Primer on Data Management

Data Management Plans

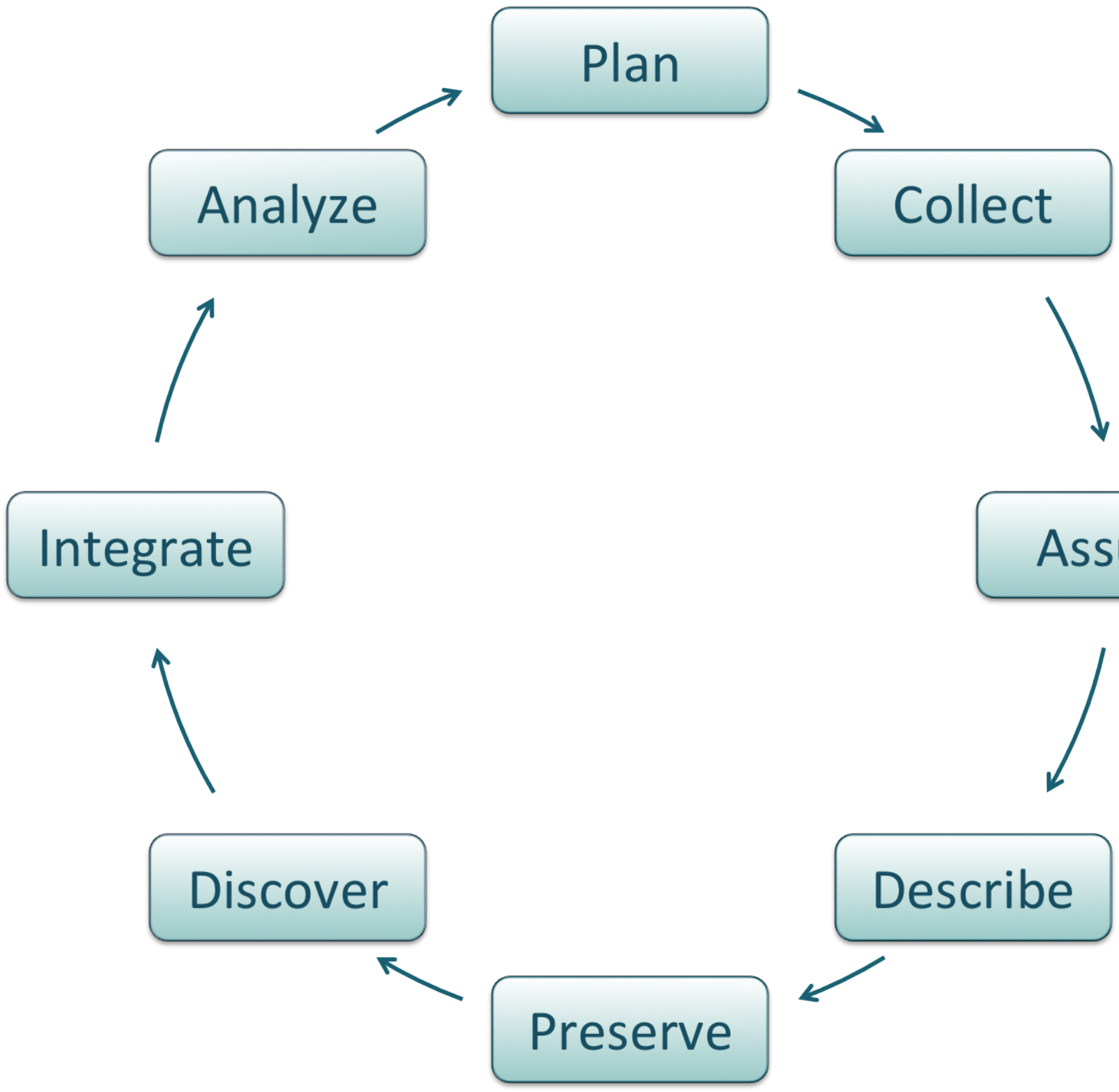
Robert Cook
Environmental Sciences Division
Oak Ridge National Laboratory

American Meteorological Society
New Orleans
January 22, 2012

Data Management Planning Topics

1. What is a data management plan (DMP)?
2. Why prepare a DMP?
3. Components of a DMP
4. Example of an NSF DMP
5. Resources

Planning for Data Management



Data Life Cycle

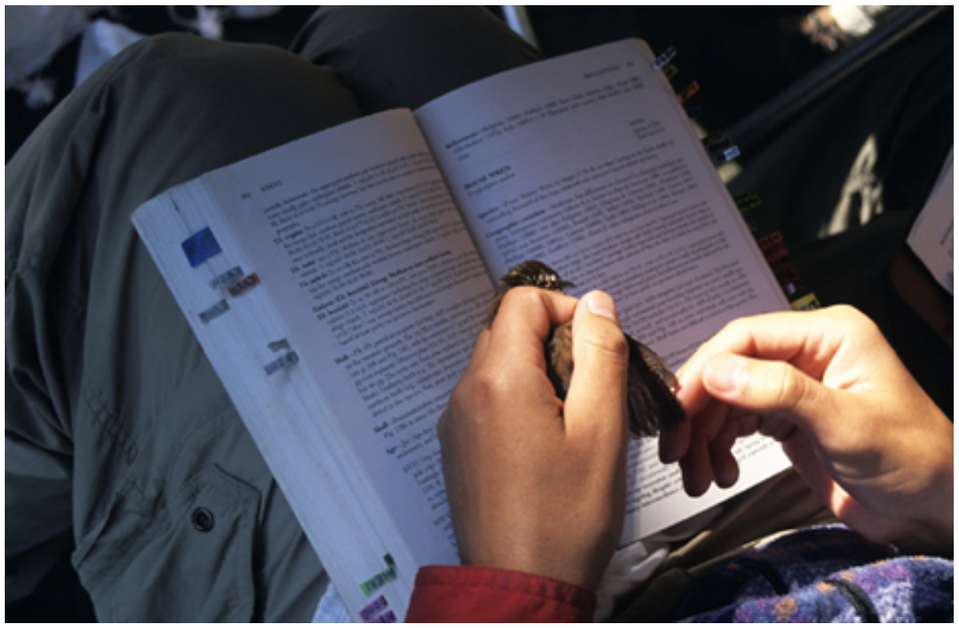
"A goal without a plan is just a wish."

Antoine de Saint-Exupery (1900 – 1944)

What is a DMP?

- A document that describes what you will do with your data **during** and **after** your research

- Workshop describes how to create a DMP and practical for managing



Value of a Data Management Plan

- National funding agencies have data sharing polices and requirements
- Scientific journals (Nature, Science, and PLoS) have sharing requirements.
- Shared, common data may help researchers collaborate and accelerate discoveries (NY Times, 2010).

For the researcher:

- helps organize data
- cultivate quality and efficiency
- help with sharing and preserving data

Thanks to Jeffery Loo, Chemical Informatics Librarian, UCB

NSF DMP Requirements

From Grant Proposal Guidelines:

Plans for data management and sharing of the products of research. **Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan".** This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (in AAG), and may include:

- the **types of data**, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
- the **standards to be used for data and metadata** format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)
- **policies for access and sharing** including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements
- policies and **provisions for re-use**, re-distribution, and the production of derivatives
- **plans for archiving** data, samples, and other research products, and for preservation of access to them

1. Information About Data & Data Format

1.1 **Description of data to be produced**

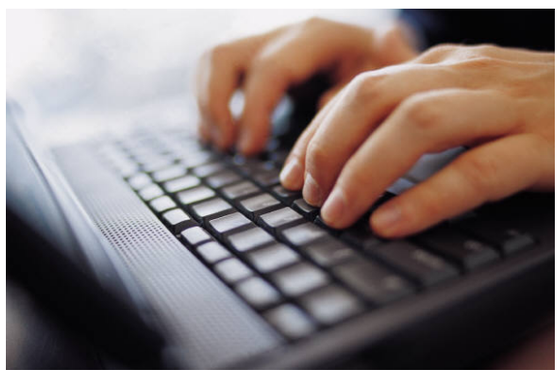
Experimental, Observational, Raw or derived,
Physical collections, Models, Images, etc.

1.2 **How data will be acquired?**

When? Where? Methods?

1.3 **How data will be processed?**

Software used, Algorithms, Workflows



1.4 **File formats**

csv, spatial data

1.5 **Quality assurance & quality control**

1.6 **Existing data**

If existing data are used, what are its origins?
Will your data be combined with existing data?
What is the relationship between your data and existing data?

1.7 **How data will be managed in short-term**

Version control, Backing up, Security & protection,
Who will be responsible?

2. Metadata Content & Format

Metadata is the documentation describing all aspects of the data (e.g., who, why, what, when, and where)

2.1 What metadata are needed

Any details that make data understandable and usable

2.2 How metadata will be created and/or captured

Lab notebooks? GPS units?

Auto-saved on instrument? Manually entered?

2.3 What format will be used for the metadata

Standards for community (EML, ISO 19115, etc.)

Justification for format chosen

3. Policies for Access, Sharing, & Reuse

3.1 Obligations for sharing

- Funding agency, institution

3.2 Details of data sharing

- How long?
- When?
- How access can be gained?

3.3 Ethical/privacy issues with data sharing

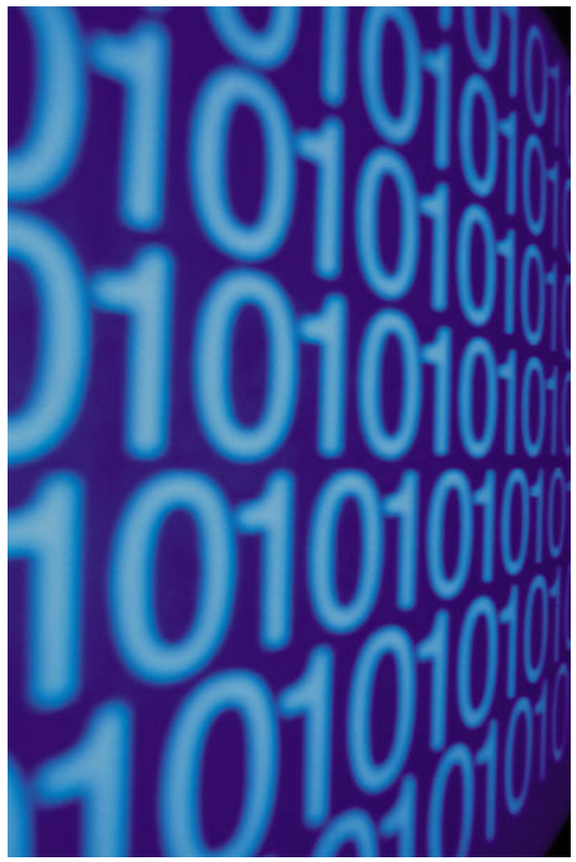
3.4 Intellectual property & copyright issues

- Institutional policies
- Funding agency policies
- Embargos for political/commercial reasons

3.5 Intended future uses/users for data

3.6 Citation

- How should data be cited when used?
- Persistent citation?



4. Long-term Storage & Data Management

4.1 What data will be preserved

4.2 Where will it be preserved

- Most appropriate archive for data
- Community standards

4.3 Data transformations/formats needed

- Consider archive policies

4.4 Who will be responsible

- Contact person for archive

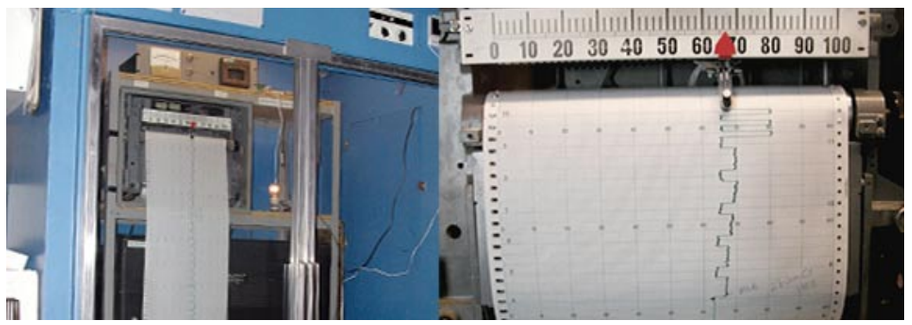
National Climatic Data Center

ATMOSPHERIC SCIENCE DATA CENTER

Example Data Management Plan

Mauna Loa CO_2 Record

- Example, based on the work of CD Keeling & colleagues
- Study the controls on the concentration of atmospheric CO_2
  - high precision and accuracy measurements.



Courtesy of NOAA/ESRL,
Photographs by Forrest Mims III

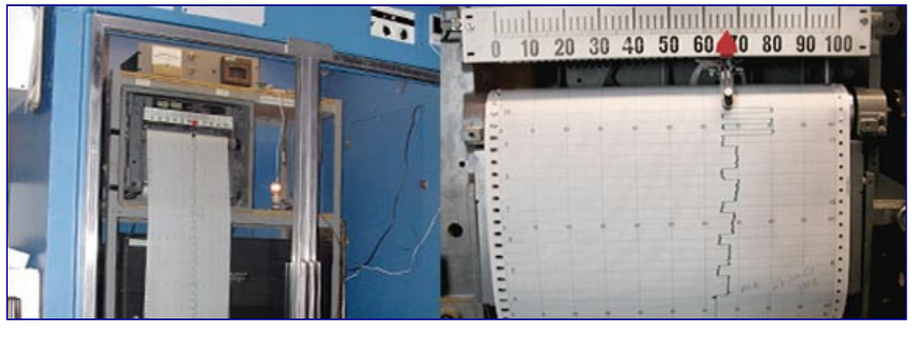
https://www.dataone.org/sites/all/documents/DMP_MaunaLoa_Formatted.pdf

Mauna Loa Example Data Management Plan

1. Information About Data & Data Format

- Collected continuously at five towers
  - a central tower and four towers located at compass quadrants.
- **Raw data files** contain continuously measured CO_2 concentrations, calibration standards, references standards, daily check standards, and blanks.
  - Site conditions will also be noted and retained.
- **Final data product** will consist of 5-minute, 15-minute, hourly, daily, and monthly average atmospheric concentration of CO_2, in mole fraction in water-vapor-free air
- Data in comma-separated-values in plain ASCII format

Courtesy of NOAA/ESRL,
Photographs by Forrest Mims III

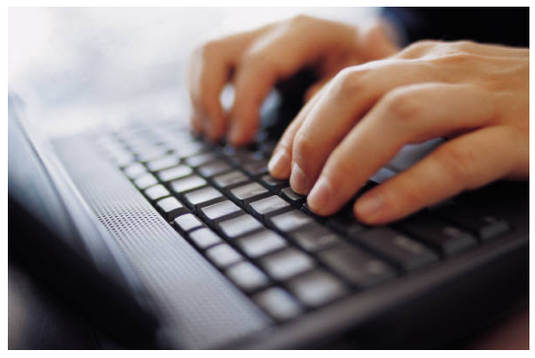


Mauna Loa Example Data Management Plan

2. Metadata Content & Format

- Metadata
  1. contextual information about the data in a text based document
  2. standard metadata (e.g., ISO 19115) in an xml file.

- Metadata formats provide a full explanation of the data (text format) and ensure compatibility with international standards (xml format).



Mauna Loa Example Data Management Plan

3. Policies for Access, Sharing, & Reuse

- The final data product released when the recalibration of standard gasses has been completed and the data have been prepared (~six months).
- No period of exclusive use by the data collectors.
- Users can access documentation and final aggregated CO_2 data files via the Scripps CO_2 Program website (http://scrippsco2.ucsd.edu).
- Raw data will be maintained and made available on request at no charge

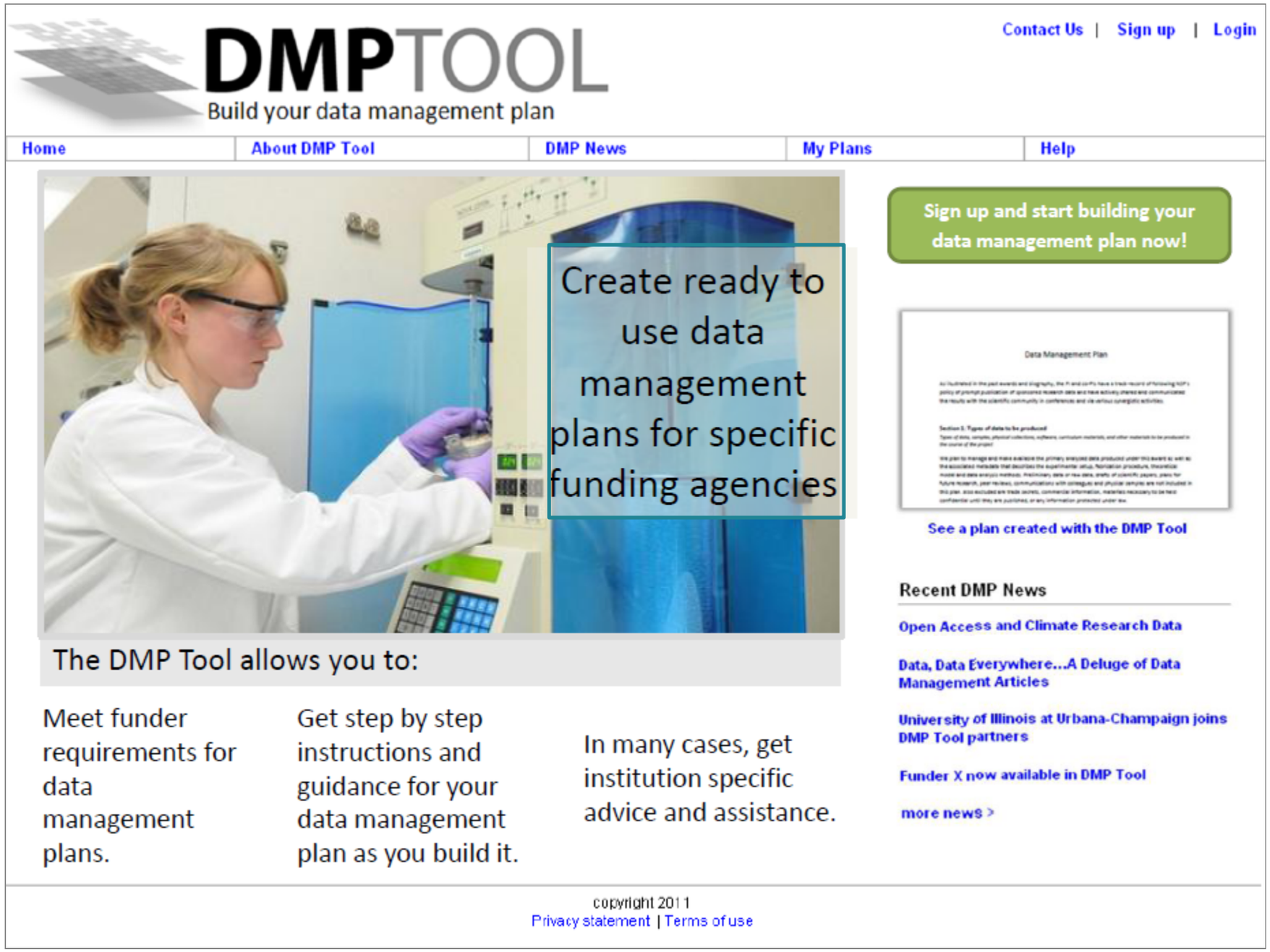
Mauna Loa Example Data Management Plan

4. Long-term Storage & Data Management

- Final data product will be available for use by the research and policy communities in perpetuity.
- Raw supporting data will be available for use by researchers to confirm the quality of the Mauna Loa Record.
- Long-term stewardship and curation at the Carbon Dioxide Information and Analysis Center (CDIAC), Oak Ridge National Laboratory.
- Standardized metadata record will be added to the archive metadata database at CDIAC
- Data product citation, including DOI:

Keeling, CD, 2004. Atmospheric CO2 Concentrations - Mauna Loa Observatory, Hawaii, 1958-2003. Numeric Data Package. Available on-line [http://cdiac.ornl.gov] Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory, Oak Ridge, TN, USA. doi: 10.3334/CDIAC/atg.ndp001

Data Management Planning Tool

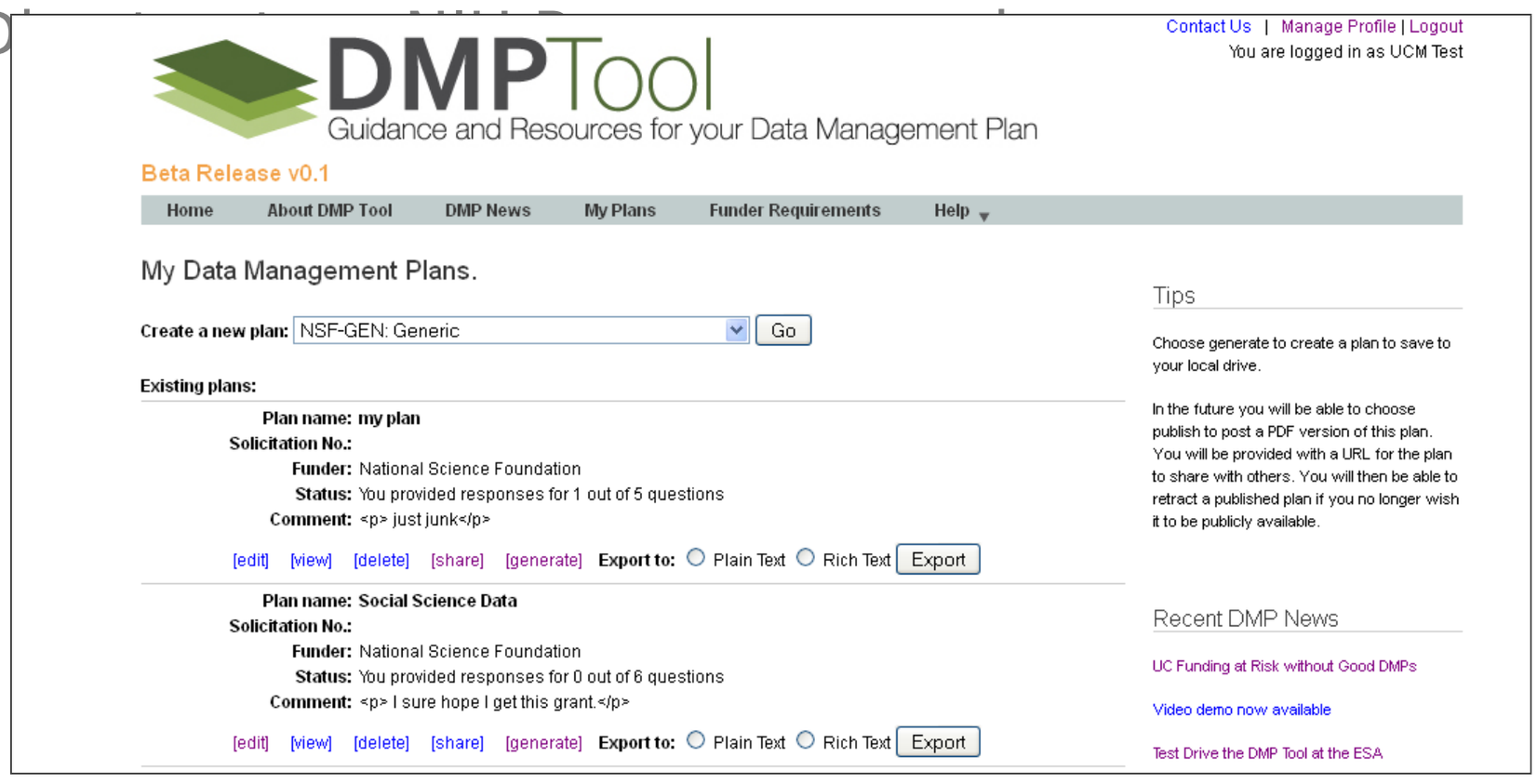


https://dmp.cdlib.org/

DMP Tool Out-of-the-Box

Open to the community

- **Step-by-step ‘wizard’ for generating data management plans**
- General guidance for each section: help text and resources relevant to all
- Resources for NSF DMP: create, edit, reuse, share, save, generate a plan
- Use templates
- Connect users with your institution’s resources



Highly Developed Barbarism

THE ALGEBRAIST

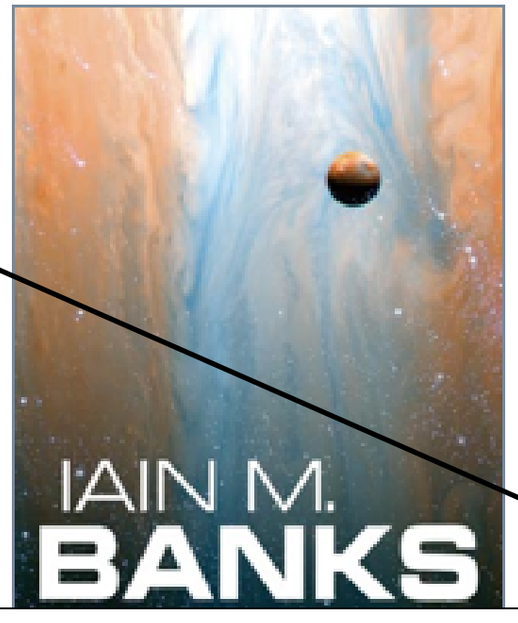
About the book | Reviews | More information

About the book:

It is 4034 AD. Humanity has made it to the stars. Fassin Taak, a Slow Seer at the Court of the Nasqueron Dwellers, will be fortunate if he makes it to the end of the year.

The Nasqueron Dwellers inhabit a gas giant on the outskirts of the galaxy, in a system awaiting its wormhole connection to the rest of civilisation. In the meantime, they are dismissed as decadents living in a state of highly developed barbarism, hoarding data without order, hunting their own young and fighting pointless formal wars.

Seconded to a military-religious order he's barely heard of – part of the baroque hierarchy of the Mercatoria, the latest galactic hegemony – Fassin Taak has to travel again amongst the Dwellers. He is in search of a secret hidden for half a billion years. But with each day that passes a war draws closer – a war that threatens to overwhelm



dismissed as decadents living in a state of highly developed barbarism, hoarding data without order, hunting their own young and fighting pointless formal wars.

References and Resources

- For a detailed questionnaire with most of the issues you may need to address in an NSF data management plan see: http://dmp.data.jhu.edu/sites/default/files/Questionnaire.doc
- MIT Libraries have a fairly comprehensive site about data management and publishing in general at http://libraries.mit.edu/guides/subjects/data-management/index.html
- The Digital Curation Center in the UK has a variety of information on data management plans at http://www.dcc.ac.uk/resources/data-management-plans
- Data Management Planning Tool: