**Module 6**

**Data Sharing & Reuse Policies**

University of Massachusetts at Amherst   
Matt Sheridan, Laura Quilter, Aaron Rubinstein

Marine Biological Laboratory and Woods Hole Oceanographic Institute

Eliza Coburn, John Furfey, Jen Walton

University of Massachusetts Medical School

Lisa Palmer

|  |
| --- |
| Learner Objectives:   1. Explain the benefits of sharing data 2. Identify issues/obstacles related to reuse and sharing of data 3. Understand open access/open science/open data 4. Understand data reuse/sharing/access policies from funding agencies, institutions and publishers 5. Recognize the need to use standard data formats, consistent file naming conventions, etc. 6. Contrast sharing data within and outside of a research team 7. Understand the need for data attribution and citation |

**History of sharing data**

**What’s old?**

* Sharing one on one
* Sharing as part of a small lab team
* Sharing between faculty and students
* Sharing a few compiled results (tables, diagrams) within the context of a publication

**What’s new?**

* Sharing with large numbers of researchers outside a research team
* Sharing data as a distinct entity not as a supplement to a paper
* Broad dissemination via the internet
* Sharing with the public

**Why share data?**

* Data as a public investment
* Required by publishers
* Required by government funding agencies
* Informs new research
* Maximizes transparency, accountability and scrutiny of research findings
* Increases the impact and visibility of research
* Provides credit to the researcher as a research output in its own right
* Critical to the success of collaborative research
* Reduces duplication of effort
* Provides great resources for education and training

**Who benefits?**

* Researcher and research team
* Scientific communities (including citizen science)
* Students
* Public
* Funding agencies

**Overview of issues/obstacles related to reuse and sharing**

Data sharing refers to the practice of making research data available to others for validation and replication of results. Major national funding agencies such as the NSF, NEH, and NIH require all requests for funding contain a data management plan (DMP) addressing how the proposed project will comply with the agency's Data Sharing Policy. In addition to meeting funder requirements, sharing data benefits the scientific community and society by providing more opportunities for collaboration, better science, and well-informed policy decisions.

The NSF requires that the following question be addressed in all grant proposals’ data management plans:

“What will be the policies for data sharing and public access (including provisions for protection of privacy, confidentiality, security, intellectual property rights and other rights as appropriate)?”

When establishing data sharing and access policies and provisions, consider *whom*you will share your data with, *how*it will be shared, and *when*in the research process you will share it. For example:

* Is the data shared with other researchers or the general public?
* Whom will you share it with? Who may be interested in your data in the future and what might it be used for?
* Are there ethical issues or privacy concerns? Do any regulations apply to the data (e.g., HIPAA)?
* Do you have the right to share the data if it is not produced by you?
* Will you make the data available before or after you formally publish your results?
* Is your data understandable by other researchers?
* Should the data be restricted or embargoed for intellectual property reasons?
* Will the data be licensed? Will there be any conditions on its reuse?

There may be times when you may consider restricting access to your data if there are ethical issues, legal issues, or time constraints. There may be conditions of confidentiality associated with the research as well. Personal data or sensitive data may not be suitable for sharing with other researcher depending on whether informed consent has been obtained from participants. You may wish to consider anonymization techniques or data aggregation for numeric data, editing of video or sound recordings, use of pseudonyms in qualitative data, etc. In any event, consultation with your university’s research ethics office is advisable. In a case where you expect your research data has the potential to become commercially valuable or exploitable by a third party, you may want to speak with your university’s intellectual property expert.

There are several ways to share research data; the appropriate method will depend on the nature of your research and the content of the data.

Informal sharing: provide access to research data upon request

Supplemental information: provide research data in support of published articles

Institutional repository: deposit research data in local repository

Disciplinary data repository: deposit research data in an appropriate community-based repository

For an excellent review of data sharing see the United States Geologic Survey (USGS) Data Management “Why Share Your Data?” (see http://www.usgs.gov/datamanagement/share/guidance.php).

**Review potential legal restrictions on reuse**

Sharing data may implicate *third-party proprietary rights*, including patents, copyrights, database rights, trade secrets or other information protected by license or non-disclosure agreements (NDAs); *data subject rights*, relating to the subject of the information, such as privacy, defamation, etc.; or *content-related legal issues* specific to various countries you source or provide the data in, such as import restrictions, obscenity, blasphemy, or other rules.

You should assess your project, generally, for the types of information involved, and understand if those types of information might cause problems in places where the information would be disseminated.

While many of these are specific to the types of information involved, proprietary rights may crop up in almost any field. Consider the major categories of proprietary rights to be (a) patent, (b) copyright, (c) trademarks, and (d) trade secrets and other proprietary rights governed by contracts (NDAs).

As noted in module 5, **copyrights** provide rights holders the right to authorize, for their copyrighted works, the reproduction, dissemination, public display and performance, and preparation of derivative works. *Any original work created and "fixed" in a medium such as paper, or electronically, is automatically copyrighted.* You should assume that all textual, graphic, audio, and video materials are copyrighted, and maintain rights metadata for these materials -- even for works you yourself create, for which you are the copyright holder. This should include, at a minimum, (1) the author; (2) dates of creation or publication; and (3) any permissions you have or know of regarding the information, including general licensing (such as Creative Commons), user-specific licensing (such as a signed contract between you and the rights holder), or other legal authorizations (such as fair use).

While copyright is likely to touch many if not all materials you collect, you will not necessarily need to get permission to use the works in your project. Many countries provide specific exemptions for certain research and educational uses of copyrighted works. In the United States and some other jurisdictions, the **fair use** doctrine provides a broad and flexible framework that privileges many educational and scholarly uses of copyrighted works. "Facts" are generally free for reuse in the US and most other jurisdictions; however, if your project involves the wholesale reproduction or dissemination of a third-party collection of data, then the European Union and some other jurisdictions provide **database protection statutes** that apply.

**Patents** provide strong rights to the patent-holder against any non-permissive (unlicensed) use of the patented invention itself. Patents can cover intangible inventions, such as algorithms or business methods. In general, dissemination of information *about* the invention -- such as a patent application -- is not protected. If your project includes the dissemination of actual patented inventions -- such as software embodying patented algorithms, or tangible products such as organisms embodying constructed DNA, then you should investigate the exceptions available for research use, and the licensing available.

**Trademarks** can include almost any kind of "mark" -- a logo or slogan or name, even a scent or color or song -- that identifies the source of a product to consumers. Trademark law is intended to protect consumers as to source confusion. So use of trademarks should typically be done accurately to represent trademark owners, and if there is any possible confusion as to source or sponsorship, to make it clear that the trademark is the *subject* of the material, not the *sponsor* of it. For instance, a project involving assessment of a variety of trademarked substances could use those trademarks to refer to the relevant substances. So long as it is done in a way to minimize any confusion as to sponsorship, trademarks can be used to refer to their relevant products or services.

**Trade secrets, NDAs, and other licenses**

While legislative rights such as copyright, patent, and trademark form the background on which information is shared, contracts and licenses shape numerous specific relationships. If you have access to information through a licensing arrangement, or an individually negotiated and signed agreement (such as a "non-disclosure agreement" or a "confidentiality agreement"), you are in most cases bound to the terms of those agreements. It is important, when acquiring access to information, that you consider the future uses you may wish to make of it, and negotiate for those rights. If you would have rights under fair use or some other general legal exception, then it may be preferable to have your contract not mention a particular right or use at all, rather than have it provided in a very narrow way or excluded altogether.

**Subject-rights such as privacy, defamation**

Rights of individuals whose information is stored in your collection vary significantly across jurisdictions. In particular, if you are operating in the European Union, a comprehensive privacy framework governs the kinds of information you can collect, and what you can do with it. Other countries have similar laws. In the United States, privacy laws are well-described as a "patchwork", and if you work with personally-identifiable information about individuals, then you should apprise yourself of federal *and* state laws. Prioritize this issue if you work with financial or medical information, or information about children.

**Jurisdiction-specific rules about content**

Every country has its own rules that govern particular types of content. In the United States, the First Amendment is broadly protective of sharing information, but export-control rules may still apply to cryptographic research or other types of information affecting national security, and both state and federal laws may also restrict particular subjects, so long as the restriction comports with the First Amendment. In the European Union and many other jurisdictions, content defaming certain ethnic groups or promoting Nazism may be regulated. Numerous governments around the world police content around sexuality, religion, and/or particular state-supported religions. You should be broadly familiar with whether your kinds of content might be regulated anywhere, and in particular any rules for jurisdictions you are working in, sourcing information from, or disseminating information to.

**Review requirements & restrictions from funding agencies**

Since 1999, in compliance with the Freedom of Information Act (FOIA), federal funding agencies require grant recipients to make data produced during the awarded research available to the public. In order to ensure the availability of this data, many federal funding agencies have emphasized the importance of data management and storage, encouraging grant applicants to plan for a sustainable model of access. More recently, the White House Office of Science and Technology Policy has issues a [directive](http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research) in support of open access to research, requiring funding agencies “with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research.”

Below is a list of the data-related requirements where appropriate for each major science funding agency and their policies on data sharing:

**National Science Foundation (NSF)**

**Requirements:**

* Data management plan <http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp>

**Data sharing:**

“[NSF] expects PIs to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.”

**National Institutes of Health (NIH)**

**Requirements:**

* Data sharing plan for grants over $500,000 <http://grants.nih.gov/grants/policy/data_sharing/>
* Published research papers must be open access according to the NIH Public Access Mandate.

**Data sharing:**

The data produced in the course of the funded research “should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data.”

**Centers for Disease Control and Prevention (CDC)**

**Data sharing:**

“The purpose of CDC’s data release/sharing policy is to ensure that (1) CDC routinely provides data to its partners for appropriate public health purposes and (2) all data are released and/or shared as soon as feasible without compromising privacy concerns, federal and state confidentiality concerns, proprietary interests, national security interests, or law enforcement activities.”

<http://www.cdc.gov/maso/policy/releasingdata.pdf>

**Department of Defense (DoD)**

**Data sharing:**

“It is DoD policy under DoD Directive 3200.12 to establish and maintain a coordinated and comprehensive program to document the results and outcome of DoD-sponsored and/or performed research and engineering (R&E) and studies efforts and provide access to those efforts in an effective manner consistent with the DoD mission.”

<http://www.dtic.mil/whs/directives/corres/pdf/320014p.pdf>

**NASA**

**Data sharing:**

“The data collected by NASA represent a significant public investment in research. NASA holds these data in a public trust to promote comprehensive, long-term Earth science research. Consequently, NASA developed policy consistent with existing international polices to maximize access to data and to keep user costs as low as possible. These policies apply to all data archived, maintained, distributed or produced by NASA data systems.”

<http://science.nasa.gov/earth-science/earth-science-data/data-management-plan-guidance/>

**Publishers’ requirements vs. open access policies/controversies surrounding open science and open data**

There have long been individual scholars who questioned the system of scholarly publishing, though it wasn’t until the 1990’s, as the internet began to change publishing and the dissemination of information, that a movement arose challenging the traditional publishing model and espousing a more open and equal method for the public good. The Budapest Open Access Initiative (BOAI) in 2001 created the most commonly accepted definition of open access based on the publication of journal articles:

*By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.” [[1]](#footnote-1)*

The open science movement seeks to expand beyond the open access initiative to ensure that other types of research output such as uncurated data, lab notebooks, software code, methodologies, are easily accessible.[[2]](#footnote-2) Open science is conducted and communicated in ways that allow others to contribute, collaborate, and add to the research effort, with all kinds of data, results, and protocols made freely available at different stages of the research process. Proponents of open science argue that promoting collaboration in science will lead to more efﬁcient research and advance scientific innovations.

These are the four goals of open science:

* Transparency in experimental methodology, observation, and collection of data
* Public availability and reusability of scientific data
* Public accessibility and transparency of scientific communication
* Uses web-based tools to facilitate scientific collaboration.

*Open science is carried out and communicated in ways that allows others to contribute, collaborate and add to the research effort, with all kinds of data, results and protocols made freely available at different stages of the research process.[[3]](#footnote-3)*

There are many web-based tools that facilitate scientific collaboration. Such tools include online lab notebooks, Google Docs, YouTube videos, and Benchfly.

There is, however, concern about the potential for individuals, institutions, or entire nations profiting from the work of others. Additionally many researchers worry that their work may be misinterpreted or edited to diminish their contributions to science, rather than enhance it. Open science strives to make all scientific output open, so that the creators will be given credit and the research can be reproduced to decrease the incidence of unsubstantiated claims.

In 2009, a group of concerned scholars drafted the Panton Principles. In their opening statement on their website they assert:

*Science is based on building on, reusing and openly criticising the published body of scientific knowledge. For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open.* [[4]](#footnote-4)

**Open data and public policy**

There are four Panton Principles, and they prescribe ways for making the data underlying published science open. The rapidly evolving open data movement has attracted a growing number of advocates, supporters, and practitioners, and has made a significant impact on government policy. Remember since 1999, in compliance with the Freedom of Information Act (FOIA), federal funding agencies have required grant recipients to make data produced during the awarded research available to the public. In order to ensure the availability of this data, many federal funding agencies have emphasized the importance of data management and storage, encouraging grant applicants to plan for a sustainable model of access. In an effort to expand public access to the results of federally funded research, in February 2013 the Office of Science and Technology Policy released a policy directing “Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research.” [[5]](#footnote-5)

In May 2013 President Barack Obama signed an Executive Order requiring that “going forward, data generated by the government be made available in open, machine-readable formats, while appropriately safeguarding privacy, confidentiality, and security.” In conjunction with this Executive Order, the Office of Management and Budget and the Office of Science and Technology released an [Open Data Policy](http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf) that stated “information is a valuable national asset whose value is multiplied when it is made easily accessible to the public.” [[6]](#footnote-6)

Prior to the recent release of these federal policies, many research funders stopped short of requiring that research data be made openly accessible. These funders simply asked researchers that their “data be made available upon request.” This requisite has given researchers more control over who accesses their data. In theory this requisite should have prompted researchers to ensure that their data remain available long into the future. However, most researchers have ultimately failed to comply with these requirements. With the recent government mandates for open access to publically funded data, agencies are developing stronger policies for data management and sharing. Making data accessible by depositing it in an open access repository (institutional, or otherwise) actually minimizes the burden on the researcher in the long run and ensures compliance.

Requiring the data supporting published articles to be open is in the interest of publishers to preserve the integrity of the publication as well as the public good; however the logistics of sharing data can be more complicated. The Journal Data Archiving Policy (see JDAP, http://datadryad.org/pages/jdap) was created to provide a standard policy for journals to adopt and a list of the journals that have done so. The JDAP policy consists of the following text:

*<< Journal >> requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as << list of approved archives here <<. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species.*

Authors publishing in a journal that has adopted the JDAP policy or a similar standard will be required to supply a persistent link to the data in an appropriate archive as part of the article submission process. Including access to data supporting published articles allows the research results to be verified or disputed. Archiving the data in an appropriate repository or archive ensures that the data is citable if it is to be reused.

Publishers, funding agencies, and institutions may all have separate requirements and restrictions concerning the sharing of research data. It is the responsibility of the researcher to comply with these restrictions, which can be in conflict. These requirements are subject to change, and have been changing rapidly over the past decade. Researchers that are granted funds are required to keep abreast of these changes, and comply. Information about data management and data sharing requirements can usually be found on the funders’ websites, though they are not always easily found:

* National Institutes of Health (NIH):
  + http://grants.nih.gov/grants/sharing.htm
* Alfred P. Sloan Foundation:
  + http://www.ur.umich.edu/update/archives/121002/sloan
* National Oceanic and Atmospheric Association (NOAA):
  + http://www.corporateservices.noaa.gov/ames/administrative\_orders/chapter\_212/212-15.html
* National Science Foundation (NSF):
  + http://www.nsf.gov/eng/general/dmp.jsp

The open data movement is rapidly evolving, and the number of advocates, supporters, and practitioners is growing. Though many funders stop short of requiring data to be openly accessible (and this is all beginning to change), they usually ask that data be made “available upon request.” This gives researchers more control over who accesses their data, but it also theoretically requires that they ensure their data remain available long into the future.

**Resources:**

1. Wikipedia page on “Open Science”: http://en.wikipedia.org/wiki/Open\_science. Accessed on 29 March 2013.
2. Dryad JDAP: http://datadryad.org/pages/jdap. Accessed on 29 March 2013.
3. OKFN Open Science Working Group’s Open Science: http://science.okfn.org/about-us/open-science-overview/. Accessed on 29 March 2013.

**Standard data formats and records management practices that facilitate reuse**

There are different ways of defining *format*; two are listed below:

* *packages of information that can be stored as data files or sent via network as data streams (aka bitstreams, byte streams).* (<http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml>); and
* *a set of semantic and syntactic rules governing the mapping between abstract information and its representation in digital form*. (http://www.udfr.org/).

Both of these definitions are correct, though both are perhaps overly complicated for our purposes here. We are typically interested in *file formats*, which are simply defined by Wikipedia (http://en.wikipedia.org/wiki/File\_format) as being:

*a standard way that information is encoded for storage in a computer file. A file format specifies how bits are used to encode information in a digital storage medium.*

It’s important to store data (textual, image, dataset, etc.) in as open a format as possible, while still keeping the characteristics of the data and its representation intact. If data, or its documentation is held in a proprietary, platform-dependent format, it is likely that these data will be inaccessible in the future, or will need to undergo a costly migration process before it is accessible and useable. Examples of problematic formats include Microsoft Word, or Microsoft Excel. It is far better to convert data (files) to an open format (Open Document Format/ODF) or a more sustainable, platform-independent format such as CSV, or PDF/A.

The Library of Congress, National Digital Information Infrastructure and Preservation Program (NDIIPP) has created a comprehensive resource on formats for preservation. It makes very specific recommendations on file formats.

The Library of Congress recommendations are listed in the table below.

|  |  |  |
| --- | --- | --- |
| **Date type** | **Format** | **More information** |
| Data Sets | CSV, DBF, CDF, HDF | http://www.digitalpreservation.gov/formats/fdd/dataset\_fdd.shtml |
| Text-based documents | PDF, PDF/A, XML | http://www.digitalpreservation.gov/formats/fdd/text\_fdd.shtml |
| Still images | TIFF, JP2 (JPEG2000) | http://www.digitalpreservation.gov/formats/fdd/still\_fdd.shtml |

For more in-depth coverage on formats and their place in the data management life cycle we suggest referring to the Digital Curation Centre’s Digital Curation Manual, especially their *Installment on File Formats*, a chapter written by Stephen Abrams of the Harvard Library (see http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/file-formats/file-formats.pdf). An excerpt from this chapter, in the “Conclusions” section, is reproduced below:

*An increasingly rich set of preservation-friendly formats, tools that can process these formats, and best practices concerning the selection, creation, use, and preservation of these formats is evolving within the digital curation community. However, the ubiquity, variety, and sophistication of new content forms are growing at perhaps an even greater pace. As much of this new content is expressed in new formats, or new applications of existing formats, it is incumbent upon curation practitioners to understand the central position format plays in preservation and access environments.* (pg. 38-39)

Pay particular attention also to the section called “Development of preservation friendly formats” beginning on page 18.

One can quickly become overwhelmed by the various format specifications, and as Abrams notes, new formats and content forms are being created. But there are many resources at our disposal. Just keep in mind that it is best to use open, non-proprietary formats, like those listed above, if, wherever, and whenever possible.

**Resources**

1. Abrams, Stephen. 200.7 DCC Digital Curation Manual Instalment on “File Formats”. Version 1.0. http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/file-formats/file-formats.pdf
2. Arms, Caroline R. and Carl Fleischhauer. 2005. Digital Formats: Factors for Sustainability, Functionality, and Quality. Paper for IS&T Archiving 2005 Conference, Washington, D.C. http://memory.loc.gov/ammem/techdocs/digform/Formats\_IST05\_paper.pdf
3. ISO 14721:2012 Reference Model for an Open Archival Information System. The Consultative Committee for Space Data Systems. Magenta Book. June 2012. <http://public.ccsds.org/publications/archive/650x0m2.pdf>

**Sharing within and outside your research team**

**Sharing within**

You may want to share data among other team members that are working on different aspects of a project. These are some considerations to aid you in being organized and making sharing data easier.

Location: where is it stored?

Folder structure: how easily can you locate folders?

File naming standards: are files logically labeled?

Versioning: how do you know if this data set represents the latest data?

Formats: can this file be easily opened?

Responsibility: whose job is it to oversee this organizational structure? Who is documenting and performing quality assurance?

Communication: does everyone know where information is stored and how to access it?

Documentation and Back up: how do you record information and share information in the laboratory notebook? Are paper notebooks scanned? Where are these stored and how are they shared?

**Sharing with others outside the team**

* Consult your university data policy/intellectual property policy
* Consult funders’ policies
* Consult your PI
* Consult your IRB office

Be aware of patient confidentiality– does the data set need to be anonymized?

Beware data ownership--do you have the right to share these data?

Timing: At which point do you share data? What is your purpose for sharing data? Do you share raw data or analyzed data? Should data be “peer reviewed”? How do you version this data? Which data set will you submit to a repository? What type/how much to share for different steps: publication, archival and preservation, retention requirements by funders, institution, IRB, etc.?

Traditionally, research papers have been published as a single package, containing both the presentation and analysis of ideas along with the supporting data. There has been a recent shift towards a requirement that all relevant data, including preliminary data (unpublished) be made publicly available at the time of publication, or earlier in the research phase of the grant cycle. Considered to be an inevitable trend, many have called for funding agencies to require preliminary data to be released for projects when data more rapidly advances research, especially data that will benefit the public good.

Researchers face an increasingly competitive funding and publication environment, and data produced from competitive grants is highly valued. The question of whether or not to share preliminary data can create tension between the data-producing scientists who have the expectation of the right to publish first and the scientists who wish to publish their own analysis of that data. Additionally, these valuable data sets can also become a stepping-stone to the success of the next funding opportunity, which is another reason data producers may seek to keep their data private.

Outlined below, are the pros and cons to consider in this debate:

Pros

* The sharing of any and all data with those in your own field can better advance this research
* Sharing of preliminary data can lead to a co-authorship for the data producer if that shared data becomes an essential part of another’s analysis
* Thinking about data sharing early in the process can lead to more robust data description and quality at the time of publication, and beyond for long-term preservation
* Publishing preliminary data can lead to coordinated publications and new insights

Cons

* Preparing data for sharing and/or responding to requests for data can be a very time consuming activity for data producers
* Other researchers may misuse or misinterpret the shared data
* Preliminary data are likely to be made available before thorough quality assurance can be performed
* Other researchers may end up publishing results that overlap with the plans of the data producer

As the mandated sharing of all research data is trending toward the inevitable, funders, research institutions and publishers must clearly state their requirements for the sharing of data, including preliminary data.

**Using data created by others**

When making a request to use another’s data, be specific. Are you looking for an entire data set that includes multiple variables or are you looking for a subset of this data? Are you looking for data from a specific time frame, in a specific geographic area? Knowing and communicating to the data creator what the “boundaries” are of the data you want to access facilitates the sharing process.

**Citation / Attribution of Data**

Acknowledgement of the use of someone else’s information or work is a long-accepted practice in scholarly communication. This attribution is usually accomplished via a “citation”: when you publish a paper or do a presentation that makes use of someone else’s information, you include in that paper or presentation a reference to the work of that other person or group. It is important to cite not only the literature consulted but also the data files used, including your own. Plagiarism takes place when you use someone else’s data or words and do not acknowledge that you have done so.

Citing data files in publications based on those data serves several purposes:

* Provides appropriate credit to the data producers and publishers
* Enables other researchers to access the data for their own use or to replicate research findings
* Assists in measuring the impact of a data set by tracking references to it in the scientific literature and in online conversations (“altmetrics”)
* Helps data producers know how their data is being utilized

Data citation is evolving and there is currently no acknowledged standard on how to cite a dataset or construct a data citation. However, several organizations have developed their own practices, and international groups are creating formal guidelines for the scientific community. The following elements are generally considered the core elements of a data citation:

* Author/Creator(s): the creators of the data; can be one or more people or organizations
* Title: the title of the data set
* Version: the exact version or edition of the data set used
* Publication Date: the date when the data set was published or released
* Publisher/Archive: the data center or repository that is archiving and distributing the data
* Identifier/Locator: URL or other linkable locator for the data; a persistent, permanent URL such as a DOI (Digital Object Identifier) or a handle is preferred

Authors should cite data they use according to established journal or professional guidelines. Examples of data citations from style manuals:

APA Style: Publication Manual of the American Psychological Association, 6thEdition (Pages 210-211).

Pew Hispanic Center. (2004). Changing channels and crisscrossing cultures: A survey of Latinos on the news media [Data file and code book]. Retrieved from <http://pewhispanic.org/datasets/>

Citing Medicine: The NLM Style Guide for Authors, Editors, and Publishers, National Library of Medicine, 2nd Edition, [http://www.ncbi.nlm.nih.gov/books/NBK7273/#A57719](#http://www.ncbi.nlm.nih.gov/books/NBK7273/)

WormBase: the Biology and Genome of C. elegans [Internet]. Release WS170. [Pasadena(CA)]: California Institute of Technology. c1999 - . Pedigree browser for MSaa; [cited 2007 Feb1]. Available from: <http://www.wormbase.org/about/citing_wormbase#012--10>

The Digital Curation Centre (DCC) in the UK has released comprehensive data citation guidelines that can be applied to almost any discipline, available at <http://www.dcc.ac.uk/resources/how-guides>.

Repositories often provide guidance on how to cite their data sets. Here are some specific guidelines and practices for data citation:

DataCite consortium: “Why Cite Data?” http://datacite.org/whycitedata and metadata schema of core elements of a data citation <http://www.datacite.org/whycitedata>

Dataverse Network Project: “Data Citation Standard” <http://thedata.org/book/standard>

Dryad, a nonprofit repository for data underlying the international scientific and medical literature: “How do I cite data from Dryad?” [http://datadryad.org/pages/faq#using](#http://datadryad.org/pages/faq)

Federation of Earth Science Information Partners: “Interagency Data Stewardship/Citations/Provider Guidelines” <http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines>

ICPSR, Inter-university Consortium for Political and Social Research: “Citing Data” <http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp>

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | ACTIVITY: | | Have students identify the components of the following data citation: | | Garbarino S, Cuomo G, Chiorri C, Magnavita N (2013) Data from: Association of work-related stress with mental health problems in a special police force unit. Dryad Digital Repository. [doi:10.5061/dryad.558pm](http://dx.doi.org/10.5061/dryad.558pm) | |

**Resources for Citing Data**

1. Ball, A. & Duke, M. (2012). How to Cite Datasets and Link to Publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. <http://www.dcc.ac.uk/resources/how-guides/cite-datasets>
2. Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). <http://www.dlib.org/dlib/march07/altman/03altman.html>
3. Green, T. (2009). We need publishing standards for datasets and data tables. OECD Publishing White Papers, OECD Publishing. <http://www.oecd.org/document/25/0,3746,en_21571361_33915056_42600857_1_1_1_1,00.html>
4. Mayernik, M.S. (2012). Data citation initiatives and issues. Bul. Am. Soc. Info. Sci.Tech., 38: 23–28. doi: 10.1002/bult.2012.1720380508. <http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.html>

Readings

1. Guttmacher, A.E., Nabel, E., and Collins, F.S. (2009). Why data-sharing policies matter. [Why data-sharing policies matter](http://www.pnas.org/content/106/40/16894.short) <http://www.pnas.org/content/106/40/16894.full>
2. Curtin, C. (2010). Data Sharing and Consent. [http://www.genomeweb.com/data-sharing-and-consent](http://www.genomeweb.com/data-sharing-and-consent%20)
3. Lathrop, R.H. and Burkhard, R. (2011). ISCB Public Policy Statement on Open Access to Scientific and Technical Research Literature. <http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002014>
4. National Academies. 2010. The Value of Shared Access and Reuse of Publicly Funded Scientific Data. <http://sites.nationalacademies.org/PGA/brdi/PGA_059258>
5. Vision, T. (2010). Open Data and the Social Contract of Scientific Publishing . OECD Publishing White Papers, OECD Publishing. <http://www.jstor.org/stable/10.1525/bio.2010.60.5.2>
6. Birney, E., Hudson, T.J., Green, E.D., Gunter, C., Eddy, S., Rogers, J., Harris, J.R., et al. “Prepublication Data Sharing.” *Nature* 461, no. 7261 (2009): 168–170. <http://www.ncbi.nlm.nih.gov/pubmed/19741685>
7. Cragin, M. H., Palmer, C.L., Carlson, J.R., and Witt, M. “Data Sharing, Small Science and Institutional Repositories.” *Philosophical Transactions of the Royal Society - Series A: Mathematical, Physical and Engineering Sciences* 368, no. 1926 (2010): 4023–4038. <http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2010.0165>
8. Piwowar, H.A., M.J. Becich, H. Bilofsky, R.S. Crowley, Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. *PLoS Med*. 5(9), e183 (2008), [doi:10.1371/journal.pmed.0050183](http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0050183).
9. Ball, A. & Duke, M. (2012). How to Cite Datasets and Link to Publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. <http://www.dcc.ac.uk/resources/how-guides/cite-datasets>
10. Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). <http://www.dlib.org/dlib/march07/altman/03altman.html>
11. Green, T. (2009). We need publishing standards for datasets and data tables. OECD Publishing White Papers, OECD Publishing. <http://www.oecd.org/document/25/0,3746,en_21571361_33915056_42600857_1_1_1_1,00.html>
12. Mayernik, M.S. (2012). Data citation initiatives and issues. Bul. Am. Soc. Info. Sci.Tech., 38: 23–28. doi: 10.1002/bult.2012.1720380508. <http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.html>

1. Budapest Open Access Initiative. “BOAI 10”. <http://www.opensocietyfoundations.org/openaccess/boai-10-recommendations>

   Accessed on 12 August 2013. [↑](#footnote-ref-1)
2. Wikipedia. page on “Open Science”. http://en.wikipedia.org/wiki/Open\_science. Accessed on 29 March 2013. [↑](#footnote-ref-2)
3. The Open Science Project. “What exactly is Open Science?” <http://www.openscience.org/blog/?p=269> Accessed 12 August 2013. [↑](#footnote-ref-3)
4. Panton Principles. <http://pantonprinciples.org/>). Accessed 12 August 2013. [↑](#footnote-ref-4)
5. Office of Science and Technology Policy. “Expanding Public Access to the Results of Federally Funded Research.” <http://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research> Accessed on 12 August 2013. [↑](#footnote-ref-5)
6. White House. “Obama Administration Releases Historic Open Data Rules to Enhance 5. Government Efficiency and Fuel Economic Growth.” <http://www.whitehouse.gov/the-press-office/2013/05/09/obama-administration-releases-historic-open-data-rules-enhance-government> Accessed on 12 August 2013. [↑](#footnote-ref-6)