Development Data Overview: Managing, protecting, and sharing your PEER data

Table of Contents

Introduction to the PEER Data Management Program

What is the PEER Data Management Program?

The PEER Data Management Program is an initiative to support PEER Principal Investigators (PIs) in applying best practices for managing scientific research data.

The goal of the PEER data management program is to help PEER researchers be great data managers. The program also seeks to make data as open as possible while ensuring sensitive information is protected and PIs get credit for the work they do.

Why does PEER have a Data Management Program?

Most science funding agencies, including USAID, require that research results, including data, be made publicly available. And journals are also requesting that authors post their data along with their articles. Being a good data manager is an important skill for researchers' career advancement. Just as important, good data management, particularly when it makes data accessible and open, facilitates new scientific discoveries.

Many in the scientific community have been embracing the philosophy of open data and transparency in research. Open data is the idea that data should be freely available for everyone to use. When research data is freely available, scientists do not need to collect all the needed data themselves; they can use previously collected and shared data instead. This allows them to focus their attention and resources on only collecting new data that the world has never collected before. Making research data accessible, discoverable, and usable to the international community fuels entrepreneurship, innovation, scientific discovery, and enhanced development outcomes. It also makes science more efficient and cost-effective because it removes duplicative work. For instance, someone might collect data on mosquitos to study disease transmission, but other researchers might want to use the same data to study biodiversity or the impacts of climate change. If you properly share your data, so that it can be easily found and cited, *you will get credit* when other researchers or government officials use your data!

Applying best practices for managing and sharing scientific research data benefits PEER PIs. You can demonstrate your commitment to high-quality research and to research integrity through strong data management. Recently, the US National Academies of Science, Engineering, and Medicine released the report "Fostering Integrity in Research". A primary recommendation is that researchers work with sponsors and with publishers to ensure that all information and data necessary to verify, reproduce, and otherwise reuse research information is available. In addition, proper data management and sharing means you receive credit for all your work, your publication, but also for the high quality data you

produce. At the same time open data increases your potential for high-quality, frequently viewed journal publications (see article JAMA. 2018;319(4):410. doi:10.1001/jama.2017.20650).

Another motivation for the PEER data management program is that data best practices increase the potential for future collaborations. If you demonstrate a commitment to data curation and data publication, then other researchers will become familiar with your high-quality data. You can more easily engage others around the world in meaningful, productive collaborations. It is a way to demonstrate your commitment to cutting edge research best practices and to technologies that drive modern, digital research.

What does the PEER Data Management Program ask me to do as part of my USAID-funded research project?

USAID and the PEER program want PEER researchers to be great data managers, produce high quality data, and make that data as open as possible so others can also benefit.

Steps for PEER data management and sharing:

1. **Choose a repository for your data**. A data repository is a place where data sets are kept for long-term storage and sharing. You should select a repository for your data by the end of year 1. We recommend collaborating closely with your US partner on this.
    - For social science and public health research, we suggest using the USAID Development Data Library: https://www.usaid.gov/data.
    - For other types of research, you can start looking for repositories here: https://www.coretrustseal.org/why-certification/certified-repositories/, http://www.sciencemag.org/authors/science-editorial-policies#data-deposition, and https://www.nature.com/sdata/policies/repositories

2. **Create a short data management plan.** A data management plan is a formal document that outlines what you will do with your data during and after a research project. This must be completed by the end of year 1 of your PEER award. You will create your data management plan in the PEER grant management system Foundant.

3. **Submit data to a repository and to the USAID Development Data Library.** As you produce research deliverables you must submit a copy of any dataset created or obtained in performance of this award in a machine-readable, non-proprietary format to an online data repository of your choosing. The repository must be trustworthy and support data citation so that you get credit. The submission must include supporting documentation describing the dataset, such as code books, data dictionaries, data gathering tools, notes on data quality, and explanations of any redactions. You must **also** submit your data to USAID's Development Data Library, called the DDL. The data can be embargoed, meaning stored privately, so that it does

not become public before your paper is published.

4. **Upload publications to USAID's document platform**. Most PEER principal investigators will author journal articles or other intellectual works. PEER awardees must submit a copy of intellectual works, particularly peer-reviewed publications, to USAID's publication library, the Development Experience Clearinghouse (DEC).

These activities are described in more detail in Section 3 of this document, PEER Data Management Implementation.

Structure of this Guide

The first section of this guide, Introduction to the PEER Data Management Program, provides an overview of the PEER data management program and explains why USAID emphasizes data management, data sharing, and public access to research results.

The second section, Key Concepts, explains data management concepts and defines terminology. Understanding these key concepts will help you as you create your PEER data management plan and ensures that you are able to share your data and publications.

The 3rd section of this document, PEER Data Management Implementation, details the steps to implementing good data management and complying with PEER data and publication requirements.

Key Concepts

The field of data management uses a set of common concepts and terminology. The objective of this section is to define key concepts and address frequently asked questions. Understanding these concepts will help you create your PEER data management plan and ensure that you are able to share your data and publications.

Metadata

Frequently asked question: Will someone else use my data? If so, might that person or organization misunderstand or mis-use my data?

Answer: Yes, one of the goals of research data management is to facilitate future re-use. There are several steps you can take to ensure that others will understand the data and use it properly. Above all, you should thoroughly document your data and your procedures:

- Provide **metadata** that explains your research data.
- Create codebooks for your data.

- Get informed consent to collect data (where required).
- Document the methodologies used to collect your data.
- Include source code and analysis workflow! (if applicable to your project).

Key concept: Metadata. Every researcher knows what 'data' is, but what is metadata? Metadata provides descriptive information (content, context, quality, structure, and accessibility) about a data product and *enables others to search for and use the data product*. A classic example of metadata is a library card catalog, seen in Figure 1. Each card within the catalog contains metadata about a single book, such as its title, author, publisher, and a location identifier (e.g. Dewey Decimal Number). In other words, each card contains enough descriptive data about the content and location of the book, that an interested reader can know if the book is relevant to their interests and where to find the book on the library shelves. This example will be expanded upon in the section on Metadata Standards. Without proper collection and storage of the books' metadata, finding a specific book within a large library would be almost impossible.

Figure 1: A library card catalog. Every drawer holds hundreds of cards and each card contains the metadata for a single book. Credit: wiseGEEK.

Similar to the example of book metadata, the metadata that you provide for your own research data is important for others to be able to find, use, and cite your work. The metadata should include all information needed for your data's future use and it should be attached to the data set itself, usually as separate files. Here is a list of what your metadata might contain:

- a brief or detailed description of the data itself;
- names, labels and descriptions for variables, records and their values;
- explanation of codes and classification schemes used;
- codes of, and reasons for, missing values;

- derived data created after collection, with code, algorithm or command file used to create them;
- weighting and grossing variables created and how they should be used;
- data listing with descriptions for cases, individuals or items studied, for example for logging qualitative interviews;
- descriptions of applications (commercial or open-source) were used to run analyses, and the versions of those applications;
- descriptions of file formats used to store the data;
- documentation of experimental protocols;
- documentation of the code written for statistical and other analyses.

Figures 2 and 3 provide a very simple example. Figure 2 shows the data collected by a researcher from their fieldwork. As you can see, this data is completely unuseable because the columns are not labeled. The numbers are meaningless. We could try to search for this data set to find more information about it, but what would we search for? Maybe you recognize that Gelisols are a type of soil, but that's not enough information to track down where this data came from.

36	KU-T1-01	72.29293	126.211	Degraded	Gelisol	Histel	9.8	9.8	4.4	0.2	6.1	0.1	72.6
37	KU-T1-02	72.29223	126.2128	Degraded	Gelisol	Turbel	9.6	16.3	4.8	0.2	12.2	0.3	52.6
38	KU-T1-05	72.28994	126.2176	Degraded	Gelisol	Turbel	9.9	28.4	3.9			0.8	10.9
39	KU-T1-06	72.28914	126.2187	Non-degr	Gelisol	Turbel	9.9	29.3	7.7	0.3	18.7	0.6	8
40	KU-T1-07	72.28832	126.2201	Non-degr	Gelisol	Orthel	8.2	34.7	6.6			0.5	5
41	KU-T1-08	72.28748	126.2212	Non-degr	Gelisol	Orthel	7.6	37.3	5.3	0.3	16.7	0.7	0
42	KU-T1-09	72.28671	126.2226	Non-degr	Gelisol	Orthel	13.5	37.7	6.9			0.5	24.8
43	KU-T1-10	72.28581	126.2238	Degraded	Gelisol	Turbel	9.4	46.8	3.5	0.2	19.3	1.4	0
44	KU-T2-01	72.29025	126.1936	Non-degr	Gelisol	Turbel	17.8	63.4	10.9	0.6	17.8	0.7	12.6
45	KU-T2-02	72.28947	126.1922	Non-degr	Gelisol	Turbel	9.6	30.7	5.5	0.3	15.7	0.7	11.1
46	KU-T2-03	72.28868	126.1908	Degraded	Gelisol	Turbel	7	17.6	4.9			0.4	41.1
47	KU-T2-04	72.28788	126.1893	Degraded	Gelisol	Turbel	8.2	24.4	2.7	0.2	14.9	1.1	5.7
48	KU-T2-05	72.28705	126.188	Degraded	Gelisol	Turbel	11.3	19.5	2.9	0.2	9.1	0.6	36.9

Figure 2: Data set without metadata. Taken from here.

Figure 3 shows the exact same data, but with an appropriate amount of metadata included. The metadata on line 35 gives unique and meaningful names to the columns in the data table below. This allows us to understand the structure of the data. From lines 7-24, those column names are described in more detail. For instance, line 20 explains that 'TN [%]', the name of Column O, stands for the total amount of nitrogen found in the soil, to a depth of 100 cm, measured as a percentage. With just the metadata from lines 7-24 and 35, an independent researcher could begin to use the data set. However, more metadata is required for independent researchers to be able to easily *find* the data set. For instance, line 6 provides GPS coordinates for a box that surrounds the all of the field sites. This way, a researcher can search for soil data samples by geographic location. Similarly, lines 3 and 4 provide the names of journal papers associated with the data set, so it is possible to search for the data set by searching for those journal articles. If another researcher wants to use this data for their own analysis, line 2 provides the format for how to cite the data set. This is important because consistent citations of the data set makes it easier for the original researcher (Matthias Siewart) to track the impact of his work.

	A	B	C	D	E	F	L	M	N	O	P	Q	R
1	/* DATA DESCRIPTION:												
2	Citation:	Siewert, Matthias Benjamin; Hugelius, Gustaf; Heim, Birgit; Faucherre, Samuel (2016): Soil organic carbon (SOC) storage in the											
3		In supplement to: Siewert, Matthias Benjamin; Hugelius, Gustaf; Heim, Birgit; Faucherre, Samuel (2016): Landscape controls a											
4	Related to:	Schneider, Julia; Grosse, Guido; Wagner, Dirk (2009): Land cover classification of tundra environments in the Arctic Lena Delta											
5	Project(s):	Changing Permafrost in the Arctic and its Global Effects in the 21st Century (PAGE21) (URI: http://www.page21.eu)											
6	Coverage:	MEDIAN LATITUDE: 72.357409 * MEDIAN LONGITUDE: 126.335384 * SOUTH-BOUND LATITUDE: 72.283111 * WEST-BOUND LONGI											
7	Parameter(s):	Event label (Event)											
8		Latitude of event (Latitude)											
9		Longitude of event (Longitude)											
10		Landform (Landform)											
11		Soil order (Soil order)											
12		Soil suborder (Soil suborder)											
18		Carbon, organic, per area [kg/m**2] (Corg area) * COMMENT: Depth in soil: 0-30 cm											
19		Carbon, organic, per area [kg/m**2] (Corg area) * COMMENT: Depth in soil: 0-100 cm											
20		Carbon, total [%] (TC) * COMMENT: Depth in soil: 0-100 cm											
21		Nitrogen, total [%] (TN) * COMMENT: Depth in soil: 0-100 cm											
22		Carbon/Nitrogen ratio (C/N) * COMMENT: Depth in soil: 0-100 cm											
23		Density, dry bulk [g/cm**3] (DBD)											
24		Ice content [%] (Ice) * COMMENT: Visible											
32	License:	Creative Commons Attribution 3.0 Unported (CC-BY)											
33	Size:	1040 data points											
34	*/												
35	Event	Latitude	Longitude	Landform	Soil order	Soil subor	Corg area	Corg area	TC [%]	TN [%]	C/N	DBD [g/cn	Ice [%]
36	KU-T1-01	72.29293	126.211	Degraded	Gelisol	Histel	9.8	9.8	4.4	0.2	6.1	0.1	72.6
37	KU-T1-02	72.29223	126.2128	Degraded	Gelisol	Turbel	9.6	16.3	4.8	0.2	12.2	0.3	52.6
38	KU-T1-05	72.28994	126.2176	Degraded	Gelisol	Turbel	9.9	28.4	3.9			0.8	10.9
39	KU-T1-06	72.28914	126.2187	Non-degr	Gelisol	Turbel	9.9	29.3	7.7	0.3	18.7	0.6	8
40	KU-T1-07	72.28832	126.2201	Non-degr	Gelisol	Orthel	8.2	34.7	6.6			0.5	5
41	KU-T1-08	72.28748	126.2212	Non-degr	Gelisol	Orthel	7.6	37.3	5.3	0.3	16.7	0.7	0
42	KU-T1-09	72.28671	126.2226	Non-degr	Gelisol	Orthel	13.5	37.7	6.9			0.5	24.8
43	KU-T1-10	72.28581	126.2238	Degraded	Gelisol	Turbel	9.4	46.8	3.5	0.2	19.3	1.4	0
44	KU-T2-01	72.29025	126.1936	Non-degr	Gelisol	Turbel	17.8	63.4	10.9	0.6	17.8	0.7	12.6
45	KU-T2-02	72.28947	126.1922	Non-degr	Gelisol	Turbel	9.6	30.7	5.5	0.3	15.7	0.7	11.1
46	KU-T2-03	72.28868	126.1908	Degraded	Gelisol	Turbel	7	17.6	4.9			0.4	41.1
47	KU-T2-04	72.28788	126.1893	Degraded	Gelisol	Turbel	8.2	24.4	2.7	0.2	14.9	1.1	5.7
48	KU-T2-05	72.28705	126.188	Degraded	Gelisol	Turbel	11.3	19.5	2.9	0.2	9.1	0.6	36.9

METADATA

DATA

Figure 3: The same data set at the previous figure, but with metadata included.

Notice that when you could only see the 'data' section, you had no idea what the numbers and categories corresponded to. You also did not have any way to search for this data even if you are told what it is. That is why metadata is necessary; it gives the information required to discover new data and to understand the data that has been found.

Metadata Standards

Metadata can take many different forms, from free text to highly structured content that uses a "metadata standard". A **metadata standard** is a format or structure for organizing metadata that is widely used by a research community. When many researchers use the same metadata standard, it is easier to search for new data sets and to merge multiple data sets together. Each academic discipline has their own metadata standards and there may be many metadata standards in a single research field.

Let's look at the metadata shown in Figure 3 again. The column names are given on line 35, but where did those names come from? Most likely, the researcher just made them up himself. He chose to label the total carbon content of the soil in column N as "TC [%]"; however, a different researcher might have named this column "C [%]", "totalCarbon", "total_carbon", "TotCar", or even just "C". The list goes on. Having so many different possible names for the *exact same quantity* makes comparing similar data sets very difficult. It would be better if all of the soil scientists could agree on a single name for this quantity, which they would then all use when labeling their own data sets. The name that they agreed upon would be part of a 'metadata standard' -- a consistent and widely used format for documenting research data.

In addition to specifying common names to use for column headers, a metadata standard may also *define the structure of your data and metadata*. Returning to the library catalog example, Figure 4 shows an individual card from within the catalog. Every card in the library's catalog system will have this exact same format. The author name (metadata 1) will always be at the top of the card on the left. The proper title of the book (metadata 2a and 2b) will always be below that. Along the bottom of the card, the LC classification number, Dewey number, Holding library, and LC control number will always appear in that exact order. The fixed locations of these pieces of metadata on the card are a part of the library's metadata standard. Further, the physical size of the card is part of the metadata standard; all cards must be the same size. Within the catalog, the cards are also ordered alphabetically by the last name of the author; cards could have been organized in alphabetical order according to the author's first name, the title of the book, or by the unique ISBN number (metadata 6), but librarians all over the world have agreed to organize their catalogs alphabetically by the author's last name as part of their metadata standard. This example demonstrates how a metadata standard can define more than just the *type* of information required, but also how the information should be stored and organized.

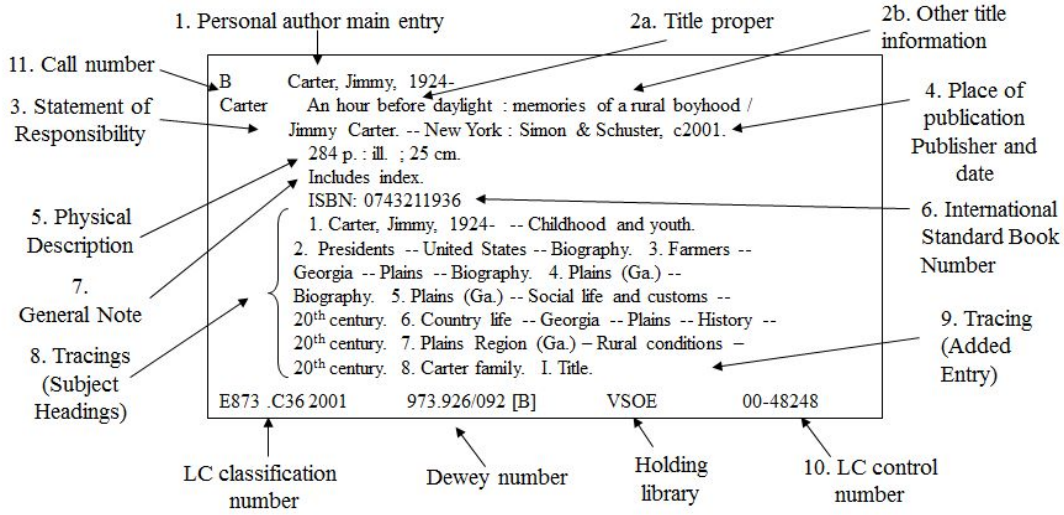

Figure 4: Elements of a library catalog card. Credit: Lorraine Lanius.

Similar to the library card catalog example, many data repositories suggest or even require specific metadata standards, especially about the file and folder structure of how your data is organized. If you know that your chosen repository requires a specific metadata standard, that makes it easy for you to choose how to organize your data and metadata. Because creation of standardized metadata can be time consuming, another consideration when selecting a standard is the availability of tools that can help generate the metadata. If your chosen repository does not require or suggest a metadata standard and you do not know which one you should use, here is a brief list of recommendations based on your research discipline.

Selected Recommendations:

- Social, Behavioral, and Economic Sciences
  - Consider using the DDI standard
  - The tool Nesstar Publisher (Windows only) can help manage DDI metadata
- Earth, Environmental, and Ecological Sciences
  - Consider using the EML standard
  - The tool Morpho can help manage EML metadata
- Geospatial data
  - You will almost surely use the ISO 19115 standard
  - Whatever tool you are currently using to manage your data will probably already use this standard (ArgGIS, QGIS, etc.)
- Other
  - Check this list to get ideas for an appropriate metadata standard
  - Ask your USG partner!
  - Ask me

Unfortunately, not every academic discipline has a good metadata standard. Do not worry if you cannot find a standard that works for your data. Instead, you can write "readme" files for your data. **Cornell University has a fantastic discussion on how to write "readme" metadata that you should follow if you will not be using a metadata standard.** They also provide a template for writing a good readme file. To supplement the Cornell discussion, please see the Minimum Metadata To Include section of the Appendix.

Frequently asked question: This is a lot of work. What is the benefit for my work?

Answer: The benefits to you include:

- Getting credit for your hard work (e..g, citations for your data!)
- Potential for increased probability of journal publication acceptance (see article JAMA. 2018;319(4):410. doi:10.1001/jama.2017.20650)
- Demonstrating a concrete commitment to high-quality research results and procedures and to scientific reproducibility (i.e., a trusted researcher with transparent, trusted results!)

- Increase the potential for future collaborations. If you demonstrate a commitment to data curation and data publication, then other researchers will become familiar with your high-quality data. You can more easily engage others around the world in meaningful, productive collaborations.
- Embracing new best practices at the cutting edge of the digital research world

The benefits to future researchers (including YOU as a future researcher) include:

- Available, accessible data to verify past results
- Available, accessible data to jump-start a new project (e.g., build off past results quickly!)
- Available, accessible data for even larger-scale, interdisciplinary or comparative research
- Available, accessible data to pioneer new, innovative ideas at the edge of scientific research (e.g., training data for artificial intelligence [neural network] initiatives, growing geospatial and real time traffic data for automated vehicle research)

Data Repository

Frequently asked question: There are many steps to use research data management best practices and to ensure that I store and share my data properly. How do I ensure my data are well curated and properly shared/published?

Answer: The scientific community recognizes that research data management is a team sport that requires support and resources. Many scientific disciplines and many institutions, including Government Agencies like USAID, have created digital **data repositories** to help curate data for the long-term and to make it accessible responsibly.

Key concept: Data repository. A data repository is a place where data sets are kept for the long-term storage and sharing. A public or university library is an example of a data repository, where the data sets are books, magazines, maps, videos, music, etc. Repositories for research data are complex organizations, usually run by government or research institutions, with large amounts of computing infrastructure and formal plans for how data will be managed. A digital data repository will have a website that a researcher can use to upload or download data.

You should use the resources and expertise that these repositories provide to their communities. Many repositories offer support and resources to help you manage your data with best practices throughout a complete lifecycle and to ensure that you submit high-quality data ready for long-term preservation and sharing. In other words, most repository and data management experts want to help you manage your data well and preserve it for others to use in the future.

Some data repositories are for specific types of data. For example, VectorBase only stores biological data on invertebrate animals that can spread human diseases. On the other hand, some repositories like the Harvard Dataverse store research data from any discipline. By storing research data on a data repository, researchers like you and me can easily search for and find data that we would like to use in our own research projects. Many journals provide a list of recommended repositories. For example, the journal Nature provides a list organized by academic discipline; PLOS ONE journals also recommend a set of repositories.

As a researcher, you should identify an appropriate, widely used repository in your discipline. You should also reach out to USAID's Development Data Library(DDL) repository. The DDL is USAID's central data repository for Agency-funded data. It is a good place to store social science, health, and agricultural data. If you put your data on another repository, you should link it to the to the DDL.

Frequently asked question:
How do I know a repository is trustworthy?

Answer: There are an incredible number of different data repositories, but only a few of them are high quality. There are three things that you should look for when choosing a data repository.

1. Make sure that the data is open and freely available to the public.
2. Make sure that the repository will assign your data set with a "persistent identifier" and a citation, so that your data set can be cited like a journal paper.
3. Try to choose a repository that is certified as being trustworthy. This may not always be possible.

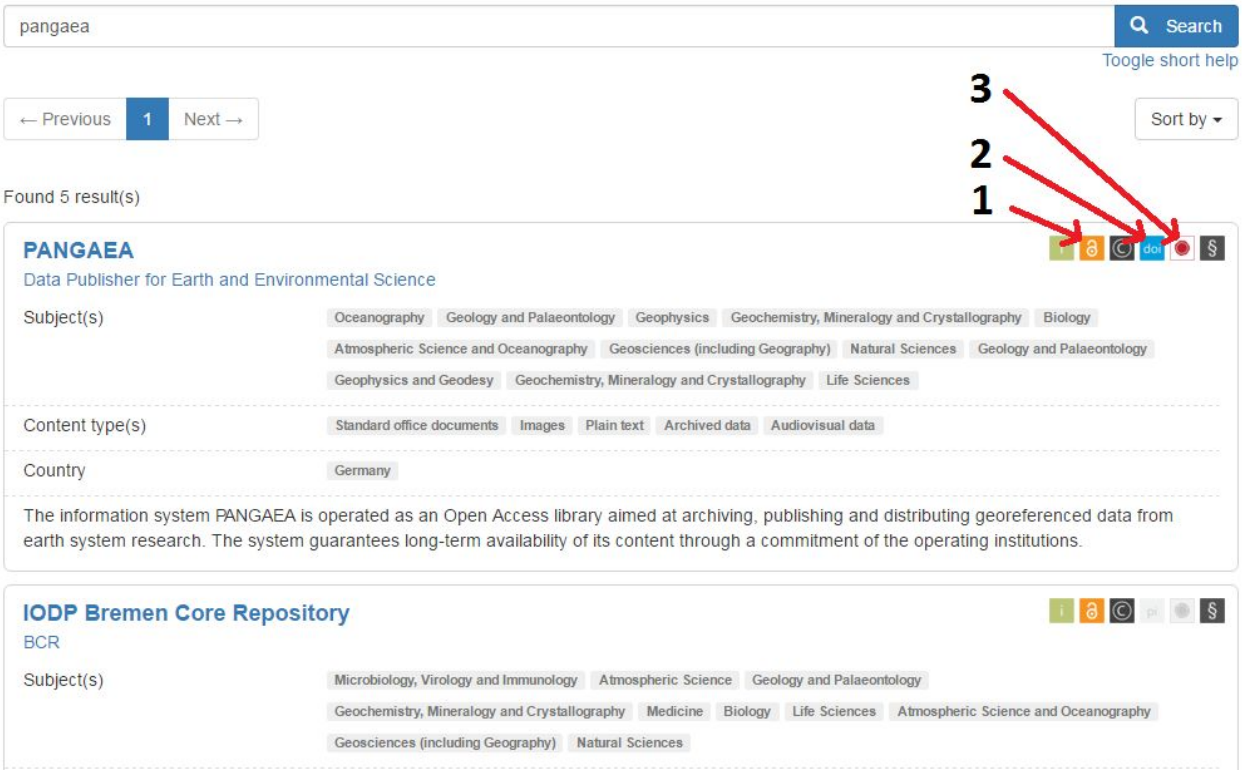

Figure 5: The result from searching for Pangaea at www.re3data.org. Arrow 1 points to the open lock symbol, which means that Pangaea's data is freely available to the public. Arrow 2 points to the blue box, which means that Pangaea issues a 'persistent identifier' for each data set. Arrow 3 points to the red dot, which means that Pangaea has been certified as a trustworthy data repository.

To determine if a repository meets these three criteria, consult the Registry of Research Data Repositories or you can check the CoreTrustSeal map for a list of repositories with a World Data System, Data Seal of Approval, or the new CoreTrustSeal certification.
For instance, a search for the repository Pangaea gives the result shown in Figure 5. The highlighted boxes in the upper right corner of each search result indicate certain desired features of the repository. All of the boxes are highlighted for Pangaea, so it is definitely a high quality repository. The search result directly below, for IODP Bremen Core Repository, shows that two of the boxes are not highlighted. This repository does not issue a persistent identifier nor has it been certified as trustworthy.

As discussed earlier, some repositories will require the data to follow a specific metadata standard. A repository also might require you to enter into legal agreements with them or charge you a one-time fee for storing your data. When choosing a repository, be sure to investigate any requirements that the repository has.

Persistent Identifier

Frequently Asked Question: How will I ensure that I get credit and that my data are properly shared and published?

Answer:
1. Get a data citation! Good data repositories will have requirements for data citation.
2. Ensure that the citation uses a persistent identifier (e.g., a DOI).

Key concept: Persistent Identifier. A **persistent identifier** is a unique combination of letters and numbers that is assigned to a digital resource, such as a webpage, journal article, or an uploaded data set. Persistent identifiers

- provide a long-term reference to a digital object.
- provide a persistent link to the digital resource.
- allow datasets to be tracked and cited.
- encourage access, discovery, and potential reuse of datasets.

Just like a person is assigned a unique number that identifies them to their government (i.e. national identification number), which remains the same no matter where the person moves, a persistent identifier is attached to a digital object and remains the same even if the object moves to different internet servers or property rights owners. The most widely used persistent identifier is the Digital Object Identifier, commonly known as DOI. The location of all digital objects that have a DOI can be immediately found by entering their DOI number at http://dx.doi.org/. Persistent identifiers try to solve the problem of finding resources that are cited, especially in academic literature.

For example, look at the following data set created by Xavier Crosta hosted on the Pangaea repository. Its persistent identifier is "doi:10.1594/PANGAEA.846541". The first three letters are "doi", which tells you that this is a DOI code. The remaining letters and numbers are used to uniquely identify the data set. The link provided in the first sentence works to guide you to the data set, but will the link still work in ten years? It is possible that a data repository might lose its funding and be forced to shut down. If it must shut down, a certified trustworthy repository will move the data in its collection to another trustworthy repository; thus, saving all of the data. This action will break the web link, but the data set itself will keep its DOI number and will still be easily found. Further, the use of a DOI improves the ability of citation trackers like Crossref to follow your work and properly count all of the citations that your work receives. This highlights the importance of having DOI numbers for your journal papers and online data sets, as well as using the DOI in your citations.

For more information on why and how to cite data, please see the USGS's primer on data citation.

Embargoes, Data Licenses & Data Rights

Frequently asked question: How do I ensure that I properly curate and store my data and yet avoid getting scooped?

Answer: PEER's Development Data program recognizes that researchers need to begin the curation and long-term management process before they are ready to publish the data to a wide audience. USAID's Development Data Policy and Public Access Plan, in line with other Federal Agency funding policies, provide the opportunity to apply for and set an **embargo** for your data submissions,

You can work with your chosen repository to implement the embargo and to set important sharing and publication procedures. **Data licenses** and data user agreements ensure you get credit for your work.

Key concept: Embargo. An **embargo** is when data is uploaded with *the restriction that the general public is not allowed to access the data* for up to one year. This allows a researcher to upload their data, as required by their grant, without letting the rest of the research community see it immediately.

An embargo is desired if the researcher plans to publish another journal article based on the data; they don't want other researchers to see the data and have a chance to publish a similar paper first. In this case, the person who uploaded the data may decide to keep the data hidden for up to one year while they work on their next journal article. Similarly, if the researcher is applying for a patent, they may want to keep the data secret until their patent application is completed.

Key concept: Data Licenses & Data Rights.
When you upload your data to a repository, you must also give other researchers the *legal rights* to use the data. This is done by choosing a "**license**" to apply to your data set. A license is legal language that describes the ways you wish to allow your data to be used. Most licenses will require that the original creator of the data be cited if the data is used. The license can either allow or forbid commercial uses of the data. The license can allow others to modify the data or forbid them from making any changes to the data set. It is up to you which license you prefer to choose. Be aware that some data repositories will only accept data if it uses their preferred license.

There are two main organizations that write the language for the licenses: Creative Commons (CC) and the Open Data Commons (ODC). There are three types of license that apply to research data, which both CC and ODC provide. The license types are:

- Attribution

    - This license lets others distribute, change, and build upon your work, even commercially, as long as they credit you for the original creation.
    - The CC-BY and ODC-By licenses are this type.
- Attribution - Share Alike
    - You let others copy, distribute, display, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first. This is the license used by Wikipedia.
    - The CC-BY-SA and ODC-ODbL licenses are this type.
- Public Domain
    - All rights worldwide to this work are waived by the author. Others may copy, modify, and distribute the work, even for commercial purposes, all without asking permission.
    - The CC0 and ODC-PPDL licenses are this type.

The Cornell University website hosts a good discussion of data rights and contains the following explanation for why you should consider using a Public Domain license:

> There is no single right answer as to which license to assign to a database or content. Note, however, that anything other than an ODC PDDL or CC0 license may cause serious problems for subsequent scientists and other users. This is because of the problem of attribution stacking. It may be possible to extract data from a data set, use it in a research project, and still maintain information as to the source of that data. It is possible to create a data set derived from hundreds of sources with each source requiring acknowledgement. Furthermore, the data in the other databases may not have originated with it, but instead sourced from other databases that also demand attribution. Rather than legally require that everyone provide attribution to the data, it might be enough to have a community norm that says "if you make extensive use of data from this data set, please credit the authors."

Sensitive Data & Data Anonymization

Frequently Asked Question: I work with sensitive information and/or vulnerable populations or resources. I am concerned about the safety and privacy of the people I work with/the security of the resources I work with. How can I ensure that my research data are protected?

Answer:
1. PLAN! From the beginning of your research project, you should assess the risk that your research information and data can present to research subjects/environment/resources.
- plan for data collection with protections in mind

- plan for any necessary informed consent
- plan for proper short-term data storage, processing, and analysis
- plan to submit proper documentation to a repository and (if necessary) plan to work with the repository to submit protected copies of data

2. Assess the risks and benefits during the course of your project
- MAKE SURE TO CONSIDER YOUR COUNTRY GOVERNMENT'S LEGAL/ETHICAL REQUIREMENTS
- document the risks that collected/generated data may present

3. Use best practice risk mitigation techniques and procedures
- protect sensitive data and information - use passwords, use access control, use strong encryption to store and transmit digital data. For physical records use access controls and a secure location.
- use statistical disclosure limitation techniques to create de-identified datasets
- create pseudonyms and codes for participants
- redact or use statistics to de-identify

4. Provide a repository with the best information to ensure data protection
- a risk assessment (or disclosure analysis plan) that identifies risks, documents how you mitigated risks, and makes recommendations about protecting and sharing data.
-Include informed consent documents
- You may opt to provide two data sets: 1. a dataset of micro-data for restricted access. 2. a data set with mitigations, such as redacted data, and data protections that can be shared more widely.

Key Concept: Sensitive Data. Some data sets contain "sensitive" information, particularly those involving human subjects. **Data is sensitive** if it can be connected to individual people (e.g. health records, financial records, political opinions), or if it identifies locations and resources that might put people in danger if they were publicly known (e.g. safe houses for women escaping from abuse, or illegal immigrant and refugee camps), etc.

Raw research data may include direct identifiers or links to direct identifiers and should be well-protected during collection, cleaning, and editing. Processed data may or may not contain disclosure risk and should be secured in keeping with the level of disclosure risk inherent in the data. Secure work and storage environments may include access restrictions (e.g., passwords), encryption, power supply backup, and virus and intruder protection.

Key Concept: Data Anonymization. These data sets can often still be shared with the research community, but extra steps must be taken to make the data safe for public viewing. The process of **data anonymization** is when sensitive information is removed from the data (or sometimes encrypted) so that the human subjects can no longer be

identified and the data may be released to the public. In the United States, data sets may never be shared openly if they contain any of the 18 "Safe Harbor" elements. Removing these elements is one possible method of data anonymization. Please see the attached Appendix "Data Anonymization: The Safe Harbor Method" for the 18 elements that must be removed from the data set. This method may not be enough, however, to fully remove sensitive information. Also, the Safe Harbor Method may not be acceptable in your country. **If you deal with sensitive data, you must find out what method of data anonymization your country requires before sharing.**

In some cases, the data cannot be anonymized and so it cannot be made public. It is still possible to share the data with specific research groups, if they ask you for permission. This will require careful planning to ensure that the data repository you are using supports this type of restricted access. If you believe your data requires restricted access, please consider using USAID's data repository and contact them for assistance.

PEER Data Management Implementation

The PEER program recognizes the importance of data and requires that awardees store data in a digital repository and share that data whenever possible.This section describes the 4 main steps, or activities, that are part of the PEER Data Management Program. Following these steps will help you manage your data throughout the lifecycle of your PEER project and ensures you comply with data and public access requirements in your PEER terms of award.

Step #1: Choose a Repository

Choose a repository for your data. You should select this by the end of year 1. We recommend collaborating closely with your US partner on this.

- For social science and public health research, we suggest using the USAID Development Data Library: https://www.usaid.gov/data.
- For other types of research, you can start looking for repositories here: https://www.coretrustseal.org/why-certification/certified-repositories/, http://www.sciencemag.org/authors/science-editorial-policies#data-deposition, and https://www.nature.com/sdata/policies/repositories
- PLOS ONE data requirements and recommended repositories: http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories

Step #2: Data Management Plan

A data management plan is a formal document that outlines what you will do with your data during and after a research project. Most researchers collect data with some form of plan in mind, but it's often inadequately documented and incompletely thought out. Many data management issues can be handled easily or avoided entirely by planning ahead. With the right process and framework it doesn't take too long and can pay off enormously in the long run. This skill will be useful for the rest of your research career and is independent of USAID-related work.

In this step you will write a Data Management Plan (DMP) that describes the data collected or generated by your PEER project and how it will be prepared for sharing. For your PEER project, you may create multiple data sets. If that is true, then your data management plan should cover all of the data sets. You may choose to upload only one of the data sets if you wish, however you are encouraged to upload them all. Open data is best when all of the data associated with the research project is made freely available.

Step #3: Data Repository

Upload at least one data set resulting from your PEER-supported research, to an online open data repository that is trustworthy and well known in your research area. This step will require you to have your data properly documented, following the format discussed in your DMP, before it can be uploaded. Once your data is uploaded, your work will be visible to other researchers in your field. They will use your data and cite your work to give you credit! This skill will be useful for the rest of your research career and is independent of USAID-related work.

The Development Data Library (DDL) is a repository of USAID-funded, machine readable data created or collected by the Agency and its partners. The DDL collects research *data*. Current policy (ADS 303 M25) requires that all USAID-funded research data be submitted to the DDL or a similar data repository. If you chose not to submit your data directly to the DDL, you must still submit your metadata to the DDL and provide a link to where the data is uploaded.

For this step, you will report the metadata and the location of your uploaded data to the DDL. This step is specific to USAID's data requirements and will be useful if you receive future funding from USAID. It will also be useful if your PEER award is still active, because you will need to do this eventually when closing your award.

Step #4: Development Experience Clearinghouse

USAID's Development Experience Clearinghouse (DEC) is the primary online resource for USAID-funded technical and project materials. The DEC holds USAID's institutional memory, spanning over 50 years; including documents, images, video and audio materials. The DEC collects research reports, evaluations and assessments, contract information, tutorials, policy and planning documents, activity information sheets, and training materials. It has nearly 200,000 items available for review or download, and adds more than 1,000 items each month.

Academic journal articles funded by USAID are considered technical documents that should be uploaded to the DEC. In this step, you will upload a single journal paper, resulting from your PEER-supported research, to the DEC. This activity is specific to USAID's data requirements and will be helpful if you receive future funding from USAID. It will also be useful if your PEER award is still active, because you will need to do this eventually when closing your award.

Activities 3 and 4 are shown in Figure 8. Your journal paper/s will be uploaded to USAID's Development Experience Clearinghouse (DEC). The data that was analyzed in the paper, and all of the metadata that you have prepared, will be uploaded to a trusted data

repository. If you are not using USAID's DDL as your repository, you will need to upload your metadata along with a web link to the location of the data to USAID's DDL.

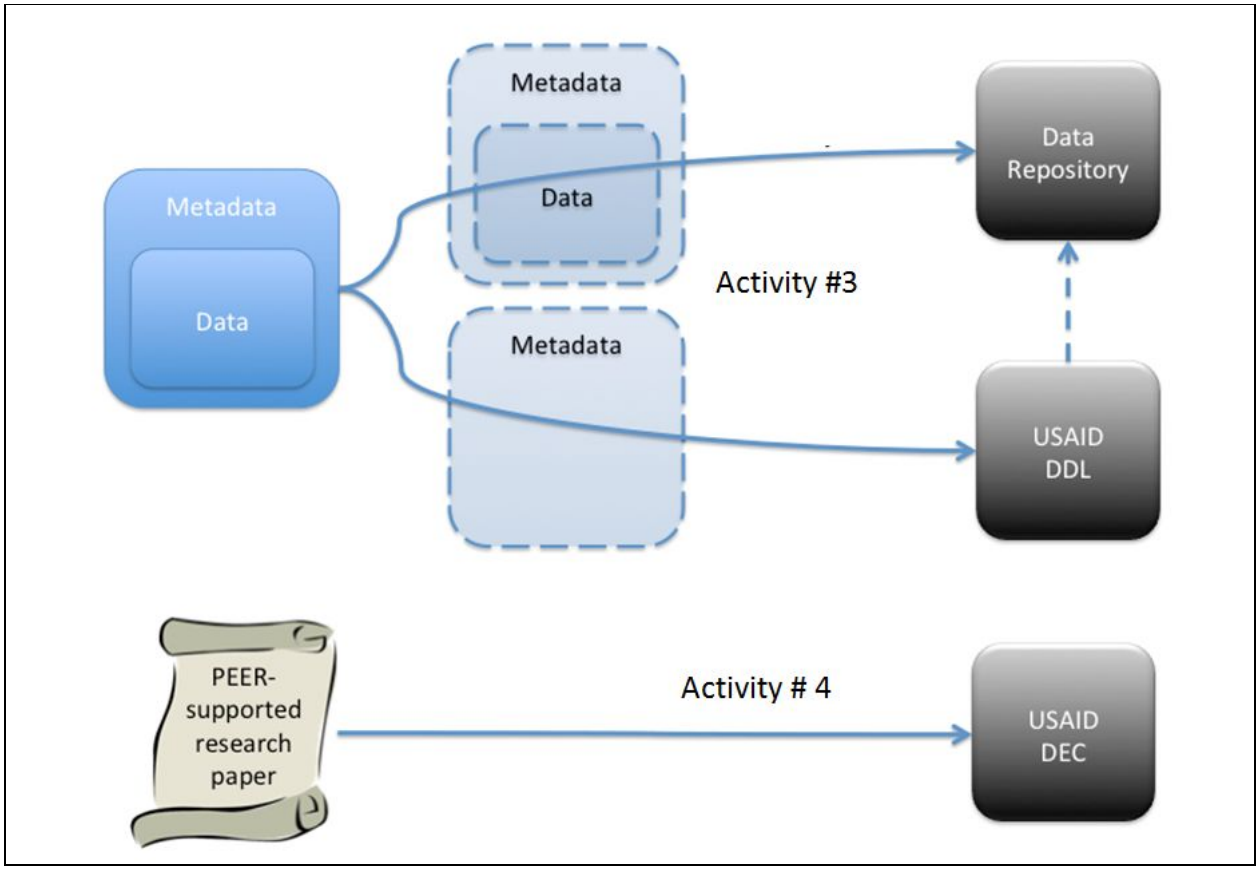

Figure 8

Appendix

Minimum Metadata to Include

Here is a good list of the recommended minimum amount of metadata that you should include in your documentation. This was adapted from the US Government Metadata Standards.

Label	Definition
Title	Human-readable name of the data set. Should be in plain English.

Description	A description (e.g., an abstract) with enough detail to allow a user to quickly understand whether the data is of interest.
Tags	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.
Last Update	Most recent date on which the dataset was changed, updated or modified.
Publisher	The publishing entity and optionally their parent organization(s).
Creator	The name and institution of the people responsible for collecting the data.
Contact Information	The name, institution, and email address for the person who should be contacted for questions about the data. This is the same idea as the 'corresponding author' on a journal paper.
Unique Identifier	A unique / persistent identifier for the dataset.
Citation	Provide the specific citation format that other researchers should use to cite your data. A good data repository will generate this for you.
License	The license or non-license (i.e. Public Domain) status with which the dataset has been published.
Public Access Level	The degree to which this dataset could be made publicly-available, regardless of whether it has been made available. Choices: public (Data asset is or could be made publicly available to all without restrictions), restricted public (Data asset is available under certain use restrictions), or non-public (Data asset is not available to members of the public).
Rights	This may include information regarding access or restrictions based on privacy, security, or other policies. This should also serve as an explanation for the selected "Public Access Level" including instructions for how to access a restricted file, if applicable, or explanation for why a "non-public" or "restricted public" data asset is not "public," if applicable.
Spatial	The range of spatial applicability of a dataset. Could include a spatial region like a bounding box or a named place.
Temporal	The range of temporal applicability of a dataset (i.e., a start and end date of applicability for the data).

Language	The language of the dataset.
Related Documents	Related documents such as technical information about a dataset, developer documentation, etc.

Data Anonymization: The Safe Harbor Method

Reproduced from Universal Patient Key, Inc.

The Safe Harbor method of anonymization and de-identification under the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule eliminates 18 patient identifiers in healthcare data. These identifiers are also known as protected health information (PHI). The Safe Harbor rule is defined in 45 CFR 164.514b(2) by the US Department of Health and Human Services. It is the hope that by manipulating or eliminating PHI in compliance to the Safe Harbor rule that the patient's identity cannot be traced back to an original data set. These 18 identifiers include:

1. Names
2. All geographic subdivisions smaller than a state usually except for the initial three digits of the ZIP code
3. All elements of dates except years
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers including license plates
13. Device identifiers and serial numbers
14. Web URLs
15. Internet protocol addresses
16. Biometric identifies (i.e. retinal scans, fingerprints)
17. Photos
18. Any unique identifying number, characteristic or code (What does this mean?)

Safe Harbor regulations require that no parts or derivatives of any of the listed identifiers be disclosed in healthcare data. For example, a data set that contains a patient's initials or the last four digits of a Social Security Number would not meet the requirement of the Safe Harbor method for de-identification. There are a few exceptions, however:

- The first three digits of any ZIP code do not have to be de-identified except in locations where the population is less than 20,000. There are 17 restricted area codes.
- Dates specifying only the year except n dates of service or other events that imply age. Ages that are explicitly stated, or implied, as over 89 years old must be recoded as 90 or above. For example, if the patient's year of birth is 1910 and the year of healthcare service is reported as 2010, then in the de-identified data set the year of birth should be reported as "on or before 1920." Otherwise, a recipient of the data set would learn that the age of the patient is approximately 100.

- The names of the patient's healthcare provider (e.g. physician, nurse, etc).