All Together Now: Collaborating to Capture and Digitally Archive Aerospace History Research Data

April Gage, NASA Ames Research Center History Office Archives

Abstract: This discussion will touch upon some key issues in the area of digital archiving, and provide an overview of goals and objectives of the NASA History Program Digital Archives Working Group. The working group was recently formed to confront challenges associated with digital archiving, from improving the quality of submissions from historical content creators to exploring tools and methodologies, with an eye toward collective knowledge sharing, problem solving, and identifying opportunities for collaboration.

The amount of information in the digital universe is expanding dramatically. According to the International Data Corporation's studies of digital information growth, the digital universe was estimated at 281 exabytes in 2007,[1] 1.8 zettabytes in 2011,[2] and will swell to 44 zettabytes by 2020.[3] Given that one zettabyte is a thousand exabytes, that's 281 to 44,000 in just over a dozen years. Portions of this rising tide of information are flooding into archives at an accelerated rate and many repositories are not adequately resourced to handle it. From an archives perspective, if the challenge were reduced to how to store a superabundance of incoming data, then some might assume that the answer might be to increase the

---

[1] Gantz, J. F., et al. (March 2008). "The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011," International Data Corporation. Framingham, MA. Retrieved from http://www.calvin.edu/~dsc8/documents/diverse-exploding-digital-universe.pdf

[2] Gantz, J. F. and Reinsel, D. (June 2011). "Extracting value from chaos." International Data Corporation. Framingham, MA. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[3] International Data Corporation (April 2014). "The digital universe of opportunities: Rich data and the increasing value of the internet of things (Executive summary)" Framingham, MA. Retrieved from https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

amount of storage space and decrease storage costs. But simply storing something is not archiving. There seems to be a pervasive lack of understanding about the complexity of digital archiving, from the tools, systems, methodologies, and resources needed, to what the business of preservation entails.

We still face a digital dark age, which can be briefly characterized as a situation in which portions of the human record are being lost, trapped on obsolete physical and digital formats, and where there is an increasing scarcity of software, hardware, and the technical knowhow to extract the information contained therein.[4]

In my experience, there is widespread complacency toward and lack of understanding about archiving, especially with respect to long-term preservation of historical digital material. It seems understandable that archives, which often reside in the attics and basements of the institutional mindset, may not be top of mind. Traditionally, archives have been where operational information retires, receding into the historical record while the organization springs forward to confront present challenges and pursue future goals. A popular perception of an archive in computing--sometimes simply a folder where old or unused files are stored--is often confused with an actual digital repository, with its systems and associated methodologies. It is not unusual to regularly encounter assumptions that data are preserved when placed in a in a folder called "archive" on a server, cloud storage area, DVD, external hard drive, or email client. To be sure, this type of complacency will prolong the digital dark age. Digital formats are inherently

---

[4] The field of data archaeology focuses on recovering data trapped on obsolete formats and making it intelligible. One such project in the aerospace realm that took place at NASA Ames Research Center was the Lunar Orbiter Image Recovery Project headed up by Dennis Wingo (see the project website at http://www.moonviews.com). Or, consider the missing Apollo 11 tapes containing the telemetry of the SSTV signal moon walk footage. Restoration took place in 2009 from derivatives on videotape and super 8 film (See https://www.nasa.gov/mission_pages/apollo/apollo_tapes.html and https://www.nasa.gov/mission_pages/apollo/apollo11_hd-video.html).

fragile and unstable. Digital preservation in archives is a complex and unfolding field unto itself.[5] In addition to ensuring stability and usability over time, digital preservation encompasses much more, such as ensuring authenticity, reliability, and accuracy of data as evidence. In other words, the material is must be trusted as being the "real McCoy," it must not be tampered with, it must be retained in a robust and secure environment, it must come from a reliable source, there must be a chain of custody, and it must remain intact as it is handled.[6] In this context, digital preservation intersects with collection, ingest, description, analysis, storage, and access activities. Putting historical material in a basic digital holding area is inadequate for long-term preservation. Expending resources to create and perpetuate this type of storage model is a short-sighted, false economy that can lead to loss of investment and loss of information.

As we reconsider archives in this forum, it may be beneficial to reexamine stakeholder interactions and resources. How can historians, archivists, and users collaborate to advance the cause for preserving aerospace history? How can archives become better equipped with the trained staff and tools to meet the challenges of our age? Can this community find new pathways for policy makers and leaders to increase advocacy and support for archives? To be sure, even in the narrow field of aerospace history, a collaborative approach may be the only realistic way forward if we are to succeed in capturing and preserving the next fifty years of historical data and making it accessible. As the title of this conference suggests, President Kennedy's September 12, 1962 Rice Stadium moon speech still resonates today as a tremendous source of inspiration for setting our sights on accomplishing this hard task, and as a call to action. Indeed, to succeed, we must be bold.

---

[5] See For an example, see the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) project under the direction of Dr. Luciana Duranti. http://www.interpares.org/

[6] John, Jeremy L. (2012). Digital forensics and preservation. DPC Technology Watch Report 12-03. Digital Preservation Coalition. Great Britain. Retrieved from https://www.dpconline.org/docs/technology-watch-reports/810-dpctw12-03-pdf/file

The NASA History Program recently formed a Digital Archives Working Group (DAWG) to tackle some of the challenges associated with digital archiving, from establishing guidelines for submissions from historical content creators to exploring tools and methodologies, with an eye toward collective knowledge sharing, problem solving, and identifying opportunities for collaboration. The group's membership comprises representatives of NASA history program archives from headquarters, field centers, and associated universities,[7] and is co-chaired by Holly McIntyre-DeWitt and myself, under the oversight of NASA's acting chief archivist Steve Garber. DAWG has two primary objectives. The first is to formulate recommendations, general guidelines, and good practices for capturing historical research data and products commissioned by NASA, with a focus on digital forms. The plan is to complete this work in a relatively short timeframe, then evaluate successes and failures, and incorporate improvements over time. The second and larger objective, which is not yet fully defined, is to tackle the technical nuts and bolts of building and preserving digital archives, processing materials, and providing access to them, and formulate recommendations for toolsets, standards, policies, and practices. Combined, these two efforts may also be instrumental in forming a third objective to develop awareness and institutional support for increasing resources to history program archives.

The first objective is driven by recent lessons learned that point to the need for a dialogue between content creators and archivists at the onset of a project, in order to ensure that appropriate data are captured, organized, and transferred in workable form to the archives, along with adequate supporting documentation. Issues encountered with submission packages are wide-ranging, and can greatly increase the time and expense required to process the material, compromise research value and usability, and delay access by the research community.

---

[7] Ames Research Center, Goddard Spaceflight Center, Jet Propulsion Laboratory, Langley Research Center, NASA Headquarters, Stennis Space Center, and University of Houston-Clear Lake.

Examples of types of data submissions that trigger additional processing activities are provided in the table below.

Table 1: Types of submissions that can increase processing time and expense

Issue	Example	Additional Processing Required to Remedy
Missing documentation	Identification of interview dates and/or participants in recorded oral interviews; oral interview release forms	Obtain permission to release and copyright waivers from narrator and interviewer (entails extra steps for deceased participants). Endeavor to identify participants and time of interview.
Incomplete data submission	Research notes, timelines, spreadsheets, associated correspondence, or supporting files for finished products	Context and supporting evidence for research products is unclear or lost.
Material is transferred on obsolete carriers	VHS tapes, audiocassette tapes, optical media (CDs, DVDs)	Evaluate technical characteristics. Migrate to digital or to current, acceptable format and verify. May entail information loss. Can require third party assistance, special equipment, or purchase of software. Document everything done to the materials.
Files captured or compressed into lossy file formats or in unstable, obsolete, or proprietary formats	MP3, JPG; specialized software programs, formats not widely used, old software, abandonware	Not applicable for lossy formats. Information is irretrievably lost. Evaluate technical characteristics. Migrate to stable, widely-used format and verify. May entail information loss. Can require third party assistance, or purchase of software. Document everything done to the materials.

Filenames contain blank spaces and non-conforming characters	# % & { } < > : " / \ | ? * $ @ ~ . (except preceding the file extension)	Remove spaces and replace non-conforming characters.
Material is transferred piecemeal	Data transferred in more than one batch contains duplication, generating confusion about which files are final	Identify a complete set of final versions. Can require opening, analyzing, and comparing individual files.
Restricted or controlled content is not identified	Items subject to International Traffic in Arms Regulation, personally identifiable information, copyrighted material	Perform an item-level review and identify and document rights and restrictions. Can require research and assistance from an export control officer or other specialist.
Content is disorganized	File folders entitled "stuff" or "stray powerpoints"	Survey and arrange data. Endeavor to discern context.
Insufficient descriptive information about submissions	No inventory, file key; other descriptive metadata missing or insufficient; NASA catalog numbers are not identified for images	Generate descriptive information. Can require research and in-depth review.

Problematic submissions can be minimized if data archiving requirements are discussed up front at the project formulation stage rather than at the end of a project, just before submission to the archive, when historians likely don't have the time or resources for re-working their submissions. Archivists can help shape an archiving plan, such as a set of minimum requirements suited to a given project, advise on how to structure and prepare data, and supply good practices and documentation templates that are vetted and approved by the agency's legal department. This investment of time up front will aid citizen historians who may be unaware of standard documentation practices as well as professional historians, who might not be conversant in the latest good practices for digital submissions. Further, this initial discussion may uncover ways in which the creators, who best understand the content, can explore possibilities for partnering with archivists to enhance description of the collection beyond what is contained in the submission package. With this process in mind, the working

group will consider modifying an approach used with principal investigators in NASA's Life Sciences Data Archive, in which a data submission agreement (DSA) is used as a mechanism for planning for the collection of experiment and other project data. The DSA establishes an agreement between investigators and the agency regarding the transfer of research related data, data formats, scope of data to be submitted, and timeframes for submittal.

The guidelines and supporting examples the working group produces might be used to formulate submission requirements for historical work packages. The goal is to formulate guidelines that will eliminate guesswork, improve the quality of submissions, reduce processing costs, minimize data loss, and ultimately, speed up accessibility.

The second DAWG aim, to explore the technical nuts and bolts of digital archiving and formulate recommendations for common approaches to methods and means, is motivated in part by a desire to pool resources to collectively investigate and form plans to address some of these complex issues. Expertise varies across the agency and resourcing for this area of archiving has not kept pace with the need for expanded capabilities, systems, and trained staff. In considering approaches for this second objective, the group's first step is to compile information about the current landscape in History Program archives with respect to resources, capabilities, and goals. Another step is to examine existing capabilities within the agency that can inform the group's efforts and steer it away from the business of reinventing wheels. A key example of an area where this is possible is in conceptualizing digital archives using the Open Archival Information System (OAIS) as a functional model. OAIS was developed by an international body, the Consultative Committee for Space Data Systems (CCSDS), to standardize space information and data systems. It was designed to be adapted to archiving a variety of information, including documentation, data, objects, and biological specimens. It has emerged as the gold standard for

conceptualizing digital archives and a fundamental design requirement for organizations and institutions around the world, including major universities, the United States Library of Congress and National Archives and Records Administration, the British Library, the Bibliothèque nationale de France, and elsewhere. Among the original four OAIS-based data archive implementations were three NASA nodes at Ames Research Center and Johnson Space Center (now part of NASA's Life Sciences Data Archive) and Goddard Space Flight Center (now NASA's Space Science Data Coordinated Archive, NSSDC). Moving from conceptualization to implementation of an actual repository, another pioneering effort in which the CCSDS participated, and a possible key area of existing capability within the agency to leverage for historical archives, is the formulation of Trustworthy Digital Repository (TDR) as per ISO standards (ISO 16363: 2012[8]). DAWG might consider investigating archival repositories within the agency to identify those based on this standard (NSSDC is one) and whether such a repository can be scaled to accommodate other collections that serve different designated communities, or whether a shared TDR could be formed for historical archives. A solid, proven, standards-based TDR capability is fundamental to long-term preservation. Building or expanding a Trustworthy Digital Repository shared across centers may prove to be achievable and more cost-effective than attempting to stand up several individual ones.

In addition to examining existing capabilities and systems, the group can consider a collaborative approach to testing, implementing, and sharing the expense of using a common toolset, such as digital forensics and other processing appliances. The field of digital forensics in archives utilizes tools and approaches to preserve and repeatedly analyze data without altering it. Forensic tools can,

---

[8] International Organization for Standardization. (2017). ISO 16363:2012 (CCSDS 652.0-R-1): Space data and information transfer systems: Audit and certification of trustworthy digital repositories. Geneva, Switzerland. (See also, The Consultative Committee for Space Data Systems Recommendation for Space Data System Practices for Audit and Certification of Trustworthy digital repositories, CCSDS 652.0-M-1 at https://public.ccsds.org/Pubs/652x0m1.pdf)

among other things, facilitate such activities as checking for fixity, protecting from alteration during capture and transfer, and detecting forgery and tampering. Further, these tools can automate processing activities such as locating personally identifiable information, extracting metadata, mining for subject content, or finding and removing duplicate files.[9] Forensic tools can benefit information producers and users in addition to archivists. They are vital for digital archiving. In fact, approaches to some of the issues outlined in Table 1 above, such as bulk renaming of files or and removing duplicates, can be automated using these appliances. Some members of the working group are already moving forward with implementing portions of the BitCurator suite of open source digital processing and forensics tools. A BitCurator installation is already in place at Goddard Space Flight Center and planned for installation at Ames Research Center. Because information technology staff at Goddard had already performed a review of the suite to ensure its suitability with respect to security and other agency requirements, information technology staff at Ames were able to consult with them and expedite approvals for installation, rather than separately expending resources to review and approve the tool. Moving forward with common approaches such as BitCurator should also increase opportunities for knowledge sharing and collaborative problem solving.

Another challenge for the working group, and for this gathering of historians and archivists, is to find ways to raise awareness and stimulate action for advocating for resources to form collaborative strategies to better collect and broadly disseminate the historical record, and truly preserve it lest all be lost. In other words, I am talking about increased funding. While resource models for current historical archiving activities at the agency are not trivial, they are inadequate

---

[9] BitCurator. (2018). BitCurator and Archival Workflows and How do these tools address archival concerns? Retrieved at http://wiki.bitcurator.net/index.php?title=BitCurator_and_Archival_Workflows and http://wiki.bitcurator.net/index.php?title=How_do_these_tools_address_archival_concerns%3F

considering the rising tide of the digital universe and the nature of the information being preserved.

More trained staff should be brought to bear on the problem. In this area most of all, the managers, policy makers, and leaders of the historical community can make a difference. They can educate and inform those who control resources, and advocate for more resources to be devoted to this daunting task. To succeed, we must be bold, and we must boldly demand what we need in order to make success possible. We cannot afford to fail.