Colorado State University Libraries
CSU Libraries
Training and Instruction
Transcription of The impact of data management, 9/29/2016

Collection: Training and Instruction (10217/195518)
Title: The impact of data management
Date: 9/29/2016
File Name: FACFLIBR_DaD-ImpDataMgmt_TM_20160929.mp4
Date Transcribed: November 2024
Transcription Platform: Konch AI

BEGIN TRANSCRIPTION

[00:00 - 01:43] Tobin Magle: Hi, and welcome to Data and Donuts. I am Tobin Magle, the data management specialist at the Morgan Library at Colorado State University. I'm trained as a research scientist, and it's my job to help researchers manage their data. I organize this monthly workshop series to raise awareness about research data management on campus. All materials will be available online and are open for use, as long as you attribute the source. But, before we jump into the details, let's talk about why data management is such a hot topic to begin with. I'm going to start out with a disclaimer. Data management does not necessarily imply data sharing. However, the same principles apply to both. I'll often describe data management best practices in the context of allowing someone who has never seen your data set before to be able to use it, based on how it's described in the supplementary material. This mindset is useful even if the data are never shared widely, because you are the future user of your own data. Not to mention others in your research group. So, why should you care about data management? Well, you already care about your data. The image on the slide is a researcher's submission from the day of data at Brown University, where researchers were asked to submit postcards with their feelings about data. It illustrates that data elicit a range of strong emotions, from fear to happiness. This response is understandable because data dictate your success or failure as a researcher. But taking care of your research data has changed a lot recently because everything is digital. Preserving digital data takes a different set of skills than preserving a physical object, like a lab notebook. This is partially because digital objects are a lot easier to lose. A hard drive failure is a lot more common than the types of natural disasters that can wipe out a physical record.

[01:43 - 03:00] Tobin Magle: Additionally, because so much is digital and we have the Internet, it's a lot easier to share primary data than it used to be. Also, the number of researchers is increasing. The number of PhDs granted has been growing linearly since the mid 2000. Thus, the amount of scientific output as measured by publication doubles about every nine years. However, the availability of research data degrades over time. By 20 years post-publication, about only half of the corresponding authors on manuscripts have valid email addresses listed. Even if the email address works, the author only responds about half of the time. Even if you get a response, only about three quarters of the authors will even mention the status of their data, and even then, only a quarter of the data is accessible 20 years after publication. And so, we are losing vast amounts of research data. In the midst of all this, research funding is tight. Despite the number of PhDs increasing every year, research funding is essentially flat. And funding agencies want to do more with less. They started by recommending that scholarly papers and digital research should be publicly available. Open access to papers for NSF and NIH funded research is now mandated through PubMed central and the NSF Public Access Repository. NSF has also mandated data sharing, and NIH is following suit. Eventually, sharing these forms of research output will be enforced. The open academic tidal wave has momentum, and it's not going to stop. Open access mandates are even backed up by the white House. A 2013 statement indicated that taxpayers should have access to the research they fund with their tax dollars. To enact this statement, NSF has a strong data sharing policy and requires grantees to submit a data management plan with their grant applications. In other words, we do what we must because we can. But even if you're not subject to any of these funder requirements, data management still has its benefits.

[03:00 - 06:03] Tobin Magle: It's good for science. It improves research reproducibility by providing transparency. It improves efficiency through organization and data reuse, and it spurs innovation by allowing ideas to be shared more freely. [pauses] Good data management practices are also good for you because you are the future user of your own data, and the you of five years ago is really bad at answering email. Your data also get reused inside it along with your papers. Collaborators get to see your work firsthand and grant reviewers take DMPs really seriously. [pauses] The idea of sharing research data openly is a very recent phenomenon. Unsurprisingly, some researchers are resistant to data sharing. I'd like to quickly address a couple of common objections to sharing research data and ways to overcome these barriers. Some researchers don't want to share their data because they see it as their personal property. The reality is that many research projects are supported by public funds, and the US government has already indicated its stance in open data. Additionally, if you work for CSU, the university likely owns your research data and you are the steward of these data because both the funding agencies and CSU promote open data. Data sharing is somewhat a fact of life now. Some people don't think sharing their data is important

because their data are too small for anyone to find useful. However, the number of research projects that are "big enough to warrant data sharing are small, and the number of small data sets is large." Thus, there is a larger volume of small data than big data, making it essential to preserve small data too. Some research involves private information that cannot be openly shared. However, de-identified patient level data and summary level data can be shared. Also, not all data must be open data. You can set up controlled access systems for data that can't be de-identified. A good example of this type of system is dbGaP, which houses human genomic data.

[06:03 - 08:28] Tobin Magle: These data are in a controlled access system because a research group determined that they could match a genome to an individual using only publicly available data. These systems let researchers know the data exist and be made available upon request. [pauses] Some researchers use their data to file patent applications. Sharing research data before a patent application is accepted as not advisable. However, good data management practices have benefits for the patent process and the data can be shared after the patent is accepted. Now that we know why data management is important, let's investigate what it is. Data management is the policies, practices and procedures needed to manage the storage, access and preservation of data produced from a research project. But mostly, it's about having a strategy for how to keep your data safe. [pauses] And where does data management fit into research? Basically always. It's not something that should be left to the end of the project. It occurs continuously throughout the research cycle. Generally, the research process goes as follows. You start with a hypothesis, then you design experiments to assess the hypothesis. Then you collect the data. Analyze the data to produce results. Publish these results in a research article. And these findings can be used to generate new hypotheses. This simple cycle is more complicated now that we have the ability and mandate to share our research data. We need to write data management plans even before data collection begins. And with the large amount of data collected, cleaning and analysis of the data are bigger jobs that are made easier by automation. We also have to take special care to archive the data properly and share digital data. [pauses] On the bright side, technology also allows us to publish and get credit for not only research articles, but associated research data and code. This is central to the concept of reproducible research. Now, all of these outputs can be used to generate new hypotheses. Data and donuts will cover topics from across the research cycle, starting with how to write a data management plan.

END TRANSCRIPTION