“Immediate access” is not the same as “Open access”

Response to: “The fight for control over virtual fossils” [1] - https://www.nature.com/articles/d41586-019-00739-0

Kate Webbink*, Janeen Jones*, Sharon Grant*

*Field Museum of Natural History

Everyone is in favor of accelerating the pace of science.

Whether we are involved in the production, use, or archival stages of these projects, we can all agree that accelerating the pace of science and improving long-term preservation benefits the broader community as well as museums. However, each group in this process has a different understanding of the technology infrastructure and costs necessary for responsibly handling data generated by digitization projects.

The Field Museum does not insist on being assigned ownership of data. We do insist on *documenting* the negotiated agreement of ownership. We always discuss and are willing to accommodate--to the extent we are able--the original data owner’s or creator’s needs. If we do not have clearly documented permission to archive or publish the data, we will not archive or publish it. This is simply a continuation of practices we already follow for physical collections.

Relative to collections-based institutions, many digital infrastructure-focused projects like Phenome10k, Morphosource, and the Open Science Framework (OSF) are young. Inherent in their goals is prevention of data loss as well as the preservation of the context in which it was created -- i.e., not forgetting why or how that data was created. These projects have vastly improved frameworks for immediate and open access to data by being flexible and building on the successes of past projects. We do not want to lock ourselves into inflexible monolithic systems, but in order to be viable long-term, infrastructure-projects need to be able to plug into a stable, flexible environment that covers lost costs.

Museums and institutions often end up covering less glamorous lost but vital project-costs related to data creation, curation and staffing. In a given digitization project, these costs are frequently hidden or lost because they either don’t make it into the project grant due to tight deadlines for submission, or are cut in response to a funder’s subsequent budget feedback.

On the data-side, lost costs include:

- Storage - including hardware maintenance.
    - **Short-term** - need more storage to process than to store the data (exacerbated by scope creep)
    - **Long-term** - most funding bodies won’t fund this, but institutions must store data afterwards (and maintain the hardware for doing so)
- Preservation - maintaining current archival standards for file formats, validation and fixity checks.
    - **Short-term** - handling formats that the institution did not know how to handle before hand (requires figuring out workflows and processes for doing so)
    - **Long-term** - format preservation/migration

- Transfer/Bandwidth - setting up network infrastructure for moving giant files and sets of files between repositories.
    - **Short-term** - the shock of sudden/new/massive files puts a burden on existing infrastructure. Underestimation of the size of the research products further exacerbates that shock.
    - **Long-term** - technology acclimates to the size/variety of the files [actually a reduction]
        - ...But the files keep getting bigger.
- Standards - ensuring assets have appropriate identifiers and contextual data for the field of study, and are permanently associated with their voucher specimens.
    - **Short-term** - Expedited workflow
    - **Long-term** - Loss of data & context; broken connections to vouchers and research products affects reproducibility; Solutions to fixing and cleaning up become slower if not impossible to implement.

On the people-side, lost costs include:
- IT Professionals (Systems, Network, Informatics, Web)
- Collections Professionals (Collection Management, Assistants, Interns, Trainers)
- Project Management (Team coordination between fields, departments, and institutions to maintain scope, deliverables, and timelines)

These lost costs will be incurred by younger projects; they might not feel it yet, but as they scale up, they will. Many institutions have yet to fully embrace managing their changing digital collections; however, libraries, archives, museums and other memory-institutions exist for the very purpose of providing such stable, flexible environments focused on long-term preservation (see the Field Museum's Data Norms [2]). While planning around ever-changing technology is difficult even 5 years in advance (let alone 100), these institutions have at the core of their mission preservation in perpetuity, disaster management, legal documentation and ethical practices.

As a result of addressing these costs, institutional repositories might seem more bureaucratic, but this is essential. Furthermore, it is done in order to provide proper attribution and validation of content and in the spirit of creative commons. Consider less rigorously maintained repositories like the Internet Archive (https://archive.org), which is an 'open' trove of rich but questionably-licensed content, making it unfit for use (inaccessible) for projects aiming to follow legal and ethical guidelines. Orphan works comprise many of the assets on archive.org, and not all orphans are mysterious accidents of the past; many are newly generated whenever digital collectors do not take the time to document or transfer existing copyright. Unverifiable information spreads more quickly when repositories and their users fail to critically assess content and its sources--not just in terms of the recent "fake news" fervor, but in terms of dubious versioning, and transparency- and replication-crises that organizations like Public Library of Science (PLoS) and the Center for Open Science (COS) were established to address.

Without spending time properly managing and documenting the data in context (who did what, why, how, when, and where), repositories run the risk of untrackable licensing and distribution of sensitive data. Better management practices need adoption beyond the scope of short-term projects and grant-funding cycles. They are fundamental for the public and scientific communities to successfully link open data and see global patterns. Documentation is necessary to comply with federal grants (e.g., NSF's PAPPG Chapter II.C.2.j [3]). Governmental funding programs run by agencies such as the NSF and IMLS set complicated and sometimes conflicting requirements between their data management policies and their unrealistic assumptions of available institutional long-term support. As an institution that accepts funds from the NSF, we are obligated to comply with these requirements.

Another disconcerting issue underlying the current state of access/storage for 3D data is the lack of clarity on its copyrightability. Until that issue is clarified, it is all the more important to explicitly document who created what why/how/when/where in the meantime. Creative Commons licenses provide an excellent framework for better-understood copyright scenarios (traditional photography which is a large portion of the storage cost from earlier digitization projects). We look forward to the out-comes of 3D community standards projects like Community Standards for 3D Data Preservation (CS3DP, http://cs3dp.org) and Developing Library Strategy for 3D and Virtual Reality (LIB3DVR, http://lib3dvr.org).

When archival or institutional repositories such as the Field Museum request that researchers transfer ownership and copyright of work they want to archive, the purpose is to enable preservation of data and media associated with specimens and objects in as thorough and stable a manner as possible. Transferring copyright to institutions that are stable and easy to find makes it easier for users to figure out how to license works in the future. Taking the time to sort out these aspects might feel like a burden in the short term, but it is critical in the long term.

Leveraging community resources to build better infrastructure is the only realistic solution, but as it stands, few repositories are fully equipped to deal with the volume and structure of new data types being generated today (let alone tomorrow). Ephemeral consumer-grade harddrives or websites are not a stable solution. Given time, money, and clarity on project scopes and timelines, memory institutions could provide more supportive foundations in the way of:

- Accommodating for current standards (e.g., Darwin Core [4] and Audubon Core [5])
- Following best practices (e.g., those recommended by the Magenta Book “ISO16363:2012” [6] or SPEC Kit 354 [7]):
    - Few data repositories have achieved “Trustworthy” certification by organizations like CoreTrustSeal, but many are trying.

Enforcing any preservation policy is “cumbersome” if it slows the fever-pace of research, but the value of doing so always becomes apparent in the wake of disaster. Berlin-Dahlem 1943, Christchurch 2016, Brazil’s National Museum in 2018, Notre Dame 2019 happen on small scales daily -- every time a harddrive tips over. (Approximately 30 times at the Field Museum since 2015.)

Research is all about uncharted territory. We understand that, but please pardon our bureaucracy whilst in an already stressed system we try to chart it for you so we don’t all run aground. After all, scientific research has to be reproducible.

Citations:

1. Lewis, Dyani. (2019). “The fight for control over virtual fossils”. Nature 567, 20-23. doi: 10.1038/d41586-019-00739-0.

2. “Use of Collections Data and Images”. Field Museum Data Norms. https://www.fieldmuseum.org/field-museum-natural-history-conditions-and-suggested-norms-use-collections-data-and-images (Accessed Sep 6, 2019)

3. NSF PAPPG Chapter II.C.2.j (Feb 25, 2019) https://www.nsf.gov/pubs/policydocs/pappg19_1/pappg_2.jsp#IIC2j (Accessed Jul 25, 2019)

4. Wieczorek, John; D. Bloom; R. Guralnick; S. Blum; M. Döring; R. De Giovanni; T. Robertson; D. Vieglais (2012). "Darwin Core: An Evolving Community-developed Biodiversity Data Standard". PLoS ONE. 7 (1): e29715. doi:10.1371/journal.pone.0029715.

5. Audubon Core. (Oct 23, 2013). http://terms.gbif.org/w/index.php?oldid=10756

6. ISO16363:2012 Magenta Book (2011)
- https://public.ccsds.org/Pubs/652x0m1.pdf (2011 / Audit and Certification of Trustwothy Dig Repos)
- https://public.ccsds.org/pubs/650x0m2.pdf (2012 / Reference Model for an OAIS)
- https://public.ccsds.org/Pubs/652x1m2.pdf (2014 / Requirements for Audit & Certification of TDR)
- “Space Data Systems.” If it’s good enough for them, it’s good enough for Earth Data Systems?

7. Hudson-Vitale et al. (2017) Data Curation. SPEC Kit 354. Washington, DC: Association of Research Libraries. https://doi.org/10.29242/spec.354

Scraps/Further Detail/Background in this doc