

WRANGLING YOUR RESEARCH DATA: THE DATA MANAGEMENT PLAN (DMP)

RALEIGH L. MARTIN - AAAS S&T POLICY FELLOW

W/ THE SEN TEAM: LESLIE HSU, WONSUCK KIM, KIM MILLER, BRANDON MCELROY

MAY 22, 2018

SEDIMENT EXPERIMENTALIST NETWORK (SEN) CLINIC AT CSDMS



This work is licensed under a Creative Commons Attribution 4.0 International License. It does not reflect an official position of AAAS or of the U. S. Federal Government. SEN is supported through NSF award EAR-1324760.

SURVEY RESPONSES

- Background: Equal mixture of experimentalists, field scientists, and modelers (many doing all three!)
- DMP experience: **7/16 respondents have prepared a DMP**. 1 uses IEDA DMP tool, 1 tried IEDA but gave up, 1 uses university library template
- Most common file formats: .csv, .mat, .xlsx, .txt, .mat, .nc, GIS
- Data curation issues: **storage** (including what to keep), **documentation/organization** (for collaboration, sharing, and integration)
- Data access issues: **Context** (biggest issue), **data import** (e.g., formats), **data heterogeneity, pre-processing time**

INTRODUCTION: WHAT IS A DATA MANAGEMENT PLAN (DMP)?

- Plan for generating, sharing, and archiving data
- Required 2 page supplementary document for NSF proposals
- Also required by other funding agencies and some data repositories



www.dataone.org

Data Management Plan Rio Grande Basin Hydrologic Geodatabase Compendium.

1. Types of Data Produced

This project will result in the production of a relational spatially-enabled database integrating all known surface water, ground water, and water quality data for the middle Rio Grande basin study area. Additionally, Visual Basic for Applications (VBA) code and Structured Query Language (SQL) code are products of the project. All updateable datasets are acquired from the original data source (for example, EPA websites). Updateable data sources are acquired at specified intervals - quarterly, or as needed. As new static data sources are discovered, they will be integrated into the proposed compendium. Data will be processed using dataset-specific VBA programs. Program file comment headers will be included in the code to explain required input variables, purpose of the program, and requirements needed by the user. Code will be annotated to promote code readability.

2. Data and Metadata Standards

Microsoft Access Database format will be used since it is readily-accessible and it is compatible with ESRI ArcGIS (<http://www.esri.com/software/arcgis/index.html>), a Geographic Information System software package used by the stakeholders. Naming conventions will be consistent - no spaces will be used in table names or field names. The file naming convention will consist of the data source, data type format for raw data files. Data reporting functionality will be built into the VBA processing programs to provide output in .txt file format for number of records per source when updateable data sources are refreshed.

Every effort will be made to go back to the authoritative source for an identified dataset. Quality control of the database will be performed using SQL statements that capitalize on the database structure to ensure relational database integrity. Appropriate primary keys will be assigned to manage possible data duplicates. Potential duplicate site IDs, will be handled through automated procedures and the creation of alternate ID tables.

A data dictionary will be created that defines the table definition, table fields, and table field data types. An entity-relationship diagram will be created that defines the relational structure of the database. A metadata record will be produced using the FGDC standard that describes the entire geodatabase.

The FGDC standard was chosen due to required Federal government standards.

3. Policies for Access and Sharing

The data are public and will be obtainable thru the New Mexico Interstate Stream Commission (NMISC). Users of the data will primarily be water resource managers in the Rio Grande Basin. USGS publications will be released describing the methods and data sources and can be used as documentation for the data and to cite the data.

4. Policies for Re-use, Distribution

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Materials generated under the project will be disseminated in accordance with University/Participating institutional and NSF policies. Depending on such policies, materials may be

2 Example DMP - Rio Grande hydrology.
© DataONE 2011

transferred to others under the terms of a material transfer agreement.

The data files have a suggested citation, which will be described in the metadata in addition to the USGS publications.

5. Plans for Archiving and Preservation

All original raw data files and data source processing programs will be versioned over time and maintained in a date-stamped file structure with text files documenting the provenance. The database will be preserved in perpetuity, housed initially at the New Mexico Interstate Stream Commission Central Office in addition to an off-site copy maintained at an NMISC field office and mirrored at the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI). We will also identify appropriate archiving institutions that might serve as a mirror repository. A data policy and stewardship plan will be established. In addition to archiving, each database table will be exported to a delimited text format to ensure accessibility of the data by other software programs. The data manager at the NMISC will be responsible for the management of long-term storage and archived data.

Example DMP - Rio Grande hydrology.
© DataONE 2011

3

ELEMENTS OF A DMP FOR NSF'S DIVISION OF EARTH SCIENCES (NSF-EAR)

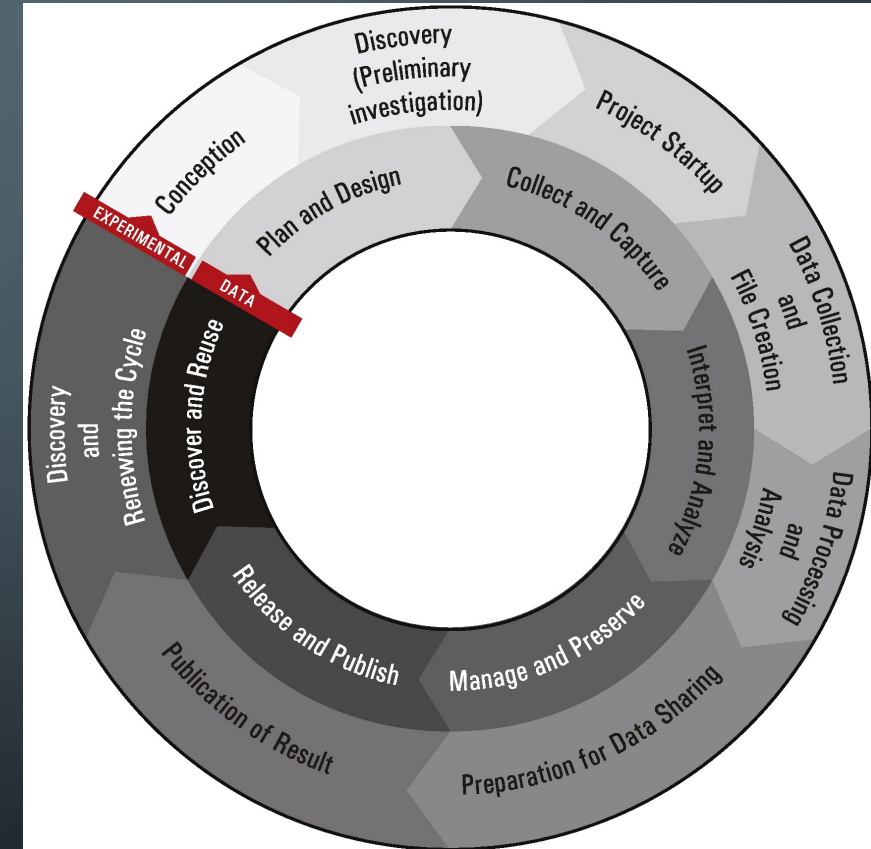
NSF DMP requirement ¹	EAR DMP guidance ²
1. Types of data produced	<ul style="list-style-type: none">• Data includes “full data sets, derived data products (e.g., model results, output, and workflows), software, and physical collections”
2. Data and metadata standards	<ul style="list-style-type: none">• Identify when data are considered “final”• NSF doesn't specify data standards – consult community and repository standards
3. Policies for access and sharing	<ul style="list-style-type: none">• Data available within 2 years of final collection (except for “exceptional circumstances”)• For continuing observations or long-term deployments, make data public in real or near-real time, as technically feasible
4. Policies for re-use and distribution	<ul style="list-style-type: none">• Data should be accessible at “no more than incremental cost”• Assign DOIs or other persistent identifiers to data products• NSF doesn't specify licensing requirements – consider community and institutional policies
5. Plans for archiving and preservation	<ul style="list-style-type: none">• Policy appendix provides “preferred” data and physical collection archives and centers• If no appropriate domain repository exists, identify and justify alternative preservation plan (e.g., museum- or university-hosted repository for long-term curation)

¹NSF 18-1 Proposal & Award Policies & Procedures Guide (Jan 2018)

²EAR Division Data Sharing Policy (April 2018)

MOTIVATION: WHY CREATE A DMP?

- A. Fulfill funding proposal requirements
- B. Plan for journal requirements
- C. Organize your plan for the research data lifecycle



A. DMP: REQUIRED FOR PROPOSALS / PROJECTS

- **NSF:** 2-page supplementary document required for all proposals
- **NASA (ROSES):** <8000 character webpage entry required for most proposals
- **USDA (NIFA):** 2-page document required for certain programs
- **USGS:** Required for all projects. Specific templates by project type.
- **NOAA (EDM):** Required for all projects. Specific templates.

SPARC provides a useful tool for data policy comparison:

<http://datasharing.sparcopen.org/>

B. DMP: KEEP UP WITH JOURNAL REQUIREMENTS

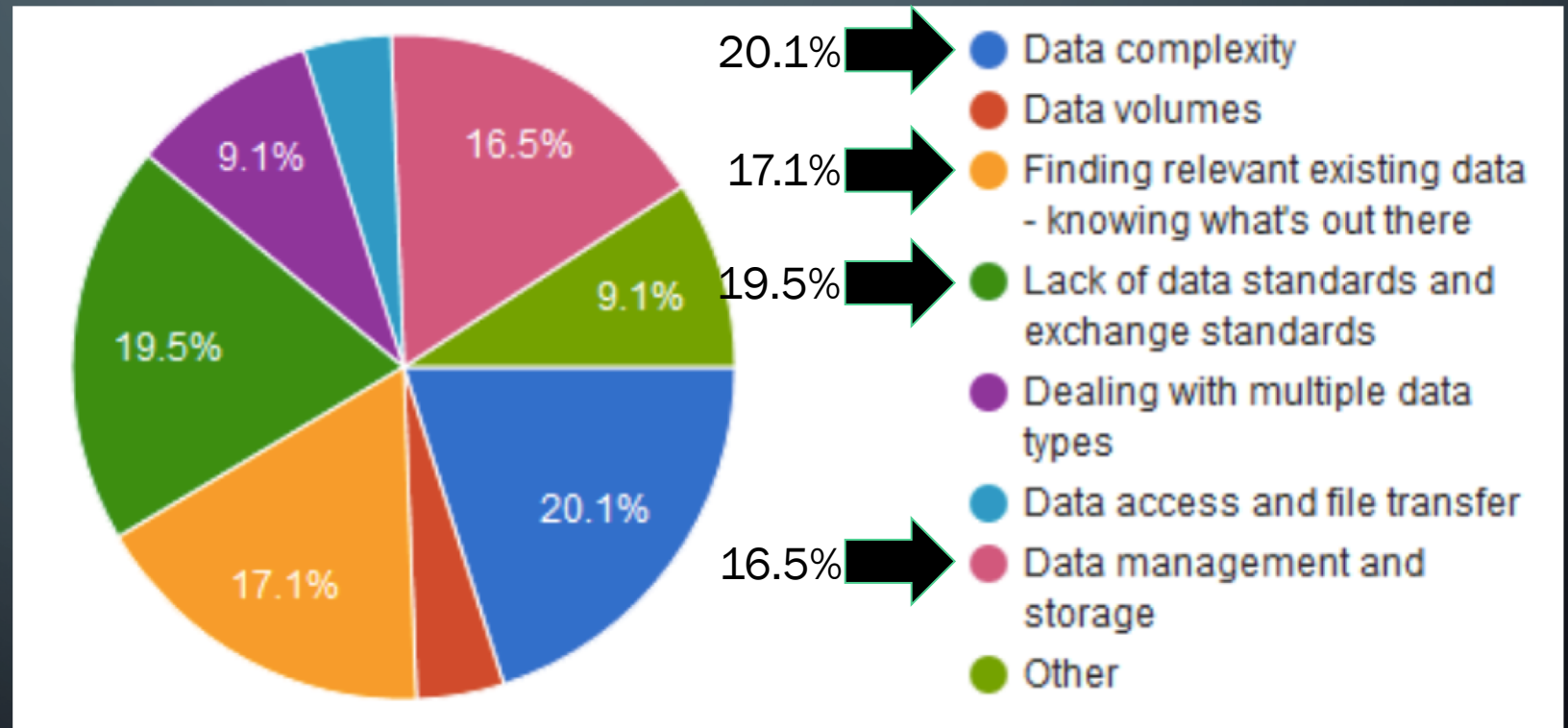
- Journals increasingly require availability of research data associated with publications.
- For AGU journals, “Data available from authors” is no longer allowed.
- Movement toward “FAIR” data
 - Findable
 - Accessible
 - Interoperable
 - Reusable



Source: Shelley Stall, AGU's “Enabling FAIR data” project

C. DMP: SUPPORTING THE RESEARCH DATA LIFECYCLE

Data preparation is a huge time sink. Can better planning reduce this burden?



Data Management Skills Gap Analysis, April 7, 2017
<http://bfe-inf.org/document/skills-gap-analysis>

WHAT IS “DATA”?

- Full data sets
- Derived data products (e.g., model results, output, workflows)
- Software
- Physical samples
- Curriculum materials

PRINCIPLES FOR DATA DEPOSIT: “FAIR”

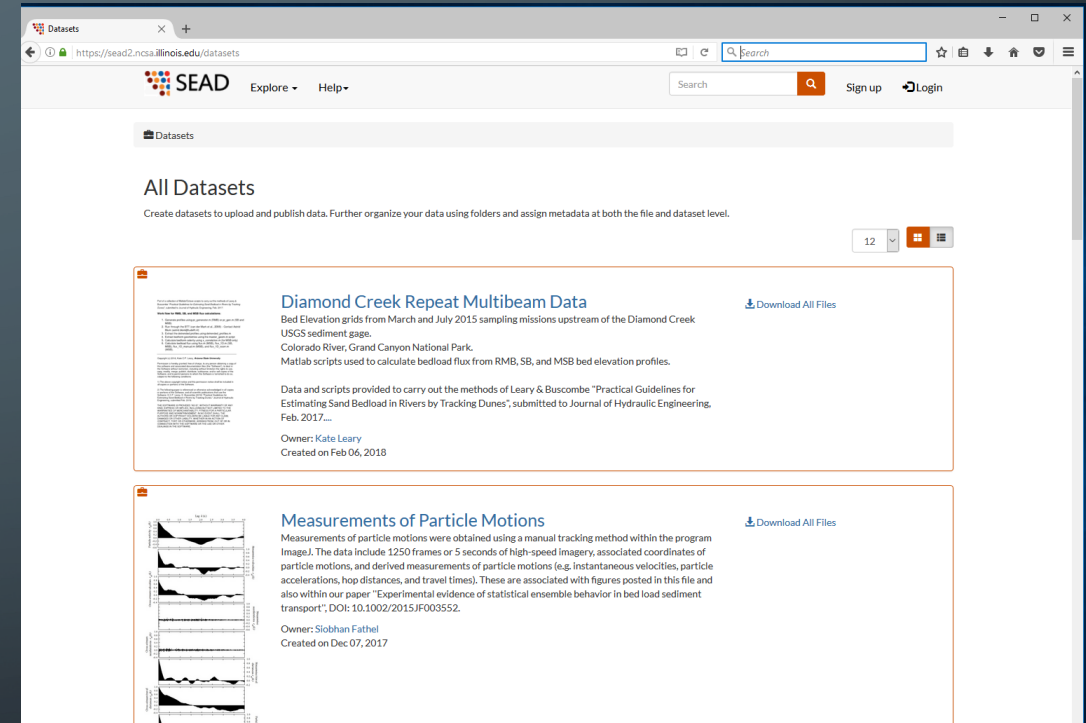
- **Findable** - Publicly available on the web (i.e., not behind a paywal) with persistent identifier (e.g., DOI)
- **Accessible** - Open access to dataset (as possible), not behind paywall (e.g., not in article supplement), long-term support (e.g., not on unmanaged website)
- **Interoperable** - Data described with appropriate machine-readable metadata
- **Reusable** - Citable with open license (e.g., CC-BY)

WHERE CAN THE DATA GO?

- **“Domain” repository** (e.g., SEN, SEAD, CSDMS) – best option when possible
- **Institutional repository** (e.g., university, museum) – check requirements
 - May store data with institutional repository but list metadata in domain repository
- **General repository** (e.g., Figshare, Zenodo, Dryad) – backup plan
- **Journal article**
 - Tables and figures – good for very high-level information, but not machine readable
 - Article supplements – **not advised**: very difficult to access, often behind paywall

PRIMARY DATASET RESOURCES

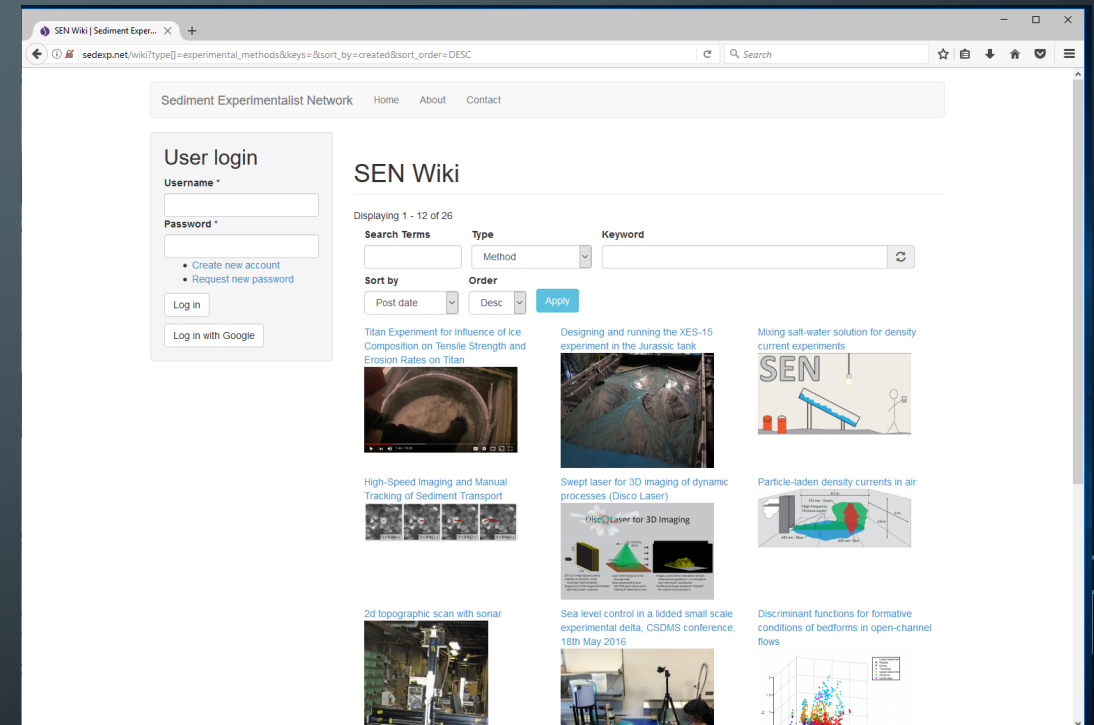
- Open resource examples – anyone can deposit, limited curation, DOIs
 - **SEAD** – Environmental data
 - **HydroShare** (CUAHSI) – Hydrologic data
 - **EarthChem** (IEDA) – Geochemical data
 - **Arctic Data Center** – Arctic data
- Restricted resource examples – require planning and curation, DOIs
 - **BCO-DMO** – Biological and chemical oceanography
 - **IRIS DMC** – Seismological data



<https://sead2.ncsa.illinois.edu/datasets>

“DERIVED” DATASET RESOURCES

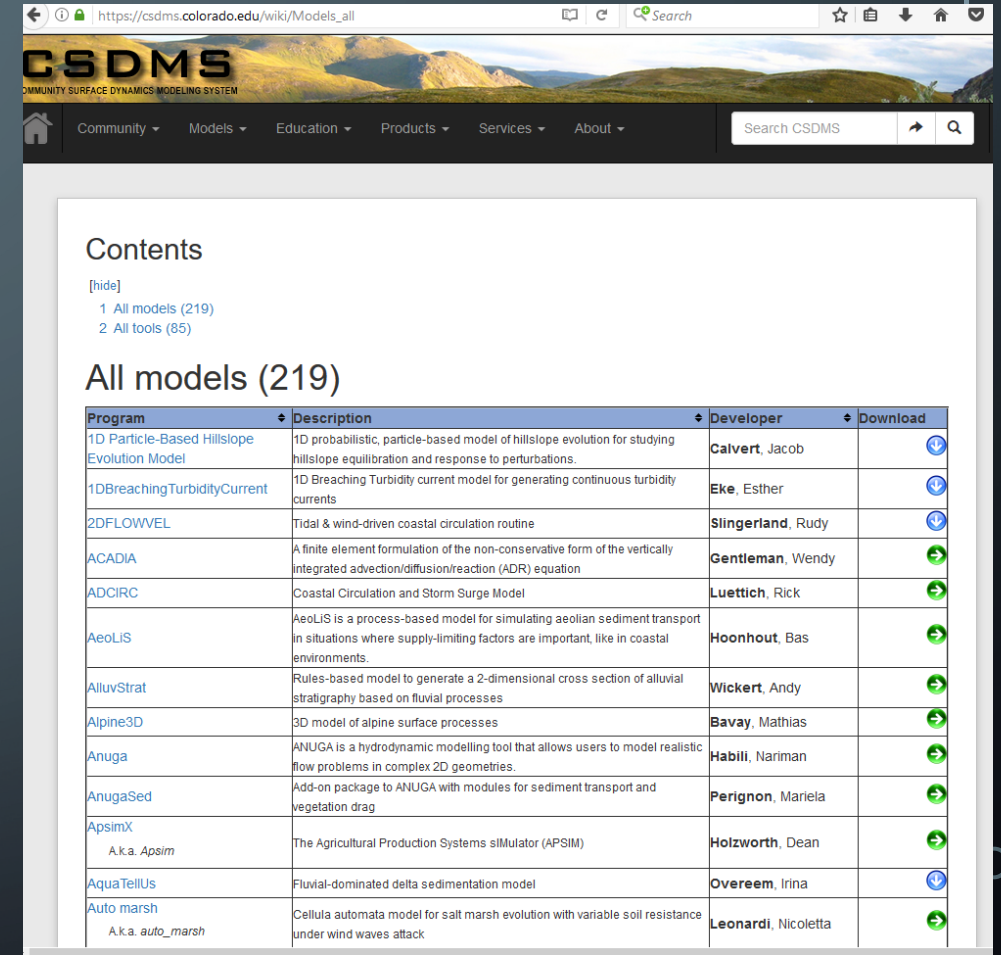
- Similar resources as for “primary” data
- **SEN Knowledge Base (SEN-KB)** – experimental sedimentology workflows and methods, no DOIs (yet?)
- **HydroShare** – Jupyter notebook workflow documentation tools



<http://sedexp.net/wiki>

SOFTWARE RESOURCES

- **GitHub** – Open collaboration resource, no domain curation or DOIs (but easy export to general purpose repos)
- **CSDMS Model Repository** – earth-surface models, multiple curation levels, DOIs
- Many data repositories will also store software, but discovery may be difficult



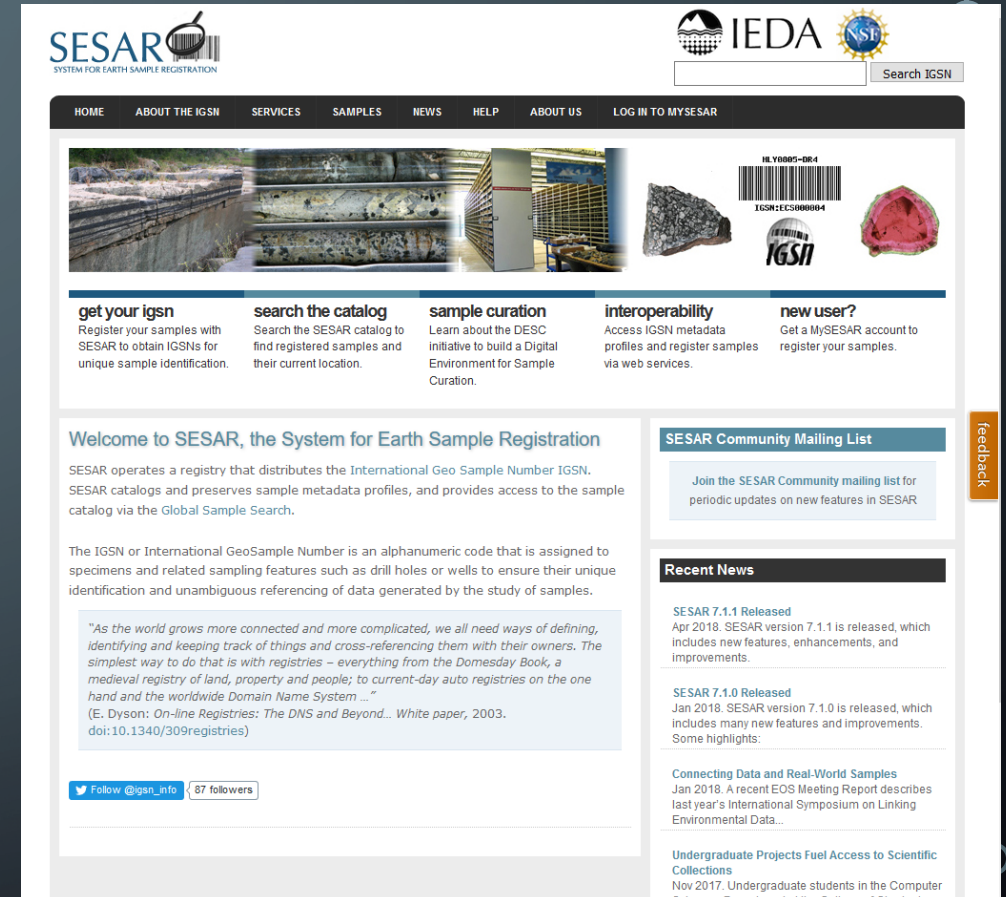
The screenshot shows the CSDMS Model Repository website. The header includes the CSDMS logo and navigation links: Community, Models, Education, Products, Services, and About. A search bar is also present. The main content area is titled 'Contents' and lists 'All models (219)' and 'All tools (85)'. Below this is a table of models with columns for Program, Description, Developer, and Download.

Program	Description	Developer	Download
1D Particle-Based Hillslope Evolution Model	1D probabilistic, particle-based model of hillslope evolution for studying hillslope equilibration and response to perturbations.	Calvert, Jacob	Download
1DBreachingTurbidityCurrent	1D Breaching Turbidity current model for generating continuous turbidity currents	Eke, Esther	Download
2DFLOWVEL	Tidal & wind-driven coastal circulation routine	Slingerland, Rudy	Download
ACADIA	A finite element formulation of the non-conservative form of the vertically integrated advection/diffusion/reaction (ADR) equation	Gentleman, Wendy	Download
ADCIRC	Coastal Circulation and Storm Surge Model	Luettich, Rick	Download
AeoLIS	AeoLIS is a process-based model for simulating aeolian sediment transport in situations where supply-limiting factors are important, like in coastal environments.	Hoonhout, Bas	Download
AlluvStrat	Rules-based model to generate a 2-dimensional cross section of alluvial stratigraphy based on fluvial processes	Wickert, Andy	Download
Alpine3D	3D model of alpine surface processes	Bavay, Mathias	Download
Anuga	ANUGA is a hydrodynamic modelling tool that allows users to model realistic flow problems in complex 2D geometries.	Habibi, Nariman	Download
AnugaSed	Add-on package to ANUGA with modules for sediment transport and vegetation drag	Perignon, Mariela	Download
ApsimX A.k.a. Apsim	The Agricultural Production Systems sImulator (APSIM)	Holzworth, Dean	Download
AquaTeliUs	Fluvial-dominated delta sedimentation model	Overeem, Irina	Download
Auto marsh A.k.a. auto_marsh	Cellula automata model for salt marsh evolution with variable soil resistance under wind waves attack	Leonardi, Nicoletta	Download

https://csdms.colorado.edu/wiki/Models_all

PHYSICAL SAMPLES

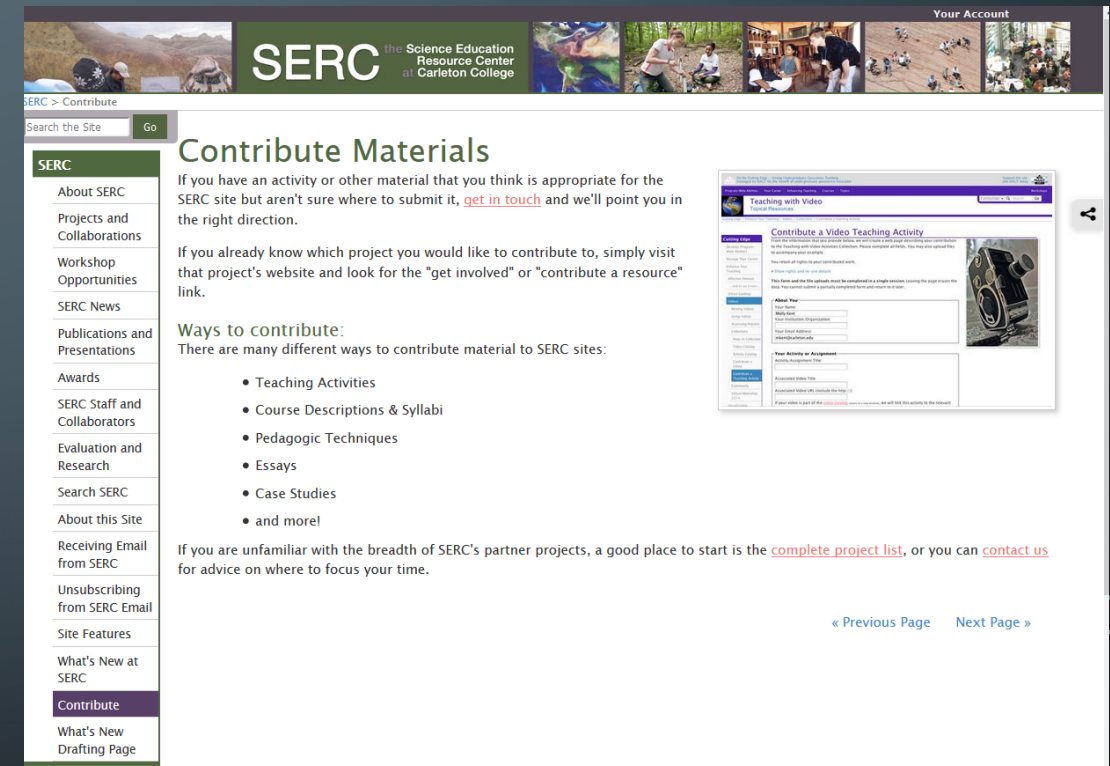
- Sample registry: SESAR (System for Earth Sample Registration)
 - Issues IGSNs (International Geo Sample Numbers)
 - Maintains geographic sample search tools
 - Does NOT store samples
- Sample storage
 - LacCore – lake core samples
 - Smithsonian NMNH Biorepository – natural history and genomic samples



<http://www.geosamples.org/>

CURRICULA, OTHER MATERIALS

- **SERC** (Science Education Resource Center) at Carleton College accepts educational materials, but lacks a one-stop repository
- Can deposit other materials in appropriate domain, institutional, or general purpose repository
- Personal websites are usually not discoverable or sustainable



SERC the Science Education Resource Center at Carleton College

Contribute Materials

If you have an activity or other material that you think is appropriate for the SERC site but aren't sure where to submit it, [get in touch](#) and we'll point you in the right direction.

If you already know which project you would like to contribute to, simply visit that project's website and look for the "get involved" or "contribute a resource" link.

Ways to contribute:
There are many different ways to contribute material to SERC sites:

- Teaching Activities
- Course Descriptions & Syllabi
- Pedagogic Techniques
- Essays
- Case Studies
- and more!

If you are unfamiliar with the breadth of SERC's partner projects, a good place to start is the [complete project list](#), or you can [contact us](#) for advice on where to focus your time.

« Previous Page Next Page »

<https://serc.carleton.edu/serc/contribute.html>

PROPOSAL DMP CREATION PRINCIPLES

- Think of the reviewers - keep it succinct, answering key questions:
 - What data will be produced?
 - Where will data be shared?
 - When will data be available?
 - Is the DMP feasible?
 - Does the DMP support the research and broader impacts?
- Organization matters
 - I suggest organizing DMP by data type
 - DMP templates can help (IEDA DMP Tool, CDL dmptool.org)

The screenshot displays the DMP (Data Management Plan) creation interface. The top section, 'DMP OVERVIEW', shows a form for creating a new DMP. It includes fields for 'DMP Template' (set to 'NSF-EAR: Earth Sciences'), 'DMP Title' (set to 'My DMP'), 'Proposal Solicitation Number', and 'Proposal Submission Deadline'. There are also sections for 'Add Co-owners' and 'Existing Co-owners'. The 'Current Status' is 'New', and there are links for 'Show History' and 'View' for 'Reviewer Comments' and 'Owner Comments'. A 'Visibility' section indicates the DMP is set to 'Test' with a 'Change Settings' button. At the bottom of the overview are 'Save', 'Cancel', and 'Save and Next >>' buttons.

The bottom section, 'DMP DETAILS', is for the 'NSF-EAR: Earth Sciences' template. It includes a 'Template Outline' with sections like 'Types of data', 'Data and metadata standards', 'Policies for access and sharing', and 'Plans for archiving and preservation of access'. The 'Instructions' tab is active, showing a detailed guide on data management and sharing. A 'Guidance' section provides specific advice on data sharing and repository selection. At the bottom of the details section are 'Cancel Changes', 'Save Response', and 'Save and Next' buttons.

<https://dmptool.org/>

RESOURCES

- DataOne: Data management education modules:
<https://www.dataone.org/education-modules>
- Coursera: “Research Data Management and Sharing”:
<https://www.coursera.org/learn/data-management>
- Earth Science Information Partners (ESIP): Data Management Training Clearinghouse: <http://dmtclearinghouse.esipfed.org>

EXERCISE: DMP CREATION FOR AN NSF PROPOSAL

- A. Think of a research project you are planning (or doing now) – ideally one related to sediment experiments (Note: we will be demonstrating SEN and SEAD later...)
- B. Pick one dataset you expect to produce (or have produced) for this project
- C. Answer these 5 questions (associated with five areas of NSF DMP):
 1. *What type of dataset is this?* (e.g., primary, workflow, model, etc.)
 2. *Which formats and standards will be used to describe this data and its metadata?*
 3. *How and when will you share this data?* (Which repository? How long after data collection?)
 4. *What are provisions for re-use?* (e.g., citing dataset with DOI, assigning a permissive license like CC-BY)
 5. *How will data / samples be preserved and for how long?* (Will data be accessible indefinitely or for a finite period of time? This will depend data volume and type.)