JLSC

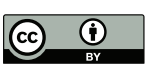
ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 4, General Issue (2016)

The Journal Article as a Means to Share Data: a Content Analysis of Supplementary Materials from Two Disciplines

Jeremy Kenyon, Nancy Sprague, Edward Flathers

Kenyon, J., Sprague, N., & Flathers, E. (2016). The Journal Article as a means to Share Data: a Content Analysis of Supplementary Materials from Two Disciplines. *Journal of Librarianship and Scholarly Communication,* 4(General Issue), eP2112. http://dx.doi.org/10.7710/2162-3309.2112



ISSN 2162-3309 10.7710/2162-3309.2112

RESEARCH

The Journal Article as a Means to Share Data: a Content Analysis of Supplementary Materials from Two Disciplines

Jeremy Kenyon
Research Librarian, University of Idaho
Nancy Sprague
Science Librarian, University of Idaho
Edward Flathers
PhD Candidate, University of Idaho

INTRODUCTION The practice of publishing supplementary materials with journal articles is becoming increasingly prevalent across the sciences. We sought to understand better the content of these materials by investigating the differences between the supplementary materials published by authors in the geosciences and plant sciences. **METHODS** We conducted a random stratified sampling of four articles from each of 30 journals published in 2013. In total, we examined 297 supplementary data files for a range of different factors. **RESULTS** We identified many similarities between the practices of authors in the two fields, including the formats used (Word documents, Excel spreadsheets, PDFs) and the small size of the files. There were differences identified in the content of the supplementary materials: the geology materials contained more maps and machine-readable data; the plant science materials included much more tabular data and multimedia content. **DISCUSSION** Our results suggest that the data shared through supplementary files in these fields may not lend itself to reuse. Code and related scripts are not often shared, nor is much 'raw' data. Instead, the files often contain summary data, modified for human reading and use. **CONCLUSION** Given these and other differences, our results suggest implications for publishers, librarians, and authors, and may require shifts in behavior if effective data sharing is to be realized.

External Data or Supplements:

Kenyon, J.; Sprague, N.; Flathers, E., 2016, "Data from: The journal article as a means to share data: a content analysis of supplementary materials from two disciplines", http://dx.doi.org/10.7910/DVN/DCSMGP, Harvard Dataverse.

Received: 10/22/2015 Accepted: 02/29/2016

Correspondence: Jeremy Kenyon, 875 Perimeter Drive, Moscow, ID 83844-2350, jkenyon@uidaho.edu

IMPLICATIONS FOR PRACTICE

1. While there are differences between the supplementary materials of authors in the geosciences and plant sciences, it may be possible to standardize approaches and policies regarding these materials across different fields.
2. Few supplementary files take the form of machine-readable data, i.e. data that require no human intervention to use, even if they are often considered data. Using supplementary materials as a vehicle for data sharing should be done critically, employing best practices similar to those utilized by data repositories.
3. File size and file type are not good indicators of the actual content of supplementary materials. In most cases, files must be manually opened to know what and how much content is stored there.

INTRODUCTION

Imagine two objects in front of you. On the left, a flowering plant. On the right, a piece of granite. Consider the study of each as a scientific phenomenon. The plant and the rock possess fundamental differences in structure and composition, and they tie into natural systems in different ways. While similarities exist at the chemical and physical levels, the study of botany and geology has ultimately led to specialized instrumentation and methods. From these material differences come normative differences in scientific behavior—practices, protocols, and communication—which formalize as distinct communities through journals and conferences. Further still, as the scientific network increases, the fields codify institutionally as separate academic disciplines (Lemaine, Macleod, Mulkay, & Weingart, 1976).

If disciplines vary materially and methodologically, does it not follow that the differences between fields like plant sciences and geology may extend to the form of their research products? After all, DNA sequencing is a technique used in the study of plants, yet not so much in the study of the earth. Conversely, volcanology and studies of plate tectonics utilize different instrumentation than is found in plant sciences. The growing area of research data management services has focused on identifying and clarifying these differences, in part due to the implications that differences in data have for management, curation, and re-use (Carlson & Brandt, 2015; Palmer, Weber, Munoz, & Renea, 2013). Supplementary materials—the catch-all term for those "extra" items provided along with journal articles—provide an opportunity to look for these differences. Authors often lack editorial guidance on the format and preparation of supplementary materials. Subsequently, the materials can vary widely, including extensive tables of data, illustrated figures, text, videos, and computer code. Increasingly, authors use supplementary files as a vehicle for sharing research data

(Pop & Salzberg, 2015). This point in particular raises questions about the role of these materials in the scholarly communication environment.

To investigate, we pursued the following research question: what are the differences between the supplementary materials published by earth scientists and life scientists? This line of questioning emerged from a previous project that showed that journal authors in both the earth and life sciences are increasingly adding supplementary materials to their journal articles (Kenyon & Sprague, 2014). To better characterize the differences in supplementary materials, we conducted a content analysis of a sample from fifteen high impact journals in two representative areas—geology and plant sciences. We selected these two fields primarily because of the very high tendency of authors publishing within them to use supplementary materials, as shown in our prior study. Further, since geology and plant sciences represent binary aspects of the natural world—the biotic and abiotic; we found the comparison curious —how might this rising tide of materials differentiate? Our findings have implications for the development of journal policies on supplementary materials as well as on data sharing and data re-use, specifically when the vehicle for sharing is the journal article.

LITERATURE REVIEW

The past decade has brought a series of new funding agency requirements for data management plans, and policies mandating public access to research data from federally funded projects. These developments have resulted in more journals requiring and facilitating the publication of online supplements of data and methods, as well as the increasing use of data repositories (Bishoff & Johnston, 2015; Ray, 2014; Tenopir et al, 2015). While the challenges and complexities of sharing scientific research data through data repositories have been discussed in depth in numerous recent papers (Borgman, 2012; Caetano & Aisenberg, 2014; MacMillan, 2014; Tenopir et al., 2011; Thessen & Patterson, 2011), relatively few studies have focused specifically on the role of journal supplementary materials as a data sharing option. It should be noted for the present discussion that supplementary data sharing is done primarily through the journal website, with few exceptions. A link in a paper to an external data repository would be considered sharing through a data repository.

Pham-Kanter, Zinner, and Campbell (2014) surveyed 1165 life science researchers about data sharing, and found that online journal supplements, along with third-party data repositories and funding agency mandates, were perceived as having had a significant effect on facilitating data sharing. While funding agency policies were found to have the greatest impact, 35% of respondents reported that journal publication policies had a major impact on increasing data sharing. The authors also reported that 58% of survey respondents valued online data and methods supplements as being helpful for their research. One of this study's

most surprising findings was a high degree of compliance with journals' requirements related to sharing of methods and data: 92% of respondents reported always having submitted, when required, a detailed description of their methods as an online supplement. Among the benefits mentioned for researchers, supplementary materials were valued for helping to decrease the cost of data sharing. Researchers could make the data available once, rather than responding to and fulfilling multiple individual requests.

A recent study by Womack (2015) compared data sharing practices in articles from the top ten journals (ranked by 5-year impact factor) in the disciplines of biology, chemistry, mathematics, and physics. Using a sample of 50 articles containing original data from each of the four disciplines, he investigated what types of data were included and whether they were publicly available. Womack reported some significant data sharing differences among these disciplines, with biology having the highest rate of data sharing (42.9%), and physics having the lowest (8%). He also noted that many of the articles in the sample included supplementary data in PDF format, which is not easily re-usable by other researchers. In other cases, graphics were used to summarize the research data, but authors didn't provide access to the underlying data. Overall, he found that only 13% of the articles in his sample made the data available in a format that would be useful to other researchers.

Another recent study that highlighted the role of supplementary materials in data sharing was Wiley's Researcher Data Insights Survey (Ferguson, 2014), which included responses from more than 2,250 researchers world-wide. Of the 52% of total respondents who said they make their research data publicly available, 67% claimed that they share their data as supplementary material in a journal, while 57% use data repositories (either institutional, discipline-specific, or general purpose) and 37% share their data via webpages. This survey also provided a comparison of the ways different disciplines share their data. Supplementary material in a journal was a key data sharing option reported by both life scientists (76%) and physical scientists (69%). Journal requirements were found to be among the top motivating factors for data sharing, along with increasing the impact and visibility of their research, and public benefit.

The Wiley survey also covered the file formats that the researchers shared (Meadow, 2014). Most respondents (82%) reported producing data in spreadsheets and CSV files; 38% created two-dimensional images, and 12% used 3D images; 22% created executable code or models; and 11% generated video or audio recordings. File sizes were smaller than expected, with 60% less than 10GB. These findings were similar to what we discovered in our prior studies of supplementary materials: researchers tend to use common workplace productivity tools for data sharing.

METHODS

Data Collection

Our study seeks to identify the differences in character of supplementary materials between articles published in geology and the plant sciences. To begin, we used Journal Citation Reports to generate a list of the top fifteen journals in 2013 in the two corresponding areas defined by JCR: geology and plant sciences. No single measure of impact is considered sufficient to completely describe the importance of a journal, as impact is a multi-dimensional phenomenon (Bollen, Van de Sompel, Hagberg, & Chute, 2009). With that in mind, we selected the Eigenfactor Score, another citation impact metric, for obtaining lists of significant journals in both fields (Tables 1 and 2). The Eigenfactor Score is a sufficient selection method for this study in that it adjusts for the variability in citations across different disciplines and uses the entire citation network across many disciplines to generate its ranking (Kim & Hong, 2015). From this set of 30 key journals, we mapped those articles that contained supplementary material in 2013 by working through each issue and volume of each journal for the year and recording whether or not the article contained supplementary materials. This 'map' of which articles contained supplementary material was used to define the population from which our sample was drawn (see supplementary CSV files for complete data).

Journal Title	Rank	Eigenfactor Score
Plant Physiology	1	0.09982
Plant Cell	2	0.08843
Plant Journal	3	0.07599
New Phytologist	4	0.06492
Journal of Experimental Botany	5	0.05215
Plant and Soil	6	0.02775
Journal of Ecology	7	0.02769
Journal of Natural Products	8	0.02694
Journal of Ethnopharmacology	9	0.0265
Annals of Botany	10	0.02593
Plant, Cell, and Environment	11	0.02446
Phytochemistry	12	0.02413
Theoretical and Applied Genetics	13	0.02143
Plant and Cell Physiology	14	0.02135
American Journal of Botany	15	0.02051

Table 1. Plant Sciences Journals used in the study

Journal Title	Rank	Eigenfactor Score
Journal of Geophysical Research	1	0.34302
Geophysical Research Letters	2	0.23074
Nature Geoscience	3	0.07616
Geology	4	0.05779
Journal of Hydrology	5	0.04608
Quaternary Science Reviews	6	0.04288
Palaeogeography, Palaeoclimatology, Palaeoecology	7	0.03582
Biogeosciences	8	0.02781
Geomorphology	9	0.02672
Global Biogeochemical Cycles	10	0.02238
Journal of Volcanology and Geothermal Research	11	0.02064
Hydrology and Earth System Sciences	12	0.02048
Geological Society of America Bulletin	13	0.01858
Annales Geophysicae	14	0.01756
Global and Planetary Change	15	0.01733

Table 2. Geology Journals used in the study

Using data from a pilot study performed in 2013, we performed several power analyses to determine how many articles we would need to sample from each group. Our goal was to achieve a significance level of 0.05 and a power of 0.95 when comparing values such as the number of supplementary files per article and the size of supplementary files across the groups. We determined that sixty articles per field were needed for a 95% confidence rate. Based upon the number of journals available to us, we conducted a random stratified sampling of four articles from each journal in 2013 to obtain sixty articles for each discipline (geology/plant sciences). We stratified our sampling method in order to avoid skewing the results towards a specific journal and the potential sub-disciplines represented by it. The first two authors each reviewed one of the two categories of journals. We reviewed each article, downloaded and opened all of the accompanying supplementary material files, and recorded detailed data for each file. Upon completion of the initial data collection, we switched categories, then reviewed and verified the other's results. In the few instances of dispute or uncertainty about categorization of data types, we highlighted those records and met later to discuss them, coming to mutual agreement in order to maintain consistency.

For each individual supplementary material file, we recorded several key variables:

- File extension. File extensions were recorded as listed on the filenames themselves

in the form in which they were downloaded. If uncertain, we checked with the National Digital Information Infrastructure and Preservation Program Format Descriptions (Library of Congress, 2015) or examined the file in a text editor to verify the type where possible.

- File size. File size was recorded at the kilobyte level.
- Multiplicity of types. A given file often contained more than one type of content, such as both a chart and a list of references. Single or multiple content type was recorded as a binary category of zero for single and one for multiple.
- Categories. The categories of content required the most preparation. We searched for and did not identify a known taxonomy of supplementary material content types, and so we derived a taxonomy based on our samples. The major category types included data, multimedia, tables, text, and charts. 'Data' was differentiated from 'table' to represent content that was purely machine-readable (proprietary software files, spatial data, code) versus those that were structured for human consumption (often containing human-oriented formatting, such as merged cells, highlighting, text emphasis). Within each of these broad categories, we also identified more specific data types, such as illustrations, maps, references, and code. We consulted an array of sources, including the Oxford English Dictionary, various dictionaries of statistics and other disciplines to develop a list of definitions for the categories. Those definitions are listed in the associated supplementary readme file.

Analysis

We tested several categories based on assessment of their distributions. For the number of files per article, we conducted a two-sample Kolmogorov-Smirnov (K-S) test to look for a difference in the location or shape of the distributions. The purpose was to identify whether the apparent range in the number of files per article was significant. For the major categories of content types, we conducted Pearson's chi-squared test for categorical datatypes to identify whether there were significant differences between the two fields. Lastly, for a test of file size, we again used a two-sample K-S test to look for difference. Given that the distributions were not normally distributed (see Figure 5 in results), we log-transformed the data, verified normal distributions using a Shapiro-Wilk test for normality ($p < 0.01$), then conducted a Welch two-sample t-test to compare the means, and an F-test to compare variance. Test results are discussed below, and the details and data are available in the associated supplementary R file.

LIMITATIONS

We recognize numerous limitations evident in our methods. First, our sample of fifteen journals in two disciplines does not represent all supplemental materials in the fields, nor do they represent the practices of every author. Rather, our study identifies patterns that may be indicators of trends, which need to be confirmed or rejected with larger or more comprehensive studies.

Also, we did not distinguish between 'raw' and 'summary' data. We recognize that the distinction can introduce complications to our results, e.g. that raw data forms may be larger or may be in specialized formats, while summary data is more likely to be smaller and formatted explicitly for presentation. Differentiating between the two would have greatly expanded our methodology, our data collection efforts, and the length of the study.

RESULTS

Supplementary materials sections often contain multiple files; therefore, we analyzed 297 individual files in our sample of 120 articles. The number of supplementary material files accompanying a given article in our study ranged from one to twenty-three (Figure 1). Over half of the articles contained only one supplemental file, with just a slight variation between disciplines. However, many of these single files were PDFs that actually contained several different types of supplemental materials, such as charts, imagery, and text combined. For the geology articles in our sample the maximum number of supplements was 9, while the plant science articles had as many as 23 supplements per article.

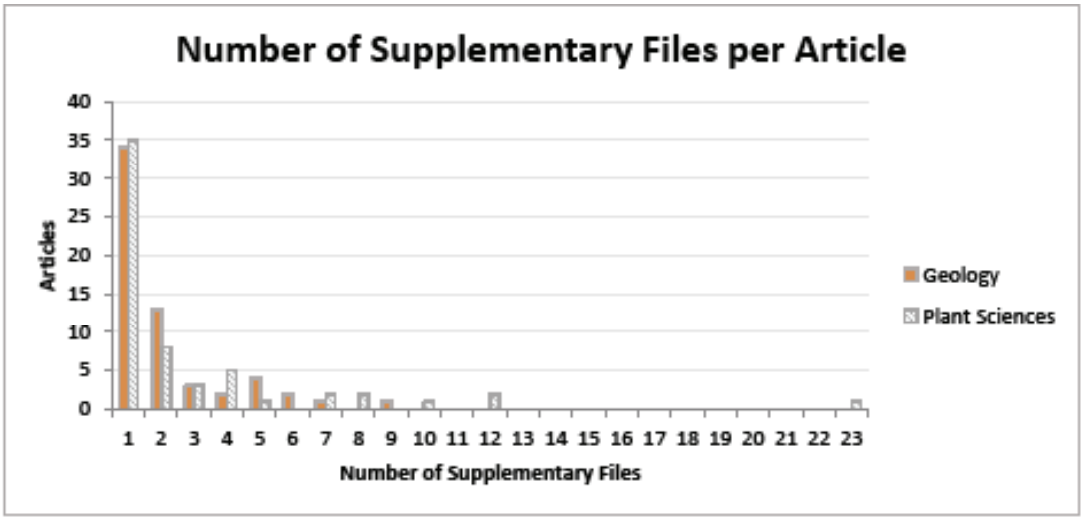


Figure 1. The number of files contained in an article's supplementary materials section, by discipline.

Given the two similar distributions, it appears plant sciences may generally contain more files on average—particularly with articles that contained 10, 12, and 23 supplementary files. We recorded a mean of 2.08 files per article for geology and 2.87 files per article for plant sciences. A two-sample K-S test determined that there is not a significant difference between the number of files per article between the two fields as represented by our journals ($D = 0.1$, $p = 0.9251$).

In the 297 files, we identified approximately 848 different supplementary content items, which we classified into categories. Figure 2 presents the major content type categories of supplementary materials identified for both disciplines. In this case, a Pearson's chi-squared test ($p = 7.6e\text{-}09$) rejected the hypothesis that there was no difference between the two fields. The graph clearly shows that authors in plant science journals produced nearly twice the formatted tables than authors in geology. Multimedia, a category that contains numerous content types (including microscopic imagery) contained more than twice as many items for plant sciences as for geology. One category where geology dominated was data—largely due to the presence of GIS files, computer code, and other machine-readable formats. Geologists also frequently used supplementary materials to add textual information. These were observed primarily as references and methods sections, as seen in Figure 3.

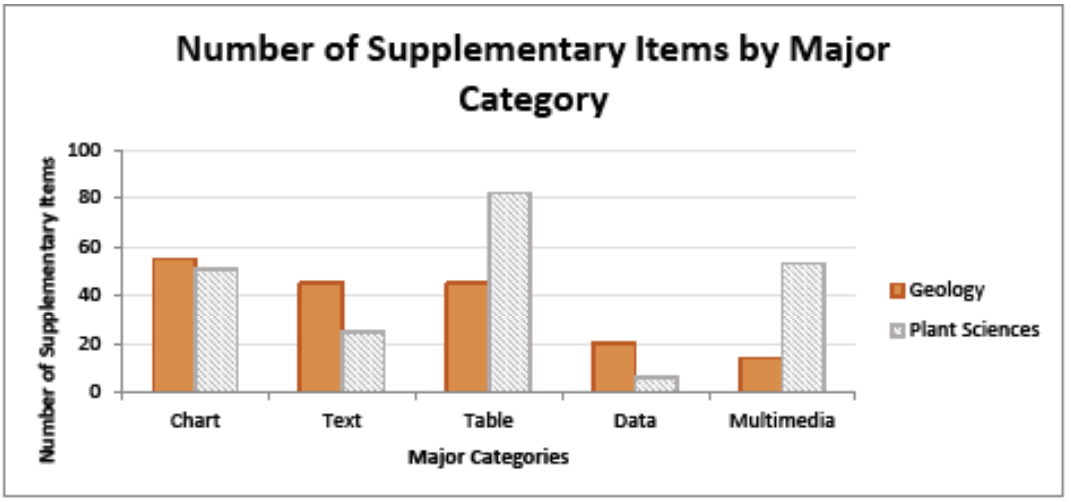
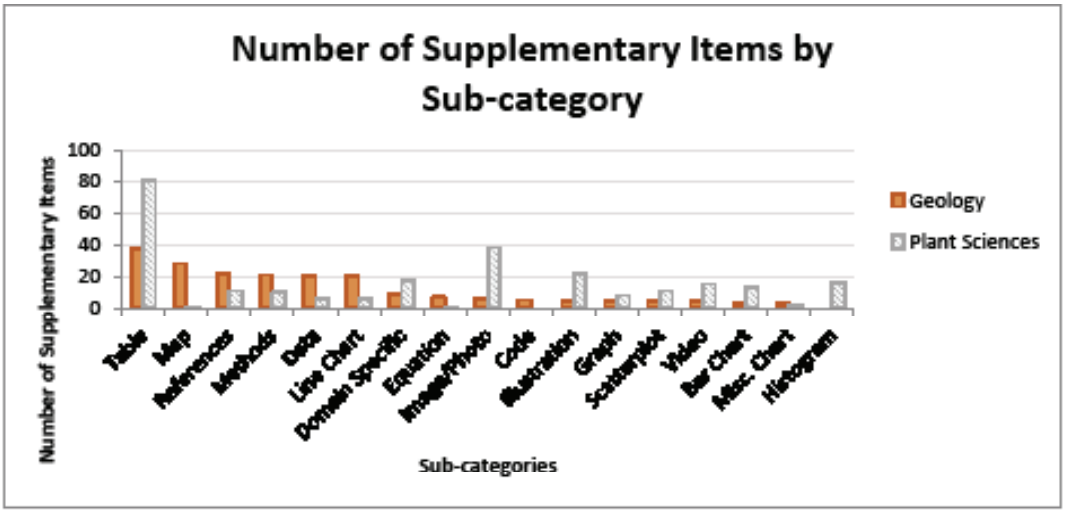


Figure 2. The major categories present in the supplementary files, by discipline. The chart is ordered from largest to smallest (by geology).

Figure 3 provides a comparison of more specific types of content that we found in the supplemental materials, such as maps, images, and videos. Each of the sub-categories was associated with a major category, which are detailed in the readme file accompanying the

supplementary files. However, at this more granular level we can see some of the possible reasons for the differences detected in Figure 2. Illustrations, photos, and images were used more frequently in plant science articles, while maps were much more prevalent as supplements to the geology articles. The category "domain-specific" refers to those types of visualizations that are derived from laboratory or field instrumentation, e.g. chromatograph or seismometer readings. 'Graph' was narrowly defined as a visualization representing a network, such as a dendrological tree.



to smallest values in geology.

Figure 4 details the set of file extensions for all the supplementary materials in our study. Many content types can be found in a variety of file formats. Tables, the most popular type observed, may be present as Excel spreadsheets or in PDFs. The PDF is the most popular format for either discipline. Combined, however, the Microsoft Office products of Excel, Word, and PowerPoint dominate plant sciences and show strongly in geology. Since some of the individual PDF files we examined contained multiple types of data (such as tables, text, and images), the file extension was not always a good indicator of the type of content present in the file. We noted some difference between the two disciplines in the use of more specialized file extensions, such as .kmz files, which were used only in geology articles and .mov files which were found only with the plant science articles.

File sizes for the supplementary materials in our study ranged widely: from 1 kilobyte to over 60 megabytes for the geology supplementary files and from 9 kilobytes to over 29 megabytes for the plant science files (Figure 5). The majority of the 172 files in plant sciences clustered around the 100kB-1MB range, while the majority of the 125 geology files skewed toward smaller files sizes than plant science. Our geology sample had a median file size of 87kB, while plant sciences possessed a median of 256kB. We found that the two fields yielded different means via the Welch t-test ($p = 0.0003$), with plant sciences the larger of the two. The F-test also reported a difference in variance ($p = 0.003$), where geology indicates a larger variance between the two.

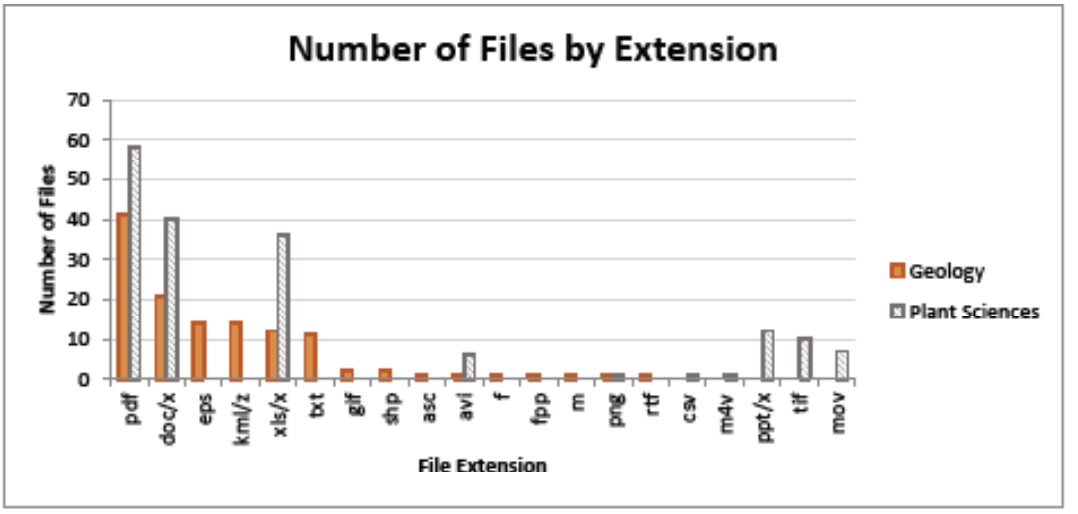
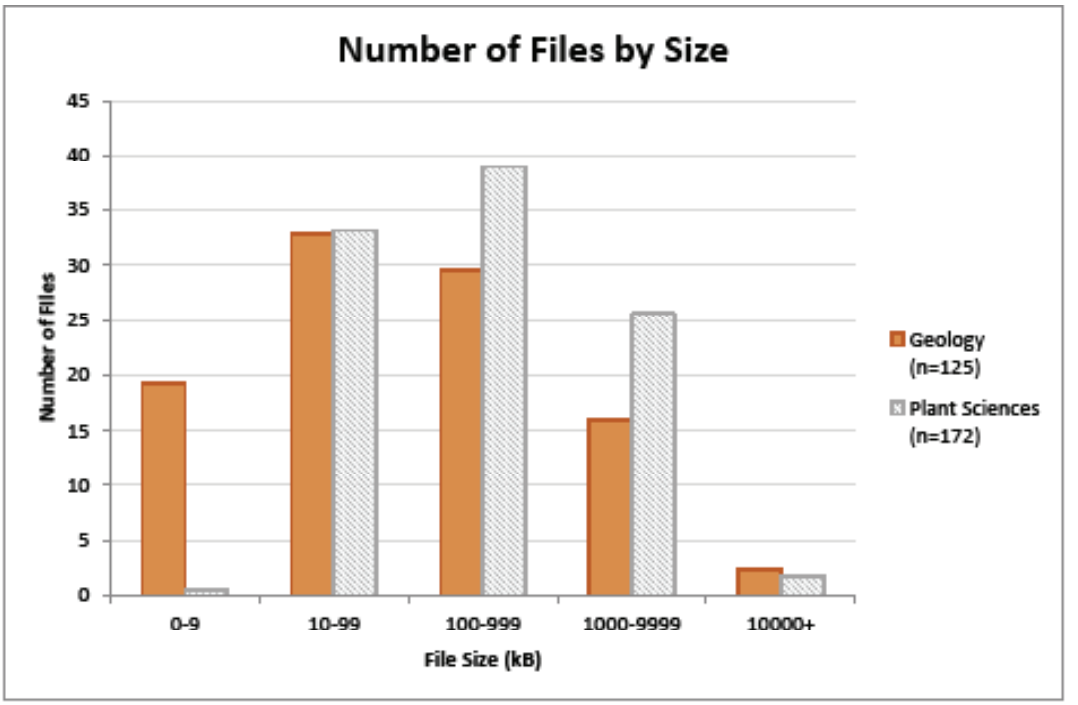


Figure 4. File extensions of the supplementary materials, by discipline. The chart is ordered from largest to smallest in geology.

File sizes for the supplementary materials in our study ranged widely: from 1 kilobyte to over 60 megabytes for the geology supplementary files and from 9 kilobytes to over 29 megabytes for the plant science files (Figure 5). The majority of the 172 files in plant sciences clustered around the 100kB-1MB range, while the majority of the 125 geology files skewed toward smaller files sizes than plant science. Our geology sample had a median file size of 87kB, while plant sciences possessed a median of 256kB. We found that the two fields yielded different means via the Welch t-test ($p = 0.0003$), with plant sciences the larger of the two. The F-test also reported a difference in variance ($p = 0.003$), where geology indicates a larger variance between the two.



DISCUSSION

Our research question was: what are the differences between the supplementary materials of these two fields? In the number of files distributed with an article, our sample failed to identify significant difference. Often, we observed the bundling of different content types into single files, resulting in the logical conclusion that while most supplementary materials sections are limited to only 1-3 files (see Figure 1) that does not necessarily prevent an author from delivering a diverse array of content. The occasional outlier, such as one article that had 23 files, still follows this pattern, with some of the files containing multiple content types.

The array of content types provides a much more interesting set of considerations. First, our data does not record a prevalence of sharing code or instrumentation outputs, perhaps supporting the rhetoric on the need for better data and code sharing. Geology authors appeared to share code more often, but not at such a rate that appears to support any notion of frequent or consistent sharing. Instead, it is much more likely that scripts and code are not generally shared. This is particularly notable in plant sciences, where these content types were absent from our sample.

Second, we saw a significant amount of tabular data, most of which was configured primarily for human consumption (thus appearing in the table sub-category, not the data sub-category). This suggests that much of it may be summary data, rather than raw data. The purpose of summary data is to present an easier-to-digest form of the data, so this would support the idea that plant sciences are extensively using formats amenable to presentation, such as Excel spreadsheets or PDFs. In both disciplines, only a small amount of data was presented in machine-readable form, while the rest would require manipulation simply to extract it from its original source. The supplementary materials contained a wealth of information, but it could be difficult to find, access, and re-use.

Third, file size poses a potential challenge to any infrastructure that seeks to store, preserve, and maintain access to information. From our results, it appears that neither field is sharing massive data files through supplementary materials. The global mean for both fields was approximately 1.4MB. Notably, plant sciences appears to deliver larger files, but for geology, there is a wider range of file sizes. This may be due to the fact that the file extensions (Figure 4) for plant scientists show a higher use of formats that carry significant format bloat, i.e. Excel, Word, and PDFs. Geology, on the other hand, used code formats and text more often, which can pack information into little more than ASCII encoded characters.

The largest files we encountered, 57 and 60MB, were stored in a separate supplementary materials data repository for the Geological Society of America. Very few of the journals stored supplementary material on a separate repository website, so both these large data files and their storage location are an unusual case. A majority of the journals presented, and as far we could tell, stored, supplementary materials through the journal's website. This differentiation between the treatment of large and small files may suggest that institutional or disciplinary data repositories could provide a niche service that would complement data sharing through journal supplementary material.

While we found expected differences, we also noted substantial similarity between the two fields, particularly in the formats used. Our findings suggest that the orthodoxy of contemporary scholarly writing and publishing dominates the mode of formally published data. In retrospect, this makes sense. Authors often submit manuscripts produced with word processing software. Microsoft Excel and PowerPoint still dominate the environments for managing tabular data and creating presentation-ready summaries of information, respectively. It is logical then, that as earth and life scientists navigate the publishing process, they submit material for publication that defers to established modes of written communication. Our results show that whether the scientist produces chromatograph readings, a map, or an illustration of plate tectonics, these materials will likely be shared via common file formats that do not require specialized software to read.

What this means for research data, and the use of supplementary materials as a method of sharing data, depends on one's view of the utility of various data formats and types. Is a Word document the best way to share computer code? Should maps be shared as PowerPoint slides? Should tables be shared within a non-editable PDF? There is a reason why data scientists and technologists are developing tools to deal with the 'problem' of tables buried in PDFs or using image classification techniques to identify discrete document types within digitized images (Aristan, Tigas, & Merrill,2015; Lorang, Soh, Datla, & Kulwicki, 2015). These tools represent an effort to impose machine-readability on preservation-friendly formats (such as PDF), which are difficult to use for computational purposes, such as indexing, processing, and analysis (Library of Congress, 2015).

CONCLUSION

Our analysis of 297 supplementary data files from the earth sciences and plant sciences in 2013 shows some differences between the two fields. In each case, the differences are most apparent in the types of content contained within the files. Earth scientists have a tendency to provide access to more machine-readable data—such as GIS files—whereas life scientists have a tendency to share tabular material. In both disciplines, the files are usually fairly small, and the formats are Microsoft Office files or PDFs. Whether a researcher is studying flowering plants or rocks, the fundamental differences in the disciplines do not extend significantly to the form of the research products presented as supplementary materials.

This places a burden on publishers, data managers, librarians, and others who support researchers in sharing their research products as functional data. For publishers, these results suggest that data sharing through supplementary materials is not likely to pose problems for infrastructure or of usability. A user capable of reading common file formats will likely be able to open files shared via supplementary materials. The size of the files does not appear to be a problem, either, although it could be a consideration for high-volume publishing. Size could also be an issue in the case of a large number of files per article, but it appears that is not often the case. All of this suggests that a publisher of a multi-disciplinary journal, or one that publishes journals in multiple fields, can very likely use the same policies governing file formats and sizes.

For librarians or data managers, there are two primary implications. First, the storage for these materials is unlikely to pose much problem. Organizations hosting open-access journals may not encounter storage problems, although there may be other considerations regarding whether or not to allow supplementary materials. Second, discovery of these materials is a problem. Our analysis shows that content packed into unrecognizable formats, without sufficient guidance, can be virtually impossible to identify without manually opening and

reviewing the files. If a user is looking for examples of high-resolution imagery in plant pathology research, those images may be difficult to find if they are shared as a relatively small PDF file. Others have noted the problem as well, specifically on "hidden" lists of references, which we observed most frequently in geology (Garcia-Perez, 2015). Hidden content poses a challenge to systems designed to count, index, or provide discovery of scholarly information, and therefore to the users of such systems.

The underlying principles of data sharing suggest that its goal is to make data discoverable, preservable, and ultimately, re-usable (ICSU-WDS, 2015). Otherwise, what would be the point of sharing datasets in addition to a write-up of the research results? One of the key constraints in re-using data is the time and labor required to prepare the data for other uses. Estimates suggest that between 50% and 80% of the time spent on data reuse is devoted to this problem (Lohr, 2014). Sharing data in a more purposeful and careful manner will not completely obviate the need for human intervention (activities such as integration, fitting data into a specific model, etc.), but it can remove some of the "friction" (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011). The publishing system can accommodate this need, but only by supporting the shift from accepting supplementary materials without review to imposing expectations for formats, documentation, and ultimately, functional data.

REFERENCES

Aristan, M., Tigas, M., & Merrill, J. B. (2015). Tabula: Extract tables from PDFs. Retrieved from http://www.tabula.technology

Bishoff, C., & Johnston, L. (2015). Approaches to data sharing: An analysis of NSF Data Management Plans from a large research university. *Journal of Librarianship and Scholarly Communication*, 3(2), 1-27. http://dx.doi.org/10.7710/2162-3309.1231

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS One*. doi: http://dx.doi.org/10.1371/journal.pone.0006022

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059-1078. http://dx.doi.org/10.1002/asi.22634

Caetano, D. S., & Aisenberg, A. (2014). Forgotten treasures: the fate of data in animal behaviour studies. *Animal Behaviour*, 98, 1-5. http://dx.doi.org/10.1016/j.anbehav.2014.09.025

Carlson, J., & Brandt, D.S. (2015). Data Curation Profiles Directory. Retrieved from http://docs.lib.purdue.edu/dcp/

Edwards, P. N., Mayernik, M., Batcheller, A., Bowker, G. C. & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41(5), 667-690. http://dx.doi.org/10.1177/0306312711413314

Ferguson, L. (2014). How and why researchers share data (and why they don't). Retrieved from http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/

Garcia-Perez, M. A. (2015). Online supplemental information: a sizeable black hole for citations. *Scientometrics*, 102: 1655-1659. http://dx.doi.org/10.1007/s11192-014-1348-x

International Council for Science – World Data System (ICSU-WDS). (2015). *WDS Data Sharing Principles*. http://dx.doi.org/10.5281/zenodo.34354

Kenyon, J., & Sprague, N. (2014). Trends in the use of supplementary materials in environmental science journals. *Issues in Science and Technology Librarianship,* 75. http://dx.doi.org/10.5062/f40z717z

Kim, Y. S., & Hong, R. (2015). About the Eigenfactor Project. Eigenfactor.org. Retrieved from http://www.eigenfactor.org/about.php

Lemaine, G., Macleod, R., Mulkay, M. & Weingart, P. (1976). *Perspectives on the emergence of scientific disciplines*. Berlin: Walter de Gruyter. http://dx.doi.org/10.1515/9783110819038

Library of Congress. (2015). National Digital Information Infrastructure Preservation Program. *Sustainability of Digital Formats Planning for Library of Congress Collections*. Retrieved from: http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml

Lohr, S. (2014). For big-data scientists, 'janitor work' is key hurdle to insights. *The New York Times*, Aug 18, p. B4, New York Edition.

Lorang, E., Soh, L. K., Datla, M. V., & Kulwicki, S. (2015). Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Magazine*, 21(7/8). http://dx.doi.org/10.1045/july2015-lorang

MacMillan, D. (2014). Data sharing and discovery: What librarians need to know. J*ournal of Academic Librarianship*, 40(5), 541-549. http://dx.doi.org/10.1016/j.acalib.2014.06.011

Meadow, A. (2014). To share or not to share? That is the (research data) question. *The Scholarly Kitchen*, November 11. Retrieved from http://scholarlykitchen.sspnet.org/2014/11/11/to-share-or-not-to-share-that-is-the-research-data-question/

Palmer, C. L., Weber, N. M., Munoz, T., & Renear, A. H. (2013). Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data. *Archives Remixed*, 3 (Summer). Retrieved from: http://www.archivejournal.net/issue/3/archives-remixed/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/

Pham-Kanter, G., Zinner, D. E., & Campbell, E. G. (2014). Codifying collegiality: Recent developments in data sharing policy in the life sciences. *Plos One*, 9(9). http://dx.doi.org/10.1371/journal.pone.0108451

Pop, M. & Salzberg, S. L. (2015). Use and mis-use of supplementary material in science publications. *BMC Bioinformatics*, 16, 237-240. http://dx.doi.org/10.1186/s12859-015-0668-z

Ray, J. M. (2014). Research data management: practical strategies for information professionals (Charleston insights in library, archival, and information sciences). West Lafayette, IN: Purdue University Press.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *Plos One*, 6(6), 21. http://dx.doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *Plos ONE*, 10(8), 1-24. http://dx.doi.org/10.1371/journal.pone.0134826

Thessen, A. E., & Patterson, D. J. (2011). Data issues in the life sciences. *ZOOKEYS*, 150(S1), 15-51. http://dx.doi.org/10.3897/zookeys.150.1766

Womack, R. P. (2015). Research data in core journals in biology, chemistry, mathematics, and physics. *Plos ONE*, 10(12), 1-22. doi:10.1371/journal.pone