RESEARCH DATA ALLIANCE

Data Rescue IG

Steve Diggs @ ESIP Winter 2019
Bethesda, MD
2019.01.16

research data sharing without barriers
rd-alliance.org

Data Rescue IG

- **Status**: **Recognised & Endorsed**
- **Background**
  - Origins in CODATA's Data-At-Risk TG
  - **~110 members**
  - Nominal inactivity required reinvigoration
- **4** co-chairs
  - Lesley Wyborn, Steve Diggs, David Gallaher, Denise Hills
  - We have directly addressed the Co-Chair rotation issue (Emeritus)
  - With new co-chairs comes a change in direction, interests, operating modes, and connections (EDGI, Data Refuge, **ESIP**)
  -
- **CODATA's IDAR-TG (Hills, Diggs, Fleischer)**

Data Rescue IG

Revised Definition

- ~~Data Rescue seeks to recover older datasets and metadata from the trappings of dated media and recording hardware~~

- Data-at-Risk seeks to identify data sets that need to be preserved and/or replicated while anticipating changes in curation personnel and evolving technologies, so that they are always available as a global resources for future research, particularly research that involves longitudinal aspects.

RDA/US DataShare Fellows for RD-IG

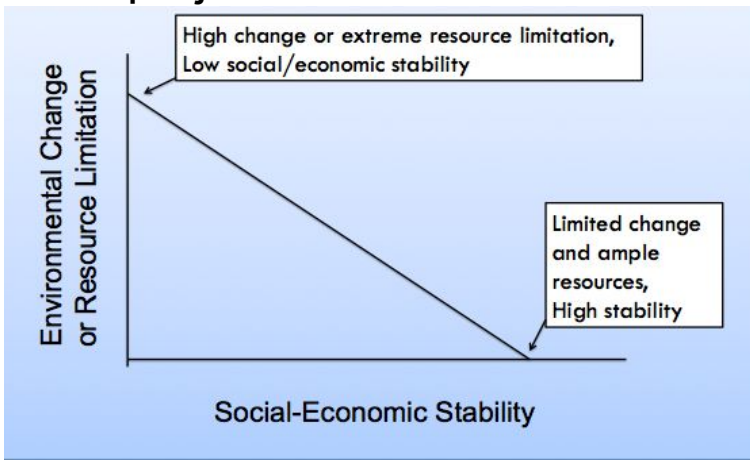
- Morgan Currie 
  - PhD: UCLA - Information Studies
  - UCLA / Stanford Postdoc
  - Focus: **Policy** / Contact: Steve Diggs

- Alia Khan
  - PhD: Univ.  Colorado - Biogeochemistry
  - NSIDC (Na now and Ice Data Center)
  - Focus: **Geoscience** / Contact: Dave Gallaher

RDA/US DataShare Fellows

Khan Survey Project

- Rescue of physical environmental/climate data to explore the relationships between climate and snow/ice/ecosystem response.
- Currently conducting a **survey** of data-at risk in the **polar**/environmental field, collecting examples of data-rescue projects, and identifying potential future needs. The results will be listed on the RDA Data Rescue IG page.
- Key organizations possessing data-at risk have been targeted in addition to the survey, as well as potential funding agencies.
- The long-term aim is to develop a foundation for **funding proposals** to develop a Data-Rescue project.



RDA/US DataShare Fellows

Khan Survey Project

Data Rescue/At-Risk Survey for Polar/Environmental Research

The natural sciences possess a rich heritage of data spanning the entire era of research, and encompassing both modern electronic formats and older analogue ones. Many of the data in analogue form cannot be accessed by present-day research, to the serious detriment of research that analyses long-term changes and attempts to predict future ones. Even data that have been digitized rarely come in formats that are interoperable, easily readable and readily accessible.

The purpose of this survey is to a) list and understand the vulnerabilities and risks of scientific research data, and b) to catalogue known data-rescue efforts as exemplars of what can be achieved.

* Required

Email address *

Your email

Name:

Your answer

Institute:

Your answer

Geographic scope of data (Arctic, Antarctic, Third-pole, etc):

Your answer

What type of data do you collect or work with? (e.g. climate, hydrology, snow cover, bio-diversity, space, oceanographic, geology, etc.)

Your answer

Is this data:

- ○ Observational Data
- ○ Computational Data
- ○ N/A

RDA/US DataShare Fellows

Currie Project

Endangered Data? DataRescue for the Geosciences

Dr. Morgan Currie

*This research is in collaboration with the Data Rescue Interest Group

Decentralized, volunteer-led DataRescue efforts launched after the election of Trump.

L.A. Times, *Quartz* and *WaPo* headlines on deleted data and rescue efforts.

DataRescues contributed to mirror repositories of federal environmental data for public access.

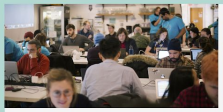
These voluntary organizations are focused on the very real possibility of federal environmental data being less accessible in the near future due to:

- **Changes in federal funding and policy**
- **Compromised access to online information resources**

Both outcomes could severely compromise environmental advocacy and the earth science enterprise.

Environmental Data Governance Initiative

- An international network that archives publicly accessible scientific data and web pages from EPA, DOE, NOAA, OSHA, NASA, USDA, DOI, and USGS.
- Monitors changes in federal agencies' governance of scientific information.
- Confidentially interviews long-time employees at EPA and OSHA.
- Monitors changes to tens of thousands of federal environmental agency web pages.



"One of the major challenges for the development of a scientific infrastructure collaboration and data sharing through information networks is to ensure the frictionless circulation of data across diverse technical platforms, organizational environments, disciplines and institutions."

– Millerand, F. & Bowker, G. "Metadata standards, trajectories and enactment in the life of an ontology"

Can we widen the scope of environmental data rescue beyond government research?

Is climate research in the geosciences publically accessible and easy to duplicate? How much is technically vulnerable? How vulnerable?

Goal: Create and populate a data-at-risk index for the RDA community.

Step 1: Identifying the Data

What geoscience research has - or could have - a role to play in federal or state level environmental policy?

Examples:
- Data on rising sea levels.
- Data showing permafrost vanishing in indigenous hunting grounds.
- Data that ties melting polar ice caps to sea level rise.

Step 2: Evaluating access

Once we identify data with clear policy implications, how accessible is this data?

Possible questions to evaluate risk:
- Who curates the data?
- How is the research and data preservation funded and for how long?
- How robust are the data management systems?
- Does the data reside somewhere else and is it easily transferrable?
- Do metadata standards encourage the data's public circulation on the web?

Research Tasks

- **Identify datasets perceived of value to environmental policy.**
- **Formulate and circulate a set of questions about the levels of accessibility and risk to those datasets.**
- **Create 'data-at-risk' scale to test at research institutions.**

See: Mayernik, et al. "Stronger Together: the case for cross-sector collaboration in identifying and preserving at-risk data." http://www.esipfed.org/press-releases/stronger-together; EDGI, "EDGI Introduction and Accomplishments" https://envirodatagov.org/publication/accomplishments-intro-may/; EDGI, "The EPA Under Siege" https://envirodatagov.org/publication/the-epa-under-siege/; Millerand & Bowker, (2009), "Metadata standards, trajectories and enactment in the life of an ontology." In S. L. Star & M. Lampland (Eds.), Formalizing practices: Reckoning with standards, numbers and models in science and everyday life (pp. 149–167). Ithaca, NY: Cornell University Press.

Special Thanks: Dr. Steve Diggs and Britt S. Paris

Goal: Create and populate a data-at-risk index for the RDA community.

Research Tasks

- **Identify datasets perceived of value to environmental policy.**
- **Formulate and circulate a set of questions about the levels of accessibility and risk to those datasets.**
- **Create 'data-at-risk' scale to test at research institutions.**

RDA
RESEARCH DATA ALLIANCE

CODATA: IDAR-TG

Reviewed and highly recommended in Gaborone

Task Group Proposal for Presentation to the 31ST CODATA General Assembly

Gaborone, Botswana, 9-10 November 2018

1. **Name of the Proposed Task Group**

Improving Data Access and Reusability (IDAR-TG)

2. **Short summary of objective(s) of the Proposed Task Group**

This Task Group's objective is to play a central role in improving scientific data access and reusability. Specifically, our main focus will be on full-stack data rescue, which includes a wide array of risks for completely dark data, non-digital datasets, and data born digital. This group will build on the original DAR-TG charter and its focus on raising awareness of data susceptible to loss throughout the research life cycle. Archivists and researchers alike would benefit from having readily available accepted principles, references, and guidelines to expose and make accessible a wide range of data and relevant services.

This task group will focus on collection and development of such principles and guidelines, in consultation with subject matter and domain experts, in order to form and continually update a common decision framework that will improve the access to scientific data. This framework will include a classification matrix to help assess a particular dataset's susceptibility to loss in the future. Additionally, this task group may eventually evolve into a consultation entity, available to institutions during their decision-making processes regarding the prioritisation of specific tasks for their data curation activities.

Finally, our task group intends to address a new and unexpected type of data rescue resulting from observations and/or survey data taken in regions populated by indigenous peoples, primarily in low and middle income countries. Rescuing these data involves ensuring that this information is available to and usable by the communities themselves as well as local researchers and policy makers.

3. **Name and Contact Details of the Principal Proposer(s)**

Dirk Fleischer
Data and Information Management for Kiel Marine Science

Christian-Albrechts University in Kiel
Kiel
Germany

Stephen Diggs
Technical Director, Hydrographic Data Office (CCHDO)

Scripps Institution of Oceanography
UC San Diego
California
USA

Denise Hills
Program Director, Energy Investigations

Geological Survey of Alabama Tuscaloosa
Alabama
USA

4. **What is the scientific merit of the proposed work and how does it contribute to CODATA's mission and objectives as laid out in the Strategic Plan;**

While only a portion of international scientific data meet the definition of "big data," there has been significant investment in data acquisition and management, and reliable projections predict that this expenditure will more than double in the next decade. It is imperative that strategic data stewardship is approached from the perspective of Return on Investment (ROI) and begin with an accurate quantification of every aspect of data acquisition, use, and on-going curation.

Data Rescue IG is *transitioning*

- **Data Rescue**
  - Applied for a session at P13/Philadelphia
  - Chairs will rotate (Hills and Diggs will stay on)
  - Change in case statement reflecting and alignment with 21st century data curation best practices