Research Data Management and Sharing
Guidance for Meeting Sponsor Requirements

Table of Contents

I. Introduction

Data management and sharing are now required by many research sponsors to:

(1) expedite scientific progress through collaborative efforts among research communities, and within research teams;
(2) increase confidence and integrity in research by enabling others to test and validate research; and
(3) share research results for public benefit.

These efforts fall within the scope of *Open Science* and *Open Scholarship,* which might be viewed as an extension of *Open Access* policies, requiring the sharing of research publications for public benefit.

At USC, careful data sharing and data management are strongly encouraged. These are key elements for achieving rigor and transparency in research, and for supporting research reproducibility. These are also essential for supporting collaborative scholarship, as reflected in USC's standards for attribution and authorship and the USC initiative for creativity and collaboration in scholarship.

Because effective data management and sharing require effort, it is important to develop plans that are realistic and achievable, and to provide the maximum opportunity to leverage cooperative efforts within research communities. Data management and sharing should adhere to the "FAIR Principles" (see Section IV), meaning:

- Findable
- Accessible
- Interoperable, and
- Reusable

Fulfilling these principles requires more than simply storing large quantities of experimental data on servers, which may have little value to others. Data management also requires curation through metadata that helps others use and understand primary data. Ideally, the metadata follows standards created within research communities. It is also important to fully document experimental methodologies, the hypotheses as they are created (not just retrospectively after

conclusion of experiments) and to link data to publications, so that readers can understand and evaluate the bases of research articles. Last, data privacy, security, long-term archival and intellectual property ownership should all be considered.

Just as importantly, data management must be communicated within the project team as soon as the project commences so that everyone knows how to execute the plan, beginning with data capture through defined standards. Throughout the project, the plan must be monitored to ensure that it is followed. And once the project is completed, the data must be archived for persistent availability.

As general references, refer to:

- Open Science Framework
- Data Management Plan Tool
- SPARC
- Digital.usc.edu
- ICPSR Guidelines

To summarize, a well-crafted data management and sharing plan provides these benefits:

- More positive assessment of proposals, increasing the likelihood of sponsor funding
- Advancement of research through collaborative efforts
- Scientific integrity, enabling validation and replication of research.

II. Forces Driving Data Management and Data Sharing

The year 2002 was estimated to be when more information was stored in digital than in analog format (the "beginning of the digital age").[1] Reflecting this shift in information generation, transmission, storage and processing, agencies of the U.S. government - as well as professional organizations, publishers, libraries, and international organizations – have been working toward new ways to share research data. Examples of the growth in research data follow:

- The Large-aperture Synoptic Survey Telescope (LSST) with 1.2 petabytes/year
- CyroEMs with terabytes/day
- Some DOE Office of Science's user facilities with >1 terabyte/second

The evolving digital capability in data transmission, processing and storage is depicted in Table 1.

[1] IBM has estimated digital data generation as doubling in 20 years in the 1950s, in 10 years in 70s, in 8 years in the 90s, in 1 year in 1917 and projected it to be every 12 hours with full implementation of the internet of things.

Table 1: Evolution of digital data capability

	Transmission (bits/sec)	Processing (FLOPS)	Storage (Bytes)
1940		0.5 Kilo (10^3)	
1950			
1960	0.3 Kilo (10^3)		
1970		Mega (10^6)	
1980	20 Kilo	Giga (10^9)	3 Exa (10^{18})
1990		Tera (10^{12})	
2000	100 Kilo		
2010	1 Giga	Peta (10^{15})	300 Exa
2020	20 Giga	Exa (10^{18})	Cloud (virtually unlimited)

In parallel to the hardware evolution, continuing progress in machine learning, deep learning, artificial intelligence and virtual reality technologies will contribute to transformational changes in data utilization. For instance, toward that goal NIH sponsored a workshop "Harnessing Artificial Intelligence and Machine Learning to Advance Biomedical Research" in June 2018 and NSF has instituted its "Big Idea" on Harnessing the Data Revolution.

National and International science and engineering organizations are working to manage the transition to digital data. In the U.S., the development of standards for data reporting and storage are being addressed by the National information Standards Organization (created in 1939). It identifies, develops, maintains, and publishes technical standards to manage information in today's continually changing digital environment. NISO is accredited by the American National Standards Institute (ANSI) and serves as the U.S. technical advisory group to the ISO Technical Committee 46.

When it was instituted in 1947, the International Standards Organization (ISO), established a Technical Committee 46: Information and Documentation. TC46 addresses standardization of practices relating to libraries, documentation and information centres, publishing, archives, records management, museum documentation, indexing and abstracting services, and information science. It has working groups addressing a) document storage and conditions for preservation, b) archives/records management and c) technical interoperability. Example standards for digital data are:

- ISO 13008:2012 Information and documentation

Digital records conversion and migration process

- ISO/TR 13028:2010 Information and documentation

  Implementation guidelines for digitization of records

- ISO 17068:2017 Information and documentation

  Trusted third party repository for digital records

Simultaneous to the push from the science and engineering publishing communities, governments have been responding to the transition to digital data. In the U.S. the White House Office of Science and Technology Policy released a memorandum in 2013 entitled "Increasing Access to the Results of Federally Funded Scientific Research." In the same year the European Commission established the Research Data Alliance (RDA) - an international community-driven organization building the social and technical infrastructure to enable data sharing.

The U.S. science and engineering data management efforts have been influenced by U.S. Federal government interest in public access to the results of federal funding. The Office of Management and Budget (OMB) issued an Open Data Policy and a Project Open Data in 2013. The 2014 DATA Act (Digital Accountability and Transparency Act, Public Law No. 113-101) directed that government expenditure data be made more available (i.e., the grant funding "input"), but by implication, it also motivated attention to the "output" – science / engineering / medicine papers and data.

III. Data Management Plan (DMP) Requirements by U.S. Federal Funding Agencies

Science funders, publishers and governmental agencies are frequently requiring data management and stewardship goals. Under the Research Data Alliance, a DMP Common Standards Working Group was launched in 2017; the specific focus of this working group is on developing a common information model and specifying access mechanisms that make DMPs machine-actionable. A set of recommendations is expected in 2019. Also in 2017, the California Digital Library received an NSF Grant for Actionable Data Management Plans.

Good data management is a conduit to knowledge discovery and innovation and, therefore, publishers and research sponsors require data management and stewardship.[2] The following provides general sponsor requirements. Please read individual solicitations for unique expectations.

National Science Foundation

In 2010 NSF announced it would start requiring data management plans for proposals. The NSF data management plan requirements can be summarized by:

- Types of Data Produced
  - What types of data (experimental, computational, or text-based), metadata, samples, physical collections, models, software, curriculum materials, and other materials will be collected and/or generated in the course of the project?
  - What descriptions of the metadata are needed to make the actual data products useful and reproducible
  - For collaborative proposals, describe the roles and responsibilities of all parties with respect to the management of data
- Data and Metadata Standards
  - In what format and/or media will the data or products be stored
  - Where data are stored in not generally accessible formats, how may the data be converted to more accessible formats
  - Solutions and remedies to providing data in an accessible format should be offered with minimal added cost.
- Policy for Re-use, Re-distribution, Derivatives
  - What are your policies regarding the use of data provided via general access or sharing?
  - Practices for appropriate protection of privacy, confidentiality, security, intellectual property, and other rights
- Policies for Access and Sharing
  - What approaches will be used to make data available and accessible to others, including any pertinent metadata
  - What plans, if any, are in place for providing access to data
  - Where no data or sample repository, metadata should be prepared and made publicly available
- Plans for Archiving and Preservation
  - When and how will data be archived and how will access be preserved over time?
  - A plan to transfer digitized information to new storage media or devices as technological standards or practices change?

[2] Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Mark D. Wilkinson et.al., Scientific Data 3:1600018 | DOE: 10.1038/sdata.2016.18

    - Will there be an easily accessible index that documents where all archived data are stored and how they can be accessed?
- Post Award Management
    - After an award is made, the PI(s) must manage their data as described in the DMP and will be monitored primarily through the normal Annual and Final Report process and through evaluation of subsequent proposals.

Many of the requirements listed above reflect the international efforts toward developing machine actionable data repositories (see Section IV).

National Institutes of Health

The National Institutes of Health (NIH) developed a data sharing policy that went into effect beginning in 2003 for applicants seeking NIH funding of $500K or more in direct costs in any one year. The policy expects final research data from NIH-supported research efforts to be made available to other investigators. It includes data from: basic research, clinical studies, surveys, and other types of research. Subsequently NIH has released a Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research in Feb 2015 and a Strategic Plan for Data Science in 2018. In the fall of 2018, NIH issued a request for information on data sharing that is expected to result in broadened requirements, likely requiring at a minimum that all primary data associated with charts and tables in publications be made available to others.

NIH applicants who are planning to share data are instructed to describe briefly:[3]

- expected schedule for data sharing
- format of the final dataset
- documentation to be provided
- whether or not any analytic tools also will be provided
- mode of data sharing (e.g., personal website, data archive,...)
- whether a data sharing agreement be required, including criteria for who can receive the data and any conditions on its use
- precise content and level of detail to be included in a data-sharing plan depends on several factors, such as whether or not the investigator is planning to share data, the size and complexity of the dataset, and the like.

The NIH Website shows three examples, each a paragraph long.

[3] https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

While the elements of a data management plan tend to be consistent across agencies, the length (see Table 2) and the organization of the content varies.

Table 2. Length of Data Management Plan for Various US Federal Agency

Agency	Length
National Institutes of Health (NIH)	paragraph
National Science Foundation (NSF)	2 pages
Department of Energy (DOE)	2 pages
Dept of Education, Inst of Education Sciences (IES)	5 pages
National Aeronautics and Space Admin (NASA)	8000 characters
National Oceanographic and Atmosphere Admin (NOAA)	2 pages
USDA, National Inst Food and Agriculture (NIFA)	2 pages
National Endowment of Humanities (NEH)	2 pages
Dept of Justice, National Institute of Justice (NIJ)	2 pages
Dept of Def (DOD, mostly ignored, but when included)	2 pages
Dept of Interior, Joint Fire Science Program	2 pages
Dept of Interior, US Geological Survey	unspecified
Dept of Homeland Security (no mention of DMP)	

A number of resources are available to assist in the development of data management plans, some with examples of successful plans. Prime resources, such as DMPTool, are provided in Appendix A. Websites with a number of example DMPs are provfded in Appendix B

IV. Data Repositories

For certain types of important digital objects, there are already well-curated, deeply-integrated, special-purpose repositories such as Genbank3, Worldwide Protein Data Bank (wwPDB4), and UniProt5 in the life sciences; Space Physics Data Facility (SPDF), and Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD6) in the space sciences.[4]

Looking to extend this to other data archives, a 2014 Workshop was organized by a Future of Research Communications group.[5] It focused on 'machine actionable' to indicate a continuum of possible states wherein a digital object provides increasingly more detailed information to an autonomously acting, computational data explorer. The workshop developed a recommended suite of requirements for data repositories, with the acronym FAIR:

To be **F**indable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be **A**ccessible**:**

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be **I**nteroperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

---

[4] Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Mark D. Wilkinson et.al., Scientific Data 3:1600018 | DOE: 10.1038/sdata.2016.18

[5] The organizers of this workshop included Eduard Hovy who, at the time, was working for ISI at USC

I2. (meta)data use vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be **R**eusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

These requirements have been endorsed by the National Institutes of Health (NIH), the American Geophysical Union (AGU), and the International Union of Pure and Applied Chemistry (IUPAC).[6]

Data archives take many forms, ranging from data explicitly included in journal publications to data stored in personal devices. A 2014 presentation by Martinsen of the American Chemical Society delineated those various depositories as depicted in Figure 1.

[6] Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Mark D. Wilkinson et.al., Scientific Data 3:1600018 | DOE: 10.1038/sdata.2016.18

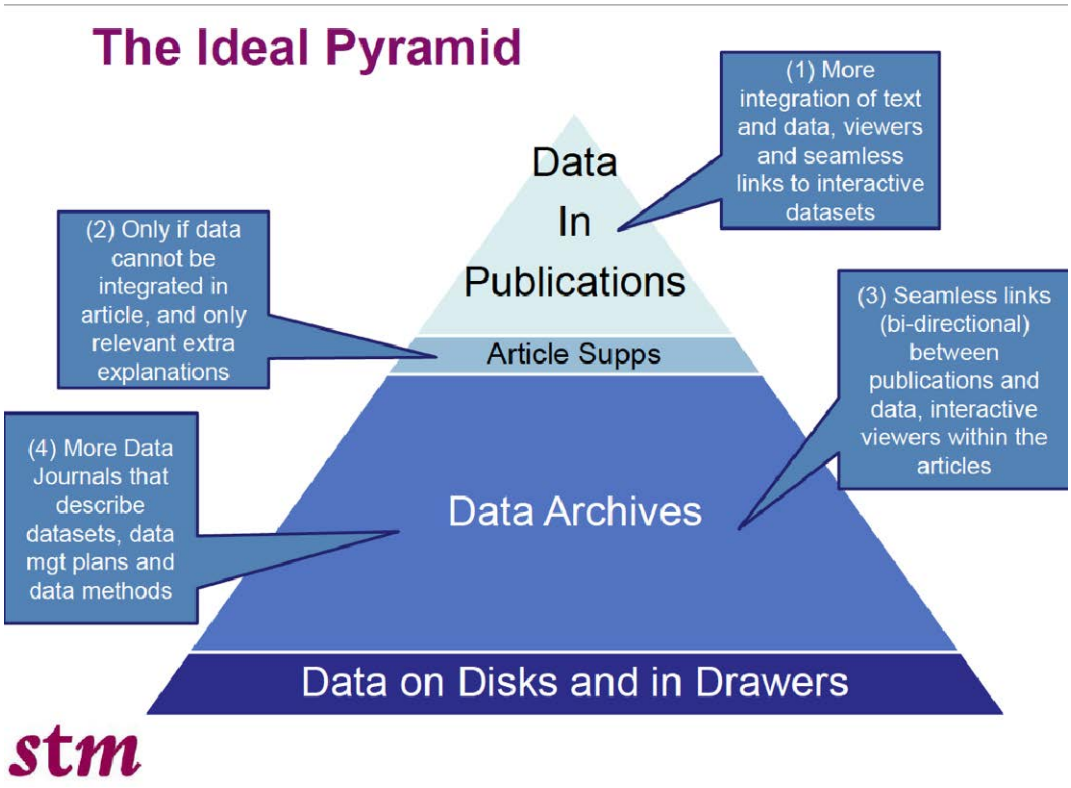

Figure 1. The Ideal Pyramid

Some examples of these repository classes are shown in Table 4.

Table 4. Examples of Data Archives

Institutional

Arch Data Deposit	Northwestern
DSpace@MIT	MIT
JHU Data Archive	JHU

Data articles

Elsevier Journal *Data in Brief*

Data repositories

over 2000 of them exist – see websites listed below

Federated data systems (open access)

Data.gov (U.S. Government's open data ~250K datasets)

National Center for Biotechnology Information

Inter-Univ. Consortium for Political and Social Research at the Univ. of Michigan

Data enclaves (controlled access)

Research Data Center (CDC National Center for Health Statistics)

Mixed mode (open/controlled)

National Institute of Mental Health Human Genetics Initiative

National Longitudinal Study of Adolescent Health

There are websites to assist in the discovery of an appropriate data repository for one's data:

- Open Access Directory
- Data Science Central
- Registry of Research Data Repositories (re3data)
- NIH Designated Repositories
- NASA Prognostics Data Repositories (materials properties)

In addition to data, an author must also pay attention to the repository of his/her publications. Most agencies are requiring articles resulting from the agency funding support also be posted to the agency designated site (in addition to whatever the journal publisher provides).

Agency	Archive website
NIH	PubMed Central
NSF	Public Access Repository
DOD	DTIC PubDefense
DOE	PAGES
ED/IES	ERIC
NASA	PubSpace
NOAA	Repository
NEH	ICPSR
NIJ	ICPSR

V. Data Resource Ecosystem

Beyond proper collection, annotation and archiving, data stewardship includes the notion of "long-term care" of valuable digital assets, with the goal that they should be readily discovered and re-used for downstream investigation. Accomplishing this goal will require a collaborative effort on the part of the funding agencies, the professional science and engineering societies, the publishing communities, and the scientists/engineers who are performing the research.

One of the goals of data management is retrieving data by use of a standardized communications protocol. Science and engineering papers now have a Digital Object Identifier (DOI) to make them more findable. For larger suites of data, equivalent identifiers are *de rigueur*; this practice will likely be extended to all data stored in archives. Resources to help in establishing an identifier are:

- DataCite is a leading global non-profit organization (USC is a member) that provides persistent identifiers (DOIs) for research data. Its goal is to help the research community locate, identify, and cite research data with confidence.
- Crossref makes research outputs easy to find, cite, link, and assess.

Many of the repositories will provide an appropriate identifier (including USC's IT)

National Science Foundation

NSF's investments in Harnessing the Data Revolution (HDR) are to generate new knowledge and understanding and to accelerate discovery and innovation. The HDR vision is to be realized through an interrelated set of efforts in:

- The foundation of data science;
- Algorithms and systems for data science;
- Data-intensive science and engineering;
- Data cyberinfrastructure; and
- Education and workforce development.

Each of these efforts is designed to amplify the intrinsically multidisciplinary nature of the emerging field of data science. The HDR Big Idea will establish theoretical, technical, and ethical frameworks that will be applied to tackle data-intensive problems in science and engineering,

contributing to data-driven decision-making that impacts society. Specifically, beginning in FY 2019, and building on past investments by nearly all NSF directorates and offices, HDR is offering two conceptualization pathways to develop institutes to accelerate discovery and innovation in data-intensive science and engineering:

- Institutes for Data-Intensive Research in Science and Engineering - Ideas Labs (I-DIRSE-IL) - aims to bring together scientists and engineers working on important data-intensive problems with data scientists and systems/cyberinfrastructure specialists; and
- Institutes for Data-Intensive Research in Science and Engineering - Frameworks (I-DIRSE-FW) - encourages applications from teams of researchers proposing frameworks for integrated sets of science and engineering problems and data science solutions.

In addition, in FY 2019, NSF is offering:

- Transdisciplinary Research in Principles of Data Science - supports developing the theoretical foundations of data science through integrated research and training activities; and
- Data Science Corps (DSC) - aims to build capacity at the local, state, national, and international levels to help unleash the power of data in the service of science and society.

National Institutes of Health

NIH has instituted a major effort to better exploit the evolving capabilities, including a webpage DataScience@NIH. It has created the position of Chief Data Strategist, who in close collaboration with the NIH Scientific Data Council and NIH Data Science Policy Council, will guide the development and implementation of NIH's data-science activities and provide leadership within the broader biomedical research data ecosystem.

In September 2018 NIH released a Strategic Plan for Data Science with five goals:

- Support a Highly Efficient and Effective Biomedical Research Data Infrastructure
- Promote Modernization of the Data-Resources Ecosystem
- Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools
- Enhance Workforce Development for Biomedical Data Science
- Enact Appropriate Policies to Promote Stewardship and Sustainability

Toward those goals, the NIH Office of Strategic Coordination has established a Data Commons pilot (DCP) which seeks to accelerate biomedical discovery by providing a cloud-based platform where investigators can store, share, access, and compute on digital objects including data, software, workflows, and more. The initial implementation is a Pilot Phase in which targeted high-value data sources will serve as test cases for the infrastructure to be developed, based on these cases:

- Genotype-Tissue Expression (GTEx)
- Trans-Omics for Precision Medicine (TOPMed)
- Model Organism Databases (MODs) that make up the Alliance of Genome Resources

The Data Commons Pilot Phase Consortium (DCPPC) is developing key capabilities to support access, use and sharing of the test case data sets. The key capabilities include:

- Guidelines and metrics for making data Findable, Accessible, Interoperable, and Reusable (FAIR)
- An approach to Global Unique Identifiers (GUIDs)
- Application Program Interfaces (APIs) based on open standards
- Architecture independent of a specific cloud platform or provider
- Workspaces to find and interact with data and associated tools
- Research ethics, privacy, and security (including authentication and authorization)
- Indexing and search functionality
- Use cases that demonstrate how the NIH Data Commons Pilot Phase can advance biomedical research
- Coordination, training, and outreach

The cloud-based repositories are being implemented through the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. It will establish partnerships with commercial cloud service providers (CSPs) to reduce economic and technological barriers to accessing and computing on large biomedical data sets to accelerate biomedical advances. The New Models of Data Stewardship program has announced initial agreements with Google Cloud and Amazon Web Services.

In addition, through its Biomedical Data Translator program, the NIH National Center for Advancing Translational Sciences (NCATS) is supporting research to develop ways to connect conventionally separated data types to one another to make them more useful for researchers and the public.

VI. USC Specific Guidance

Intellectual Property and Publication Considerations

Data produced by USC employees in the course of research is owned by the university under applicable law and USC's intellectual property policy, unless the university has assigned ownership, e.g., to the sponsor of a clinical trial or research services project. It is USC's practice to have the principal investigator consent to any agreement to assign data ownership.

Unlimited sharing of data is typically not required as part of a data sharing plan. However, USC encourages data sharing for public benefit, specifically for on-going research and other non-commercial use. Whenever research data is to be made available outside the university, in addition to the logistics of how data access is controlled and monitored, it is, important to consider privacy issues, use and redistribution rights, and guarantees, if any, regarding the data.

For USC-owned data provided outside the university, it is imperative to add a license.* Data licenses exist on a spectrum from totally open to very restrictive. For help in choosing the appropriate license, please contact the USC Stevens Center for Innovation. If there are no privacy issues in the distribution of the data outside USC and you wish to allow research use of the data without redistribution or commercial rights, you can consider using the following language as a condition for download and use of the data:

> "This research data [a description of the data should be added here] is provided by the University of Southern California ("USC") on an AS-IS basis only. USC SPECIFICALLY DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITIATION, ANY WARRANTY AS TO MERCHANGIBILITY OR FITNESS FOR A PARTICULAR PURPOSE. The data may be reproduced and used for non-commercial research purposes only. No part of the data may be redistributed, in whole or in part, without the prior written consent of USC and the may not be used, directly or indirectly, for commercial purposes without a separate license from USC. By downloading the data you acknowledge and agree the the foregoing use terms for the data."

Please consult the USC Stevens Center for Innovation on intellectual property considerations surrounding data ownership and sharing, including consideration of commercial use of the data.

Data received by the university to support the conduct of research may be restricted by agreements executed with data providers. Researchers should carefully review such agreements for limitations and restrictions to be followed relating to non-disclosure, privacy, security, publication, etc. The USC Institutional Review Board has also created a policy governing use of social media data (section 14.1) that must be followed.

In most cases, research publications, in contrast to data, are owned by the authors under USC's intellectual property policy. Nevertheless, the copyright owner of a publication is often obligated to make publications available to the public as a condition of funding agreements. NIH, for example, requires posting publication in Pubmed within 12 months. In addition, many publishers require assignment of the copyright in research publications to the publisher. Researchers should carefully review publishers' copyright transfer agreements to ensure that they do not violate their obligations to sponsors for public access.

Data Security and Privacy

Research repositories derived from clinical data are governed by USC's Policy on Biorepositories , which defines the roles of data guardians, bioethics committees and patient consent for data sharing. It is important to consult this policy before the creation of a data management plan for clinical data.

---

*It is also important to develop a plan for attribution of research products, in accordance with USC's standards. This would include attribution for the data itself as a research product, and attribution by others who utilize the data.

For some data types, USC is obligated under law or contract to protect privacy. Examples include protected patient information (i.e., Health Insurance Portability and Accountability Act, HIPAA), identifiable census data, restricted data used in department of defense research, data from EU residents under the GDPR, and student data. Be sure to consult requirements and, where applicable check with the Office of Compliance or General Counsel as to whether a privacy consent needs to be addressed.

In addition, consult USC's policies on data security, as well as USC's IRB policy on social media data, if relevant to the work.

Data Storage Resources at USC

USC data storage resources can be found at https://itservices.usc.edu/storage/, and community data storage and sharing resources can be found at digital.usc.edu. Please keep in mind that simple data storage is not equivalent to data management, which requires additional attention to curation, meta-data, accessibility and so on.

VII. Summary and Check-off

To conclude, a well-crafted data management and sharing plan should provide solutions for:

1. Definition of data types that will be produced and managed.
2. Fulfillment of the FAIR principles, including adherence to accepted field specific standards and data curation, as well as identifiers and metadata.
3. Mechanism used for data storage and data backup.
4. Persistent archival and preservation, extending beyond end of project.
5. Data security, data privacy and associated ethical considerations.
6. Management of intellectual property rights associated with data and data transfer agreements (in consultation with the Stevens Center for Innovation).
7. Attribution for data sets and attribution in subsequent publications.
8. Discussion of how the data management plan will promote the advancement of future research and promote rigor and reproducibility.

Appendix A: Resources to assist in DMP development

DMPTool[7]

DMPTool is a free, open-source, online application that helps researchers create data management plans. Templates are provided for:

National Science Foundation (NSF, and for many of its Divisions)

National Institutes of Health (NIH)

Department of Energy (DOE)

Department of Energy. Basic Energy Sciences (BES)

Department of Defense (DOD)

Department of the Interior (DOI), Joint Fire Science Program

Department of Transportation (DOT)

Department of Education (ED), Institute of Education Sciences

Institute of Museum and Library Sciences (IMLS)

National Endowment for the Humanities (NEH)

National Oceanographic and Atmosphere Administration (NOAA)

US Dept. of Agriculture (USDA), National Inst of Food and Agriculture (NIFA)

US Geological Survey (USGS)

SPARC[8]

SPARC (the Scholarly Publishing and Academic Resources Coalition) works to enable the open sharing of research outputs and educational materials in order to democratize access to knowledge,

[7] DMPTool is a service of the University of California Curation Center of the California Digital Library. The original contributing institutions were: University of California Curation Center (UC3) at the California Digital Library, DataONE, Digital Curation Centre (DCC-UK), Smithsonian Institution, University of California, Los Angeles Library, University of California, San Diego Libraries, University of Illinois, Urbana-Champaign Library, and University of Virginia Library. Given the success of the first version of the DMPTool, the founding partners obtained funding from the Alfred P. Sloan Foundation to create a second version of the tool, released in 2014.

[8] SPARC and Johns Hopkins University Libraries have a joint project on a community resource for tracking, comparing, and understanding both current and future U.S. federal funder research data sharing policies.

accelerate discovery, and increase the return on our investment in research and education. The agencies being tracked are:

Health and Human Services, Agency for Healthcare Res. and Quality (AHRQ)

Health and Human Services, Asst. Sec. for Preparedness & Response (ASPR)

Center for Disease Control and Prevention (CDC)

Department of Defense (DOD)

Department of Education (ED)

Department of Energy (DOE)

Department of Transportation (DOT)

Food and Drug Administration (FDA)

National Aeronautics and Space Administration (NASA)

National Institutes of Health (NIH)

National Institute of Standards and Technology (NIST)

National Oceanic and Atmospheric Administration (NOAA)

National Science Foundation (NSF)

US Agency for International Development (USAID)

US Department of Agriculture (USDA)

US Geological Survey (USGS)

University Libraries

Univ Southern California

California Digital Library

Inter-University Consortium for Political and Social Research (ICPSR)

Duke Univ

Johns Hopkins Univ

Massachusetts Inst of Technology

North Carolina State Univ

Purdue Univ

Univ Calif San Diego

Univ Minn

Univ Mich

Univ Oregon

Univ Virginia

Digital Object Identifiers

One of the goals of data management is retrieving data by use of a standardized communications protocol. Science and engineering publications now have a Digital Object Identifier (DOI) to make them more findable. For larger suites of data, equivalent identifiers are *de rigueur*; this practice will likely be extended to all data stored in archives. Resources to help in establishing an identifier are:

- DataCite[9] is a leading global non-profit organization (USC is a member) that provides persistent identifiers (DOIs) for research data. Its goal is to help the research community locate, identify, and cite research data with confidence.
- Crossref [10] makes research outputs easy to find, cite, link, and assess

ORCID IDs

ORCID provides a persistent digital identifier that distinguishes researchers from every other researcher. All researchers are strongly encouraged to obtain an ORCID ID as their persistent identifier.

[9] https://datacite.org/

[10] https://www.crossref.org/

Appendix B: Websites with examples of Data Management Plans

University of Minnesota

- NIH provides several examples of DMPs for studies involving human subjects.
- Fire Science.gov has the best template
- NASA/JPL Mars Global Surveyor Science Data Management Plan (1995)
- USGS National Climate Change and Wildlife Science Center

Stanford University

- ICPSR: Sample Data Management Plan for Social and Political Science Data
- NEH-ODH: Data Management Plans from Successful Grant Applications
- NSF, Biology Directorate, Plant Genome Research Program (PGRP): Examples of three data management plans (pdf) from funded grants by Stanford Professor Virginia Walbot, including additional information and guidance

UC San Diego

NSF

- Office of Cyberinfrastructure (OD/OCI)

  DMP Example Allan Snavely From Allan Snavely's proposal to the Strategic Technologies for Cyberinfrastructure (STCI) program.

- Office of Integrative Activities (OD/OIA)

  DMP Example Todd Martz SIO.pdf From Professor Todd Martz's proposal entitled "MRI: Development of an instrument for testing and calibration of autonomous sensors for the marine CO2 system" to the Major Research Instrumentation Program for consideration by the Division of Ocean Sciences.

  DMP Example Chaitan Baru SDSC.pdf From Dr. Chaitan Baru's Type II proposal to the NSF-wide Cyber-Enabled Discovery and Innovation (CDI) program.

- Division of Environmental Biology (BIO/DEB)

  DMP Example Cleland.pdf From Elsa Cleland's successful proposal The influence of plant functional types on ecosystem responses to altered rainfall, to the Ecosystem Studies Program.

- Division of Integrative Organismal Systems (BIO/IOS)

  DMP Example Nitz.pdf From Douglas Nitz's successful proposal, CAREER: Parietal Cortex and the Transformation of Spatial Cognition into Action, to the Activation Program in the Neural Systems Cluster.

DMP Example Laurie Smith.pdf From Laurie Smith's successful proposal, Polarization of Plant Cell Division by Receptor-Like Proteins, to the Plant Genome Research Program.

- Division of Computing and Communication Foundations (CISE/CCF)

  DMP Example Cosman.pdf From Pamela Cosman's successful proposal, CIF: Medium: Mobile multiview video: compression, rendering, and transmission, to the Communications and Information Foundations Program.

- Division of Graduate Education (EHR/DGE)

  DMP Example Michael Kalichman.doc From Michael Kalichman's successful proposal, Integrating Ethics Education: Capacity-Building Workshops for Science and Engineering Faculty, to the Ethics Education in Science and Engineering Program.

- Division of Chemical, Bioengineering, Environmental, and Transport Systems (ENG/CBET)

  DMP Example Shah.pdf From Sameer Shah's successful proposal, Supply and Demand of Proteins During Neuronal Growth and Extension, to the Biomedical Engineering Program.

- Division of Civil Mechanical and Manufacturing Innovation (ENG/CMMI)

  DMP Example John Fontanesi SOM.doc From Professor John Fontanesi's proposal to the Service Enterprise Systems (SES) program in the Systems Engineering and Design (SED) Cluster in the Civil Mechanical and Manufacturing Innovation (CMMI) Division. The proposed research is to determine the utility of a hybrid agent-based, discrete event simulation and analysis, to understand and improve hospital functions.

- Division of Electrical, Communications and Cyber Systems (ENG/ECCS)

  DMP Example ECE.pdf From a proposal to conduct research on signal processing techniques for photonic systems.

  DMP Example Xiang,J.doc From Professor Jie Xiang's proposal to conduct research on nanoelectronics.

- Division of Ocean Sciences (GEO/OCE) Physical Oceanography (GEO/PO)

  DMP Example Pinkel.pdf From Rob Pinkel's successful proposal, Collaborative Research: Tasmanian Tidal Dissipation Experiment (T-TIDE), to the Physical Oceanography Program.

  DMP Example SIO-BO.doc From a proposal to the Biological Oceanography program in the Division of Ocean Sciences.

  DMP Example SIO-PO.doc From a proposal to the Physical Oceanography Program in the Division of Ocean Sciences to conduct research on the flux of carbon dioxide generated by breaking waves across the ocean/atmosphere boundary layer.

  DMP Example SIO OCE.pdf From a proposal to NSF's Division of Ocean Sciences for research on marine aerosol particle chemistry.

  DMP Example Jennifer MacKinnon.pdf From Professor Jennifer MacKinnon's proposal to the NSF OCE program in physical oceanography. The proposed work involves studying currents

and turbulent mixing in the deep ocean using a variety of specialized oceanographic instruments and involving co-PIs from six different institutions.

- Division of Mathematical Sciences (MPS/DMS)

  DMP Example Rogalski.pdf From Daniel Rogalski's successful proposal, Noncommutative surfaces and Calabi-Yau algebras, to the Algebra and Number Theory Program.

- Division of Social and Economic Sciences (SBE/SES)

  DMP Example Ayelet Gneezy.pdf From Ayelet Gneezy's successful proposal Social Pricing - Image Management, Social Preferences and Pay-What-You-Want, to the Decision, Risk and Management Sciences Program.

  DMP Example Wixted.pdf From John Wixted's successful proposal, Signal Detection Theory and Eyewitness Memory, to the Law and Social Sciences Program.

- Crosscutting – Multiple Directorates and Offices

  DMP Example Psych.doc From a proposal to the crosscutting NSF/NIH program, Collaborative Research in Computational Neuroscience (CRCNS): Innovative Approaches to Science and Engineering Research on Brain Function, for experimental psychology research.

  DMP Example anon.doc From a proposal to the crosscutting NSF/NIH program, Collaborative Research in Computational Neuroscience (CRCNS): Innovative Approaches to Science and Engineering Research on Brain Function.

Appendix C: Other Resources

Websites with Agency Specific Data Management Plan Guidance

DOE

ED, Institute of Education Sciences

NASA, SMD

NIH

NOAA

NSF

National Academies of Sciences, Engineering and Medicine (NASEM) Reports Pertinent to Science and Engineering Data Management

https://www.nap.edu/

19537 1982 Data Management and Computation

19463 1983 Materials Properties Data Management: Approaches to a Critical National Need\

19127 1988 Selected Issues in Space Science Data Management and Computation

5504 1997 Bits of Power: Issues in Global Access to Scientific Data

10664 2003 Government Data Centers: Meeting Increasing Demands

12615 2009 Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age

13282 2012 Communicating Science and Engineering Data in the Information Age (NCSES)

24777 2017 Data Management and Governance Practices (transportation agency)

25015 2018 International Coordination for Science Data Infrastructure: Proc of a Workshop

25214 2018 Data Matters: Ethics, Data, & International Research Collaboration in a Changing World