Comparing Data Management Techniques in the Natural Sciences

Brenda Jonsson

iSchool, University of Arizona

LIS 671: Introduction to Digital Curation and Digital Collections

Dr. Zack Lischer-Katz

December 15, 2021

Comparing Data Management Techniques in the Natural Sciences

In his book *Letters to a Young Scientist* (2014), Pulitzer prize winner Edward O. Wilson prophesies that the future of science is interdisciplinary--not just within the sciences, but also within the arts and humanities. He says, "...the attempt to make such linkages will be a key part of intellectual life in the remainder of the twenty-first century" (p. 63). It is difficult for the sciences to be interdisciplinary since it takes so many years of formal training to gain expertise in any one area, and so specialization is expected. Nonetheless, Wilson is not the only one who believes that there are great things to come with more interdisciplinary work. Are there opportunities to share data, and data management tools and techniques in a cross-disciplinary manner? To look for some of these possibilities, this paper will compare and contrast different areas of scientific research data management. The disciplines that this paper will compare include astronomy, herpetology, neuroscience, oceanography, chemistry, and materials science.

History.

The history of modern astronomical data collection starts with observations made using telescopes, although people have depended on data from the observation of the sky for agricultural, religious, and navigational purposes for millennia (Borgman, 2015, p. 84). Astronomy's projects went from being done by single observers to being collaborative, long-term, and ambitious, requiring a large investment of resources (p. 87). By the late 20th century, astronomy had switched from analog data to digitally-collected data, eventually performing to a point beyond human readability (p. 87-88).

Materials science is an interdisciplinary field that studies the properties and creation of physical materials. It has many different specialities like tomography, combinatorial chemistry, and crystallography. There is a strong industrial influence in this science.

Chemistry has arguably existed since at least the ancient Egyptians' time, but there are still many manuscripts and books preserved from the time of alchemy on.

Neuroscience data started off as anecdotes and physicians' notes about dissections of the brain and healing wounds from the battlefield, but it got more sophisticated with the invention of the microscope and the discovery that electricity affects nerves ("History of neuroscience," 2021). The discovery of neurons is generally attributed to a 19th century anatomist named Santiago Ramón y Cajal.

Oceanographic data has only been collected for about 200 years. Oceanographers collect data on the sea's depth, temperature, salinity, current, waves, color, transparency, ice, and the area's temperature, pressure, humidity, wind speed, precipitation, cloud, fog data and more (Xie et al., 2019, p. 114).

The history of herpetological data collection in large part consists of collecting specimens and documenting the animals' natural history. In the paper this paper looked at, herpetology was not found to be one of the sciences that have data-intensive research. Arguably, the three most well-known that do are astronomy, the biosciences, and particle physics (Hill et al., 2016, p. 399), but the other disciplines researched in this paper tend to describe themselves as much more data-intensive.

Current State of Research Data Management.

In astronomy today, there are many agreements about standards for data structures. One of the main tools astronomy uses are telescopes. Telescopes capture digital images. They obtain signals on a variety of wavelengths, like x-ray, microwave, ultraviolet, etc. (pp. 91-93). However, images taken at wavelengths outside of human vision are assigned colors, and the way

these images are colorized vary widely. So a lot of these popular images are rejected for publication in research papers (Borgman, 2015, p. 95).

Most astronomical observations that were acquired from government-funded missions are available to the public (but some information is unavailable due to its military applications). The CDS in France coordinates a lot of astronomy's research data. However, most astronomy researchers' data is not in an institutional archive. Most information is emailed to colleagues (p. 96-102).

Astronomical data sets are attractive to computer science researchers for its volume, consistency, and lack of human subjects. Now, much astronomical data collection has to be automated. The amount of code researchers release varies widely (pp. 100-106).

Tools used to collect and manage data in herpetology today include 3D printing, laser scanning, photogrammetry, and VR (virtual reality) or AR (augmented reality) simulations (Pesado & Aciti, 2019, p. 91). These tools are mainly designed to create replacements for the more valuable and easily damaged preserved specimens.

Oceanographic data comes from observations collected through sensors, ships, and satellites, and computer simulations (Xie et al., 2019, p. 114). Oceanography has begun to globalize and turn into a federation of institutions that are trying to share research data (Baker & Chandler, 2008, p. 2133).

Materials science uses a lot of algorithms to understand materials, and so machine learning is an important tool in the field (Hill et al., 2016. p. 400). There is an emphasis on learning how to streamline data-intensive research for the sake of maximizing profits in materials science (Himanen, 2019, p. 1).

Neuroscience has to deal with large patient samples collected from multiple hospitals, centers, and countries (Pozamantir et al., 2008, p.25). Some of the data collected includes medical history, biochemical and genetic tests, structured neurological examinations, neurological tests, and MRI data (p.26). It also makes computer simulations of the brain and processes experimental data with computers. Bouchard et al. argues that the current mismanagement of data from high-performance computing in neuroscience is holding the field back from making new discoveries (2016, p. 628). New technologies enable recording brain signals simultaneously using multiple recording techniques, and for longer periods of time, producing 100s of terabytes of data (p. 629). Bouchard emphasizes the need for managing multi-modal data since there is a need to record audio, video, movement, and haptic data alongside brain data (p. 630).

When it comes to chemistry, Chen & Wu (2017) found in a survey of 119 chemistry graduate students and researchers that (1) The most common type of data is experimental; (2) The more common formats of data include spectograms, experimental instrument test data, and test photos or pictures; (3) By far the most popular ways to record data are paper notebooks, followed closely by electronic documents, and these are easily lost or destroyed (pp. 347-348).

Themes and differences.

There are many areas in which data management between the sciences is similar, and a few in which they are different.

Algorithms.

The need to constantly develop new algorithms to analyze or interpret data is common across every discipline. Algorithms vary in their reusability. New code tailored to each situation often needs to be programmed for each new project.

Big data as a force to be reckoned with.

Data intensive science is referred to as "the fourth paradigm," with the other three being experiments, theory, simulation (Hill et al., 2016, p.399). Science today is often focused on extracting "knowledge… from datasets that are too big or complex for traditional human reasoning" (Himanen et al., 2019, p. 1).

Informatics.

Informatics is an interdisciplinary field that must always merge with another discipline outside of information science. It can be described as the data-based modeling of behaviors (Hill et al., 2016, p. 400). Informatics can also coordinate the intake of information. (Baker & Chandler, 2008, pp. 2139-2140). A lot of sciences depend on informatics to coordinate their data management efforts.

A large difference--privacy.

A discipline like neuroscience in which patients are involved has to think about how to protect private information (Pozamantir et al., 2008, p. 25). There are many legal restrictions when working with human subjects that mean that data must only be released to people at the level at which they are authorized.

Interdisciplinary potential.

The clearest potential for interdisciplinary data to come together is in education for the general public--especially on the internet. There are many scientific databases available to the public online, and there could be more, and more made in a layman-friendly fashion.

Another area of great potential is in creating science-based VR and AR environments. Multiple sensors can be combined to give more information about what's happening. It might be possible to take many different sources of data to make increasingly more accurate models of the

world. Multiple sensors generally increase the accuracy of whatever is being measured (Hall et al., 2009, section 1.2). It is not unreasonable to extrapolate that simulations could also benefit from more accurate measurements. Perhaps advances could be made in complex networks about complex adaptive systems when a more open system can be looked at (when more variables and more, more accurate data can be added).

Another area of potential for interdisciplinary data is in the ways that librarians respond to the needs of researchers. Across all of the papers reviewed, the needs were similar. A lot of training could be given to researchers for how to manage data from the very start of their project, and how to manage their data so that it can be submitted to publishers. Chen & Wu (2017) say that the library can offer researchers "stable and safe long-term storage, data locating and tracking, data association and discovery among scientific research findings…" Every discipline looked at for this paper could benefit from those things.

In 5-10 years.

The most solvable controversies about research data management revolve around open access and standardization. It is reasonable to assume that the scientific community will work on these problems in the next 5-10 years.

Regarding open access, whereas librarians have ideals about the free flow of information baked into their professional ethos, many scientists are unwilling to share their data. In materials science, manufacturing data are often regarded as trade secrets (Hill et al., 2016, p. 405). Hill also states that "it is not clear that making one's research data broadly available will lead to any of the following: (1) enhanced impact and more citations for one's work; (2) improved funding opportunities; or (3) improved chances at professional advancement and promotion." (p. 403). So not every field has the incentive to share.

Astronomy is one example of really excellent data sharing and data management in science, but even it has its challenges. Not all of the data makes it to a repository, and astronomy will withhold data for the sake of security (Borgman, 2015, p. 102). There are several other, generally less serious incentives to manage data in a less than transparent way--but they are still incentives. Hill lists such examples as researchers wanting to keep exciting results private so others don't publish those results, and researchers wanting to withhold negative results. Prestigious publishers also want to generate more revenue by holding materials behind a paywall (p. 401). Nonetheless, many sources of funding and other organizations are working towards making scientific papers available to anyone with an internet connection.

In 5-10 years, we may start seeing more cooperation between institutions--even internationally, more standardization of terminology and metadata, and more government legislation outlining how research data must be managed. Hill states that the governments of the UK and Ireland have been modifying copyright laws to facilitate text and data mining (Hill et. al., 2016, p. 404). Hill also states that the US has directed national research agencies to make federally-funded research publicly available, and that the European Commission has also been pushing for broader access to the research the government pays for (p. 404). Himanen (2019) states that all European funding for 2020 and beyond requires Open Access (p. 3).

Conclusion.

The most important interdisciplinary potential that all of these disciplines had was to borrow from each other were their ideas to create standard infrastructures for research data management: the management of the observational data, analytical data, and publication demands that arise out of scientific research. The areas in which libraries can help researchers are in the following five areas of the data lifecycle: the data management plan, data generation and

collection, data recording and processing, data preservation and backup, and data publication and sharing (Chen & Wu, 2017, pp. 352).

There are a lot of common complaints across disciplines about the way research data is managed now. Bouchard (2016) states that "the single greatest impediment to fully extracting the return on investment into neuroscience data collection is the lack of community alignment and coordination around standards for experiments, data, and metadata" (p. 630). This lack of standardization is a problem for all fields reviewed in this paper except for herpetology, but perhaps that is because the herpetologists in Pesado & Aciti's paper seemed focused on education. Biology is widely considered one of the sciences that are data-intensive. Hill's paper basically rephrases what Bouchard is saying as "five main barriers" to broader data sharing: buzzwords in informatics, idiosyncrasies in workflows, conflicting stakeholders, limited standards for data, and a lack of incentives to share (Hill et al., 2016, p. 399).

The survey of chemists yielded the following concerning results for librarians to keep in mind as they design their research data management programs: (1) Only 2% of respondents preserved their data in a repository. Almost all data was stored in personal computers, paper notebooks, USBs or hard drives, or the research group's computer; (2) 42% of respondents did not meet the requirements for submitting the data requested by academic journals; (3) About 10% of respondents either did not want to share their data or did not know if they wanted to; (4) Very few researchers knew about publishers' requirements for data; (5) Very few researchers knew about the existence of even the most common data repositories in their field; and most importantly for librarians, (6) About 70% of the researchers said that they need the library to help them with research data management (Chen & Wu, 2017, pp. 349-351).

References

Baker, K.S., & Chandler, C. L. (2008). Enabling long-term oceanographic research: Changing data practices, information management strategies and informatics. *Deep-Sea Research. Part II, Topical Studies in Oceanography, 55*(18), 2132–2142. https://doi.org/10.1016/j.dsr2.2008.05.009

Borgman, C. L. (2015). *Big data, little data, no data : scholarship in the networked world*. M.I.T. Press.

Bouchard, K. E., Aimone, J., Chun, M., Dean, T., Denker, M., Diesmann, M., Donofrio, D., Frank, L., Kasthuri, N., Koch, C., Ruebel, O., Simon, H., Sommer, F., & Prabhat. (2016). High-Performance Computing in Neuroscience for Data-Driven Discovery, Integration, and Dissemination. *Neuron (Cambridge, Mass.), 92*(3), 628–631. https://doi.org/10.1016/j.neuron.2016.10.035

Chen, X., & Wu, M. (2017). Survey on the Needs for Chemistry Research Data Management and Sharing. *The Journal of Academic Librarianship, 43*(4), 346–353. https://doi.org/10.1016/j.acalib.2017.06.006

Hall, D.L., Liggins, M. E., & Llinas, J. (2009). *Handbook of multisensor data fusion : theory and practice.* CRC Press. https://learning.oreilly.com/library/view/handbook-of-multisensor/9781351835374/xhtml/ch00_fm03_title.xhtml

Hill, J., Mulholland, G., Persson, K., Seshadri, R., Wolverton, C., & Meredig, B. (2016). Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin, 41*(5), 399–409. https://doi.org/10.1557/mrs.2016.93

History of neuroscience. (2021, November 4). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=History_of_neuroscience&oldid=1053474624

Pesado, P., & Aciti, C. (2019). *Computer Science – CACIC 2018 24th Argentine Congress, Tandil, Argentina, October 8–12, 2018, Revised Selected Papers* (1st ed. 2019.).

Pozamantir, A., Lee, H., Chapman, J., & Prohovnik, I. (2008). Web-based Multi-center Data Management System for Clinical Neuroscience Research. *Journal of Medical Systems, 34*(1), 25–33. https://doi.org/10.1007/s10916-008-9212-2

Wilson, O. (2014). *Letters to a Young Scientist* . Liveright. https://smile.amazon.com/gp/product/0871403854

Xie, C., Li, M., Wang, H., & Dong, J. (2019). A survey on visual analysis of ocean data. *Visual Informatics, 3*(3), 113–128. https://doi.org/10.1016/j.visinf.2019.08.001