Title:

Dance like nobody's watching, but version, backup, secure, save in sustainable and open formats, retain on separate local and remote drives, document with machine-readable community standards, deposit in a timely manner in a long-lived, publicly accessible repository, assign identifiers and a copyleft license, cite these in manuscripts, share online, and measure the scholarly and societal impacts of your digital research products and include the metrics in your grant proposals, CV, and P&T dossier like somebody is…

By Andrew Creamer

Today, I would like to briefly share a data management librarian's perspective on the emerging ecosystem for measuring the scholarly and societal impacts of digital research products. By digital research product, I mean this broadly as the electronic files of research data, documentation, code, protocols and methods, and broader and societal impacts materials produced by researchers, in addition to their publications over the course of a research project. In particular, I would like to acknowledge three challenges present in this ecosystem that I have come to recognize over my five years as a scientific data management librarian and member of the University's Broader Impacts committee, helping Brown University faculty and student researchers to retain, document, cite and publicly share their digital research products publicly--activities that are essential for any possibility of future impacts, let alone their measurement. I also want to start by stating that I feel that access by the public to digital research products, and the emerging tools and platforms like **Chris and John's *Discovery Engine*** that allow for the public's engagement with products of research and that provide mechanisms for capturing the many layers of this interaction, enabling the ability to communicate more complex and textured feedback to funders on the impacts of the

research they sponsor, and providing a peer-review-like potential for evaluating the ***quality*** of available data sets and code libraries in our collections, are all goals that I feel strongly libraries, funders, and publishers should support. However, as I just mentioned there are some barriers preventing the public's access to digital research products.

Challenge #1: No carrots, no sticks

Since 2008 the NIH has required that its funded researchers comply with its *Public Access Policy* requiring them to deposit a copy of their final peer-reviewed manuscripts in *PubMed Central* (PMC). Researchers' compliance with this policy is reported and monitored in progress reports and evidence of their compliance is tied to their receiving funds and future awards. By all measures it has been a success and foundation for communicating results to the public. However, while the NIH's inclusion of a data sharing plan for some awards actually predates this article-based policy, beyond clinical trial and genomic data there is still no broad system in place to monitor researchers' public sharing of digital research products or ensure their compliance with implementation and follow through of their submitted data sharing plans. [note on slide] For genomic data, is that there is both a policy and funder-supported repository in place for deposit of these data, dbGaP, in the same way for PMC is for NIH-funded publications.

We are fast approaching a decade since the NSF requirement that a data management plan (DMP) be submitted with all proposals, and researchers' expected inclusion of and the agency and reviewers' actual notice and evaluation of digital

research products listed in NSF proposal's "Results from Prior NSF Support" section relies all too much on the personal and professional values of the parties involved. So the current ecosystem is one either lacking concrete compliance policy, funder-provided archival and dissemination infrastructure, inconsistent review consideration, and little recognition. In other words, no carrots or sticks.

Challenge #2: Data sharing starts at home

Sticks aside, I argue that academic institutions and publishers can cooperate with funders on fertilizing a field for the future planting and harvesting of carrots. University Libraries and IT can cooperate to build infrastructure for retention, preservation, and sharing, and the creation of machine-readable metadata for discovery, reuse, repurposing, and metrics capture. Promotion and tenure guidelines can be adapted to acknowledge and provide criteria for the *evaluation* of the societal and scholarly impacts of digital research products and consideration of their reported metrics. This could, in turn, then help researchers to help change the current culture-- the fear of getting scooped, the fear and stigma of the "*research parasite"* referenced by Drs. Longo and Drazen, two editors for the NEJM, who recently visited Brown to describe one who uses or re-purposes the data of others without attribution and for their own academic gain or without the intentions of its creators, and instead develops a culture that encourages the citation, attribution, and co-authorship, where warranted. Publishers can continue to develop data availability policies for their journals, and to build, hopefully affordable, tools to enable digital research products' discovery and tracking of their citation and measuring their societal and scholarly impacts.

One of the challenges for setting up a system to monitor compliance with federal policies requiring public access to data and implementation of submitted and funded proposals' management and sharing plans is the *heterogeneity* of digital research outputs and sources. One size just does not fit all when it comes to the storing and sharing of digital research products. As a data management librarian my most common refrain for a majority of inquiries I receive from researchers is: "It depends."

> ***Researcher: "How long should I keep these files. Where should I keep these files? Can we deposit them in the repository"***
>
> ***Data Management Librarian: "It depends. Who funded it? When were they created? Where are they stored? Did the research involve human subjects? Is there PII or PHI--any data covered by FERPA, HIPAA? Did it involve any clinical trials? Drugs or devices? Are there any risks involved? Did participants provide consent for long-term retention, preservation, and sharing? Are there any risks? Is it aggregate or person-level? Can identifiers be removed? What countries or U.S. states were your collaborators in? Where did you publish the results? What documentation exists? What format are the data in? Is this data related to any invention disclosure?"*** **And so on.**

Thus, I argue that an important ingredient for carrot fertilizer is institutional policy. Institutional policies set the bar for the expectations for the quality of data management, guidance and best practices, and can set expectations for retention and sharing for even the most heterogeneous digital outputs. One of the tenets of archival and library

science is appraisal and selection, and this one of the reasons I argue the Library deserves a seat at the table when it comes to drafting such policies. Just have a listen to all the areas where these come into making decisions about what to keep, when, and how.

1) Endorsing data documentation and use of scientific community reporting standards, FAIR principles, and data management best practices to encourage data sets that are in a state easily retained, shared, identifiable and discoverable, accessible, useable and citable by other researchers and the public.
2) What data and digital research products ***must*** be retained and/or preserved and/or shared? How, for how long, and why (e.g., (scrutiny and replication of methods, reproducing actual results, rare, expensive, enduring value for further reuse, laws, liability, regulatory compliance, defense of intellectual property)?
3) What are the constraints for the retention and sharing of data (ethics, privacy, confidentiality, export controls, intellectual property, risks, costs)?
4) What ***must***, ***should***, or ***can*** be destroyed? Why, how, after how much time?

Challenge #3: Swimming upstream

I end now with some points I contributed to a recent white paper for the director and escience advisory board of the National Libraries of Medicine New England Region (NN/LM NER) at UMass. We wrote the white paper to help set new data management and sharing educational goals for librarians in our 6-state region. Since 2009 libraries in New England have made significant advancements in establishing and offering services to support the management, retention, preservation, and sharing of their researchers' digital data. It is now common to find in smaller academic libraries a team of librarians or in larger ones an individual librarian dedicated to supporting researchers' navigation of funders' public access policies, with the writing of data management and sharing plans, and with the location of infrastructure and assistance with carrying out these plans. Similarly, we have been actively engaged in helping researchers to meet the relatively new requirements from publishers to retain their digital research data underlying published results and to provide a citation or statement for their availability or else face the risk of retraction.

When I started at Brown, a study by Pinfield et al. (2014) found that data management librarians saw the major challenge for supporting faculty's sharing of digital research products was awareness and persuading researchers to recognize the importance of data management, and to actually seek out help, and I think these goals framed much of our early activities. So to prepare for the white paper my colleagues and I wanted to know what does the literature say about researchers' perceived needs

today? Mainly we found that over the last decade they have not changed much. For example, an often cited paper written in 2011, Tenopir et al. found among the reasons for researchers ***to not share data*** **was insufficient time and lack of funding to get data into shape for sharing**, among expected copyright and ownership issues, and lack of knowledge about metadata, funder requirements, and repositories. This year, nearly a decade later, Stuart et al. (2018) asked researchers about their challenges to sharing their data. The main challenge identified by respondents was still organizing data in a *presentable and useful way* among other reasons including lack of knowledge about funder requirements, copyright and licensing, and repositories.

The question that we now need to start asking is if researchers *were* to get this time and the funds to get there data into presentable shape, then what would they still need to do to make their data available? The answer is a lot. The reality is this level of curation is too difficult to do at the end of a study. Yet this is where most of our support services are aimed. It is unreasonable to expect anyone to take a researcher's data set from a completed study and go back and re-perform hours of experiments in order to collect the missing metadata from the sample and instrument and go back and comment and document their code and then "clean" the data set, metadata, and code files in order to get them presentable and in shape for sharing. So the researcher, and librarians and other support personnel, have to find more ways to connect **upstream** to, from the very outset of a project, set expectations that data will be made available and scrutinized, and make a plan to capture the documentation necessary at point of data collection/creation and code development necessary for reuse, repository deposit, and

publication, and organize and collocate files, and appraise and select for long-term preservation the files necessary for the validation of results and those having enduring value, and look at systems for capturing the impacts of the outputs. So researchers need us upstream to help get advice on how to create and collect metadata, use sustainable formats, and integrate organizational best practices necessary to get their data, metadata, and code presentable for sharing, as well as downstream, where we excel at helping them with selecting a copyleft license and a long-lived, publicly-accessible repository, and obtaining DOIs for the citation of their data set, metadata, and code and their locations in their publications.

I believe there is opportunity for publishers and libraries to partner to address this issue, but at present I believe our available services are not aligned with the needs in the study. For example, *Springer Nature*, which actually sponsored the Stuart et al. 2018 white paper that found organizing data into presentable shape good enough to share was the impediment to data sharing, surprised many of us in the data management library world when they rolled out a new Research Data Support service around the same time of the white paper release. For $340.00 + tax, they offer authors an “Enhancement Report” a team to “organize files into a logical structure and collections”; add keywords; perform checks for spelling and human subject identifiers; they will issue the data set a DOI for citation; and they will deposit the data set into one open repository, FigShare.” In other words, they will charge the authors publishing an article in one of their journals several hundreds of dollars for the same downstream

services that data management librarians at many universities do now and for free (and do well), but our researchers may not be aware.

Yet, what our service does not do, or similar ones I've seen developed by other publishers, is actually address the researcher need identified in the literature -- the need for a partner ***upstream* at the *outset* of a study**. Our services cannot make experimental metadata appear that were never collected in the first place or go back to re-label files and reorder directories to make a cross institution-collaboration more efficient- not even for $340.00 + tax. So the advice I would like to end on is: if data management, sharing, and measuring societal and scholarly impacts are presently mostly downstream activities (with the exception if it's a completely transparent study online, on an open science platform, allowing for registered protocols, public scrutiny, data sharing in real time, comment, and contribution), then all stakeholders have so much more to do *upstream*, and the more we can do together in cooperation the better.