Data Rescue IG meeting

David Gallaher - National Snow and Ice Data Center, University of Colorado, Boulder, Colorado

Steve Diggs - Data Curation/CI at Scripps Institution of Oceanography San Diego

RDA 9th Plenary Meeting
Barcelona
April 6, 2017

Objectives of meeting:

A. To identify and discuss the future directions of this IG

B. Get out some ideas

C. To set up a Working Group to investigate Data At Risk, and to deliver models for plausible recommen and actions.

Data Rescue IG Meeting Agenda:

1. Introduction, and report of the Data Rescue Workshop (Boulder, September 2016).
2. Definitions: Data Rescue vs Data At Risk, and how they invite an expansion to the IG's Case Statement(s).
3. Brief contributions of Data Rescue stories (successful or not).
4. Brief examples of Data At Risk, actual or feared.
5. Open discussion on aspects of Data At Risk: How big, and how real, is the problem? Short-term and long-term solutions.
6. Consider creating a Working Group for solutions to Data At Risk/Data Rescue challenges.
7. Wrap-up: Preview of the afternoon's Joint Session on Data Rescue (1400–15:30, Plenary Room), and plans for P10 in September.

Data Rescue Workshop (Boulder, September 2016).

Background on Meeting
Held in conjunction with International Data Week
Approximately 60 attendees

Data Rescue Workshop Selected Responses

How would you define “at risk” data?

“At risk” data is any data that is not accessible in a retrievable digital format.

Important to rank data by its value - how do you decide how best to target data rescue efforts?

Funding for data rescue, curation, and future planning are all on a continuum and compete for the same pool.

We need to modernize data rescue; we are stuck in the past of data curation.

Publishing metadata – metadata needs to be full and utterly

Forgotten data - only author knows it exists

The decision regarding whether it is acceptable to dispose of the “original” might have to depend on the characteristics of the “original”.

Data Rescue Workshop Selected Responses

If a set of records has been requested by no one for more than 50 years should it be kept?

- Do they know they exist?
- Storage expensive; but yes, don’t know what it will lead to.
- Value of data may change; not necessarily used for project it was collected.
- Tiers of storage based on access/use.
- If cost to save is cheaper than cost to collect, that’s a good reason to keep
- Prioritize data for discovery.
- How to not duplicate efforts
- DOI’s make data discoverable and not so likely to be duplicated
- Govt policy; vs scientists interpretation of policy
- when are copyright/embargo periods?

Data Rescue Workshop Selected Responses

To what extent do metadata records require the interpretation and involvement of the researchers originally involved in data collection?

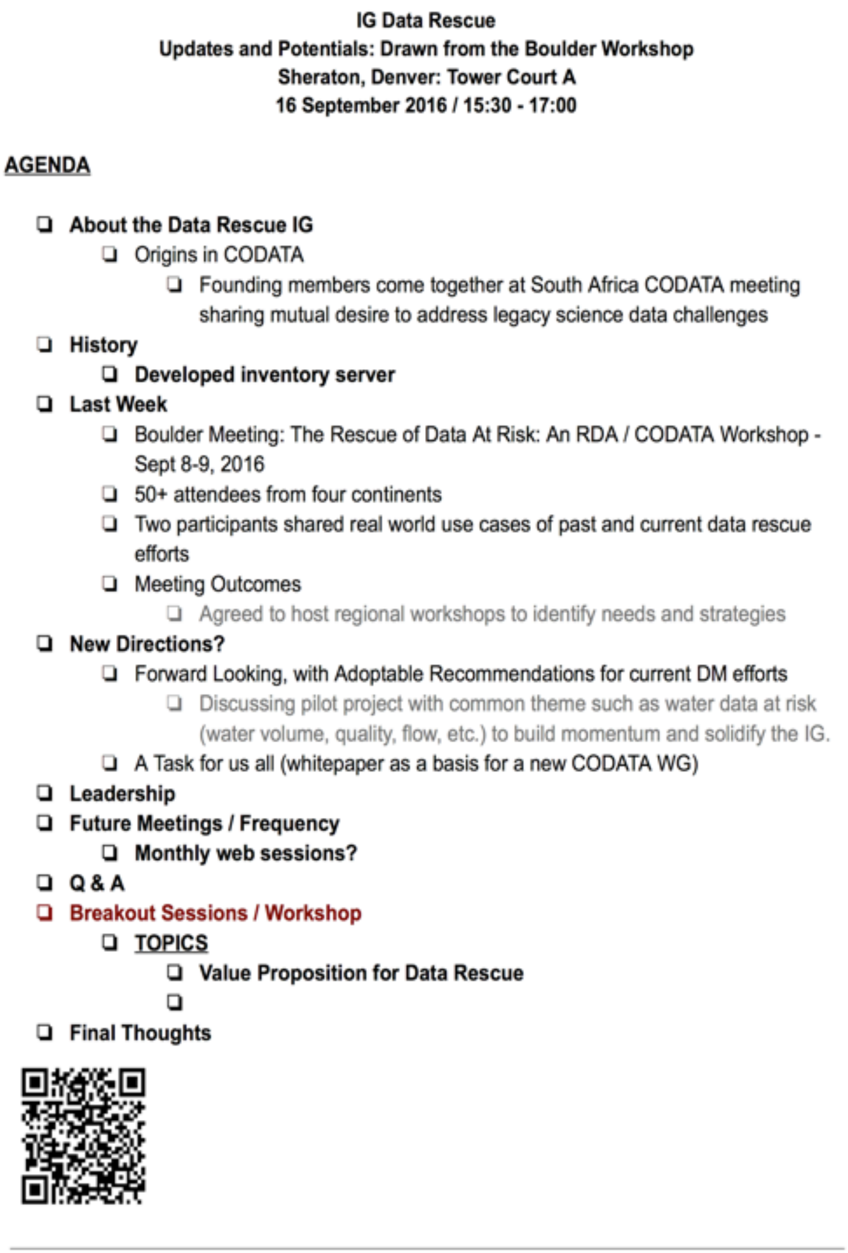
- For actually getting content, it is essential. For figuring out how to process the data.
- Original person can tell you why they did it, what it was originally designed for, what it shouldn't be used for.
- Each organization might have different approaches for what metadata is collected. Often have to take various pieces that are out there, and make sense of it.
- Need especially for the observer for with the description
- Metadata - different people talk about different things. E.g. records vs metadata for a full collection
- Go out for beer with the creators.
- Might not be able to determine accuracy and/or uncertainty

Data Rescue Workshop Selected Responses

Does your organization have a process in place for 1) defining levels of service, 2) versioning data, 3) de-accessioning data?

- Levels of service can vary based on customer usage. Are these properly scaled to respond to this variance? Ideally yes.
- Versioning: DOIs complicated this issue, one landing page for current version, with references to previous versions.
- How do you roll back datasets to ensure reproducibility of research results? One option is to produce disclaimer for users on what version to use to ensure reproducibility. Corrections can lead to large shifts in the data, which must be documented and made clearly visible.
- Subsetting data is convenient but can result in loss of information. Make sure you document what/how you subsetted.

Notes from Denver P8 Data Rescue Session



IG Data Rescue
Updates and Potentials: Drawn from the Boulder Workshop
Sheraton, Denver: Tower Court A
16 September 2016 / 15:30 - 17:00

AGENDA

- □ **About the Data Rescue IG**
  - □ Origins in CODATA
    - □ Founding members come together at South Africa CODATA meeting sharing mutual desire to address legacy science data challenges
- □ **History**
  - □ **Developed inventory server**
- □ **Last Week**
  - □ Boulder Meeting: The Rescue of Data At Risk: An RDA / CODATA Workshop - Sept 8-9, 2016
  - □ 50+ attendees from four continents
  - □ Two participants shared real world use cases of past and current data rescue efforts
  - □ Meeting Outcomes
    - □ Agreed to host regional workshops to identify needs and strategies
- □ **New Directions?**
  - □ Forward Looking, with Adoptable Recommendations for current DM efforts
    - □ Discussing pilot project with common theme such as water data at risk (water volume, quality, flow, etc.) to build momentum and solidify the IG.
  - □ A Task for us all (whitepaper as a basis for a new CODATA WG)
- □ **Leadership**
- □ **Future Meetings / Frequency**
  - □ **Monthly web sessions?**
- □ **Q & A**
- □ **Breakout Sessions / Workshop**
  - □ **TOPICS**
    - □ **Value Proposition for Data Rescue**
    - □
- □ **Final Thoughts**

Proposed Definition of: Data Rescue versus Data at Risk

Why is this important?

Data Rescue – Capture and conversion of data from an older media (paper, film, obsolete format) to a modern format with metadata

Data at Risk – Data that may be deleted or made unavailable either by neglect or design

How might these definitions invite an expansion to the IG's Case Statement(s)?

Brief contributions of Data Rescue stories (successful or not).

Tell us your story (*5 min max*)

Raider of lost ark warehouse (Imaginary)

EROS data center film storage (Real)

Japanese hydrographic survey during the WW2
- a very unsatisfactory archeology -

Shoichi Kizu
(Tohoku University, Japan)

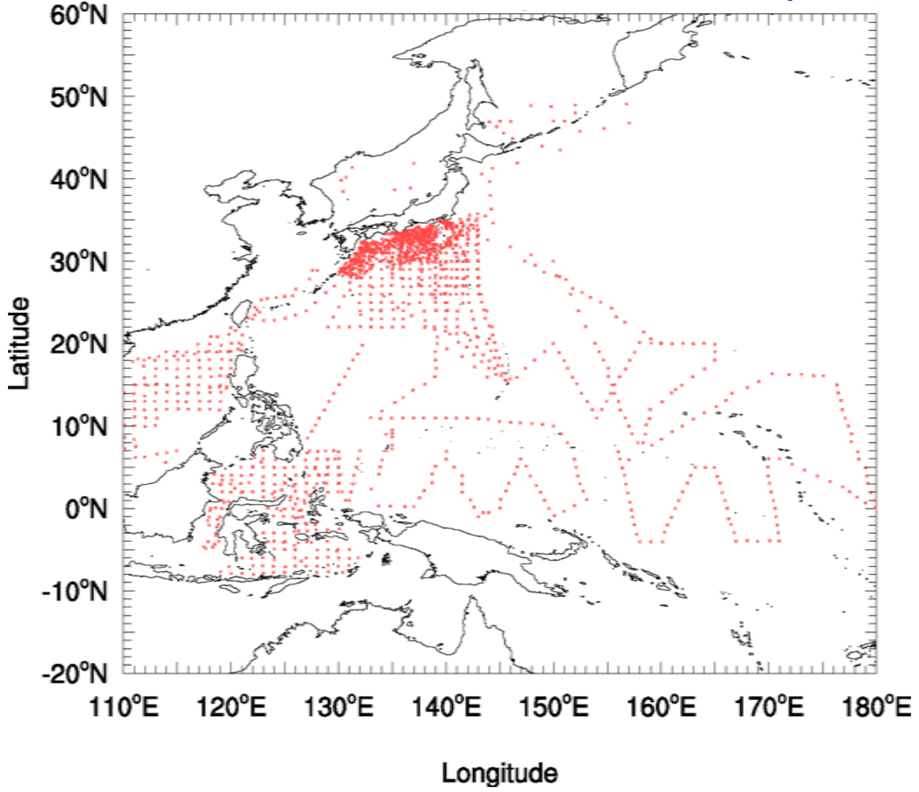
edited and updates by
S. Diggs
(Scripps Institution of Oceanography)

Specific Case:

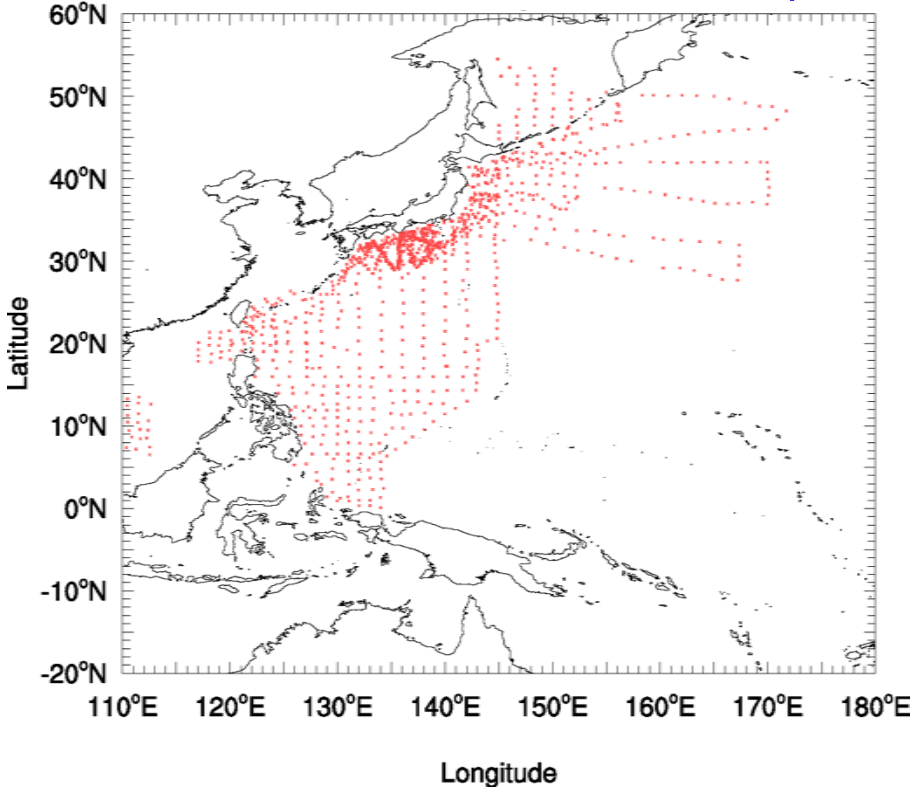
<u>Temperature</u> Measurements Sea of Japan during WWII

Deepest sampling > 1000m, country code “JP” in WOD09

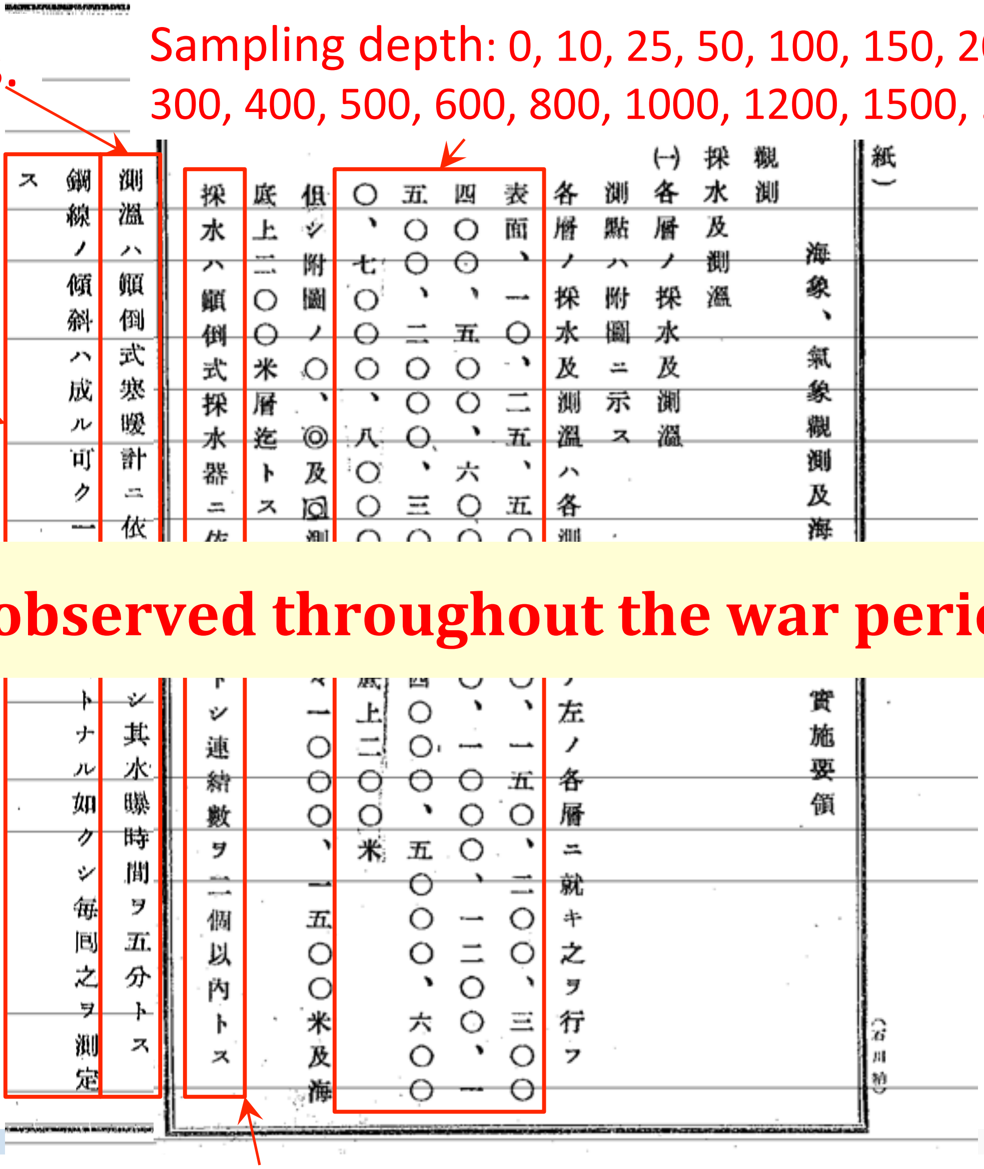
1941 1608 pts



1942 1195 pts



A sample document of “instruction” for a captain of a naval survey ship



Use reversing thermometers.

Sampling depth: 0, 10, 25, 50, 100, 150, 200, 300, 400, 500, 600, 800, 1000, 1200, 1500, …

Try best to keep wire angle smaller than 10 degrees.

紙）

海象、氣象觀測及海

觀測

採水及測溫

(一) 各層ノ採水及測溫

測點ハ附圖ニ示ス

各層ノ採水及測溫ハ各測

表面、一〇、二五、五〇

四〇〇、五〇〇、六〇〇

五〇〇、二〇〇〇、三〇

〇、七〇〇〇、八〇〇〇

但シ附圖ノ〇、◎及回測

底上二〇〇米層迄トス

採水ハ顚倒式採水器ニ依

測溫ハ顚倒式寒暖計ニ依

鋼線ノ傾斜ハ成ルヘク一

ス

Not sure if they are really observed throughout the war period.

（石川精）

實施要領

ノ左ノ各層ニ就キ之ヲ行フ

〇、一五〇、二〇〇、三〇〇

〇、一〇〇〇、一二〇〇、一

四〇〇〇、五〇〇〇、六〇〇

底上二〇〇米

一〇〇〇、一五〇〇米及海

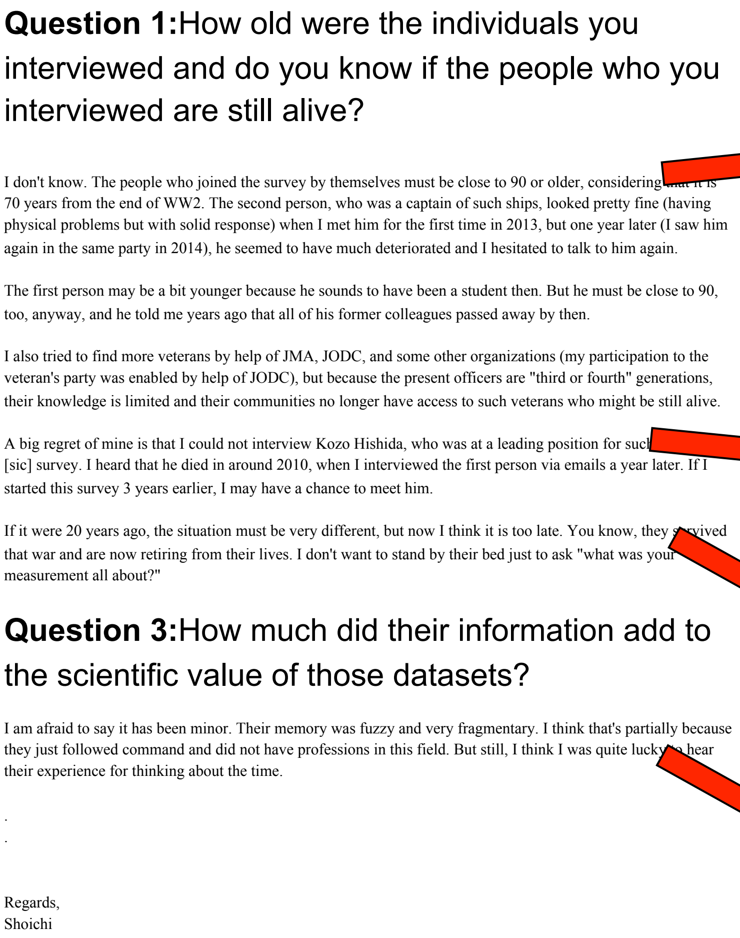
トシ連結數ヲ二個以內トス

シ其水瞬時間ヲ五分トス

トナル如クシ毎回之ヲ測定

Use reversing bottles, but do not attach more than 2 per cast.

A Race Against Time ...



Question 1: How old were the individuals you interviewed and do you know if the people who you interviewed are still alive?

I don't know. The people who joined the survey by themselves must be close to 90 or older, considering that it is 70 years from the end of WW2. The second person, who was a captain of such ships, looked pretty fine (having physical problems but with solid response) when I met him for the first time in 2013, but one year later (I saw him again in the same party in 2014), he seemed to have much deteriorated and I hesitated to talk to him again.

The first person may be a bit younger because he sounds to have been a student then. But he must be close to 90, too, anyway, and he told me years ago that all of his former colleagues passed away by then.

I also tried to find more veterans by help of JMA, JODC, and some other organizations (my participation to the veteran's party was enabled by help of JODC), but because the present officers are "third or fourth" generations, their knowledge is limited and their communities no longer have access to such veterans who might be still alive.

A big regret of mine is that I could not interview Kozo Hishida, who was at a leading position for such war-time [sic] survey. I heard that he died in around 2010, when I interviewed the first person via emails a year later. If I started this survey 3 years earlier, I may have a chance to meet him.

If it were 20 years ago, the situation must be very different, but now I think it is too late. You know, they survived that war and are now retiring from their lives. I don't want to stand by their bed just to ask "what was your measurement all about?"

Question 3: How much did their information add to the scientific value of those datasets?

I am afraid to say it has been minor. Their memory was fuzzy and very fragmentary. I think that's partially because they just followed command and did not have professions in this field. But still, I think I was quite lucky to hear their experience for thinking about the time.

.
.

Regards,
Shoichi

...The people who joined the survey by themselves must be close to 90 or older, considering that it is 70 years from the end of WW2. The second person, who was a captain of such ships, looked pretty fine (having physical problems but with solid response) when I met him for the first time in 2013, but one year later (I saw him again in the same party in 2014), he seemed to have much deteriorated and I hesitated to talk to him again.

.

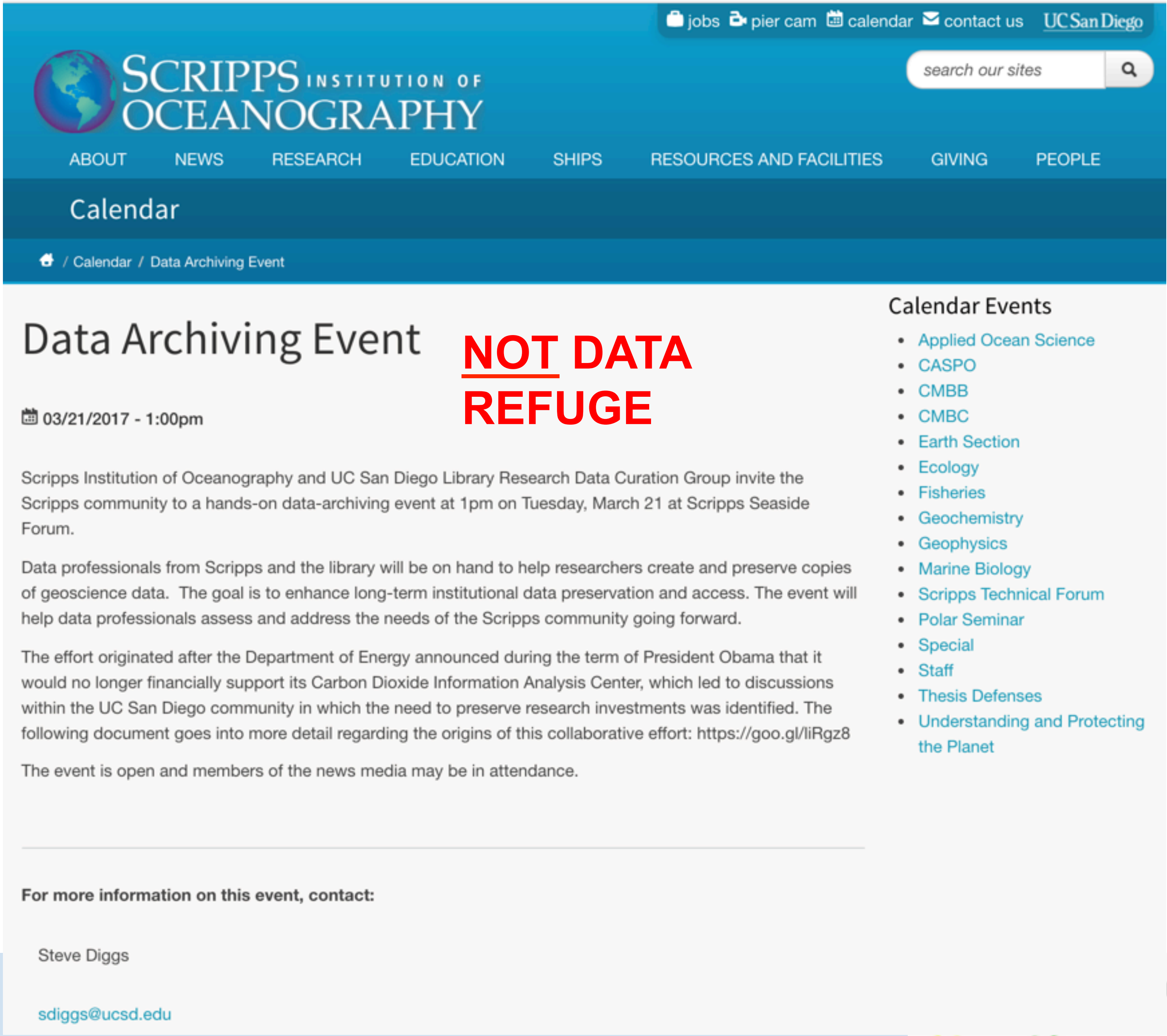
.

A big regret of mine is that I could not interview Kozo Hishida, who was at a leading position for such war-time* [sic] *survey. I heard that he died in around 2010, when I interviewed the first person via emails a year later. If I started this survey 3 years earlier, I may have a chance to meet him.

.

If it were 20 years ago, the situation must be very different, but now I think it is too late. You know, they survived that war and are now retiring from their lives. I don't want to stand by their bed just to ask "what was your measurement all about?"

I am afraid to say it has been minor. Their memory was fuzzy and very fragmentary. I think that's partially because they just followed command and did not have professions in this field. But still, I think I was quite lucky to hear their experience for thinking about the time.



Data Archiving Event

<u>NOT</u> DATA REFUGE

03/21/2017 - 1:00pm

Scripps Institution of Oceanography and UC San Diego Library Research Data Curation Group invite the Scripps community to a hands-on data-archiving event at 1pm on Tuesday, March 21 at Scripps Seaside Forum.

Data professionals from Scripps and the library will be on hand to help researchers create and preserve copies of geoscience data. The goal is to enhance long-term institutional data preservation and access. The event will help data professionals assess and address the needs of the Scripps community going forward.

The effort originated after the Department of Energy announced during the term of President Obama that it would no longer financially support its Carbon Dioxide Information Analysis Center, which led to discussions within the UC San Diego community in which the need to preserve research investments was identified. The following document goes into more detail regarding the origins of this collaborative effort: https://goo.gl/liRgz8

The event is open and members of the news media may be in attendance.

For more information on this event, contact:

Steve Diggs

sdiggs@ucsd.edu

Calendar Events

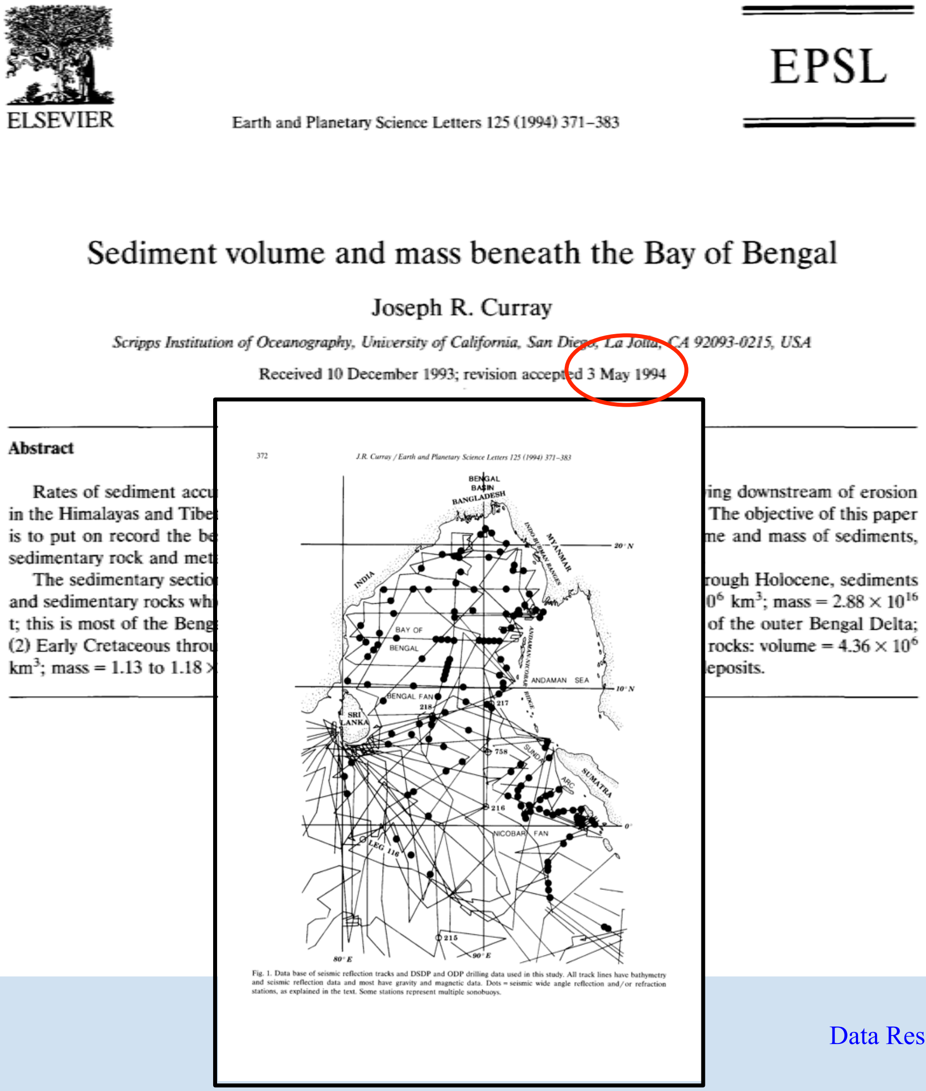
- Applied Ocean Science
- CASPO
- CMBB
- CMBC
- Earth Section
- Ecology
- Fisheries
- Geochemistry
- Geophysics
- Marine Biology
- Scripps Technical Forum
- Polar Seminar
- Special
- Staff
- Thesis Defenses
- Understanding and Protecting the Planet

JOSEPH CURRAY

EMERITUS, GEOSCIENCES RESEARCH DIVISION

Scripps Institution of Oceanography

SCRIPPS INSTITUTION OF OCEANOGRAPHY
UC San Diego



ELSEVIER

Earth and Planetary Science Letters 125 (1994) 371–383

EPSL

Sediment volume and mass beneath the Bay of Bengal

Joseph R. Curray

Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0215, USA

Received 10 December 1993; revision accepted 3 May 1994

Abstract

Rates of sediment accu
in the Himalayas and Tibe
is to put on record the be
sedimentary rock and met
The sedimentary sectio
and sedimentary rocks wh
t; this is most of the Beng
(2) Early Cretaceous throu
km^3; mass = 1.13 to 1.18 ×

ing downstream of erosion
The objective of this paper
me and mass of sediments,

rough Holocene, sediments
0^6 km^3; mass = 2.88 × 10^{16}
of the outer Bengal Delta;
rocks: volume = 4.36 × 10^6
eposits.

372 J.R. Curray / Earth and Planetary Science Letters 125 (1994) 371–383

Fig. 1. Data base of seismic reflection tracks and DSDP and ODP drilling data used in this study. All track lines have bathymetry and seismic reflection data and most have gravity and magnetic data. Dots = seismic wide angle reflection and/or refraction stations, as explained in the text. Some stations represent multiple sonobuoys.

"I wasn't able to make the SIO data meeting yesterday, but I'm not getting any younger, and I want to give the GDC my Bay of Bengal seismic data."

ACTIONS:

- Digitize Master Map
- Data Summary Table
- Raw Data
- Processing Notes

From P7: Data Rescue Steps

1. Identify high-profile datasets that need rescuing (climate, meteorological, socio-medical)
2. Identify experts associated with those datasets
3. Secure funding (by domain) for those datasets
4. Match funding to experts
5. Execute
6. Publish results, publicize through RDA / Belmont Forum, etc.

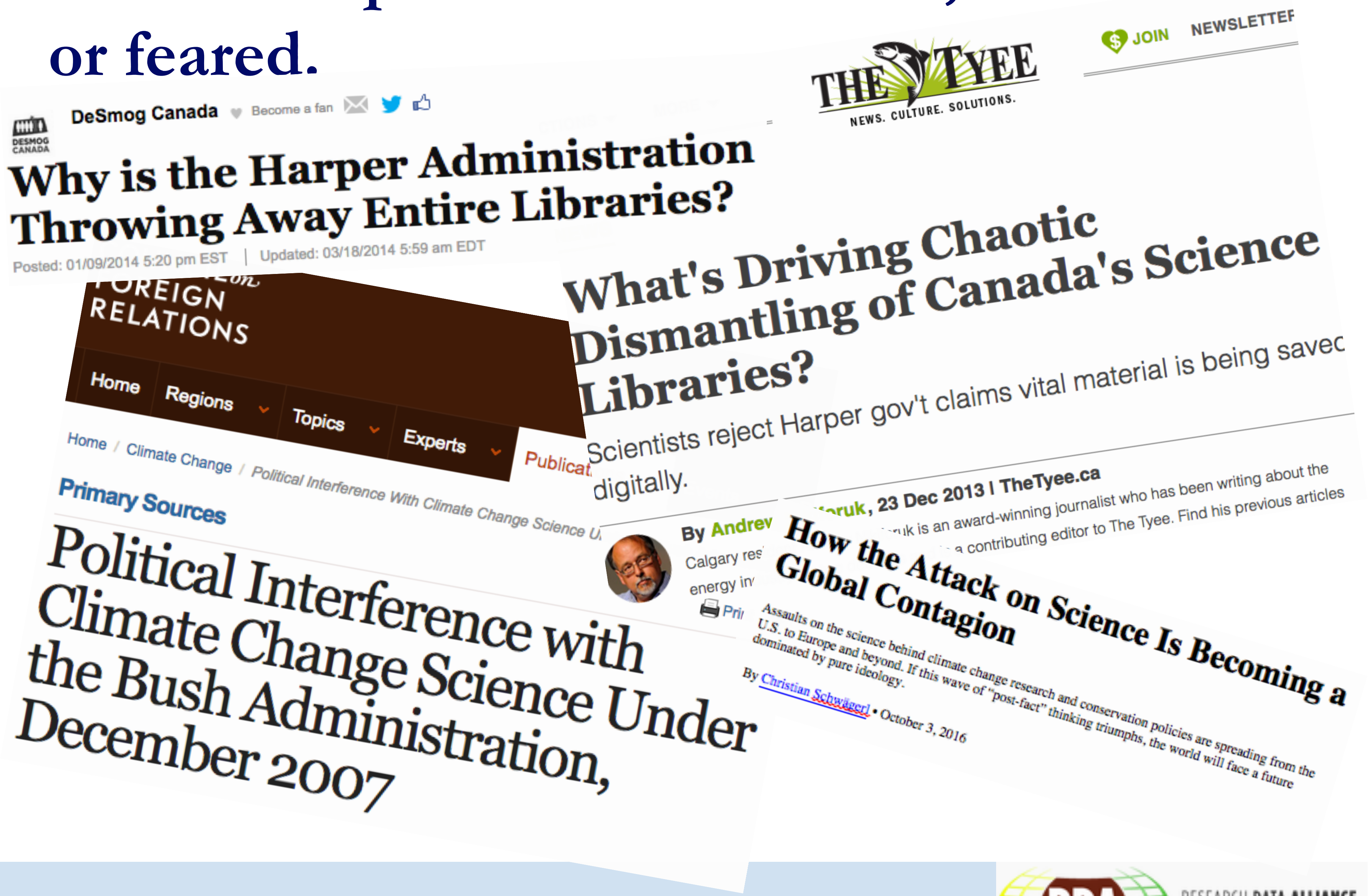
When do we know data is at risk?

1. When it is no longer available (mirror sites?)
2. When there are statements indicating it's demise
3. When there are warning from the repository (trust)
4. When the link has been removed
5. Other?

Why might data be legitimately be deleted?

1. When it is superseded by an improved version
2. When it is known to contain significant errors
3. The data is protected (archeological sites, etc)
3. Other?

Brief examples of Data At Risk, actual or feared.



DeSmog Canada Become a fan

Why is the Harper Administration Throwing Away Entire Libraries?

Posted: 01/09/2014 5:20 pm EST | Updated: 03/18/2014 5:59 am EDT

THE TYEE NEWS. CULTURE. SOLUTIONS.

JOIN NEWSLETTEF

What's Driving Chaotic Dismantling of Canada's Science Libraries?

Scientists reject Harper gov't claims vital material is being savec digitally.

By Andrew ...oruk, 23 Dec 2013 | TheTyee.ca

Calgary res... energy in... ...uk is an award-winning journalist who has been writing about the ... a contributing editor to The Tyee. Find his previous articles

...on FOREIGN RELATIONS

Home Regions Topics Experts Publicat

Home / Climate Change / Political Interference With Climate Change Science U

Primary Sources

Political Interference with Climate Change Science Under the Bush Administration, December 2007

How the Attack on Science Is Becoming a Global Contagion

Assaults on the science behind climate change research and conservation policies are spreading from the U.S. to Europe and beyond. If this wave of "post-fact" thinking triumphs, the world will face a future dominated by pure ideology.

By Christian Schwägerl • October 3, 2016

Destruction of US EPA records 2006

Truth Out, Friday 08 December 2006

Washington, DC - In defiance of Congressional requests to immediately halt closures of library collections, the U.S. Environmental Protection Agency is purging records from its library websites, making them unavailable to both agency scientists and outside researchers, according to documents released today by Public Employees for Environmental Responsibility (PEER). At the same time, EPA is taking steps to prevent the re-opening of its shuttered libraries, including the hurried auctioning off of expensive bookcases, cabinets, microfiche readers and other equipment for less than a penny on the dollar.

In a letter dated November 30, 2006, four incoming House Democratic committee chairs demanded that EPA Administrator Stephen Johnson assure them "that the destruction or disposition of all library holdings immediately ceased upon the Agency's receipt of this letter and that all records of library holdings and dispersed materials are being maintained." On the very next day, December 1st, EPA de-linked thousands of documents from the website for the Office of Prevention, Pollution and Toxic Substances (OPPTS) Library, in EPA's Washington D.C. Headquarters.

Last month without notice to its scientists or the public, EPA abruptly closed the OPPTS Library, the agency's only specialized research repository on health effects and properties of toxic chemicals and pesticides. The web purge follows reports that library staffers were ordered to destroy its holdings by throwing collections into recycling bins.

Rise of the Territorialist

The foreign policy expert Ulrich Speck of the Washington, D.C.-based Transatlantic Academy has coined a new term "territorialists," in contrast to "globalists." Territorialists find climate change suspect, in part because it could mean that Europeans or Americans have to forego material wealth in order to help other people living in faraway lands, such as the inhabitants of Pacific Islands. The climate doesn't have walls, and the science studying it is globalist by nature. The worldwide network of measuring stations that monitor temperature, water salinity, and air currents is the same kind of masterpiece of international cooperation as the retrieval of the hugely important Vostok and Dome C ice cores — with their invaluable climate data — from the Antarctic by European, Chinese, Japanese, and Russian scientists, among others.

It is precisely this global ethos of science that draws the territorialists into the fray. When asked recently why he disliked environmental thinking, Alexander Gauland, a leader of Germany's AfD, answered: "Excuse me, but 'environmental' has nothing to do with national identity." This type of thinking is on the rise, and it could have earth-changing consequences.

Open discussion on aspects of Data At Risk:

1. Don’t Panic. How big, and how real, is the problem? It is more than Trump.

2. Work with the data centers

3. What are Short-term and long-term solutions?

Recommendations From ESIP

Stronger together: the case for cross-sector collaboration in identifying and preserving at-risk data

- Confirm the current risk level of data sets in line for rescue.
- Contact the data center before rescuing their data at large scale
- Data center personnel may be able to guide you to the best mechanism for accessing the data (and associated metadata)
- Not all conditions for access are ill-intentioned (data protected).
- Gather, maintain and use all persistent identifiers (PIDs), DOIs Provenance / chain of custody – All data must be traceable back to their original sources, and validation mechanisms such as checksums.
- Plan for maintenance and versioning. Many federal data sets change over time, with new data being added, or values being changed
- Creating snapshots of data may exacerbate problems related to authoritative versioning and communication of changes.

http://www.esipfed.org/press-releases/stronger-together

Authors: Matthew S. Mayernik, Robert R. Downs, Ruth Duerr, Sophie Hou, Natalie Meyers, Nancy Ritchey, Andrea Thomer, Lynn Yarmey

Data Refuge, a nationwide volunteer effort led by librarians, scientists, and coders to discover and back up research data at risk of disappearing

Recently, the first dedicated effort to make a mirror of the 2.5 million datasets/40 TB of data contained within Data.gov was completed, with the mirror being placed on the University of California infrastructure in partnership with the California Digital Library. In addition to Data.gov, which relies on federal departments to self-report their raw data, there are hundreds of federal FTP servers that contain data and thousands of federal websites that may contain links to data over HTTP. Very few of these FTP/HTTP resources have machine-readable metadata, and many require scraping or custom data export that crawlers and bots can't do.

Is this the best approach? Is it practical?

Data-At-Risk Inventory (DARI) 2010-2013

The Data-at-Risk Initiative (DARI) is a project of the Committee on Data for Science and Technology (CODATA) Data at Risk Task Group (DARTG) attempted to create an inventory of valuable scientific data that are at risk of being lost to posterity. Examples of at-risk data include analog formats, such as loose paper documents, laboratory notebooks, photographs or glass plates, digital data stored in obsolete and deteriorating formats such as magnetic tapes or floppy disks, or any formats that face disposal due to inadequate storage. The DARI was not a repository for data.

This group appears to have gone quiet.

High latency – Small size storage - Long term DNA solution . DNA could store all of the world's data in one room – *Robert Service - Science*

Humanity has a data storage problem: More data were created in the past 2 years than in all of preceding history. Researchers encode digital data in DNA to create the highest-density large-scale data storage scheme ever invented. Capable of storing 215 petabytes in a single gram of DNA, the system could, store every bit of datum ever recorded by humans in a container equivalent to 2 pickup trucks. Cost is an issue.

DNA could store all of the world's data in one room

They started with six files, including a full computer operating system, a computer virus, an 1895 French film called *Arrival of a Train at La Ciotat*, and a 1948 study by information theorist Claude Shannon. They first converted the files into binary strings of 1s and 0s, compressed them into one master file, and then split the data into short strings of binary code.

They sent these as text files to Twist Bioscience, a San Francisco, California–based startup, which then synthesized the DNA strands encoding with their files. To decode them, the pair used modern DNA sequencing technology. A computer translated the genetic code back into binary reassemble the six original files. 100% of the data was recovered.

Next Steps?

Create a Working Group for Recommendations for Data At Risk & Data Rescue challenges

- Collaborate/partner across data centers and external organizations to protect “at risk” data.
- Build a fence at the top of the hill rather than an ambulance at the bottom of the hill
- Create non-traditional funding models for long term storage and replication (clearinghouse?).
- Develop Strategies, levels of support and guideline procedures for both Data Rescue and Data at Risk.
- Develop a page on the IG's Website where people can list data considered to be at risk.
- Expect calls from our new Data Share fellows, one for Data Rescue & one for Data at Risk (Morgan, Alia Khan)

Why Do Data Rescue?

If you don't know where you've been…
How can you tell where you are going?

Don't Forget: Publicize your successful project

Wrap Up:

- Plans for P10 in September in Montreal
- Preview of the afternoon’s Joint Session on Data Rescue (1400–15:30, Plenary Room)