (accessed November 8, 2021).

9. Julien Masanès, *Web Archiving* (Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2006), https://doi.org/10.1007/978-3-540-46332-0 (accessed November 8, 2021).

10. Niels Brügger and Ralph Schroeder, eds., *The Web as History: Using Web Archives to Understand the Past and the Present* (London: UCL Press, 2017), http://discovery.ucl.ac.uk/1542998/1/The-Web-as-History.pdf (accessed November 8, 2021).

11. P. W. Singer and Emerson T. Brooking, *LikeWar: The Weaponization of Social Media* (Boston: Houghton Mifflin Harcourt, 2018).

12. Joseph Cox, "The Inevitable Weaponization of App Data Is Here: A Substack Publication Used Location Data from Grindr to Out a Priest Without Their Consent," *Vice* (2021), https://www.vice.com/en/article/pkbxp8/grindr-location-data-priest-weaponization-app?utm_source=pocket-newtab (accessed November 8, 2021).

13. Jessica Ogden, '"Everything on the Internet Can be Saved:' Archive Team, Tumblr and the Cultural Significance of Web Archiving," *Internet Histories* (2021): 1-20, https://doi.org/10.1080/24701475.2021.1985835 (accessed November 8, 2021).

14. Madhavi Mallapragada, "Cultural Historiography of the "Homepage,'" in *The SAGE Handbook of Web History,* eds. Niels Brügger and Ian Milligan (Los Angeles: Sage, 2019).

15. Matthews S. Weber, *Web Archives: A Critical Method for The Future of Digital Research* (Aarhaus: WARCnet, 2020), https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf (accessed November 8, 2021).

16. The At-Risk Afghanistan Web Archive (ARAWA) project is available at https://archive-it.org/collections/17329.