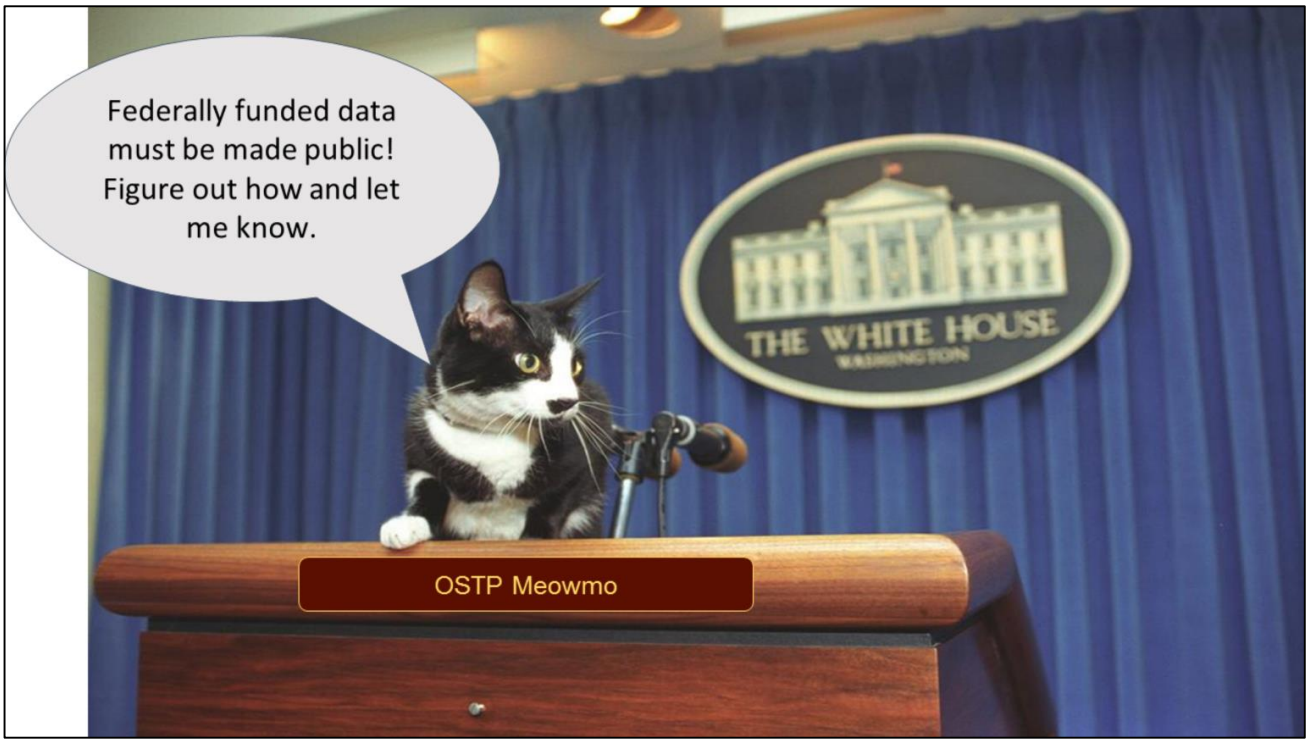
Digital Preservation and Data Repositories: just what does “long-term” mean, anyway?

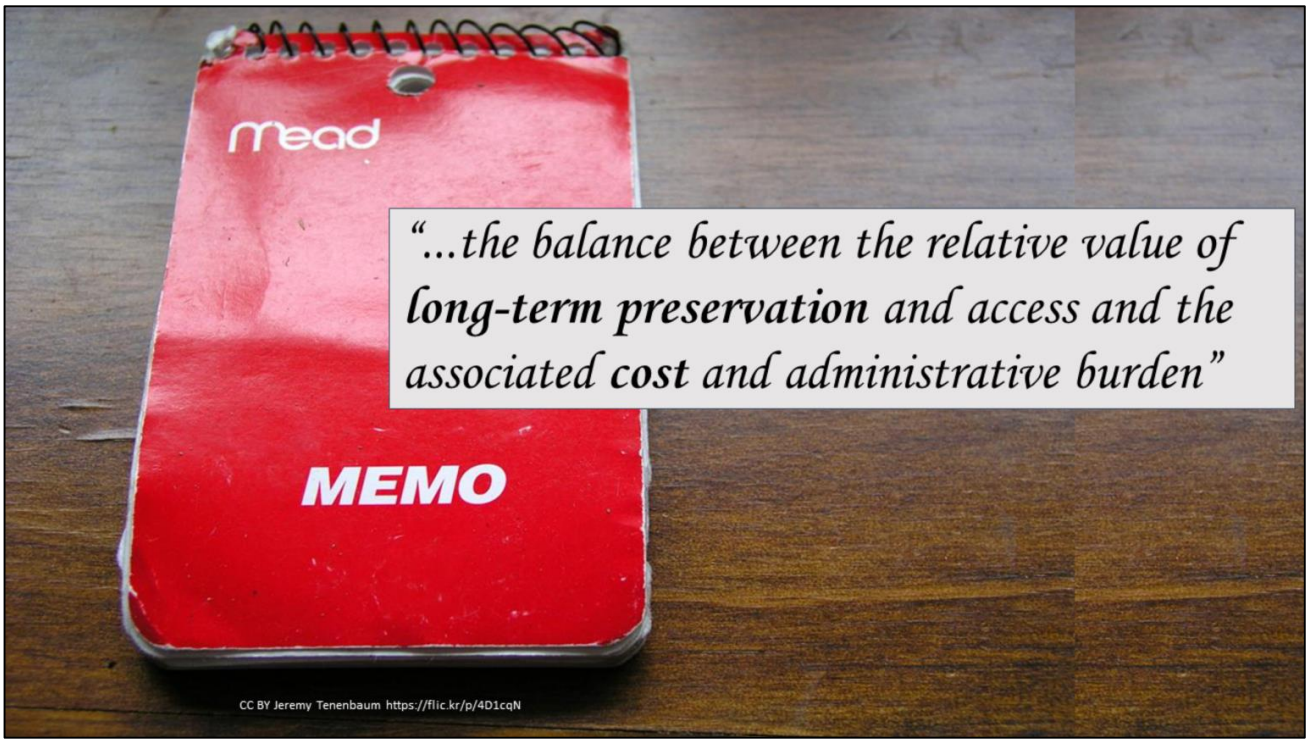
Mahria Lebow
mahria@uw.edu

Moriah Caruso
neilsm@uw.edu



Many academic libraries were ready to act after the release of the US Federal Government's OSTP memo on access to federally funded research data. There was an expectation that in the plans generated by federal agencies in responses to the memo that libraries would play an important role. Though largely the plans did not end up mentioning institutional repositories for storage and access to research data, libraries and librarian can - regardless - take a leadership position in setting norms and expectations for preservation.

Photo based on https://flic.kr/p/aQYUVM



The OSTP memo listed specific things that the agency's plans needed to address. For preservation, it said, to find "the balance between the relative value of long-term preservation and access and the associated cost and administrative burden".

This directive explains the language seen in the responses. As a part of trying to craft policy for our future data repository, I examined all of the responses released so far, hoping to ensure our policies and system design are in line with requirements. of the 17 agencies have put out a plan, 14 mention preservation. By and large they use the phrasing directly from the memo "long-term" and a few talk about the cost-benefit analysis. Only one, USGS, gets more specific that you should follow the same rules as you would for USGS record retention.

[refs:https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf, this document links to all of the individual responses: https://docs.google.com/spreadsheets/d/1PYOhBh6bglh6BkQFlpvNLOwlpzvQyguWAG8AkQMtU0s/edit#gid=0]

So, what does long-term mean in the eyes of the feds, does the absence of a definition imply forever? More importantly are researchers assuming that we will keep their stuff forever?

It is important to note that these plans are all drafts at this point. Things can, probably and hopefully will change.

Interested in clues as to the direction the drafts might head in, I looked at data sharing policies that predateding the OSTP memo. 13 of the NSF directorates had a data sharing policy. Only 3 of these mentioned preservation in any capacity. And one had a time frame, Engineering, which said that the data needs to be publicly available for 3 years. So, that's one idea. IES, part of the dept of ed, also had a pre OSTP data sharing policy and it said the data needed to be made available for at least 10 years. Quickly, other countries: Canada's TriCouncil draft plan says to follow community norms and the UK's Concordant draft says 10 years, unless directed otherwise by a funder.

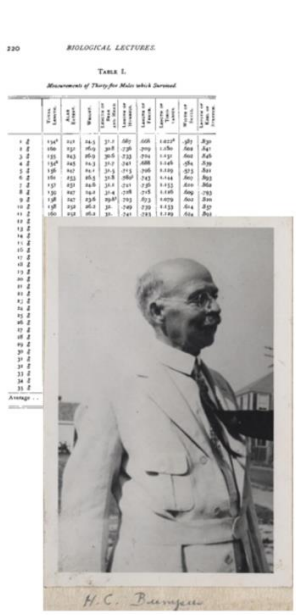
[refs: https://ies.ed.gov/funding/datasharing_implementation.asp, http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1, http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/ConcordatOpenResearchData.pdf]

The situation that Mahria has laid out is obviously less than ideal, but with little guidance from the Government, how long is long enough? Their lack of specificity does not justify ours; Libraries are currently moving to fill the perceived need and building out capacity to collect digital data. We have to make our own decisions, otherwise they will likely, unwittingly, be made for us, by our technology choices and the vagaries of funding, or lack thereof.

But let's take a quick step back and talk about data preservation—what it is, why it's important, and how the work gets done.

Like all good librarians, I'll start with a definition.
Already, I hope many of you know the challenges of digital preservation—and many come in that last clause: "beyond the limits of media failure or technological change" –and in a minute I'll get into why that is especially relevant for data.

But first, why preserve data? Why worry about ensuring persistent access to it, over a large span of time?



"H. C. Bumpus". Embryo Project Encyclopedia (undated). ISSN: 1940-5030 http://embryo.asu.edu/handle/10776/2503.

Photo credit: Billwhittaker at en.wikipedia

1. Repeatability and verification are main tenets of the scientific method. Keeping data accessible makes these actions possible.
2. Data preservation also allows for meta-analysis of datasets over time
3. Data sets are reused in new research, to answer questions unconsidered by the original researchers: my favorite case in point is a data set that is more than 115 years old, Edward H Bumpus' dataset from his 1899 publication "The elimination of the unfit as illustrated by the introduced sparrow". Papers using and citing this data span more than 100 years of science. This kind of data also falls into a whole set of research data that are unreproducible: we can't go back and measure the wing length of Bumpus' sparrows killed by a snowstorm in new england any more than we can re-do an archeological dig—you can only excavate once. Think about an ice core sample, space data observing a supernova, the genomes of species that have now become extinct, observations rooted in time.
4. Then there is the research that needs government funding to even exist, data from the large hadron collider, the hubble space telescope, on a smaller scale, even observations at a university wind tunnel, a 50 year longitudinal study of eating habits: projects of a size and scope that make them highly unlikely to be repeated by a less-well-funded researcher. If data of any of these types is lost, corrupted, or just trapped in an arcane file format and dependent on obsolete hardware, there is a virtual dead end data is lost-the time, money and talents of those researchers is wasted, and the products of that research, the published papers remain the only testament to the work. An important one! But not everything.

Photo credits not in slide:
Upper left: From Bumpus, H.C. (1899) The elimination of the unfit as illustrated by the introduced sparrow, Passer domesticus. Biological Lectures, Marine Biological Laboratory, Woods Hole,: 209-226. See digital copy at http://hdl.handle.net/2027/hvd.32044106243728?urlappend=%3Bseq=230
Bottom middle: http://www.ncbi.nlm.nih.gov/genome/guide/human/

"L'evoluzione della specie" by Aldo Cavini Benedetti on Flickr

Open Archival Information System (OAIS) Model

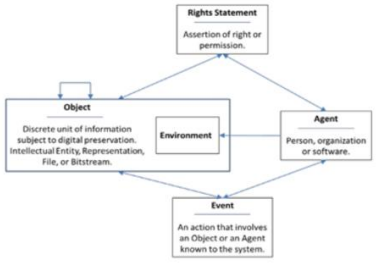
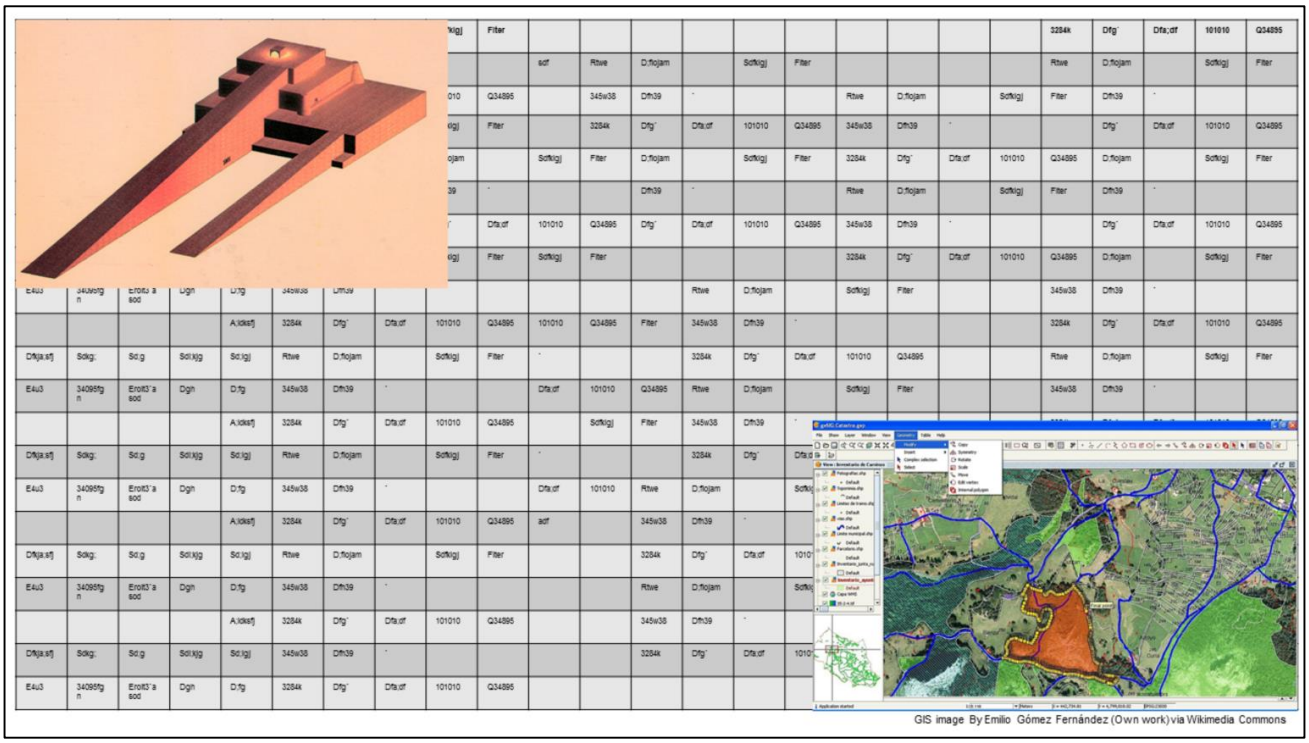


Figure 1: The PREMIS Data Model

Preserving research data is important. How do we do it? Obviously, my time is limited, but I can say you can use the same bag of tricks as other digital materials, geographically dispersed replication, checksums and monitored storage, harvesting and storing technical metadata, recording preservation metadata, migration, normalization—emulation, even. OAIS-compliant repositories, PREMIS metadata, these standards are out there. But there are two things about data that makes it a little different from other digital materials.

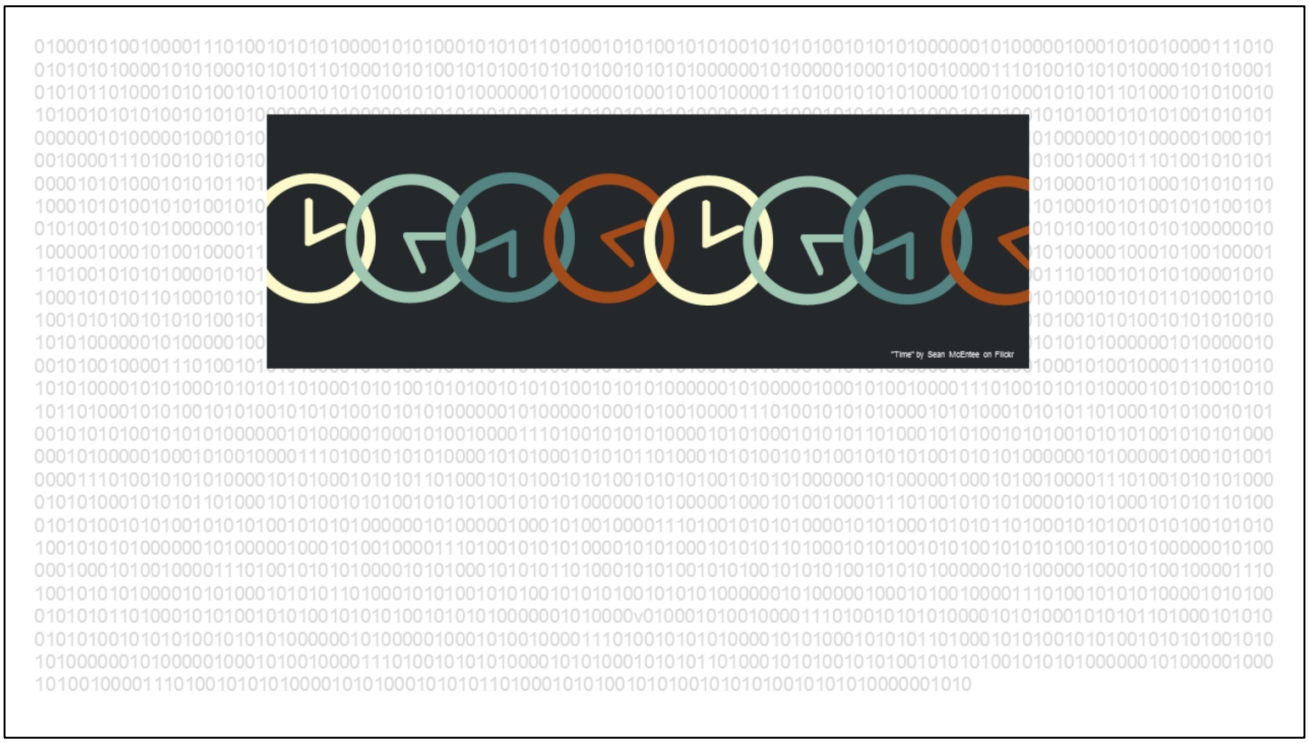
Premis data model: http://www.loc.gov/standards/premis/v3/index.html
Team digital preservation: https://youtu.be/pbBa6Oam7-w
OAIS: http://public.ccsds.org/publications/archive/650x0m2.pdf



GIS image By Emilio Gómez Fernández (Own work) via Wikimedia Commons

First, the bad news: data is a wild and wooly area, researchers use discipline specific software and hardware to collect the data and proprietary file formats abound. Researchers develop their own methods, naming conventions, code books, abbreviations. Even if the data is recorded in an excel spreadsheet, itself a fairly widely used format, well understood, the data entered in columns and rows is only as good as the headings—or the metadata supplied alongside. The good news about research data is the advent of the Data Management Plan, and the involvement of libraries and information professionals in that process—this is a digital preservation coup! The easiest and best way to preserve digital materials is to make sure they are well-thought-out from the beginning. Recommend good file formats, encourage excellent data hygiene—the help that we can provide upfront pays dividends for the lifetime of the data.

Photo credits not included in slide: upper left: CAD Ziggaraut from https://commons.wikimedia.org/wiki/File:SialkCAD.jpg



But this brings us right back—what is the “right” amount of time to preserve data? How long is long enough? My answer is : “it depends”, some might be worth more effort than others because of the factors I have already mentioned. That is why planning for selection in data repositories is my last rant of the day. We all can attest that digital storage is NOT free, and that as computing power increases, so does the size and complexity of data. But then again, who could have predicted the longevity of Bumpus’ sparrows? Is it cheating to promise or imply “long term preservation” without the people, resources, and systems in place to assure it? Accept all materials university researchers care to self-deposit and only *recommend* appropriate file formats? Say “we’ll reevaluate when we run out of storage” or “we’ll reevaluate in ten years” trusting a monitored storage system to provide bit-level preservation and increasing technological savvy to take care of the obsolete hardware and software? Questions abound, but not asking them is asking for trouble.

Questions or Comments?

Mahria Lebow
mahria@uw.edu

Moriah Caruso
neilsm@uw.edu