What Are the Critical Elements of a Successful Data Sharing Plan?

Dr. Robert Hanisch

Director, Office of Data and Informatics (ODI)

Material Measurement Laboratory

US National Institute of Standards and Technology

Data Sharing Plans, Data Management Plans

- A Data Management Plan (DMP) describes data that will be acquired or produced during research; how the data will be managed, described, and stored, what standards you will use, and how data will be handled and protected during and after the completion of the project.

  - Establish a schedule for reviewing and updating a DMP in combination with project events such as funding approval, project review, and publication.

  - Ensure DMP content contains a level of detail that enables stakeholders (funders, project staff, and repository managers) to understand the reality of the project activities.

  (https://www.usgs.gov/products/data-and-tools/data-management/data-management-plans)
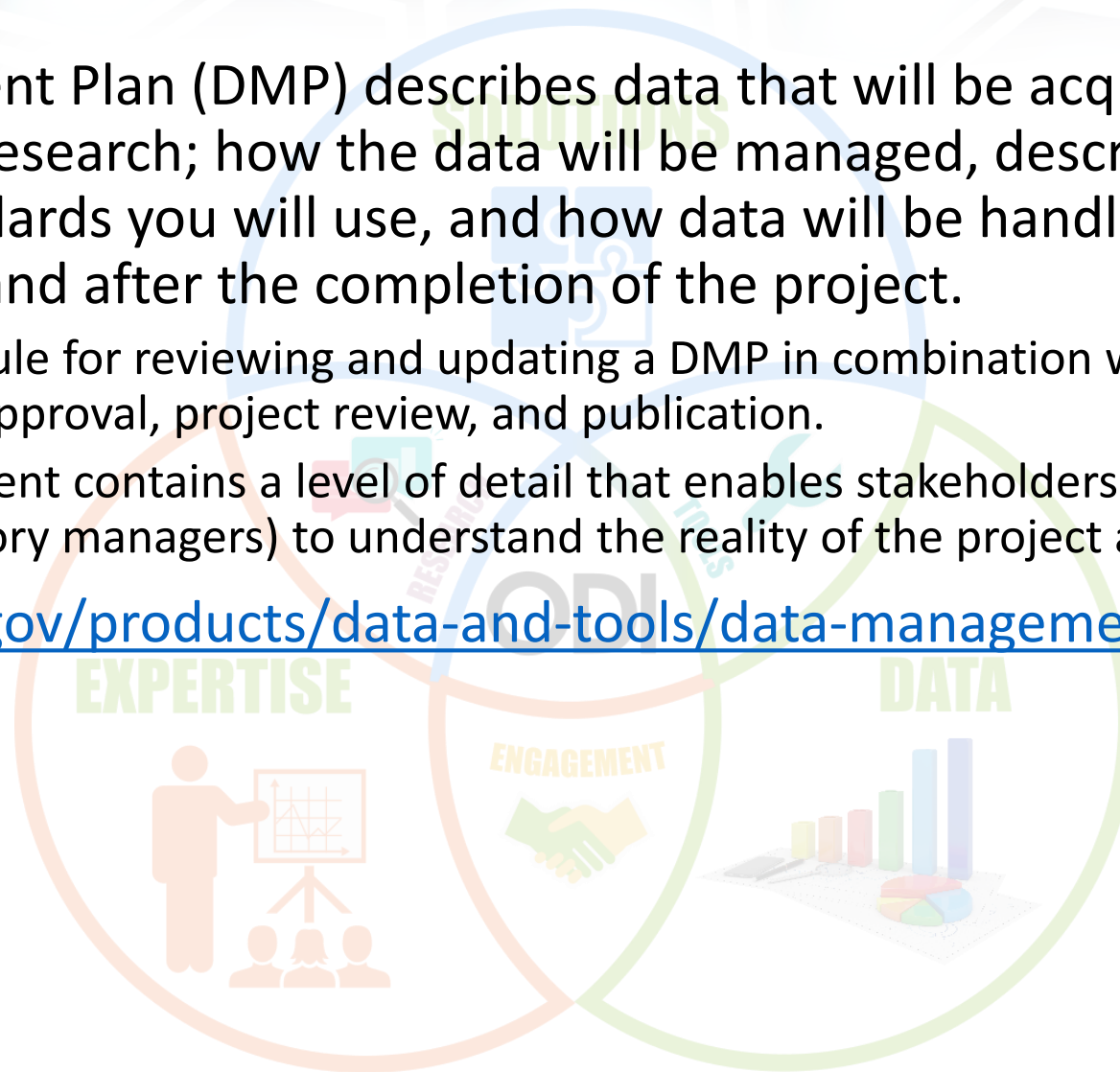
Data Sharing Plans, Data Management Plans

- In general, a data-sharing plan should address the following:
  - Data Description: What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
  - Existing Data: Will you be using existing data? Relationship between the data you are collecting and existing data.
  - Audience: Who will potentially use the data?
  - Access and Sharing: How will data files be shared? How will others access them?
  - Formats: What data formats will you be creating?
  - Metadata and Documentation: What documentation will you provide to describe the data? Metadata formats and standards?
  - Storage, backup, replication, versioning: Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
  - Security: Are the system and storage that will be used secure?
  - Budget: Any costs for preparing the data? Costs for storage and long-term access?
  - Privacy, Intellectual Property: Does the data contain private or confidential information? Any copyrights?
  - Archiving, Preservation, Long-term Access: What plans do you have to archive the data and other research products? Will it have long-term accessibility?
  - Adherence: How will you check for adherence of this plan?
- Further considerations would include:
  - Which of your data sets have long-term value to others?
  - How will you or the repository you work with ensure that data are curated to withstand changes in storage technologies and data formats?

(https://vpr.harvard.edu/data-sharing-and-management-plans)

MATERIAL MEASUREMENT LABORATORY

Benefits of DMPs/DSPs

- Gets researchers to think *in advance* about their data management and sharing strategy, helps to avoid data loss

- If available to the researcher's CDO and/or IT department, aids in capacity planning and budget control

- If available to colleagues or others, particularly in machine-readable format, allows for solution sharing
  - Search across DMPs for instrumentation, analysis software, storage platforms, repositories

- Asset for improving reproducibility and increasing transparency of research outputs

- Supports data re-use and additional discoveries, particularly if the data are FAIR

Barriers to DMPs/DSPs

- Some researchers feel embarrassed to expose poor data management practices or lack of experience

- Some researchers fear that research ideas could be scooped if DMPs/DSPs are shared

- Some researchers see DMPs/DSPs as purely a bureaucratic checkbox

- Some DMP/DSP templates and preparation tools are awkward to use and poorly adapted to the great diversity of research methods
  - Too generic
  - Too detailed

- Although many funding bodies now require DMPs/DSPs, their evaluation and bearing on the success of a proposal is far from uniform

A DMP That I Like!

- General information

  The instrument will be housed in [facility] and incorporated into the existing internal network. This isolated network <mark>ensures data security during acquisition</mark>, and until such time the data is removed from the [facility] server. Users of the instrument will be required to designate whether the data collected are proprietary or not, which dictates exclusion or inclusion into the FAIR data initiative. Data excluded from FAIR will be the responsibility of the user, who will have access to the data on the server for a period of up to one year. It is important to note that isolation of the [facility] internal networks affords greater protection of data and instruments but prohibits the possibility of virtual operation for remote users.
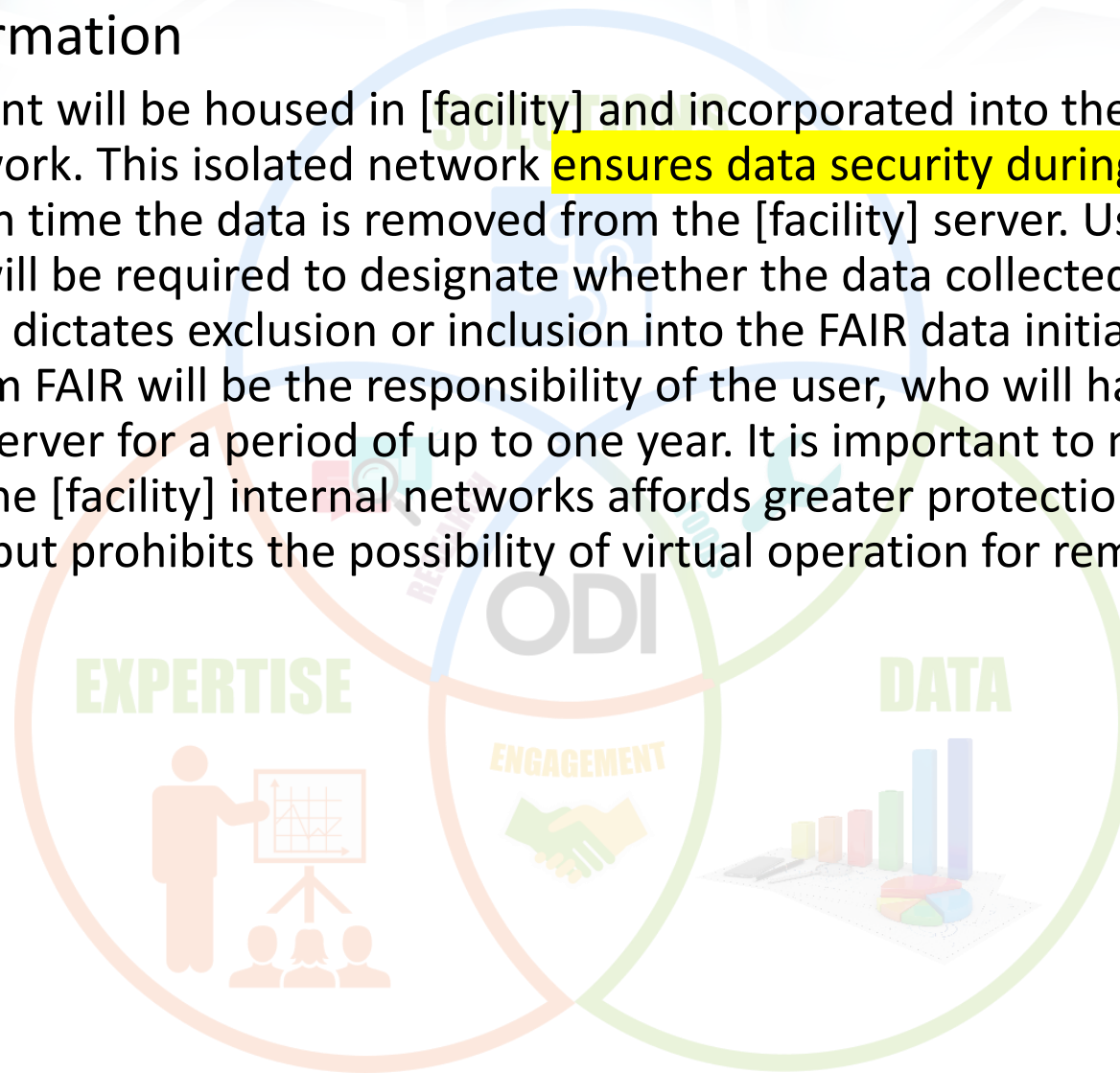
A DMP That I Like!

- Data types and data processing

The data that will be produced in the course of the project include <mark>experimental measurement data files, plots, images, graphs, tables, and text</mark>. This project will not involve the acquisition of data from live animal or human subjects.

Files containing the experimental data and metadata/spectra/etc. or any other suitable experimental data that is generated during this project will be aggregated, placed in a folder on the [facility] server with a title that includes a researcher code and date/time stamp. Data identified by users as "FAIR" data, will be copied to an external server, the FAIR data repository. This <mark>data will be processed using standard software such as Origin, SigmaPlot and Microsoft Excel, and specialized software (the Materials Data Curation System (MDCS)) for visualization, data integration, and curation</mark>. As a result, users will have access to the MDCS interface to manage their research data and, for instance, define which is most important, proprietary and/or appropriate to make open.

A DMP That I Like!

- Metadata

    The metadata of the research conducted on the instrument will be entered in a ==templated web-based digital research notebook==. For instance, users will be directed to annotate their research data with information about the research project, sample preparation, the researcher that synthesized samples, and the researcher that collected the research data. Furthermore, the images of the sample and the location of the measurement spot on the sample will be recorded and stored as necessary. ==All of this information will allow the PI to assure reproducibility and integrity of the data==.

- Format

    The data files obtained during the measurements will have a broad range of formats dictated by the format of the software that controls the instrumentation. However, the raw experimental data will also be stored in text formats (developed in conjunction with instrument vendors) to allow integration of correlated data, and subsequently uploaded to MCDS. The plots/images resulting from the data analysis will be stored in the Origin and SigmaPlot formats, as well as in image formats such as TIFF, JPEG, PDF, etc.

A DMP That I Like!

- Access

    Access to data sharing with the research community will be provided using current best practices for data delivery websites, that is, using a ==documented Representational State Transfer (REST) Application Programming Interface (API)==. This allows both human and computer access to the data and improves that accessibility and reusability of the data.

- Sharing

    It has been the PI's standard practice to share the primary data and/or finalized plots, tables, and text suitable for publication with collaborators as soon as the data are available. This sharing is routinely provided for free with the expectation that appropriate co-authorship is included and acknowledgements crediting the funding organizations (e.g., NSF in this case) are listed. In addition, ==this project will make available non-proprietary data in a 'FAIR' (Findable, Accessible, Interoperable, and Reusable) way== as outlined in the proposal.

A DMP That I Like!

- Re-use and license

  The data obtained in this project can be reused with the caveat that the original paper and the funding source will have to be cited and an appropriate acknowledgement will have to be added. A faculty member reserves the right to be a coauthor on a paper when his/her reused data constitutes a significant contribution to the collaborative research. Where possible, researchers will be encouraged to identify data that they consider to be non-proprietary (through the MCDS system) and provide input on the most important aspects of each dataset so that the format, organization, and descriptive metadata of such data is appropriate for FAIR release to the research community. Datasets available by this mechanism will be assigned a DOI via DataCite so they may be cited and will be released with a Creative Commons (CC) 0 license.

A DMP That I Like!

- Intellectual property

  The PI does not anticipate that there will be any significant intellectual property issues involved with the data. In the event that discoveries or inventions are made in direct connection with the data, access to the data will be granted upon request once appropriate invention disclosures and/or provisional patent filings are made. The data acquired and preserved in the context of this proposal will be further governed by the University of North Florida policies pertaining to intellectual property, record retention, and data management.

- Preservation

  The PI's will identify materials of possible long-term interest for archiving. It is reasonable to assume that these data will become baseline data for future projects, thus complete and secure retention will last at least five years beyond project completion. Any FAIR data released to the community will be retained for the same period and will be deposited to an appropriate domain specific repository before the end of the five years.

DMP c/o Dr. Stuart Chalk, Dept. of Chemistry, University of North Florida

Take-aways

- DMPs/DSPs express intention; not easy to confirm compliance

- DMPs/DSPs need not be exhaustive or exhausting!

- How can we better incentivize researchers to write and maintain DMPs/DSPs?
  - Confirm benefits and overcome barriers
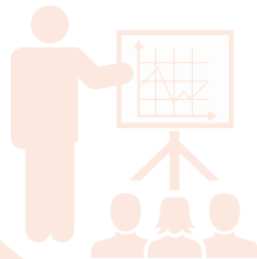 - Share success stories

Contact

Robert Hanisch

robert.hanisch@nist.gov

https://nist.gov/people/robert-hanisch