Much Ado About Data: Intellectual Property Issues Surrounding Academic Research Data

Rebekah Cummings

University of California, Los Angeles

Master of Library and Information Science, June 2013

rebekah.cummings@utah.edu

Abstract

According to a 2013 Oxford study, intellectual property (IP) is the element of data curation that causes the most confusion for academic researchers. Intellectual property is always murky, but even more so as it relates to data because "facts" are not eligible for copyright protection under United States copyright law. In this paper, the author frames the complex issues surrounding intellectual property and data curation including the unique nature of data, the motivations behind open data sharing, and the legal landscape that undergirds current data practices. This paper also demonstrates how librarians can use the four factors of fair use – purpose, nature of the work, amount used, and effect on the market – when assessing risk in data reuse.

Keywords: intellectual property, data curation, data management, fair use, four-factor test

"Now let us turn to what is emerging as the most dramatic and troubling border skirmish for both academia and the law: the propertization of scientific data." – Coryanne McSherry, 2001

Introduction

Academic libraries have traditionally been tasked with the acquisition, organization, and dissemination of scholarly research products, usually in the form of journals and monographs. Increasingly, funding agencies and research institutions are viewing the underlying research data that supports these publications as an important scholarly product in its own right, deserving of preservation and distribution. In response, a new brand of librarian has emerged to manage data sets and dismantle barriers to data access and sharing. These barriers have been lowered significantly through improved technology and communication, the scaffolding of tentative standards, attribution mechanisms for data authors, and policies to promote data sharing. A major barrier that persists, however, is the lack of clarity regarding the intellectual property rights surrounding academic research data. In order to realize the promises of open data and to fulfill their emerging role as data stewards, librarians must understand the unique nature of data, the motivations behind open data sharing, and the legal landscape that undergirds current data practices. Supported by that background, data librarians should have a clear idea of what rights researchers have over their data, how scholars can legally analyze, compare, and combine secondary data in their research, and the legal mechanisms for sharing research data.

What Are Data?

The antecedent to any conversation on data policy is a clear understanding of what is meant by the term "data." The law treats data as facts, but university policy and communities of practice have far more disparate views of data (University of Michigan Library, 2013). The U.S. Office of Budget and Management define research data as "the recorded factual material

commonly accepted in the scientific community as necessary to validate research findings." ("Circular A-110," 2012). Ann Zimmerman at the University of Michigan extends this definition even further to include "information relevant to the data that is independent of the data themselves but without which the data would be incomprehensible." (Zimmerman, 2008).

These definitions indicate is that data are not just lists of facts, but all the documentation needed to make sense of those facts. In the social sciences, for instance, research data would include spreadsheets of survey responses, as well as supplementary materials such as the surveys and questionnaires used to generate the responses and a codebook to decipher variables. Without these supplementary materials, the spreadsheets of raw data would be useless in validating findings, replicating research, combining datasets, or reanalyzing data.

Data librarians should adopt the broader definition of research data for two reasons. First, data curation requires archived datasets to have sufficient documentation for the data to be reusable and the research replicable. If the definition of "data" does not include the supplementary material used to replicate research and repurpose data for secondary analysis, then the definition of data does not comport with the actual practices of data curation or science. Second, limiting the definition of data to observations, measurements, and "facts" does not acknowledge the wide variety of data practices throughout the research community. In the humanities and social sciences, for example, data includes materials such as texts, audio-visual materials, computer codes, oral histories, interviews, and field notes. These non-traditional forms of data must also be considered in discussions of intellectual property and privacy considerations regarding data sharing and reuse.

Another important consideration is the difference between how the law treats academic research data as compared to data generated by private parties or the United States government.

Trade secret law can protect private data, such as the formula for Coke or the Google search algorithm. Generally, eligibility for protection under trade secrets law requires a party to do two things: 1) the owner of the information must take measures to protect the confidentiality of the data; and 2) the information must have independent economic value by, among other things, remaining confidential (McSherry, 2001). Competitors can attempt to replicate trade secret data through reverse engineering or trial and error, but cannot legally acquire the data through unethical means such as hacking or espionage (*Ibid*). At the other end of the spectrum resides U.S. government data, which, once generated, immediately enters the public domain (Reichman & Uhlir, 2003).[1] These open datasets are the backbone of the data commons for the social sciences, which draw heavily from government databases such as the U.S. Census and the Common Core of Data.

Academic research data occupies the space squarely between private research data that does not have to be shared with anyone and government data that must be shared with everyone. Public funding supports a great deal of academic research data, even at private universities. Academic research data, however, does not automatically enter the data commons. Instead, most academic research data are shared through a robust "gift exchange" culture between researchers, relying on the scientific or "Mertonian" norms of sharing and community (Cole, 2004; Merton, 1973). The result is a "delicate process of negotiation, in which data are traded as the result of informal compromises between public and private interests that are worked out on an ad hoc and continual basis (Reichman & Uhlir, 2003). In response to this trend, data management and sharing have gained visibility over the past decade, and in 2011, the National Science Foundation (NSF) brought data stewardship to the fore when it instituted a two-page Data Management

[1] Excluding some types of high-risk, classified, or confidential data

Plans requirement for every research proposal submitted to NSF (National Science Foundation, 2010).

The NSF had at least three motivations for implementing the Data Management Plan requirement. One motivation behind this requirement is to promote transparency and reproducibility – the gold standard of science – in data-intensive research (Jasny, Chin, Chong, & Vignieri, 2011). A second motivation, as previously noted, is to extend knowledge by making high-quality data available for combination and reanalysis (Borgman, 2012). A third motivation is to make the results of publicly-funded research publicly available, an incentive which resonates with tax-payers and policy-makers who believe that tax money should serve the broadest possible interests and not be hoarded for individual gain (*Ibid*).

Research and academic data are different than private or government-generated data in that academic research data do not automatically enter the public domain nor are they necessarily protected by trade secret statute. Moreover, academic research data is a broader concept than the traditional notion of "data" as currently recognized under the law. Managing and sharing academic research data is important in order to ensure the reproducibility of research, better scientific research, and public accountability. This background provides a solid footing under which data librarians can make informed decisions regarding the use and reuse of research data and counsel scholars on their rights as data authors and consumers.

The Legal Landscape Surrounding Research Data

The protectability of data under United States statutory and common law is questionable because courts and legal scholars view data as synonymous with facts, which are not protectable under the law. Likewise, the protectability of data worldwide is just as blurred. While United States law does not protect data, Europe has adopted database laws that extend legal protection to

databases and their underlying data. The disparate treatment of data worldwide causes researchers and academics a great deal of uncertainty in that they do not know what data is free to use, share, combine, and compare in their research.

Copyright and patent law in the United States requires a work to have some measure of originality to be eligible for copyright protection, and facts are a classic example of material that lacks such originality (McSherry, 2001). Observational and historical facts are the raw material from which scholars perform research and extend knowledge, and to restrict access to these facts would severely limit our collective ability to "promote the progress of Science and useful arts" – the underlying goal of copyright law (*U.S. Constit. Art. 1, Sec. 8, Par. 8.*, 1776). Furthermore, the unearthing of facts is thought to be less an act of creation than an act of discovery. As the preeminent authority on copyright, Melville Nimmer, stated:

> The discoverer of a scientific fact as to the nature of the physical world, an historical fact, a contemporary news event, or any other fact may not claim to be the author of that fact. If anyone may claim authorship of facts, it must be the Supreme Author of us all. The discoverer merely finds and records (Nimmer, 1977).

The United States Supreme Court has long agreed with Nimmer's sentiment. For example, the Supreme Court held as early as 1880, in Baker v. Seldon, that documents must contain a minimal amount of originality to qualify for copyright (*Baker v. Seldon 101 U.S. 99*, 1880). The Supreme Court reaffirmed that originality is a prerequisite for copyright protection as recently as 1991 in *Feist v. Rural* (*Feist Publication, Inc. v. Rural Telephone Serv. Co., 499 U.S. 340*, 1991).

The Federal Court of Appeals has even extended this standard for copyright beyond facts to aggregated research. In 1981, the Fifth Circuit in *Miller v. Universal Studios, Inc.* reversed

the decision of the lower court by holding that research is sufficiently original to invoke copyright privileges. The justification for this decision is that research is merely the gathering of facts without any degree of creativity or self-expression (*Miller v. Universal City Studios, Inc. 605 F.2d 1365*, 1981; Suich, 1982).

The decisions in *Baker*, *Feist*, and *Miller* all hold that the work in compiling facts, the "sweat of the brow," is not sufficient to establish a copyright claim *prima facie*. The logic behind all of these rulings is that facts belong in the public domain because they lack originality, are not the created by the discoverer, and would severely impair further research if held behind a statutory curtain.

Unlike copyright law for books, journals, and other tangible works in a fixed expression, intellectual property rights regarding data are not harmonized worldwide (National Science Foundation, Creative Commons, UNM, 2011). Although U.S. law and policy promotes the broadest possible dissemination of data, European Law has extended copyright protection to databases through *sui generis* database rights since 1996 (Reichman & Uhlir, 2003). The unevenness in international data laws has caused uncertainty in the scientific community about what data are freely available for use and what data use must be negotiated through contracts and licenses. These waters are further muddied by the fact that much scientific research is conducted through international cooperation. The Large Hadron Collider, for example, resides underneath the Franco-Swiss border near Switzerland, but was built in collaboration with over one hundred countries representing scientists and funders from all over the world ("Large Hadron Collider," 2012). As science becomes increasingly international, data rights will become correspondingly complex.

The response to this uncertainty is often to apply contracts or licenses to datasets, which assures the data author of their rights but hinders the open use of data for research and scholarship. Like all contracts, these agreements bind only the parties in agreement, in this case the data author and the researcher contracting with the data archive. If the secondary user chooses to publish the data elsewhere, the original user agreement does not follow the data downstream (National Science Foundation, Creative Commons, UNM, 2011). Furthermore, when combining datasets, multiple user agreements can create incompatibilities when working with datasets from disparate sources. These complexities are at the heart of the research communities' reticence to data sharing and reuse. A last distinction to consider is the difference between a database and its contents. In many cases the database may be covered by copyright, even in the U.S., while the underlying data is in the public domain.

Ownership And Property Rights Of Data

The intricacies surrounding data law make it difficult for researchers to have a clear understanding of data ownership, data rights, and data reuse. Before data evolved from being part of the research process to an actual research product, ownership of scholarly data was rarely discussed. Data seldom left the possession of the research team, and as a result, the perception was (and is) that data belongs to the researcher. In reality, ownership of datasets and other original records of research reside with the university where the research was conducted (UCLA Social Science Data Archives, 2012). Many researchers, however, believe that data ownership resides with the research team that collects the data. For example, in a recent survey at University of North Carolina, Chapel Hill, only 15% of the nearly 2,800 respondents thought that the University owns their data (University of North Carolina at Chapel Hill, 2012). Nearly 50% believed that researchers retain ownership of their data.

The murkiness of Intellectual Property as it relates to data is further exemplified in a survey given by the Data Management Rollout (DaMaRo) Project at Oxford. The DaMaRo survey measured the confidence levels of researchers in performing eleven different tasks related to data management, including preparing datasets for long-term preservation, version control, storage, and documentation. Of the eleven tasks listed, researchers said that they felt the least confident in dealing with licensing, copyright, and other intellectual property issues related to datasets ("DaMaRO Survey Results," 2012). In response to this general confusion, data librarians must be equipped to guide researchers through their rights when depositing and extracting data from repositories.

Rights Of Data Authors

Even if universities own the research data produced within their institutions, researchers have certain rights over their data such as the right to choose a repository in which to deposit their data, who has access to the data, and how secondary users can work with the data. Most university policies concede that individual disciplines are too disparate to apply standards writ large, and therefore, rely on "communities of practice" to determine best practices related to data sharing. Even the National Science Foundation states that data sharing and management policies must be "flexible and driven by communities of practice" (Task Force on Data Policies Committee on Strategy and Budget National Science Board, 2011). One of the primary responsibilities of data librarians is to work with researchers to let them know what their rights are as data producers and data consumers. Data librarians must also be familiar with emerging community standards to counsel researchers in best practices and direct them to appropriate repositories.

Analysis Regarding Use Of Datasets

Data librarians assist researchers with data management plans, help researchers prepare their data for storage and preservation, and find relevant datasets for secondary use. All three of these tasks require legal analysis of intellectual property laws surrounding datasets.

The NSF Data Management Plan mandate requires researchers to state who holds the Intellectual Property rights to the data and any constraints in the reuse of the data. One such constraint is whether or not privacy rights are attached to the data. If the data includes human subjects, it must be free of all identifying information before it can be shared. Often, a de-identified dataset can be released while the restricted counterpart is held in a "dark archive" for preservation, but not access (Gutmann, Schürer, Donakowski, & Beedham, 2004). Data Management Plans are written prior to IRB review, and therefore provide an excellent opportunity for researchers to consider in advance how they will protect the confidentiality of their research subjects and still provide access to their de-identified data.

As previously discussed, data-as-facts are not subject to copyright, but many forms of data – defined as the recorded materials needed to replicate research – would be subject to copyright. For instance, surveys and questionnaires are highly original and may be eligible for copyright protection. Additionally, datasets *might* be covered by copyright if the collection of facts is "selected, coordinated, or arranged" in a way that passes a threshold of originality (*17 U.S.C. §101. Definitions.*, n.d.). Librarians must balance the interests of the university, the researcher, and the funding agency when assisting researchers in deciding what mechanism to use when archiving and sharing data. The role of research universities is to ensure "the broadest possible access to the fruit of its work both in the short and long term by publics both local and global." (Hahn, Lowry, Lynch, Shulenberger, & Vaughn, 2009). Universities and funding

agencies are aligned in this regard because both have a vested interest in maximizing their investment and furthering their mission by making their scholarly content as visible and usable as possible. Researchers, on the other hand, may wish to retain an intellectual monopoly over their data until they have completed their research or the publications of findings. Data librarians must help researchers strike a balance between protecting their interests while making their data open to the research community and the taxpayers who funded their data collection.

The most complicated legal analysis that a data librarian will need to make is in assessing whether or not a researcher can use data from a secondary source. Unlike assisting data authors with their data management where the main concern is in balancing the interests of vested parties, counseling researchers on how they can use secondary data bears legal risk if the dataset is covered by intellectual property or privacy restrictions.

The first consideration is whether or not the data is open access data. Open data is defined as "data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and share-alike" ("Open Data – An Introduction | Open Knowledge Foundation," 2012). Examples of high-quality open data include U.S. government data, open data from trusted repositories, and data that bears a waiver such as the Creative Commons "CC0" waiver. Open access data can be used without reservation or fear of legal repercussions.

If the data are not open access, the second consideration is whether the researcher obtained the data via contract or license. Contracts often come under the banner of data use agreements or data access policies (Uhlir, 2012). Proprietary databases, such as Rand or Roper, often allow researchers to use their high-quality data under certain conditions or for a fee. Under these conditions, researchers cannot violate the terms of use by claiming fair use or academic

freedom, because they agreed to the contract or license at the outset, which is binding under the law. Whatever the terms of the agreement specify, those are the parameters under which the researcher may proceed.

There are other considerations regarding data reuse that a librarian needs to undertake beyond the legal questions of access and usability. One such consideration is the quality of the data. In other words, is the data from a trusted source? The usability of data relies almost entirely on the documentation and credibility of the original data gatherer. If a researcher cannot vouch for the methods used to collect the data and cannot access the supplementary materials needed to understand the data, then the data is not eligible for secondary analysis. In data reuse, "quality is paramount." (Gamble & Goble, 2011).

Once it has been established that the data are from a trusted source, the next consideration would be to see if the data author attached user permissions to the dataset such as a Creative Commons license ("About The Licenses - Creative Commons," 2012). If the data carry this type of license, the data consumer can use the dataset under the permissions set forth by the data author without contacting the data author directly.

If a dataset has no licenses or waivers attached, the data librarian and/or researcher should attempt to contact the rights holder. If the rights holder cannot be contacted or has passed away, the librarian can attempt to contact the institution where the research was conducted to ask for permissions. In the unusual case where the dataset is a high quality dataset without any terms of use attached and neither the librarian nor researcher can contact the data author or institution where the research was conducted, the data librarian will need to assess the risk involved in using the dataset under the umbrella of fair use.

Fair Use And Datasets

Determining fair use for datasets in the United States has an inherent benefit that does not apply to journals, books, photographs, or other works fixed in a tangible means of expression in that data are presumptively in the public domain. As previously mentioned, the law treats data as facts, and facts are not protectable by copyright. Secondary users of data, however, should still apply a Fair Use Checklist to the facts of their situation to minimize legal risk to their institution ("Fair Use Checklist — Columbia Copyright Advisory Office," 2012).

The first factor in determining fair use is the purpose of the use. Academic researchers looking to use data for scholarship, research, and teaching are at an advantage because the nature of their work is usually to extend the corpus of knowledge, not for material gain or commercial purposes. The main question here is: Does use of the data promote scientific progress or our collective knowledge? If so, the purpose favors fair use.

The second factor is the nature of the copyrighted work. The nature of most data is factual and non-fiction, which also favors fair use. Some forms of data, however, such as surveys or questionnaires, display a high level of originality. The main question here is: How much originality is present in the data? Data that comes directly off of instruments, such as weather data, are not terribly original, while a list of interview questions or hundreds of hours of video footage is more likely proprietary.

The third factor is the amount of the work used. How much of the dataset is the researcher using and how important is that secondary data to their analysis? If a data consumer relies solely on secondary analysis for their work or the data will be their primary data, that type of use would favor permissions. Likewise, using secondary data as background material or in conjunction with many other types of data would favor fair use. The main question to address

here is: How central is the secondary data to the research, and how much of the original data is being used?

The fourth factor on the Fair Use Checklist is the effect on the market for the original work. When evaluating data from this factor, consider whether there is a commercial market for the data. Although some data have a very high market value, such as computer code or data from pharmaceutical research, most data found outside of repositories has a higher scholarly value than economic value. An evaluation of the data on this factor should balance the rights of the data author with the rights of the public to scholarly information. A key question to ask on this factor is: Will the use of this data limit the original data author's (or institution's) ability to generate revenue from this dataset?

After an evaluation of the four factors, a data user should document their efforts to gain permission, write down their fair use analysis, and give credit to the data author. As a reminder, data should only be used if it is from a trusted source so anonymous citation is not an option. Lastly, data librarians should ensure that the data is from a country that does not give copyright protection to databases. If in doubt, best practices would suggest seeking data with less ambiguous provenance.

Conclusion:

Data librarians are becoming an increasingly common fixture in research institutions as funding agencies and universities require and promote the management of data as a new commodity. In order to better serve their communities and to minimize legal risk for their institutions, data librarians need a solid understanding of intellectual property law as it relates to data sharing and reuse. Although much data resides in the public domain, the expanding definition of data requires researchers and librarians, in conjunction with counsel, to do legal

evaluations of data in the absence of explicit legal mechanisms for sharing data. An understanding of data rights and intellectual property will promote the sharing of data, and ultimately, lead to improved transparency in science and a greater pool of intellectual capital from which researchers can draw.

References

17 U.S.C. §101. Definitions.

About The Licenses - Creative Commons. (2012). Retrieved December 4, 2012, from http://creativecommons.org/licenses/

Baker v. Seldon 101 U.S. 99 (1880).

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

Circular A-110 Revised 11/19/93 As Further Amended 9/30/99 | The White House. (2012). Retrieved November 23, 2012, from http://www.whitehouse.gov/omb/circulars_a110

Cole, S. (2004). Merton's contribution to the sociology of science. *Social Studies of Science*, *34*, 829–844.

DaMaRO Survey Results – Research Data Management Training for the Sciences | DaMaRO. (2012). Retrieved November 26, 2012, from http://blogs.oucs.ox.ac.uk/damaro/2012/11/21/damaro-survey-results-research-data-management-training-for-the-sciences/

Fair Use Checklist — Columbia Copyright Advisory Office. (2012). Retrieved December 4, 2012, from http://copyright.columbia.edu/copyright/fair-use/fair-use-checklist/

Feist Publication, Inc. v. Rural Telephone Serv. Co., 499 U.S. 340 (1991).

Gamble, M., & Goble, C. (2011). Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model. In *ACM WebSci'11* (pp. 1–8). Koblenz, Germany. Retrieved from http://www.websci11.org/fileadmin/websci/Papers/177_paper.pdf

Gutmann, M., Schürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. *Data Science Journal*, *3*, 209–221.

Hahn, K., Lowry, C., Lynch, C., Shulenberger, D., & Vaughn, J. (2009). *The University's Role in the Dissemination of Research and Scholarship--A Call to Action*. Association of American Universities. 1200 New York Avenue NW Suite 550, Washington, DC 20005. Tel: 202-408-7500; Fax: 202-408-8184; Web site: http://www.aau.edu. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED511357

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and Again, and Again ... *Science*, *334*(6060), 1225. doi:10.1126/science.334.6060.1225

Large Hadron Collider. (2012, November 24). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/w/index.php?title=Large_Hadron_Collider&oldid=523863447

McSherry, C. (2001). *Who Owns Academic Work?: Battling for Control of Intellectual Property*. Harvard University Press.

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

Miller v. Universal City Studios, Inc. 605 F.2d 1365 (1981).

National Science Foundation. (2010). *NSF Data Management Plans*. Washington, DC: NSF. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

National Science Foundation, Creative Commons, UNM. (2011). *Data Governance Workshop, Final Report*. Arlington, VA.

Nimmer, M. B. (1977). The Subject Matter of Copyright Under the Act of 1976. *UCLA Law Review*, *24*(5), 978–1024.

Open Data – An Introduction | Open Knowledge Foundation. (2012). Retrieved November 27, 2012, from http://okfn.org/opendata/

Reichman, J. H., & Uhlir, P. (2003). A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment. *Law and Contemporary Problems*, *66*(1), 315–462.

Suich, T. (1982). Copyright Law—Will the Denial of Copyright To An Author's Research Impede Scholarship? Miller v. Universal City Studios, Inc., 605 F. 2d 1365 (5th Cir. 1981). *Western New England Law Review*, *5*(1), 103.

Task Force on Data Policies Committee on Strategy and Budget National Science Board. (2011). *Digital Research Data Sharing and Management*. Arlington, VA: National Science Foundation.

U.S. Constit. Art. 1, Sec. 8, Par. 8. (1776).

UCLA Social Science Data Archives. (2012). Ownership, Copyright, and License Agreements. Retrieved November 26, 2012, from http://dataarchives.ss.ucla.edu/archive%20tutorial/ownership.html

Uhlir, P. F. (Ed.). (2012). *For Attribution—Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, DC: National Academies Press.

University of Michigan Library. (2013). Facts and Data. Retrieved May 24, 2014, from http://www.lib.umich.edu/copyright/facts-and-data

University of North Carolina at Chapel Hill. (2012). *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership.* Retrieved from http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf

Zimmerman, A. S. (2008). New knowledge from old data - The role of standards in the sharing and reuse of ecological data. *Science Technology & Human Values*, *33*, 631–652.