Data Practices in the Crop Sciences: A Review of Selected Faculty Publications

Sarah C. Williams

The Version of Record of this manuscript has been published and is available in the *Journal of Agricultural & Food Information*, 2012, 10.1080/10496505.2012.717846

Abstract

This bibliographic study gathered information about the data practices of crop scientists from their publications. Two recent articles were reviewed from each faculty member in the Department of Crop Sciences, University of Illinois College of Agricultural, Consumer and Environmental Sciences. The goals of this study were to learn the common data types used in crop sciences research, to describe data reuse and sharing practices in the literature, and to highlight resources for acquiring and sharing data, especially repositories with an agricultural emphasis.

Introduction

To effectively provide data services to faculty, staff and students, agricultural librarians must understand the data practices in their subject areas. Disciplines have unique data practices and data cultures, defined as "the social conventions of acquisition, curation, preservation, sharing, and reuse of data" (Thessen and Patterson, 2011, 10). Given the scope and scale of the life sciences, these disciplines do not share a single data culture (Thessen and Patterson, 2011). "Agricultural research today is both blessed and cursed with the exploding amount and diversity of scientific information" (McLaren et al., 2009, 141); as a result, data practices and cultures in the agricultural sciences also are likely to vary. Previous research used interviews to gain a better understanding of researchers' data practices (Diekmann, 2012; Swan & Brown, 2008; Witt, Carlson, Brandt, & Cragin, 2009). This literature-based study offers a different perspective on data practices in the crop sciences.

Agricultural librarians also should be aware of data repositories, metadata schemas, and ontologies in development within the disciplines, because as Bracke (2011) stated, the tools developed within the disciplines are often the most relevant for researchers. Kirlew (2011) also emphasized the importance of librarians being aware of disciplinary data resources to better support researchers.

There are several notable data initiatives developed for or applicable to the agricultural sciences. One prominent example is the International Consortium for Agricultural Systems Applications (ICASA), which develops standards "to provide a reliable, portable, flexible structure both for documenting field experiments (or their equivalents in greenhouses or growth chambers) and for specifying conditions for running dynamic simulation models" (Hunt, Hoogenboom, Jones, & White, 2006, 4). Researchers using simulation models are the primary adopters of these standards, but White and van Evert (2008) noted a developing trend toward using the standards to document other types of agricultural research. ICASA provides a data exchange system. However, as of March 2012, ICASA members could only access existing datasets; no new datasets were being accepted (ICASA Data Exchange). The Plant Ontology Consortium is another relevant initiative. The Plant Ontology is a controlled vocabulary for describing plant anatomy and morphology and plant growth and development stages (Plant Ontology Consortium, 2012). This report highlights additional agricultural data initiatives and repositories.

The goals of this study were to learn the common data types used in crop sciences research, to describe data reuse and sharing practices in the literature, and to highlight resources for acquiring and sharing data, especially repositories with an agricultural emphasis.

Background

Data is a burgeoning topic in library and information science literature. Some research reports focus on data services and data practices in specific disciplines, and two recent articles concentrate on agricultural sciences. Bracke (2011) used a case study to describe how agricultural librarians can transform their approaches and skills to meet the data needs of scientists. Suggestions included: earlier involvement in the research process; a willingness to experiment and acceptance of imperfection in this rapidly evolving stage of data curation; and an awareness of metadata schemas and vocabularies in disciplines outside of librarianship.

Between December 2008 and March 2009, Diekmann (2012) conducted an exploratory study of the data practices of agricultural scientists. Given the disciplinary similarity, the results of his interview-based study complement the results of this literature-based report. Diekmann interviewed faculty from a variety of agricultural departments who conduct field and laboratory research. The study noted data types reported by the interviewees, provided insights into attitudes and challenges of sharing and reusing data, and discussed the early stages of data management in the agricultural sciences.

Thessen and Patterson (2011) described in detail the sociological and technological issues that affect the success of data-centric research in the life sciences, but excluded the agricultural and food sciences on the basis of their applied nature and emerging data infrastructure. Other studies (Key Perspectives, 2010; Patterns of information use, 2009; Swan & Brown, 2008) have investigated data practices of researchers across a broad spectrum of disciplines, including the sciences.

The *PARSE.Insight* study (Thaesis & van der Hoeven, 2010) was another project focused on the data practices of scientists across disciplines in Europe. Smit (2011) cited statistics from this project in her exploration of data and publication integration. She highlighted the various ways that scientists store and find research data, and noted that while publishers accept a variety of supplementary data, they often do not have robust plans for validating, linking and preserving the data. Smit stated that the *PARSE.Insight* study underscored the lack of conventions and best practices for data management by researchers and publishers, and concluded that the study provided evidence that researchers and publishers want persistent links between data and publications.

Smit and Gruttemeier (2011) discussed differences in the data practices and guidelines of four scientific journals (i.e., *Journal of Neuroscience*, *Nature*, *Science*, and *Cell*). The practices illustrated attempts by publishers to address the complexities of data (e.g., large volume, numerous file formats). The authors outlined some best practices for integrating data and publications, such as bi-directional linking between data and publications and uniform data citation practices, which will be important considerations if publishers increasingly rely on repositories (disciplinary or institutional) to manage data as an alternative to accepting data as supplementary files.

While Smit and Gruttemeier represented organizations that publish and disseminate scientific research, Santos, Blake, and States (2005) wrote a letter to the editor of *Nature* that described the challenges of supplementary files from the researchers' perspective. The authors gathered supplementary gene-expression profiling data from 10,128 papers in 139 journals.

International standards for data representation exist for these data, yet the researchers found no evidence of the adoption of these standards across journals or within a single journal. Differences in file formats and data organization made it impossible to analyze the supplementary data, unlike the highly accessible data found in public repositories. As a result, the researchers encouraged scientific journals to adopt policies requiring data to be submitted to repositories when they exist, restricting submission of supplementary files to those data for which no suitable repositories exist.

Nevertheless, the trend of releasing supplementary files with journal articles continues, with no best practices to guide selection, delivery, discovery, or preservation (Beebe & McVeigh, 2012). In response, the National Information Standards Organization (NISO) and the National Federation of Advanced Information Services (NFAIS) created a joint working group (with two subgroups: business and technical) to develop recommended practices for publishers to handle supplementary files. The Business Working Group provided a draft for public comment from January to February 2012 (Beebe & McVeigh, 2012). The document included practices for the selection, discovery, and citation of supplementary files.

Recent studies of scientific data repositories also are highly relevant. Kirlew (2011) conducted a bibliographic analysis to identify life sciences data repositories, and then gathered information about the contents and features of 21 of those repositories. Kirlew's study included some of the repositories used by agricultural scientists conducting genetic research (e.g., GenBank, Gene Expression Omnibus), but it did not mention any of the agricultural-focused repositories included in this crop sciences study. Marcial and Hemming (2010) conducted a study of one hundred scientific data repositories. The authors provided a list of the one hundred repositories, but focused the study on identifying characteristics of the repositories and examining similarities across the repositories. Some of the characteristics examined were the disciplinary focus, repository size, business type, and sponsorship. The disciplines most represented by the repositories were geoscience (26%), medicine (20%), biology (15%), and astronomy (14%).

Methodology

This study is based on a thorough review of the two most recent articles of each faculty member in the Department of Crop Sciences at the University of Illinois College of Agricultural, Consumer and Environmental Sciences. The review was conducted from October 2011 through January 2012. The faculty directory of the Department of Crop Sciences (http://cropsci.illinois.edu/directory/faculty), which includes faculty with joint appointments in other university departments or units, served as the source of faculty names. Sixty-two faculty members were included in this study, and as a result, 124 articles were reviewed.
Included in the review were assistant, associate, and full professors, but not emeritus or adjunct professors.

Publications included in this study were research or review articles, with most discovered via an author search in the Web of Knowledge Science Citation Index Expanded database. In cases of common names, a faculty member's affiliation (e.g., Illinois, USDA) served as an additional search parameter in the address field, if necessary. Some articles were identified from faculty members' directory web pages.

The author reviewed the entire article for mention of resources used to acquire or share data, and for information about types of data included in the paper. Findings were recorded in five broad topics: research/publication type, data sources, data sharing, data types, and notes.

The author defined seven research/publication types: field, genetic, greenhouse/laboratory, model/method, molecular, review, and survey (Figure 1). The articles were categorized by research/publication type to help identify any differences or similarities between types of research. Every effort was made to assign a single type to each article, based on the majority of the data presented, but some articles were truly a combination of research types. For example, some studies had significant field and greenhouse components, and some studies generated genetic data and laboratory data.

The data sources and data sharing categories were the focus of this study. Data sources were defined as sources other than traditional citations to literature for information or ideas. Most articles clearly stated specific sources used. Some authors mentioned sources vaguely; this was especially true for National Center for Biotechnology Information (NCBI) sources, so these were identified as specifically as possible. Examples of data sources included data repositories, supplementary files, and weather stations. Data sharing meant publicly sharing data beyond that published in the journal article. Sharing was noted as either sharing via the journal website (i.e., supplementary files) and/or through external resources (e.g., GenBank).

The author recorded the core data generated and used by each published study. Other potentially useful or interesting information, such as open access publication, was captured in the notes category.

Results

All Articles

The 124 articles reviewed in this study appeared in 64 distinct journals. The five journals with the highest frequency of articles (number of articles in parentheses) were: *Crop Science* (16), *Weed Technology* (6), *Agronomy Journal* (5), *Journal of Experimental Botany* (5), and *Plant Physiology* (5). Appendix 1 provides an alphabetical list of all 64 journals and the number of articles published in each. All of the articles were published between 2001 and 2011, with nearly 50% published in 2011 (Table 1).

Each article was assigned a research/publication type. Of the seven types identified, the three most common were field research, greenhouse/laboratory research, and genetic research (Table 2). Because some of the studies were a combination of research types, thirteen articles were assigned two types, and one article was assigned three types.

<u>Data Source Articles and Data Sources</u>

Fifty-five of the reviewed articles (44%) used a source of data other than traditional literature citation for information or ideas. These articles appeared in 39 journals (Appendix 1). Two or more articles citing data sources appeared in the following journals (number of articles in parentheses): *Crop Science* (6), *Journal of Experimental Botany* (3), *Plant Disease* (3), *Agronomy Journal* (2), *BMC Plant Biology* (2), *Global Change Biology Bioenergy* (2), *Journal of Environmental Quality* (2), *Molecular Biology and Evolution* (2), *Plant Physiology* (2), and *PLoS ONE* (2). Over 50% of the articles using data sources were published in 2011 (Table 1).

The data sources used in these articles varied. Six articles used data from other published articles, with the data actually incorporated into the study. Three articles used supplementary files associated with other publications. Two articles utilized data from growers, and two different articles used data from weather stations. One article mentioned using data from an unpublished data file. Another study involved a geographic information system database that incorporated spatial data layers acquired from various regional centers and commissions.

Numerous repositories and organizational websites (Appendix 2) also were the source of data used in these publications. Some of these sources are widely used in the life sciences, while others focus more on agricultural sciences, such as the Census of Agriculture, FAOSTAT, the International Survey of Herbicide Resistant Weeds, and the USDA-NASS Agricultural Chemical Use Database.

Many of the data source articles were related to genetic research (Table 2). Over 40% of the 55 articles were assigned to the genetic or the genetic and greenhouse/laboratory types. Furthermore, genetic research articles frequently used data sources. Of the 124 original articles, 34 were related to genetic research (i.e., either mainly genetic or genetic combined with another research type), and 23 of these articles (68%) used data sources. Other research and publication types also used data sources. This was the case for about one-third of the articles focused on field or greenhouse/laboratory research.

Data sources varied somewhat depending on research/publication types. Nearly every research or publication type – field, genetic, greenhouse/laboratory, model/method, and review – used data from other published articles. The three articles that used supplementary files from published articles, mentioned earlier, emphasized genetic research. Field research mainly used data sources for meteorological or atmospheric data and for production data. For greenhouse/laboratory research, data sources tied to accession or identification numbers were common, such as GenBank accession numbers or USDA-ARS Germplasm Resources Information Network (GRIN) plant introduction numbers. The genetic research articles used a variety of web-based data repositories, especially those associated with NCBI and EBI (the European Bioinformatics Institute), as well as SoyBase (Soybean Genetics and Genomics Database) and MaizeGDB (Maize Genetics and Genomics Database).

Data Sharing Articles and Resources

Thirty of the 124 articles reviewed (24%) noted publically sharing data beyond what was published in the journal article. The most common sharing method was supplementary files published on the journal website. For 19 articles, this was the only sharing method used. Eight articles shared data in supplementary files and via an external resource, such as GenBank. In the cases of two articles, data sharing was through supplementary files and via a faculty or departmental website; one article used only an external resource to share data. An additional two articles did refer readers to external sources (i.e., GRIN and the Cotton Market Database) for more information about accessions and markers, but these were not included in the total of data sharing articles; the articles did not imply that the researchers actually had submitted new data to these resources. This situation could be an example of another form of data reuse.

The 30 data sharing articles appeared in 23 journals (Appendix 1). The journals publishing at least two data sharing articles (number of articles in parentheses) were: *Crop Science* (3), *Plant Physiology* (3), *Environmental Science and Technology* (2), *Molecular Biology and Evolution* (2), and *PLoS ONE* (2). A majority (67%) of the data sharing articles were published in 2011 (Table 1). Eight of these articles (36%) were open access articles. For three additional articles (published by the American Chemical Society), the supplementary files were openly available even without a subscription.

The articles that noted data sharing, whether on the journal website or via an external resource, were primarily genetic research articles (Table 2); 60% of these 30 articles were assigned to the genetic or the genetic and greenhouse/laboratory types. An analysis of all 124 articles in this study also illustrates the prominence of data sharing in genetic research. Thirty-four total articles were related to genetic research (i.e., either mainly genetic or genetic combined with another research type), and 18 of them (53%) shared additional data. Of all of the articles that focused on field research, only two (6%) noted sharing additional data. Less than a quarter of the articles focused on greenhouse/laboratory research shared additional data.

Twenty-nine articles shared data as supplementary files on the journal website, whether only on the journal website or in combination with an external resource or a faculty or departmental website. These articles appeared in journals from 14 different publishers, a mix of society, commercial, and non-profit publishers (Table 3).

The supplementary files included a variety of data and file formats. Data in supplementary files were not limited to articles that shared tables but were broadly defined to include articles that shared only figures, graphs, or photos. The most common file format of supplementary files was PDF. Next were Microsoft Word and Excel files, which were equally common. Other formats included text, zip, Encapsulated PostScript (EPS), and image (TIF, GIF) files.

The supplementary file formats of two articles were unknown, because the supplementary files were not discoverable on the journal website. Both articles were published by the Crop Science Society of America (in *Crop Science* and *Plant Genome*). In both cases, the articles referred to one or more supplementary files, which were unavailable. Similarly, for the two articles that mentioned sharing data on a faculty or departmental website, those files were not discoverable. Of the nine articles that utilized external sources, the data were discoverable. Most articles provided accession numbers within the text of the article, which facilitated the data search, although one accession number needed to be slightly modified to locate the data. In one case, a supplemental file provided direct links to the data in the external resource (Pedigrees of Oat Lines).

Since most of the data sharing articles were related to genetic research, the external resources used were mainly focused on genetic data (Appendix 2). Some of these external resources are widely used in the life sciences, such as the NCBI repositories, but the Pedigrees of Oat Lines is an example of a repository with an agricultural emphasis.

Data

The common data types are broadly categorizable in two ways. One category includes data that describe the experiment, such as site latitude, longitude and slope; tillage practices; soil pH, bulk density and moisture content; species, cultivars or populations; herbicide, fertilizer or insecticide application dates and rates; and greenhouse or laboratory growing conditions. The other category covers data generated by the experiment. Field and greenhouse/laboratory research produced similar types of data. Common data types were plant or production data (e.g., yield, biomass fresh and dry weight, grain weight, stand count, shoot height), organism data (e.g., number of nematodes, earthworms, eggs, aphids), ratings data (e.g., disease severity, root injury, herbicide injury), and physiological data (e.g., leaf chlorophyll content, photosynthetic rate, grain starch and protein concentrations). Genetic research occasionally generated these data types, but more commonly generated unique data, such as DNA or protein sequences; chromosome number and position; allele frequency; number of single nucleotide polymorphisms, and phylogenetic trees. Notably, all three research types (i.e., field, greenhouse/laboratory and genetic) had examples of digital image data, such as photos of leaf samples, leaf cross-sections, and chromosome spreads.

Discussion

This study found that crop scientists use a variety of data sources in their research- data from published articles, supplementary files, weather stations, repositories, and organizational websites. Similarly, the *PARSE.Insight* project, which studied scientists in many disciplines, reported that researchers find and access data in a variety of ways, including colleagues (over 70%), formal literature (over 60%), institutional repositories (over 50%), and disciplinary repositories (less than 30%) (Smit, 2011). In Diekmann's (2012) study of agricultural scientists, important data sources mentioned by participants were historical land use and management records, and historical image data. This bibliographic study reveals that meteorological or atmospheric data sources also are important sources for field research. Examples included the Illinois State Water Survey/Illinois State Climatologist Office, the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center, and the U.S. Naval Observatory.

While a mix of research types (e.g., field, genetic, greenhouse/laboratory) used data from other sources, the genetic research articles primarily noted additional shared data, whether on the journal website or via an external resources. These differences illustrate the diverse data practices even within the crop sciences. Of the articles that focused on field research, only two (6%) mentioned sharing data beyond what was published in the article. Similarly, Diekmann (2012) found that most participants did not typically deposit or share raw data. One interviewee said, "Yes, our main data eventually finds its way into a journal article in summary form and is presented in a table or figure, and that's typically the way we do it, like most other people in our field." (Diekmann, 2012, 27-28)

In the *PARSE.Insight* study, researchers indicated that they would be willing to use a variety of data sharing methods, including organizational repositories (over 80%), disciplinary repositories (60%), and publishers (over 50%) (Smit, 2011). Bracke (2011) noted that data repositories currently are uncommon in the agricultural sciences. It would be interesting to research the willingness of agricultural scientists, especially field researchers, to use these data sharing methods. A variety of factors might contribute to a lower use of or interest in repositories among some agricultural scientists. Diekmann (2012) described data management challenges and data sharing concerns. While many researchers might have similar challenges and concerns (e.g., competition in academic research, misinterpretation of data), agricultural scientists may face unique issues. For example, the participants in Diekmann's study (2012) noted that field research is subject to biological and spatial variation and uncertain environmental conditions that force researchers to modify their experimental design and methods. These modifications can make it difficult to manage the data and can increase the time and effort required for data annotation necessary for data sharing.

Nevertheless, there are some agricultural sciences data initiatives. As mentioned earlier, the ICASA develops standards for documenting field, greenhouse, and growth chamber experiments and for specifying conditions for simulation models (Hunt, Hoogenboom, Jones, & White, 2006). The present study identified several additional agriculturally-focused repositories in use by crop scientists for data sharing and acquisition, such as MaizeGDB, the Pedigrees of Oat Lines, and SoyBase. Interestingly, one of the articles described the Plant Structure Ontology (PSO), the first generic ontology for the anatomy and morphology of a flowering plant (Ilic et al., 2007). The PSO was submitted to the Plant Ontology website, and the PSO was used by The Arabidopsis Information Resource (TAIR) and MaizeGDB, two data repositories noted in this bibliographic study. This example illustrates the interconnectedness of data initiatives, and emphasizes the need for agricultural librarians to be aware of data repositories, metadata schemas, and ontologies being developed within the disciplines.

In this study, the most common data sharing method was supplementary files on the journal website. Twenty-nine articles shared data as supplementary files on a journal website, sometimes in combination with an external resource or a faculty or departmental website. The study identified only 30 articles that shared additional data, so 96% of the data sharing articles relied on supplementary files. There were 19 data sharing articles that only used supplementary files on the journal website, which was 63% of the total. These findings support the statement by Smit and Gruttemeier (2011) that supplementary files compose a substantial portion of shared data. Based on the *PARSE.Insight* study, Smit and Gruttemeier (2011) noted that 25% of researchers make their data publicly available, and over half of those share their data by submitting it with manuscripts.

The heavy reliance on supplementary files emphasizes the importance of the efforts of the NISO and NFAIS joint working group that is developing recommended practices for publishers to handle supplementary files (NISO, 2012). During the course of this study, the supplementary files were not always easy to discover; these were usually found through a reference within the article. As the NISO draft for public comment noted, most abstracting and indexing services do not indicate when supplementary files are available (Beebe & McVeigh, 2012). The draft recommended three areas for publishers to consistently present supplementary files: an indication in the online table of contents that supplementary files exist for an article; links to supplementary files near the top of the first page of an online article; and navigation within the supplementary files that matches the navigation provided at the article level. These three areas should be sufficient, as long as publishers follow the recommendation to provide links near the top of the first page. With link resolvers, library users often link from an abstracting and indexing database directly to an article record on a publisher's website. These users would not see an indication of supplementary files on a table of contents page, so the links near the top of the actual article become critical for notifying readers about the supplementary files.

The frequent use of supplementary files in this study also raises a new research question. It would be interesting to study whether disciplinary repositories existed for the types of data submitted as supplementary files, and if so, why the repositories were not utilized. Many of the data sharing articles were of genetic research, which has several well-established repositories. Perhaps repositories did not exist for the data in the supplementary files; perhaps the researchers were not aware of existing repositories, or perhaps the process for submitting data to the repositories was cumbersome. There are numerous possibilities that could be studied further to better understand the use of supplementary files by agricultural scientists and other researchers.

This study also draws attention to the persistence of shared data. There were only two articles that mentioned sharing data on a faculty or departmental website, but in both cases, those files were unavailable. For two of the 29 articles that mentioned supplementary files, the files could not be located on the journal website. Of the nine articles that utilized external sources, those data were all found. Given the rather small number of data sharing articles in this study, no broad statements can be made about the persistence of data shared through different methods, but this study does provide a glimpse into the issue.

This bibliographic study identified a wide variety of data types. In this study, two broad categories emerged – experimental description data and experimentally-generated data. For the field and greenhouse/laboratory research in this study, these categories parallel the ICASA standards (Hunt, Hoogenboom, Jones, & White, 2006). Corresponding to the experimental description data, the ICASA standard has an experiment dataset to describe the details of an experiment, with subsets that can be used, when applicable, to describe the chemicals, environmental modifications, fertilizers, genotypes, initial conditions, and tillage. Corresponding to the experimentally-generated data, the ICASA standard has a summary results subset for measurements or observations made once or a few times in an experiment, and a time-course results subset for measurements or observations made at intervals throughout the experiment. Given the thought and effort that has gone into this standard, it will be an important standard to track for future developments for agricultural field and greenhouse research data.

Conclusion

Complimenting interviewed-based studies, this literature-based study provides a different perspective on data practices in the crop sciences. Notable findings include the variety of data sources used and the differences in data sharing between field and genetic research. This study revealed a heavy reliance on supplementary files to share data, which suggests the NISO and NFAIS recommended practices for supplementary files will be an important development to monitor. This study also draws attention to some agriculturally-focused repositories that are often overlooked in broader studies of science or life science repositories. As suggested by Kirlew (2011) and Bracke (2011), this disciplinary information will be valuable for librarians providing data services to agricultural scientists.

This study also suggests a few areas for future research. With the variability and uncertainty of field research, it would be interesting to investigate the willingness of agricultural scientists, especially field researchers, to share data via repositories and publishers. Given the heavy reliance on supplementary files noted in this study, future research could investigate whether disciplinary repositories exist for data submitted as supplementary files, and if so, why the repositories were not utilized. This study also provides a glimpse of shared data persistence, which will be a significant issue in the future.

Acknowledgements

The author thanks Anita Foster for providing comments on the final draft of this article.

References

Beebe, L., & McVeigh, M. (2012). *Recommended practices for online supplemental journal article materials, Part A: Business working group recommendations* (Draft for public comment). National Information Standards Organization and National Federation of Advanced Information Services. Retrieved from http://www.niso.org/workrooms/supplemental

Bracke, M. S. (2011). Emerging data curation roles for librarians: A case study of agricultural data. *Journal of Agricultural & Food Information,* 12(1), 65-74. doi:10.1080/10496505.2011.539158

Diekmann, F. (2012). Data practices of agricultural scientists: Results from an exploratory study. *Journal of Agricultural & Food Information*, 13(1), 14-34. doi: 0.1080/10496505.2012.636005

Hunt, L. A., Hoogenboom, G., Jones, J. W., & White, J. W. (2006). *ICASA version 1.0 data standards for agricultural research and decision support*. Retrieved from http://www.icasa.net/DataStandards/index.html

ICASA data exchange. (n.d.). Retrieved from http://dssat.net/data/exchange

Ilic, K., Kellogg, E. A., Jaiswal, P., Zapata, F., Stevens, P. F., Vincent, L. P., … Rhee, S. Y. (2007). The plant structure ontology, a unified vocabulary of anatomy and morphology of a flower plant. *Plant Physiology*, 143(2), 587-599. doi: 10.1104/pp.106.092825

Key Perspectives. (2010). *Data dimensions: Disciplinary differences in research data sharing, reuse and long term viability. SCARP synthesis study*. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/projects/scarp

Kirlew, P. W. (2011). Life science data repositories in the publications of scientists and librarians. *Issues in Science and Technology Librarianship*, 65. Retrieved from http://www.istl.org/11-spring/refereed1.html. doi: 10.5062/F4X63JT2

Marcial, L. H., & Hemminger, B. M. (2010). Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10), 2029-2048. doi: 10.1002/asi.21339

McLaren, C. G., Metz, T., van den Berg, M., Bruskiewich, R. M., Magor, N. P., & Shires, D. (2009). Informatics in agricultural research for development. *Advances in Agronomy*, 102, 135-157. doi: 10.1016/S0065-2113(09)01004-9

NISO. (2012). *NISO/NFAIS Supplemental journal article materials project*. Retrieved from http://www.niso.org/workrooms/supplemental

Patterns of information use and exchange: Case studies of researchers in the life sciences. (2009). Research Information Network and the British Library. Retrieved from http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/patterns-information-use-and-exchange-case-studies

Plant Ontology Consortium. (2012). *Plant ontology consortium*. Retrieved from http://www.plantontology.org/

Santos, C., Blake, J., & States, D. J. (2005). Supplementary data need to be kept in public repositories. *Nature*, 438(7069), 738. doi: 10.1038/438738a

Smit, E. (2011). Abelard and Héloise: Why data and publications belong together. *D-Lib Magazine*, 17(1/2). Retrieved from http://www.dlib.org/dlib/january11/smit/01smit.html. doi:10.1045/january2011-smit

Smit, E., & Gruttemeier, H. (2011). Are scholarly publications ready for the data era? Suggestions for best practice guidelines and common standards for the integration of data and publications. *New Review of Information Networking*, 16(1), 54-70. doi: 10.1080/13614576.2011.574488

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs*. Research Information Network. Retrieved from http://www.rin.ac.uk/data-publication

Thaesis, & van der Hoeven, J. (2010). *Insight report: Insight into digital preservation of research output in Europe*. Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf

Thessen, A. E., & Patterson, D. J. (2011). *Data issues in the life sciences*. Data Conservancy. Retrieved from http://dataconservancy.org/publications

White, J. W., & van Evert, F. K. (2008). Publishing agronomic data. *Agronomy Journal*, 100(5), 1396-1400. doi:10.2134/agronj2008.0080F

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, 4(3), 93-103.

Figure 1 Broad Topics and Research/Publication Types Used in this Study

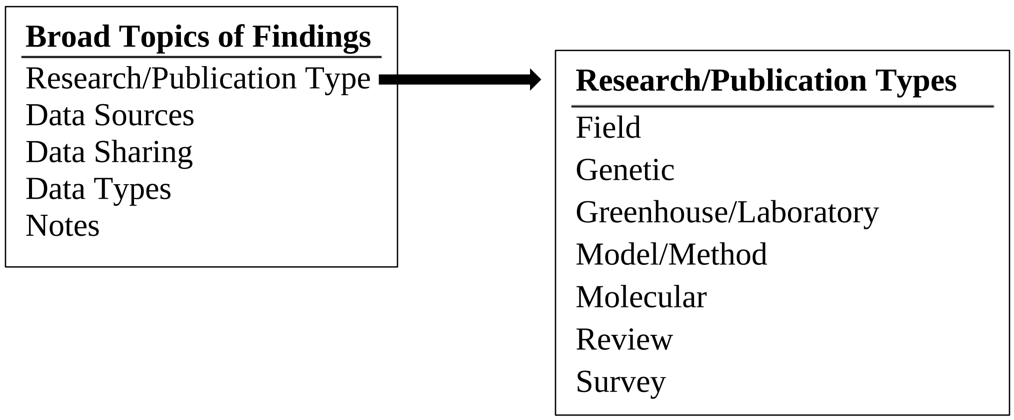

Table 1 Number of Articles by Publication Year

Publication Year	Total Number of Articles (n = 124)	Number of Data Source Articles (n = 55)	Number of Data Sharing Articles (n = 30)
2011	58	29	20
2010	23	5	2
2009	15	7	4
2008	13	6	3
2007	6	4	1
2006	4	3	0
2005	1	1	0
2004	1	0	0
2003	1	0	0
2002	1	0	0
2001	1	0	0

Table 2 Number of Articles by Research/Publication Type

Research/Publication Type	Total Number of Articles (n = 124)	Number of Data Source Articles (n = 55)	Number of Data Sharing Articles (n = 30)
Field	33	11	2
Field & Genetic	1	0	0
Field & Greenhouse/Laboratory	7	1	0
Field & Greenhouse/Laboratory & Genetic	1	0	0
Genetic	27	21	17
Genetic & Greenhouse/Laboratory	5	2	1
Greenhouse/Laboratory	33	10	7
Model/Method	4	2	2
Molecular	2	2	1
Review	8	4	0
Survey	3	2	0

Table 3 Publishers of Articles with Supplementary Files

Publisher	Number of Articles with Supplementary Files on the Journal Website (n=29)
American Chemical Society	3
American Phytopathological Society	2
American Society of Plant Biologists	2
BioMed Central	4
Botanical Society of America	1
Crop Science Society of America	4
Genetics Society of America	1
National Academy of Sciences	1
National Research Council of Canada	1
Nature Publishing	1
Oxford	3
Public Library of Science	2
Springer	2
Wiley-Blackwell	2

Appendix 1 Journals Included in This Study

Journal Title	Total Number of Articles (n = 124)	Number of Data Source Articles (n = 55)	Number of Data Sharing Articles (n = 30)
Agronomy Journal	5	2	0
American Journal of Botany	1	0	1
Analytical Chemistry	1	0	1
Annals of the Entomological Society of America	1	1	0
BioEnergy Research	1	0	0
Biology Direct	1	1	1
Biomass and Bioenergy	2	1	0
BMC Genomics	1	1	1
BMC Plant Biology	2	2	1
Crop Protection	2	0	0
Crop Science	16	6	3
Environmental Pollution	1	0	0
Environmental Science and Technology	2	1	2
Euphytica	1	1	1
Food Chemistry	1	0	0
Genetics	1	1	1
Genome	1	1	1
Genome Biology	1	1	1
Global Change Biology	1	0	0
Global Change Biology Bioenergy	4	2	1
HortScience	3	0	0
HortTechnology	1	1	0
Journal of Agricultural and Food Chemistry	2	0	0
Journal of Economic Entomology	2	0	0
Journal of Environmental Quality	3	2	0
Journal of Experimental Botany	5	3	1
Journal of General Virology	1	1	0
Journal of Heredity	1	0	0
Journal of Molecular Evolution	1	1	0
Journal of Nematology	1	1	0
Journal of Phytopathology	1	0	0
Journal of the American Society for Horticultural Science	1	0	0
Landscape and Urban Planning	1	1	0

LWT - Food Science & Technology	1	0	0
Molecular Biology and Evolution	2	2	2
Molecular Plant-Microbe Interactions	1	1	1
Nature Genetics	1	1	1
Nematropica	2	1	0
Pesticide Biochemistry and Physiology	1	0	0
Pest Management Science	3	1	0
Photosynthesis Research	1	1	0
Photosynthetica	1	0	0
Phytopathology	2	1	1
Plant and Soil	1	0	0
Plant, Cell and Environment	2	1	1
Plant Cell, Tissues and Organ Culture	1	0	0
Plant Disease	4	3	0
Plant Genome	1	0	1
Plant Pathology	1	0	0
Plant Physiology	5	2	3
Planta	1	1	1
PLoS ONE	2	2	2
Precision Agriculture	1	1	0
Proceedings of the National Academy of Sciences	1	1	1
Soil and Tillage Research	1	0	0
Soil Biology and Biochemistry	1	1	0
Soil Science	3	0	0
Soil Science Society of America Journal	2	1	0
Theoretical and Applied Genetics	1	0	0
Water, Air, and Soil Pollution	1	0	0
Weed Research	1	1	0
Weed Science	3	1	0
Weed Technology	6	1	0
World Journal of Microbiology and Biotechnology	1	0	0

Appendix 2 Data Repositories and Organizational Websites Included in This Study

Data Repositories and Organizational Websites	**Number of Articles that Used Data from these Resources**	**Number of Articles that Shared Data via these Resources**
Census of Agriculture	2	
EBI (European Bioinformatics Institute) InterPro Databases	1	
EBI Nucleotide Sequence Database		1
EBI UniProt Knowledgebase (UniProtKB)	5	
FAOSTAT	2	
Gene Ontology (Gene Ontology Consortium)	1	
Gramene	1	
Illinois State Water Survey/Illinois State Climatologist Office	5	
International Survey of Herbicide Resistant Weeds	1	
Kyoto Encyclopedia of Genes and Genomes	1	
MaizeGDB (Maize Genetics and Genomics Database)	3	
Maize HapMap	1	
Microbial Community Analysis (MiCA)	1	
National Trends Network (National Atmospheric Deposition Program)	1	
National Weather Service Regional Office	1	
NCBI (National Center for Biotechnology Information) BLAST/databases	7	
NCBI EST	1	1
NCBI GenBank	12	3
NCBI Gene Expression Omnibus		3
NCBI Genome Survey Sequences		1
NCBI RefSeq	1	
NCBI Sequence Read Archive		2
NCBI Trace Archive	2	
NOAA (National Oceanic and Atmospheric Administration) Eastern Illinois Climate Division	1	
NOAA National Climatic Data Center	1	
Pedigrees of Oat Lines (POOL)		1
Pfam (Protein Family) Database	3	
Phytozome	2	
Plant Ontology		1

Plant Repeat Databases (Michigan State University)	1	
Plant Variety Protection Office Databases	1	
RCSB (Research Collaboratory for Structural Bioinformatics) Protein Data Bank	1	
SilkDB (Silkworm Genome Database)	1	
SoyBase (Soybean Genetics and Genomics Database)	4	
SUPERFAMILY	3	
The Arabidopsis Information Resource (TAIR)	2	
TIGR Plant Transcript Assemblies	1	
University of Illinois National Soybean Pathogen Collection Center	1	
University of Illinois Sweet Corn Disease Nursery	1	
University of Illinois Variety Testing	2	
USDA-ARS (United States Department of Agriculture – Agricultural Research Service) Germplasm Resources Information Network (GRIN)	1	
USDA-ARS Maize Genetics Cooperation Stock Center	1	
USDA-ARS Systematic Mycology and Microbiology Laboratory Fungal Databases – Specimens	1	
USDA-NASS (National Agricultural Statistics Service) Agricultural Chemical Use Database	1	
U.S. Energy Information Administration	1	
U.S. Naval Observatory	1	