JESLIB 2012; 1(3): 148-158
doi:10.7191/jeslib.2012.1019

Journal of eScience Librarianship
putting the pieces together: theory and practice

Lurking in the Lab: Analysis of Data From Molecular Biology Laboratory Instruments

Jen Ferguson

Northeastern University, Boston, MA, USA

Abstract

Objective: This project examined primary research data files found on instruments in a molecular biology teaching laboratory. Experimental data files were analyzed to learn more about the types of data generated by these instruments (e.g. file formats) and to evaluate current laboratory data management practices.

Setting: This project examined experimental data files from instruments in a teaching laboratory at Brandeis University.

Methodology: Experimental data files and associated metadata on instrument hard drives were captured and analyzed using Xplorer^2 software. Formats were categorized as proprietary or open, and characteristics such as file naming conventions were noted. Discussions with the faculty member and lab staff guided the project scope and informed the findings.

Results: Files in both proprietary and open formats were found on the instrument hard drives. 62% of the experimental data files were in proprietary formats. Image files in various formats accounted for the most prevalent types of data found. Instrument users varied widely in their approaches to data management tasks such as file naming conventions.

Conclusions: This study found inconsistent approaches to managing data on laboratory instruments. Prevalence of proprietary file formats is a concern with this type of data. Students express frustration in working with these data, and files in these proprietary formats could pose curation and preservation challenges in the future. Teaching labs afford an opportunity for librarians interested in learning more about primary research data and data management practices.

Introduction

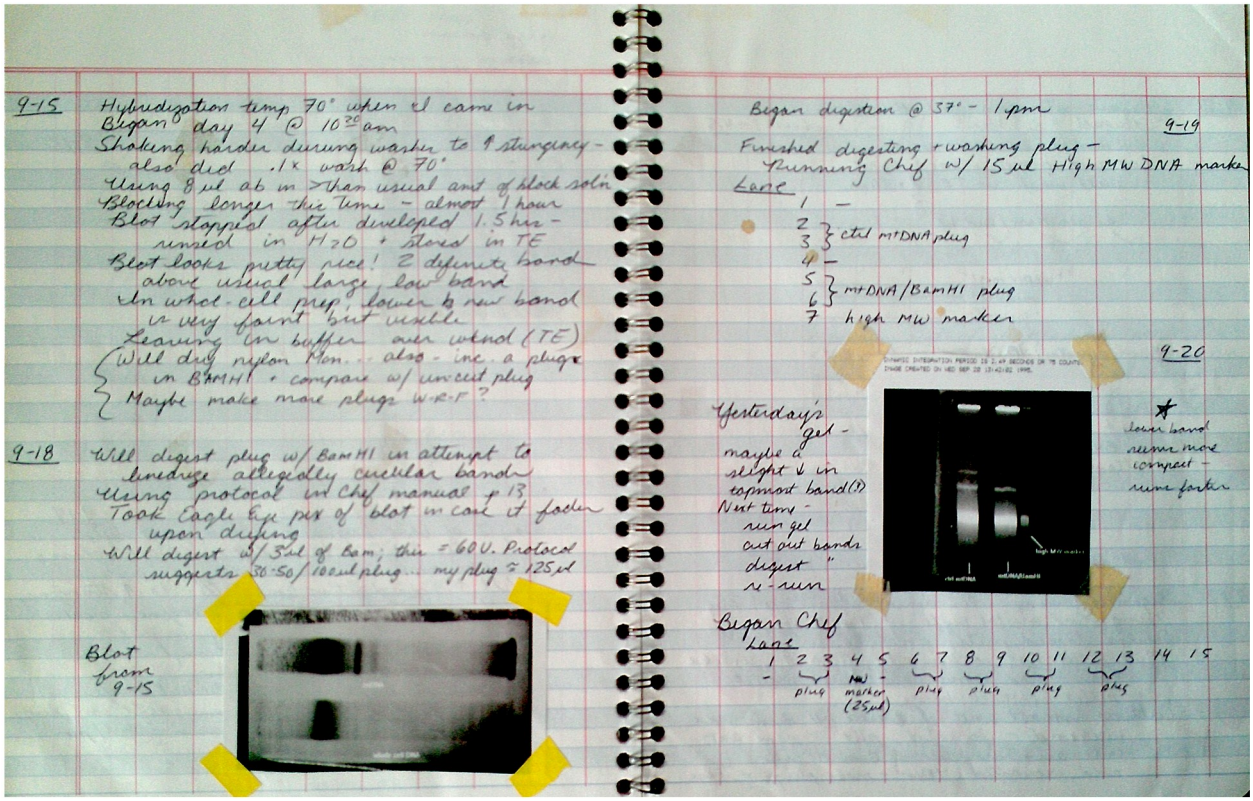
"Those digital technologies that have dramatically accelerated the process of science have often rendered the curation of data in context more difficult." -Frey 2008

Librarians and libraries are working to expand their traditional roles as caretakers of the intellectual outputs of their institutions. For example, librarians are beginning to apply their skills in information organization to managing research data. Most librarian efforts to date have focused on research data far along in the data lifecycle, such as data deposited in support of published findings. This published data typically progresses through many iterations of refinement and derivation, and can be quite different from its starting material, the primary research data. Primary research data can be thought of as the raw, measurable output that results from manipulating experimental conditions. Very little primary data will ultimately be published, but these are the data that science faculty, staff, and students are tasked with managing daily. For this reason, some familiarity with the characteristics of primary

Correspondence to Jen Ferguson: jen@jasf.us
Keywords: primary research data, data curation, data preservation, file format, proprietary, teaching, management

JESLIB 2012; 1(3): 148-158
doi:10.7191/jeslib.2012.1019

Figure 1: The state of laboratory information management, circa 1995. The author's lab notebook shows typical practice at the time, with written notes interspersed with lab instrument outputs. This represents a midpoint in the shift from entirely handwritten recording, to reference within the notebook to data stored elsewhere – i.e. in digital form on laboratory instruments.



research data is helpful for librarians interested in research data management.

This paper describes a project undertaken to learn more about the current state of primary research data at its source – the type of data lurking in the lab. The author generated and managed primary research data during previous work in molecular biology, but was uncertain if the characteristics of experimental data had changed since her days at the lab bench. Accordingly, this project was initiated to gain knowledge of the types of experimental data files generated by lab equipment, and to determine whether proprietary file formats were still common given growing calls for increased data sharing, openness, and research transparency.

The state of primary research data challenges can be viewed as a convergence of factors intersecting within the current climate of data management, openness, and preservation. One factor is researcher adherence to good lab practices and clear provenance of work. The characteristics of modern laboratory instruments are another factor. The following section describes these factors in more detail.

Good lab practice and provenance

From the time that they first step into a laboratory, science students are trained in organizational and record-keeping methods that librarians and archivists would readily recognize as practices in provenance. Good practices in keeping a lab notebook have traditionally been the best way for scientists to not only recall and record their results, but to

also demonstrate provenance of their work. A well-kept lab notebook does more than just jog the memory of the person who wrote it; ideally, it also contains enough context to allow another researcher to read the notebook in the future and replicate the experiment using the information given. The descriptive information that has traditionally been memorialized in lab notebooks can prove valuable for decades and disciplines far beyond the original work. For example, as Fox and Hendler (2011) note, this context is crucial for data visualizations derived from the original work, so that errors and variations can be understood.

Laboratory notebooks have historically been handwritten accounts. When additional technology entered the laboratory setting, lab notebooks evolved into a combination of written notes and pasted-in outputs of laboratory instruments, such as printouts and photographs (Figure 1). As laboratory technology became more sophisticated, the outputs of these instruments increasingly remained solely in digital form. In some situations – for example, a lengthy DNA sequence – data may simply be too large and unwieldy to paste into a notebook. In other cases, the instrument itself may be needed to view the data. Accordingly, many lab notebooks have shifted back to a handwritten state in which the outputs remain on the instrument, and the notebook merely contains the names of the experimental data files. When good laboratory notebook practice is followed, a link still exists between the description of the experiment (in the notebook) and the experimental results (on the hard drive). However, since those two pieces of work are no longer collocated, the link between them becomes much more tenuous. As Frey (2008) observes, this situation strips results of context, and makes reproducibility considerably more difficult. At worst, this practice can even represent a break in provenance: without the lab notebook providing the experimental context and answers to key questions of who, what, where, and when, data files on the hard drive become useless.

Laboratory instrument characteristics

Analytical and laboratory instrumentation is big business. Pharmaceutical and biotech equipment is an $11 billion per year industry, and is the largest sector within the $45 billion global instrumentation market (Thayer 2012). These instruments are crucial components of scientific research infrastructure. Labs are stocked with a variety of this analytical equipment, all measuring experimental outputs in support of a focus of study. In many cases these laboratory instruments are connected to hard drives that collect and store data generated in the course of experimental procedures.

Laboratory instruments can gather and analyze experimental data using proprietary software supplied by the instrument vendor. The experimental data generated by these machines, in turn, are often in proprietary file formats. These formats may even change with each software release (Bowen 2005). Proprietary file formats are inherently difficult to work with anywhere other than the originating instrument, not to mention share and preserve into the future. This proprietary ecosystem makes it difficult for researchers to link data from multiple instruments together, and can create barriers to data sharing and reuse. Laboratory equipment manufacturers, like those of any other industry, are constantly subject to mergers, acquisitions, and liquidations. When a lab equipment manufacturer folds, its proprietary systems often quickly become extinct. If a company acquires or merges with another manufacturer, it may or may not choose to support the proprietary systems it inherits from the transaction.

Laboratory instruments are expensive and thus are often shared between individual scientists, several lab groups, and sometimes even external researchers. Subsequently, there can be many users of an instrument, but in some cases no clear caretaker of that

instrument. Lab instruments can be connected to hard drives that are off the campus IT grid, so these computers may be overlooked by typical campus infrastructure support such as backup, maintenance, and virus protection. Additionally, this equipment is often purchased with grant funds that may vanish. As Dorothea Salo (2010) notes, a short-term, project-based approach to scientific research, as driven by the cyclical nature of grant funding, is at odds with values of long-term access and stewardship.

These situations - provenance issues posed by the separation of lab notebooks from instrument outputs, and laboratory instruments shared by many people, gathering data in formats that are often proprietary - create data management headaches for scientists. They can also make it difficult for scientists to comply with growing calls for data sharing and openness. It's helpful for librarians interested in research data management to have some knowledge of the characteristics of primary research data. Thus, this project was launched to learn more about the current state of primary research data lurking in the lab.

Methods

Project discussions and scope

A biology faculty member was approached with a proposal to examine and analyze the data files generated by instruments in her molecular biology teaching lab. The proposal was framed as a research project, with a goal of learning more about the characteristics of primary research data. After a discussion about the data management practices in the lab, the faculty member readily granted access to her laboratory instruments, noting that she could also potentially benefit from the findings. She was particularly interested in learning more about the naming conventions and file organization strategies used by her students, as this information could help address recurring issues with students unable to locate their experimental data files. Accordingly, these criteria were included in the investigation.

File captures

File management software programs were evaluated as possible tools to capture information from the laboratory instrument hard drives. The file manager Xpolorer^2 (Zabkat Software 2011) was chosen for its ease of use, customizable views, and the availability of a portable version. The portable version of Xplorer^2, installed on a flash drive, provided a way for quick and noninvasive 'snapshots' to be taken of the lab instrument hard drives for later analysis, while avoiding the need to install software directly on each hard drive. Information was captured from hard drives dedicated to four laboratory instruments: an Agilent Technologies atomic force microscope, a LI-COR NEN DNA analyzer, a Bio-Rad Gel Doc XR, and a Hitachi fluorescence spectrophotometer.[1]

Analysis of data characteristics

Xplorer^2 was used to flatten directory structures and sort the hard drive 'snapshots' by various file characteristics and metadata elements such as creation date, file name, size, and file extension. File extensions were researched using various sources. File-extensions.org (2012) was the richest source found for information on the formats encountered in this study. File formats were categorized as proprietary or open/standard based on definitions such as this one from openformats.org (2010): "A proprietary format encodes data in such a way that a file will only [*sic*] readable with the original software used to create it." Information on student data management practices, such as

1. The atomic force microscope is a very high-resolution microscope, used to examine and manipulate surface features of biologics at nanoscale. The DNA analyzer is used for DNA sequencing, PCR reactions, and DNA analysis by gel electrophoresis. The Gel Doc is an imaging system used to document and analyze nucleic acids and proteins. The fluorescence spectrophotometer is used in the teaching lab to analyze recombinant proteins.

JESLIB 2012; 1(3): 148-158
doi:10.7191/jeslib.2012.1019

Table 1: Experimental data file formats found on the lab instruments, listed in order of prevalence. Note that two instruments (the Gel Doc and LI-COR) share one hard

File Format	Instrument	Number of Experimental Data Files
.samp	Gel Doc/LI-COR	244
.jpg	Gel Doc/LI-COR	157
.mi	Atomic force microscope	109
.dx	Fluorescence spectrophotometer	93
.1sc	Gel Doc/LI-COR	70
.fds	Fluorescence spectrophotometer	65
.tif	Gel Doc/LI-COR	41

Figure 2: The most common experimental data file formats found across all four laboratory instruments.

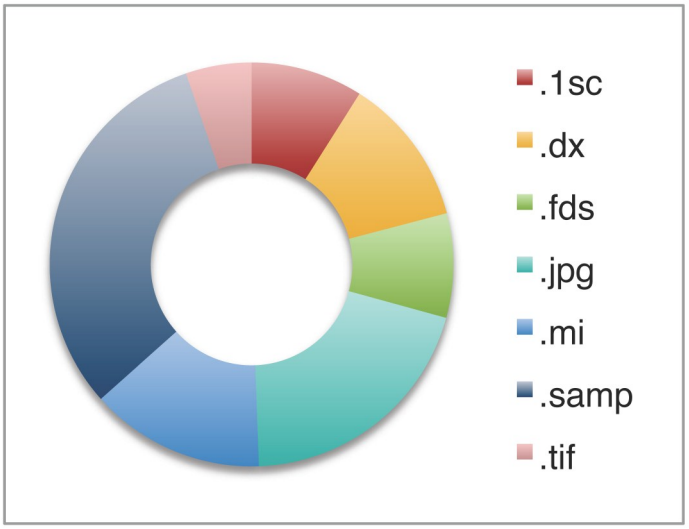
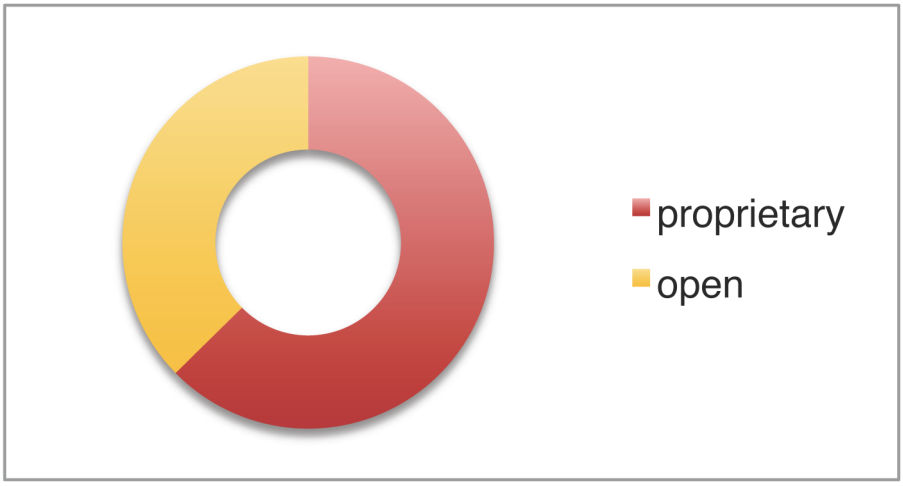


Figure 3: Percentage of proprietary vs. open formats for experimental data files.



file naming conventions and storage locations, was also noted by examining the hard drive 'snapshots.'

Results

"...collecting scientific data is often difficult and instrument-specific. As a result, most scientific data is created in a form and organization that facilitates its generation rather than focusing on its eventual use." - Fox and Hendler 2011

Experimental data file characteristics

A total of 779 experimental data files, comprising seven different file formats, were found on the lab instrument hard drives. These file formats are listed in Table 1.

The prevalence of each file format was expressed as a percentage of the total number of experimental data files found (Figure 2). Image files in various formats were the most common type of experimental data found across the four laboratory instruments.

File formats were categorized as proprietary or open based upon definitions available at openformats.org. 62% of the experimental data files found were in proprietary formats (Figure 3).

Experimental data from these instruments can be stored as packages comprised of several different file types (NMSU 2006). For example, DNA sequence data can consist of a bundle of files including a sequence text file (.txt), a curve file (.scf), and gel image files (.samp).

One instrument (the LI-COR DNA analyzer) had relatively few data files stored on its hard drive, considering how heavily the instrument is used and how long it has been in use by the teaching lab. The answer to this mystery was revealed in the operating manual for the instrument (LI-COR 2006). The LI-COR uses hard-drive space as temporary storage only, and continuously overwrites its oldest data sets. Further research revealed that certain laboratory instruments can be configured to save experimental data to cloud storage provided by the instrument vendor.

Data management practices

Faculty and staff observed that their students were sometimes frustrated when working with experimental data. The faculty member mentioned how often she received "frantic 3 a.m. emails" from students unable to insert their experimental results into PowerPoint slides, or even open the experi-

Figure 4: A snapshot of data management practices. File names given by students are shown for a sampling of .1sc files, illustrating the variety of naming conventions used. Note that in some cases, attempts are made to provide experimental context within file names - e.g. dates, organism strain ('wt706'), and number of base pairs ('60bp').

ura gel 4 Tues G and Wed R	1sc
ura Gel 5 - 3.10.11	1sc
ura Gel 6 - 3.10.11	1sc
ura Mon R	1sc
Restriction Digest 5.26.11 DNA 1 & 2	1sc
awesome 2010	1sc
awesomer 2010	1sc
Deb 2010-03-09 yeast gel	1sc
7.22.10 Gel 1=60bp wt706 Tmp Tm & cycle	1sc
7.22.10 Gel 2=60bp Temp-wt001& Neal	1sc
7.22.10 Gel 2=60bp wt001 & Neal Temp	1sc
7.22.10 Gel 3=100bp wt706 Tmp, Tm &cycle	1sc
7.22.10 Gel 4=100bp wt001 & Neal Temp	1sc
ura gel 1 Mon R starred	1sc
ura gel 2 Monday Ravenclaw un-starred	1sc
Dpn Gel 5 10 WT	1sc
a	1sc
attractive	1sc
group	1sc
joe	1sc
lab 2010-10-04 StatiionC	1sc
lab 2010-10-04 StationB	1sc

mental images at all once the students were away from the lab and the proprietary systems that generated the data. Data files could be converted to other formats in order to accomplish these tasks, but students were not always aware that this step was possible or necessary.

Teaching lab faculty and staff also noted challenges in organizing files so students could efficiently locate them. They cautioned their students to give their data meaningful names to ensure retrieval. However, they could not closely monitor each individual due to the sheer number of students in the teaching lab. Thus, students largely devised their own file naming and organization conventions. File naming conventions and management practices observed in this study were as diverse as the people who generat-

ed the data. In some cases, teaching assistants gathered all their student groups' data into dedicated folders. Some students attempted a form of version control by including the date of the experiment in their file names, and generally used naming practices that would help ensure easy rediscovery. Others took less rigorous approaches to this task (Figure 4).

Data management practices also varied from one lab instrument to the next. On some hard drives a login was necessary to store data files, which were then saved in a user's dedicated storage space. On other instruments, no user-level login was required, so data files of many users were comingled. Certain instruments dictated the path where data files were stored. These circumstances made it difficult to ensure consistent approaches to data management across instruments.

Discussion

"I have a rather old file recorded on a Biorad GelDoc unit, saved as .1sc file. As you can imagine, I have quite a problem opening the file as the old GelDoc system is already gone for a while and the new one does not support the old file formats (instead creates .scn!). Thanks Biorad! " - Johannes-P. Koch 2011

Experimental data files in both proprietary and open formats were found in analysis of lab equipment hard drives. 62% of the experimental data files found on the hard drives were in proprietary formats.

Image files made up the bulk of the experimental data and were usually in open formats, while other types of experimental data files tended to be in proprietary formats. As noted by teaching lab faculty and staff, the proprietary formats created obstacles to students working with the experimental data.

Given these findings, several factors pose potential challenges to use, sharing, management, and preservation of this type of experimental data:

- Automatic overwriting and cloud storage of data sets
- Data consisting of bundles of different file types
- Proprietary file formats

Automatic data overwriting and cloud storage of data sets are attractive options for those tasked with managing research data on a daily basis. Research data sets can be quite large, and data overwriting and cloud storage solutions ensure that storage space is always available. Most of this primary data will not become part of communicated results or publications, so perhaps the convenience of nearly unlimited storage space is worth the risk of lost data. Vendors make their data storage solutions even more enticing by offering features such as data visualization tools and data sharing options within their cloud storage (e.g., Life Technologies 2011). However, data overwriting and cloud storage are potentially worrisome from a data curation and preservation standpoint. Given a business climate in which lab equipment manufacturers are constantly being acquired, merged, and even going out of business, what happens to data stored in the cloud when a vendor ceases to exist? As grant funding requirements and institutional research data policies continue to more stringently dictate researchers' data management practices, will the practice of overwriting data sets be viewed as being in compliance with those requirements?

Bundled data consisting of a collection of different file types also present challenges. In one case, DNA sequence data can consist of a sequence text file (.txt), plus a curve file (.scf), and gel image files (.samp). In this example, some elements of the bundle are in proprietary formats (e.g., .samp) and some are open (e.g., .txt). The relationship between these files could be difficult to elucidate for a curator/preservationist, or even for a scientist attempting to locate these files on a hard drive or replicate results.

The biggest challenge is the prevalence of proprietary file formats throughout the lab instrument ecosystem. The barriers to use and re-use of these files become immediately apparent when trying to learn more about these file formats. Internet searches using these file extensions as keywords inevitably yield hits on email lists and forum posts from people asking how to open, work with, and/or convert the files. In some cases, newer versions of a vendor's proprietary ecosystem are not even backward compatible with older versions of the same instrument, as noted in the quote at the beginning of this section. Many of the barriers imposed by these proprietary file formats can be overcome with time, effort, and research, but the process is rarely simple, and information can be lost in the solution. In instances where data file conversion is possible, the tools available tend to be commercial solutions that are also proprietary and operating system-dependent. UNIX/Linux-based tools do exist, but the source code behind these tools is not open, so their methods and algorithms are unclear (Wenig and Odermatt 2010). Scientists may be reluctant to use these workarounds due to cost, operating system barriers, and/or lack of clarity regarding conversion tool algorithms. Conversion, even when possible, is no panacea, as format conversion may result in the loss of crucial provenance and contextual information regarding the experimental conditions that produced the data. Simply put, proprietary file formats create difficulties for students and researchers who work with, share, and manipulate experimental data.

Scientific researchers appear to be going to great lengths to work around the barriers created by proprietary formats. Research for this project found web resources such as Bio-Formats (2012) devoted to addressing the problem by offering guidelines for converting proprietary experimental data files into more malleable formats, and posts by graduate students writing and sharing scripts to convert one file format into another. Why are scientists tolerating this state of affairs rather than putting pressure on equipment manufacturers to change their ways? Some scientists (i.e. Linkert et al. 2010) are voicing concern about the situation, but energies largely seem devoted to finding workarounds rather than engaging directly with equipment manufacturers who use proprietary systems. One possibility is that those who purchase the equipment (faculty and administrators) tend to be less impacted by proprietary formats than those who work more directly with the instruments and data (students, postdocs, and research staff).

The lesson of JCAMP-DX

While proprietary systems are common, some laboratory instrument manufacturers have elected to use open standards for their data formats. One example is the file format .dx, or JCAMP spectroscopic data exchange format, a data type collected by the fluorescence spectrophotometer in this study. The open .dx format reflects the work of the spectral portability subcommittee of the Joint Committee on Atomic and Molecular Physical Data, or JCAMP. This group was formed in the late 1980s to address problems with accurate sharing of spectral data and, with the cooperation and assistance of infrared spectrometer manufacturers, chose a standard external form for data exchange. Their design included forward-thinking criteria such as "acceptability by a wide variety of computers, communication systems, and storage media" and "expandability of each data field to whatever length is required; ability to add new data fields as the need arises" (McDonald and Wilks, 1988). It is impressive how well those guidelines endure today, considering how drastically the state of storage media and computers has changed in the decades since the standard was created. This persistence is especially remarkable in reading the list of cooperating businesses, which includes many companies that are now defunct or are no longer manufacturing spectrometers. Still, their file portability legacy has trickled down to their successors. Will

other equipment manufacturers follow the example of JCAMP-DX, or will proprietary systems rule the lab equipment ecosystem? Could future grant funding requirements put pressure on researchers to work with open file formats, thus, in turn, pressuring equipment manufacturers to comply? Given the growing emphasis on access to, and preservation of, research data, it will be interesting to see what develops.

Implications for libraries and librarians

Despite exceptions such as the .dx format, this investigation found that a majority of primary research data files were in proprietary formats. Assuming that this situation is typical of research data in other settings and disciplines, proprietary systems are common in laboratories. These systems offer very real benefits to scientists, so it's likely that these proprietary ecosystems will persist. Unless the movement toward data openness and transparency begins to heavily impact primary research data, libraries may be best served by prioritizing the stewardship and caretaking of other types of institutional intellectual outputs.

However, there is still value in understanding the characteristics of primary research data. For librarians interested in data management, some familiarity with the qualities of research data is very helpful. This knowledge is also useful for liaison librarians to better understand the worlds of the faculty and researchers they work with, and can in turn help inform strategic planning for libraries.

Teaching laboratories can afford a great opening for librarians interested in becoming more involved in research data management. Teaching labs typically enroll freshman and sophomore level science majors and train these students in current laboratory techniques. These labs can have significant data management challenges due to high student volume and turnover. Teaching lab faculty and staff struggling to cope with this data may be receptive to advice and assistance. These labs could also provide an opportune point of instruction in data management for science students during what is often their first exposure to laboratory research. From a logistical standpoint, teaching labs also tend to have more predictable quiet cycles in the course of an academic year than research labs. These are times in which investigations into data management or curation efforts will be minimally disruptive.

References

Bowen, Andrew. "Beyond LIMS: The Integrated Data Pipeline." Tessella Scientific Software Solutions: V1.R0.M0, 2005. Accessed July 19, 2012. http://www.tessella.com/wp-content/uploads/2008/04/beyondlims.pdf

File-Extensions.org. "File Extension Library." Accessed July 18, 2012. www.file-extensions.org.

Fox, Peter and James Hendler. "Changing the Equation on Scientific Data Visualization," *Science* 331, no. 6018 (2011): 705-708, http://dx.doi.org/10.1126/science.1197654

Frey, Jeremy. "Curation of Laboratory Experimental Data as Part of the Overall Data Lifecycle," *International Journal of Digital Curation* 3, no. 1 (2008): 44-62, http://dx.doi.org/10.2218/ijdc.v3i1.41

Koch, Johannes-P. 2011. ".1sc Biorad Format." To ImageJ mailing list. [imagej@list.nih.gov]. Accessed July 19, 2012. http://imagej.1557.n6.nabble.com/1sc-biorad-format-td3682149.html

LI-COR, Inc. "Operator's Manual: NEN Model 4300 DNA analyzer." Last revised April 2012. Accessed July 17, 2012. http://biosupport.licor.com/docs/4300_OpMan_08591.pdf

Life Technologies. "AB LifeScope." 2012.

Accessed July 27, 2012. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/ngs-data-analysis-software/lifescope.html

Linkert, Melissa, Curtis T. Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, Josh Moore, Carlos Neves, Donald MacDonald, Aleksandra Tarkowska, Caitlin Sticco, Emma Hill, Mike Rossner, Kevin W. Eliceiri, and Jason R. Swedlow. "Metadata Matters: Access to Image Data in the Real World," *Journal of Cell Biology* 189, no. 5 (2010): 777-782, http://dx.doi.org/10.1083/jcb.201004104

LOCI, University of Wisconsin-Madison. "Bio-Formats." Last modified September 25, 2012. Accessed July 19, 2012. http://loci.wisc.edu/software/bio-formats

McDonald, Robert S. and Paul A. Wilks, Jr. "JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form." *Applied Spectroscopy* 42 no. 1 (1988): 151-162.

New Mexico State University. "Li-Cor DNA Seq Services." Last modified March 2006. Accessed July 27, 2012. http://research.nmsu.edu/molbio/MOLB_2_Sequence/NMSU%20%7B%7BIDNA_Seq_Services%7D%7D.htm

Openformats.org. Accessed July 17, 2012. http://www.openformats.org/

Salo, Dorothea. "Retooling Libraries for the Data Challenge," *Ariadne* 64 (2010). http://www.ariadne.ac.uk/issue64/salo

Thayer, Ann. "Top Instrument Firms," *Chemical & Engineering News* 90, no. 18 (2012): 2-17.

Wenig, Phillip and Juergen Odermatt. "OpenChrom: A Cross-Platform Open Source Software for the Mass Spectrometric Analysis of Chromatographic Data," *BMC Bioinformatics* 11 (2010): 405, http://dx.doi.org/10.1186/1471-2105-11-405

Zabkat Software. Xplorer2 portable edition (Version 1.8.1) [Software]. 2011. http://zabkat.com/x2port.htm

Acknowledgments

The author thanks M. Kosinski-Collins and D. Bordne of Brandeis University for generously providing guidance, insights, and access to their laboratory equipment.

Disclosure: The author reports no conflicts of interest.

ISSN 2161-3974