Simple and Effective Methods for Managing and Sharing Scientific Data

The CZ Hub Team

November 10, 2021

Data are first-class products of research

Data and Models as Research Products

Some principles from the FORCE11 Data Citation Synthesis Group:

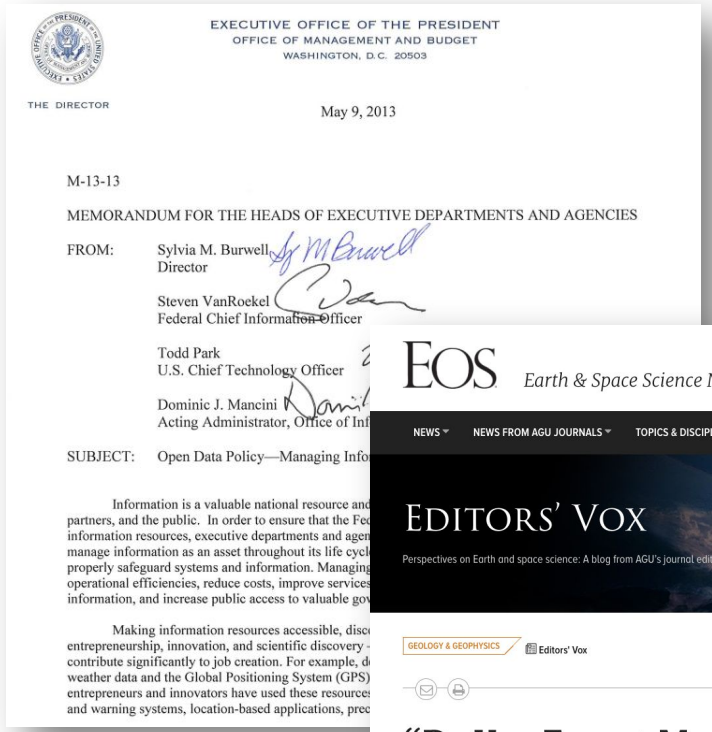
1. Data should be considered legitimate, citable, products of research
2. Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data
3. In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited
4. . . .

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

Data Sharing: Requirements

- Federal grants require data sharing and availability (data management plans)
- Many scholarly journals (all AGU journals)
- **Data** is defined broadly (per AGU policy):
  - Data used to generate, or be displayed in, figures, graphs, plots, videos, animations, or tables in a paper.
  - New protocols or methods used to generate the data in a paper.
  - New code/computer software used to generate results or analyses reported in the paper.
  - Derived data products reported or described in a paper.

"When you publish a research paper, you are also simultaneously publishing the data that supports your work. The readers of your article have equal rights to see both the words and the numbers – they are inseparable."



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF MANAGEMENT AND BUDGET
WASHINGTON, D.C. 20503

THE DIRECTOR

May 9, 2013

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Sylvia M. Burwell
Director

Steven VanRoekel
Federal Chief Information Officer

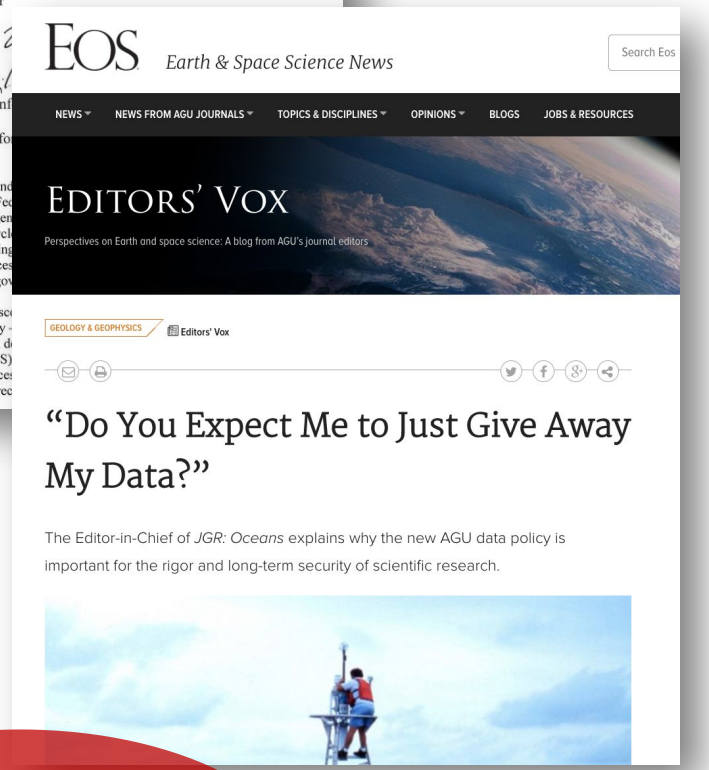
Todd Park
U.S. Chief Technology Officer

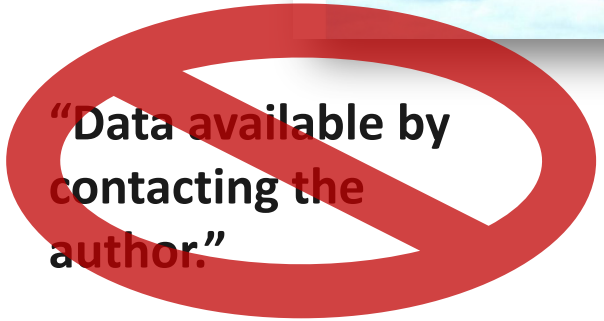
Dominic J. Mancini
Acting Administrator, Office of Inf

SUBJECT: Open Data Policy—Managing Info

Information is a valuable national resource and
partners, and the public. In order to ensure that the Fe
information resources, executive departments and agen
manage information as an asset throughout its life cycl
properly safeguard systems and information. Managing
operational efficiencies, reduce costs, improve services
information, and increase public access to valuable gov

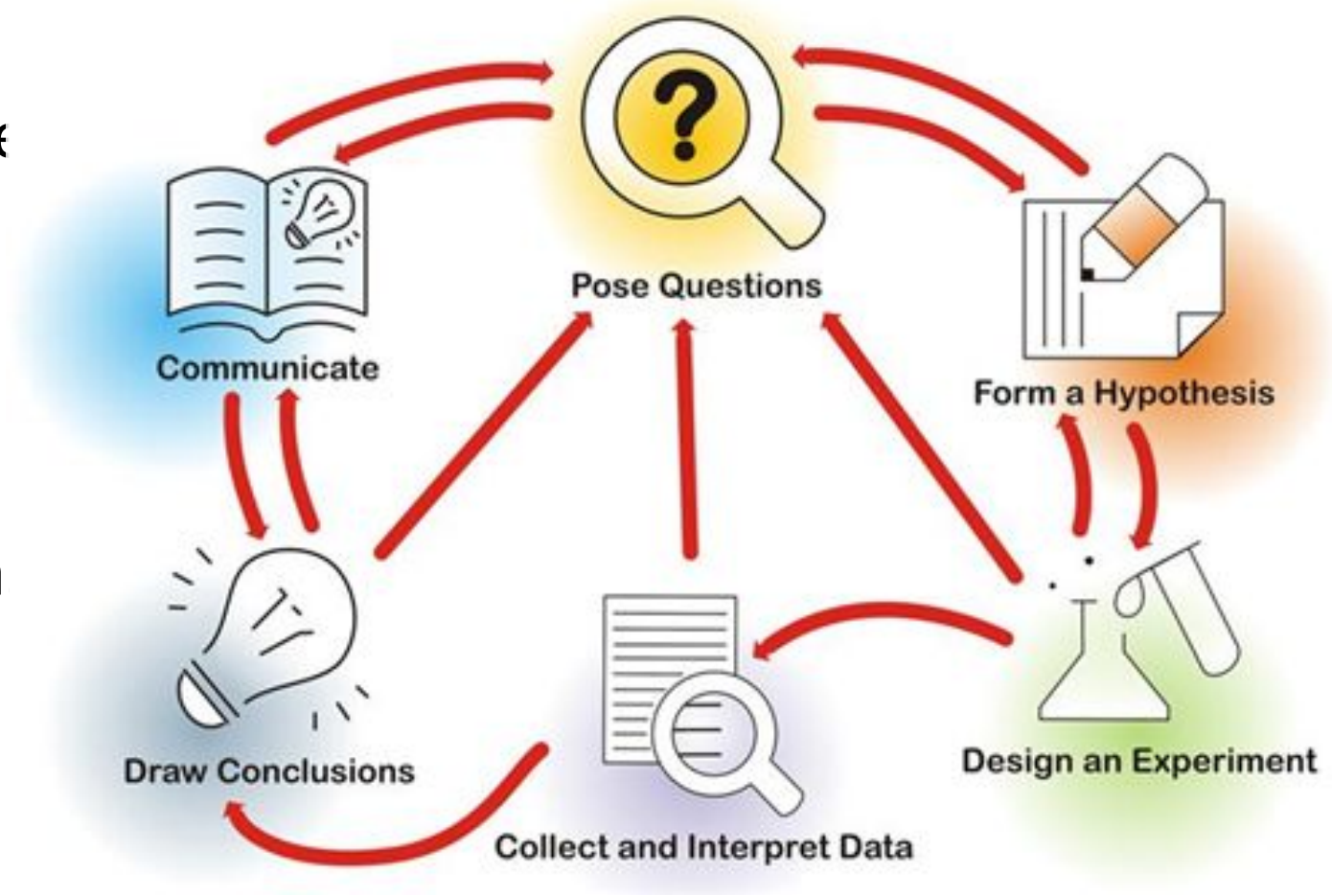
Making information resources accessible, disc
entrepreneurship, innovation, and scientific discovery
contribute significantly to job creation. For example, d
weather data and the Global Positioning System (GPS)
entrepreneurs and innovators have used these resources
and warning systems, location-based applications, prec





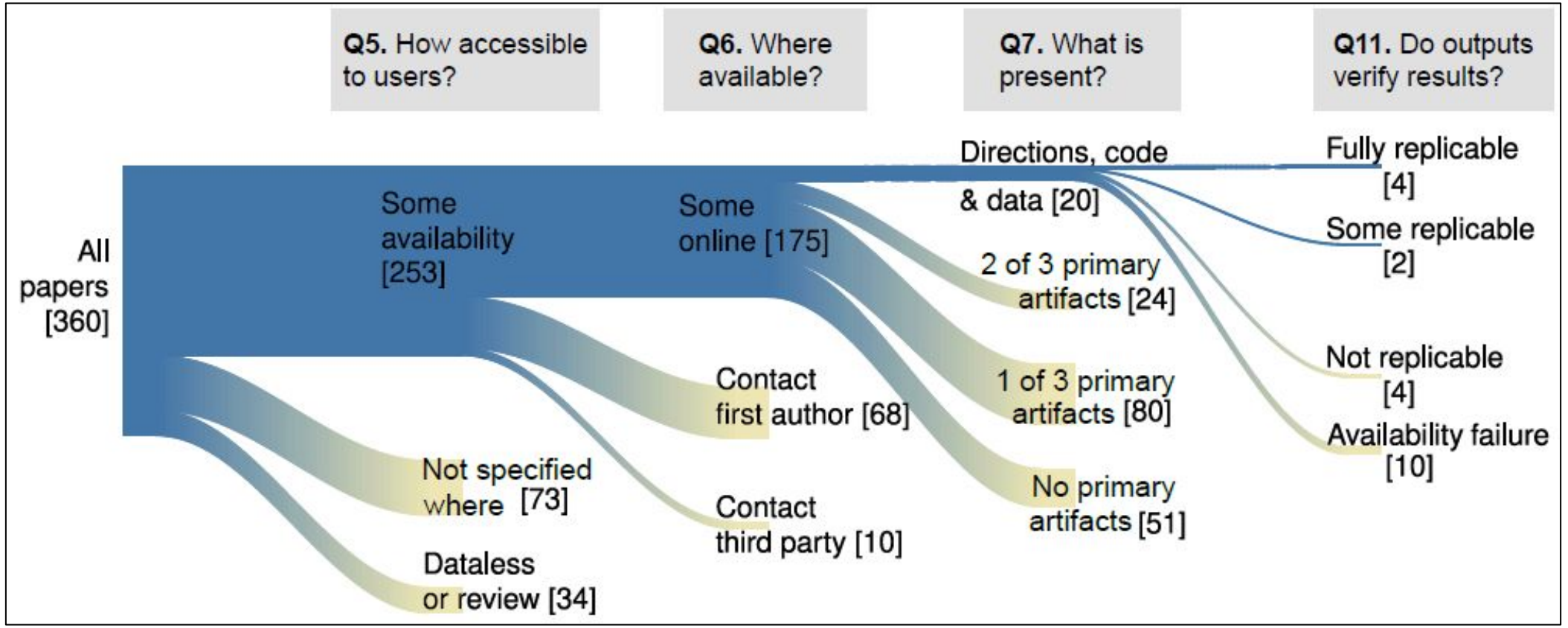
Data Sharing: Reproducibility

- Communicating and sharing data is an essential compone of the scientific process
- If our science is not reproducible, we haven't completed the loop
- The value of data sharing ha not yet been fully realized



Availability and reproducibility of 360 papers in 2017

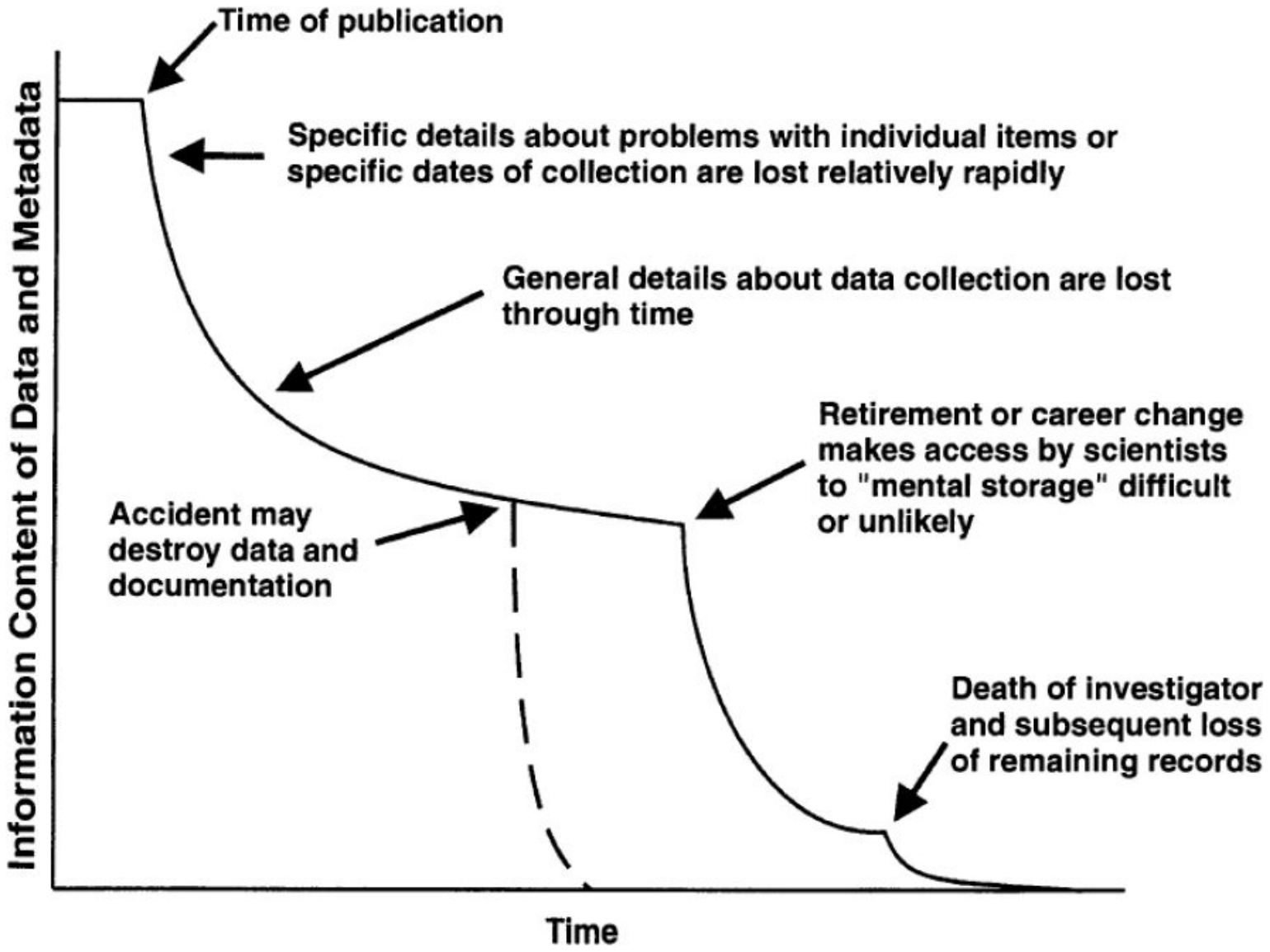
(Stagge et al., 2019 in Nature-*Scientific Data*)



Environmental Modeling & Software Hydrology and Earth System Sciences	Water Resources Research Journal of Hydrology	J. American Water Resources Association J. Water Resources Planning & Management

Stagge, J.H., Rosenberg, D. E., Abdallah, A.M., Akbar, H., Attallah, N.A., James, R. (2019). Assessing data availability and research reproducibility in hydrology and water resources, Scientific Data, 6, 190030, https://doi.org/10.1038/sdata.2019.30

Information Entropy



Example of the degradation of information content associated with data and metadata over time

Michener, W.K. (2006) Meta-information concepts for ecological data management, Ecological Informatics, 1(1):3-7. https://doi.org/10.1016/j.ecoinf.2005.08.004

Information Entropy

"Do not underestimate your ability to forget details about a study!"

Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M. (2009). Some simple guidelines for effective data management. Bulletin of the Ecological Society of America 90:205-214. https://doi.org/10.1890/0012-9623-90.2.205

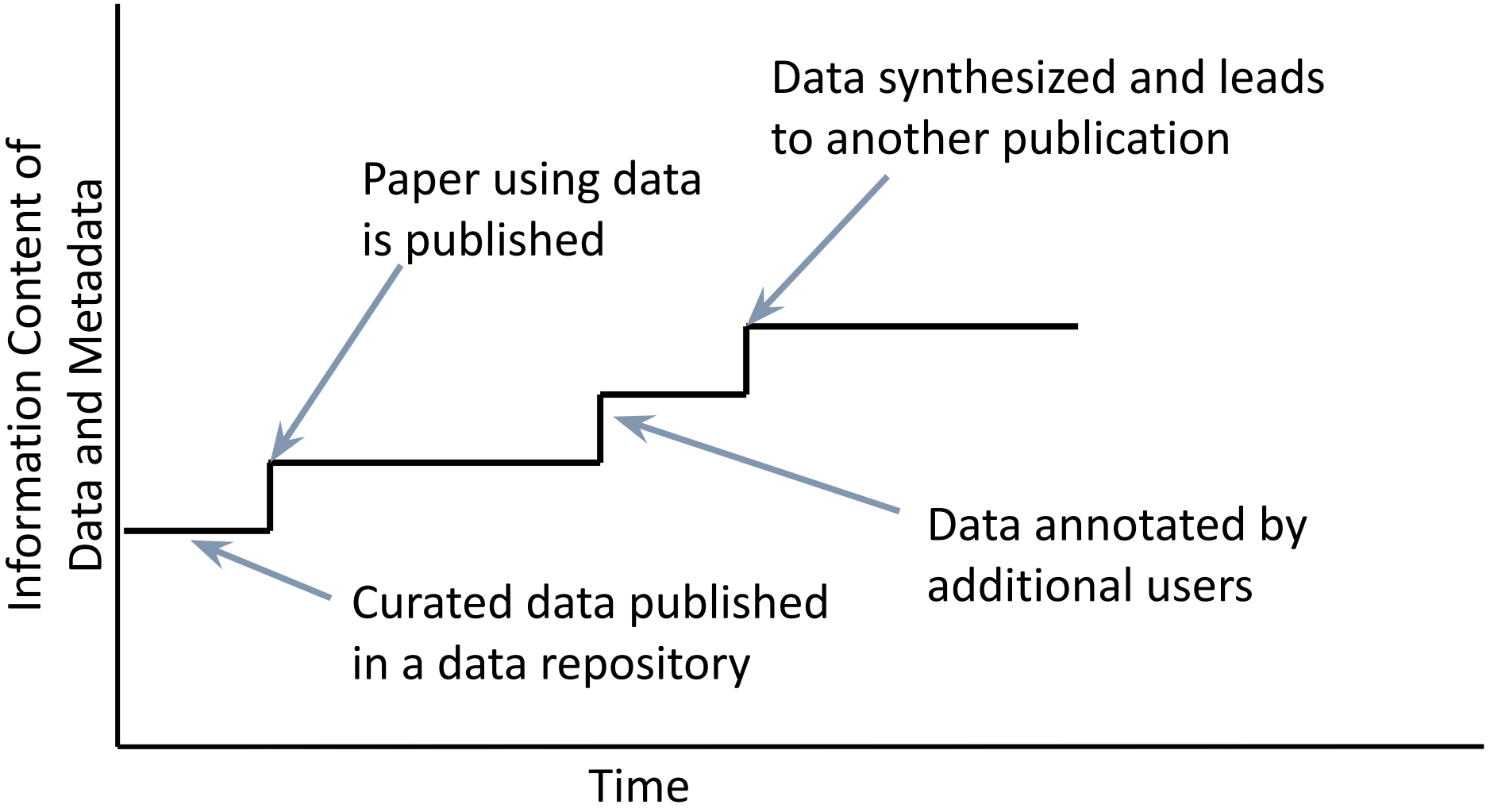
"If the information on an observation is lost, it is lost forever because it is almost impossible to measure the observation again in the original context."

Specht, A., Guru, S., Houghton, L., Keniger, L., Driver, P., Ritchie, E.G., Lai, K., Treloar, A. (2015). Data management challenges in analysis and synthesis in the ecosystem sciences. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2015.03.092

"If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data are interpretable and reusable by others."

Borgman, C.L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6): 1059-1078. https://doi.org/10.1002/asi.22634

Information Entropy: What if instead?

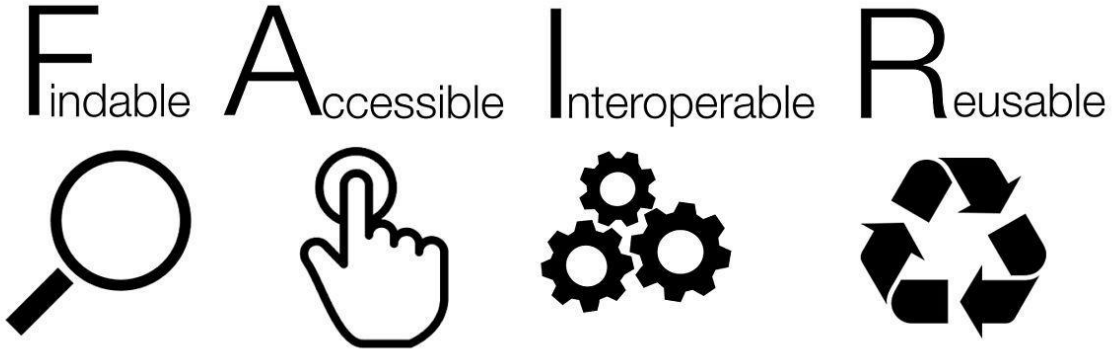


Investigator Data Workflow

- Easily create a digital instance of a dataset or model
- Quickly share it with colleagues (perhaps privately at first)
- Add value through collaboration, annotation, and iteration
- Describe with metadata
- Eventually…share publicly or formally Publish

How do you go about doing this?

What does it mean for data to be FAIR?



A set of 14 guiding principles to make data:

- **Findable**: Data have sufficient metadata and a unique, persistent identifier making data discoverable on the Web
- **Accessible**: Metadata and data are understandable to humans and machines and are available via a trusted repository
- **Interoperable**: Metadata use formal community standards
- **Reusable**: Data have clear metadata, usage license, and information about provenance

The extent to which data are FAIR affects their value and extent of reuse.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, https://doi.org/10.1038/sdata.2016.18.

Some Data Management Resources

- **Guidelines for Structuring and Formatting Data**
  - Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M. (2009). Some simple guidelines for effective data management, Bulletin Ecological Society of America, 90(2), 205-214, https://doi.org/10.1890/0012-9623-90.2.205
  - Broman, K. W., & Woo, K. H. (2018). Data Organization in Spreadsheets. The American Statistician, 72(1), 2–10. https://doi.org/10.1080/00031305.2017.1375989
  - Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., & Slavkovic, A. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Computational Biology, 10(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542
  - Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., Poisot, T., Woo, K. H., Zimmerman, N. B., & Hollister, J. W. (2016). Ten Simple Rules for Digital Data Storage. PLOS Computational Biology, 12(10), e1005097. https://doi.org/10.1371/journal.pcbi.1005097
  - Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), Article 10. https://doi.org/10.18637/jss.v059.i10
- **Guidelines for Citing Data**
  - Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. PLOS ONE, 15(4), e0230416. https://doi.org/10.1371/journal.pone.0230416
- **Guidelines for Making Data More Reusable**
  - White, E., Baldridge, E., Brym, Z., Locey, K., McGlinn, D., & Supp, S. (2013). Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution, 6(2), Article 2. https://doi.org/10.4033/iee.2013.6b.6.f
- **Guidelines for Selecting a Data Repository**
  - Sansone, S.-A., McQuilton, P., Cousijn, H., Cannon, M., et al (2020). Data Repository Selection: Criteria That Matter. Zenodo. https://doi.org/10.5281/zenodo.4084763

Data Management 101

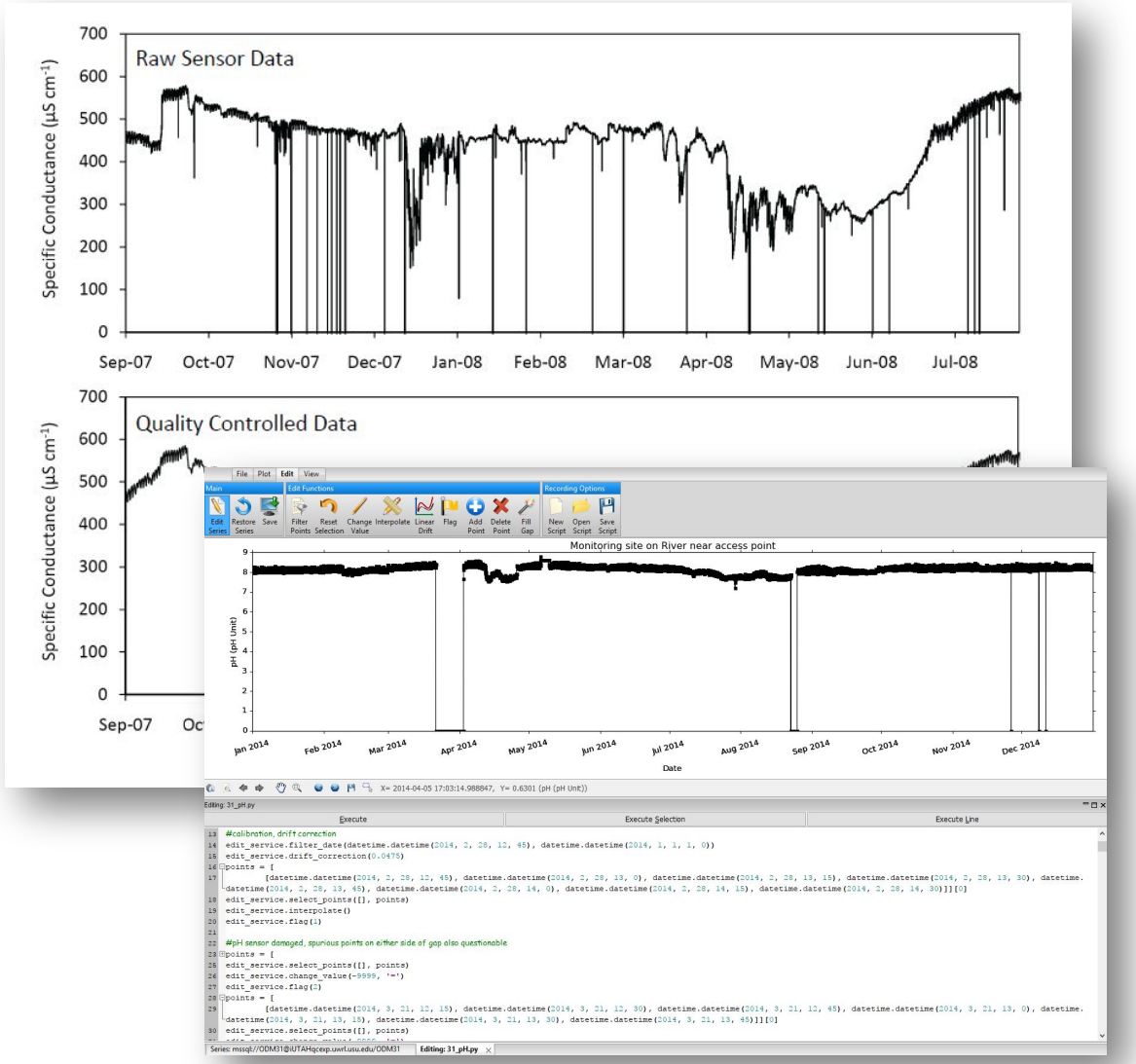
- Simple guidelines to improve data management
- Benefits
  - Improved data organization – facilitates analysis
  - Improved reproducibility
  - Improved capacity for data re-use
  - Facilitates compliance with funding sources and publishers

An oldie – but a goodie!

Borer, E.T., E.W. Seabloom, M.B. Jones, and M. Schildhauer (2009). Some simple guidelines for effective data management, *ESA Bulletin*, 90(2):205-214, https://doi.org/10.1890/0012-9623-90.2.205

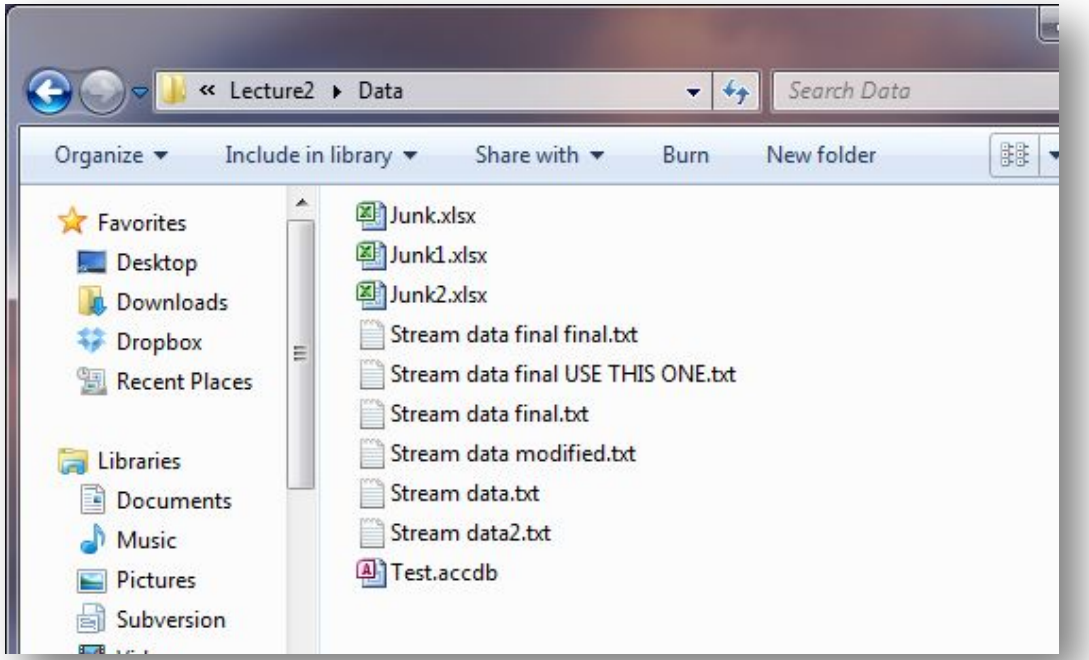
1. Don’t mess with the raw data

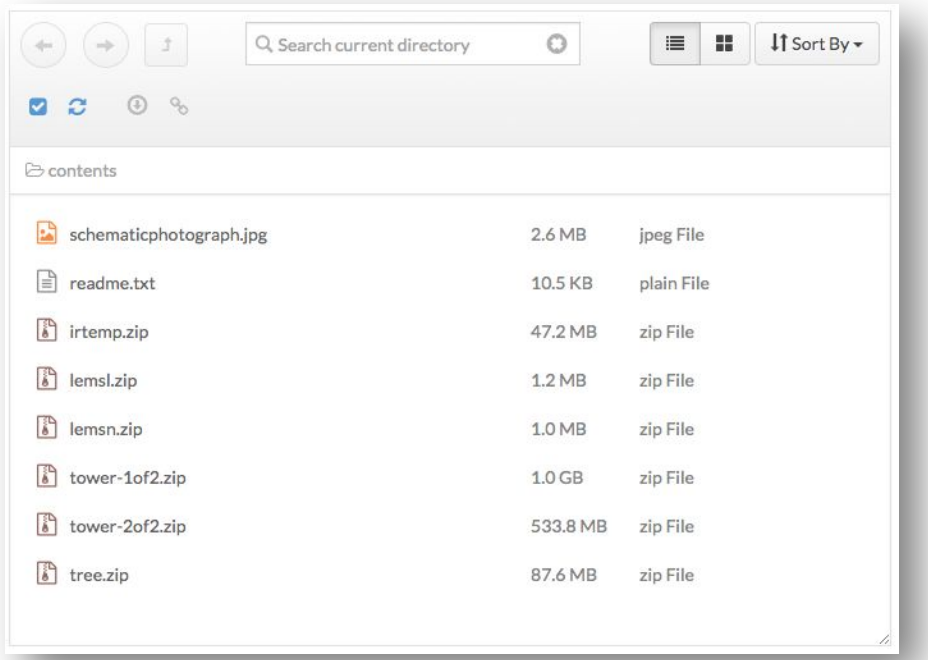
- Always store uncorrected data with all of its “bumps and warts”
  - You could change something that was actually correct
  - You could make mistakes while correcting other mistakes
- Script procedures and write results to a new file/copy of the data



2. Use descriptive file names

- Use only plain ASCII characters and avoid spaces
- Brief, but indicative of content
- Include a “readme” file when using many files in a directory





3. Use descriptive headers in files/tables

- Standard convention for many software applications
- Encapsulate data and descriptive metadata together

```
# ---------------------------------- WARNING ----------------------------------------
# The data you have obtained from this automated U.S. Geological Survey database
# have not received Director's approval and as such are provisional and subject to
# revision.  The data are released on the condition that neither the USGS nor the
# United States Government may be held liable for any damages resulting from its use.
# Additional info: http://waterdata.usgs.gov/nwis/help/?provisional
#
# File-format description:  http://waterdata.usgs.gov/nwis/?tab_delimited_format_info
# Automated-retrieval info: http://waterdata.usgs.gov/nwis/?automated_retrieval_info
#
# Contact:   gs-w_support_nwisweb@usgs.gov
# retrieved: 2012-08-28 12:03:39 EDT       (sdww01)
#
# Data for the following 1 site(s) are contained in this file
#    USGS 10109000 LOGAN RIVER ABOVE STATE DAM, NEAR LOGAN, UT
# -----------------------------------------------------------------------------------
#
# Data provided for site 10109000
#    DD parameter statistic   Description
#    01   00060     00003     Discharge, cubic feet per second (Mean)
#
# Data-value qualification codes included in this output:
#     A  Approved for publication -- Processing and review completed.
#     P  Provisional data subject to revision.
#
agency_cd       site_no datetime        01_00060_00003  01_00060_00003_cd
5s      15s     20d     14n     10s
USGS    10109000        2011-08-28      297     A
USGS    10109000        2011-08-29      294     A
USGS    10109000        2011-08-30      290     A
USGS    10109000        2011-08-31      284     A
USGS    10109000        2011-09-01      281     A
USGS    10109000        2011-09-02      276     A
USGS    10109000        2011-09-03      272     A
USGS    10109000        2011-09-04      270     A
USGS    10109000        2011-09-05      265     A
USGS    10109000        2011-09-06      267     A
USGS    10109000        2011-09-07      262     A
```

3. Use descriptive headers in files/tables

- Standard convention for many software applications
- Encapsulate data and descriptive metadata together

	Box #	Date	ID	Location	R
1	1	6/10-6/17	A1	NHMU	G
1	1	6/10-6/17	A1	NHMU	G
1	1	6/10-6/17	A1	NHMU	G
2	1	6/10-6/17	A2	NHMU	G
3	1	6/10-6/17	A3	NHMU	G
3	1	6/10-6/17	A3	NHMU	G
3	1	6/10-6/17	A3	NHMU	G
4	1	6/10-6/17	A4	NHMU	G
4	1	6/10-6/17	A4	NHMU	G

Site_ID	d18O
SLV-WS-001	-15.57
SLV-WS-003	-15.58
SLV-WS-006	-15.76
SLV-WS-007	-15.59
SLV-WS-008	-15.59
SLV-WS-009	-15.65
SLV-WS-009	-15.53
SLV-WS-010	-14.49
SLV-WS-011	-15.44

```
--------------------------------------------------------------------------------
Tower Data Tables
--------------------------------------------------------------------------------
Data table files prepended with 'Tower' correspond to measurements taken on the sonic tower.  Each file corresponds to a different day of
the year. The tower was located at cooridate position (9.10 m,6.65 m,0).

Sampling Rate: 20 Hz

Equipment:
1. CSAT3: Campbell Scientific CSAT3 three-dimensional sonic anemometer. Measurements taken at heights of 1.5m, 4.0m, 7m, 10m. Manual:
https://s.campbellsci.com/documents/us/manuals/csat3.pdf
2. EC150: Campbell Scientific EC150 open-path gas analyzer. Measurements taken at 4m (co-located with CSAT3). Manual: http://
s.campbellsci.com/documents/us/manuals/ec150.pdf
3. thermocouple: Omega engineering type-E finewire thermocouples.
4. KH20: Campbell Scientific KH20 krypton hygrometer. Measurements taken at 4m (co-located with CSAT3). Manual: https://
s.campbellsci.com/documents/us/manuals/kh20.pdf

Data Table:
Column titles in the data table are listed below along with an explanation of the value
1. year
2. day of year: Julian day of year (Jan 1 = 1, etc.)
3. time: hour and minute of measurement (local time, MDT)
4. seconds: second of measurement
5. KH20 H2O: water vapor concentration from KH20 (grams H20 per meter^3 air)
6. KH20 mV: raw KH20 voltage reading (millivolts)
7,12,17,22. x-wind @ XXm: wind speed from CSAT3 at heights of 1.5m, 4.0m, 7m, 10m in x-dir +North (meters/second)
8,13,18,23. y-wind @ XXm: wind speed from CSAT3 at heights of 1.5m, 4.0m, 7m, 10m in y-dir +East (meters/second)
9,14,19,24. z-wind @ XXm: wind speed from CSAT3 at heights of 1.5m, 4.0m, 7m, 10m in z-dir +up (meters/second)
10,15,20,25. sonic temp @ XXm: air temperature measured by CSAT3 at heights of 1.5m, 4.0m, 7m, 10m (degrees Celcius)
11,16,21,26: sonic diagnostic @ XXm: diagnostic flag from CSAT3 at heights of 1.5m, 4.0m, 7m, 10m
27. CO2 @ 4m: air CO2 concentration measured by EC150 at a height of 4m (milligrams CO2 per meter^3 air)
28. H2O @ 4m: air H2O concentration measured by EC150 at a height of 4m (grams H2O per meter^3 air)
29. gas diagnostic @ 4m: diagnostic flat from EC150 at a height of 4m
30. EC150 temp @ 4m: air temperature measured by EC150 at a height of 4m (degrees Celcius)
31. EC150 pressure @ 4m: air pressure measured by EC150 at a height of 4m (kiloPascals)
32. CO2 signal @ 4m: EC150 CO2 signal strength at a height of 4m
33. H2O signal @ 4m: EC150 H2O signal strength at a height of 4m
34,35,36,37. air temp @ XXm: air temperature from finewire thermocouples at heights of 1.5m, 4.0m, 7m, 10m (degrees Celcius)
```

4. Archive data in non-proprietary formats

- Microsoft Excel is widely available and used now, but what about in 10 years? 20 years?
- How many other software programs can open your data?
- Will your data disappear if the file format/software become obsolete?

What format to use?

- Store it in a file format that can be used by many different software programs
  - Text files – e.g., comma separated values (CSV) for tabular data
- Use a standard file format accepted by your scientific community
- Consider:
  - Format – the file type
  - Syntax – the structure within the file
  - Semantics – the values in the data

Best Practices For Tabular Data

- Each row should have a single observation
- Each column should represent a single variable/attribute
- Every cell should have a single value
- There should be only one column for each type of information
- Do not mix data types within a column
- Use standard formats within cells
  - Be consistent
  - Avoid special characters
  - Avoid using your column delimiter in the data
  - Use standard date formats (e.g., YYYY-MM-DD)
- Use null and NoData values correctly to represent empty or missing data values
- Do not repeat metadata - set up separate table

ID	Site_ID	Name	Latitude	Longitude	City	State_or_Province	Country	ID2	Site_ID	d18O	d2H	d18O_sd	d2H_sd	Type	DEX	Site_ID
SLV-15-125	SLV-WS-001	Maverick	40.54414	-111.87157	Draper	Utah	USA	15-103	SLV-WS-001	-15.57159565	-117.0815005	0.026185373	0.086739176	Tap	7.4912647	SLV-WS-001
SLV-15-129	SLV-WS-003	Seven Eleven	40.5265	-111.87182	Draper	Utah	USA	15-103	SLV-WS-003	-15.58521063	-117.0362963	0.017747305	0.087391515	Tap	7.64538874	SLV-WS-003
SLV-15-133	SLV-WS-006	Prostop	40.49896	-111.88444	Draper	Utah	USA	15-103	SLV-WS-006	-15.76595176	-117.410647	0.026016241	0.061117043	Tap	8.71696708	SLV-WS-006
SLV-15-134	SLV-WS-007	Seven Eleven	40.48592	-111.88457	Draper	Utah	USA	15-103	SLV-WS-007	-15.59920887	-119.1033377	0.028137458	0.093218217	Tap	5.69033326	SLV-WS-007
SLV-15-079	SLV-WS-008	Spring View Farms Trail	40.48642	-111.92639	Bluffdale	Utah	USA	15-103	SLV-WS-008	-15.59811322	-117.2283092	0.028781583	0.17131442	Tap	7.55659656	SLV-WS-008
SLV-15-080	SLV-WS-009	Seven Eleven	40.50505	-111.89738	Draper	Utah	USA	15-103	SLV-WS-009	-15.65822932	-117.0422576	0.014637371	0.064054855	Tap	8.22357696	SLV-WS-009
SLV-15-132	SLV-WS-009	Seven Eleven	40.50505	-111.89738	Draper	Utah	USA	15-103	SLV-WS-009	-15.53472795	-117.1648488	0.03884896	0.206939051	Tap	7.1129748	SLV-WS-009
SLV-15-076	SLV-WS-010	Holiday	40.50714	-111.98578	Riverton	Utah	USA	15-103	SLV-WS-010	-14.49605205	-113.2791354	0.02547515	0.091029684	Tap	2.689281	SLV-WS-010
SLV-15-073	SLV-WS-011	McDonalds	40.50857	-112.01044	Herriman	Utah	USA	15-103	SLV-WS-011	-15.44874428	-119.1508446	0.020055622	0.073675097	Tap	4.43910964	SLV-WS-011

White, E.P, Baldridge, E., Brym, Z.T., Locey, K.J., McGlinn, D.J., and S.R. Supp (2013). Nine simple ways to make it easier to (re)use your data, PeerJ PrePrints, http://dx.doi.org/10.7287/peerj.preprints.7v2

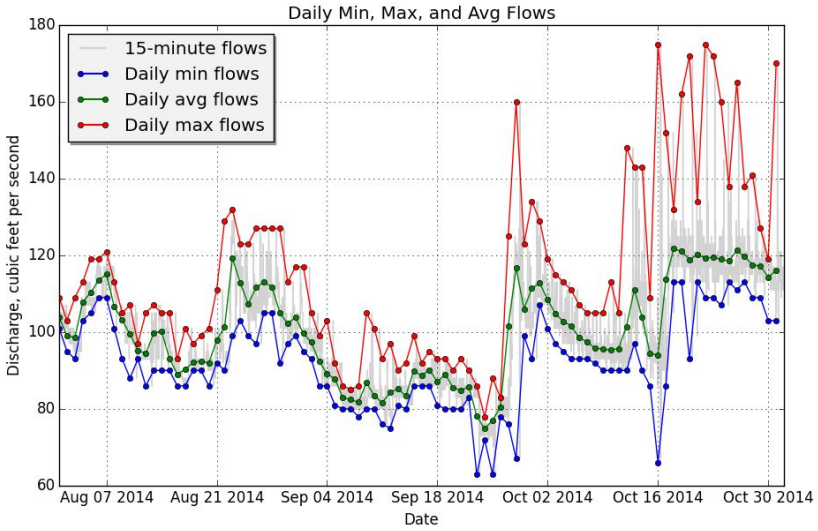
5. Consider data entry

- All data collection involves some data entry
  - Recording observations and notes in a field notebook
  - Transcribing field notebooks and sheets into digital forms
  - Automated processing of sensor data streams into a database
- When you create data entry tools:
  - Use pre-designed forms or templates (electronic or paper)
  - Use lists of valid values rather than free form text entry
    - Example: “Temperature” versus “T”, “Temp”, “Tem”
  - Use validation checks (e.g., range checks)
    - Example: pH must be between 0 and 14. If it’s not – there is a problem!

6. Automate analyses

- Code creates reproducible results
- Code is a record of the steps involved in processing and analyzing data
- Code can be shared
- Code can be re-executed at any time

```
# Create a plot of the streamflow statistics
# ------------------------------------------
fig = plt.figure()
ax = fig.add_subplot(1, 1, 1)
ax.plot(localDateTimes, dataValues, color='lightgrey', linestyle='solid', label='15-minute flows')
ax.plot(dailyFlows[0], dailyFlows[1], color='blue', linestyle='solid', marker='o', markersize=5, label = 'Daily min flows')
ax.plot(dailyFlows[0], dailyFlows[2], color='green', linestyle='solid', marker='o', markersize=5, label = 'Daily avg flows')
ax.plot(dailyFlows[0], dailyFlows[3], color='red', linestyle='solid', marker='o', markersize=5, label = 'Daily max flows')
ax.set_ylabel('Discharge, cubic feet per second')
ax.set_xlabel('Date')
ax.grid(True)
ax.set_title('Daily Min, Max, and Avg Flows')

# Add a legend with some customizations
legend = ax.legend(loc='upper left', shadow=True)

# Create a frame around the legend.
frame = legend.get_frame()
frame.set_facecolor('0.95')

# Set the fontsize in the legend
for label in legend.get_texts():
    label.set_fontsize('large')

for label in legend.get_lines():
    label.set_linewidth(1.5)  # the legend line width

fig.tight_layout()
plt.show()
```



The initial investment to learn pays off later!

7. Consider storage media

- CDs?
- DVDs?
- External hard drives?
- Don't strand your data!!!

Borer et al.: "*As hard as it is to believe today, we can foresee the day when CD-ROMs might be difficult to read.*"

2000

2000

1976

1985

1986

1994

1995

Preservation/Backup Media

How are you preserving your data now?

- Does your office look like this?
- What are the potential problems?
- What are some potential solutions?

Backups

1. Have a plan
2. Stick to your plan
3. **Test your plan** – preferably under non-failure conditions!
4. Betty Rozum's 3-2-1 rule
   - 3 copies
   - 2 types of storage media (e.g., cloud, hard drive)
   - 1 copy offsite
5. Consider hard copy data – lab notebooks, research notes, field notes, etc.

- Don't trust/rely on hardware redundancy

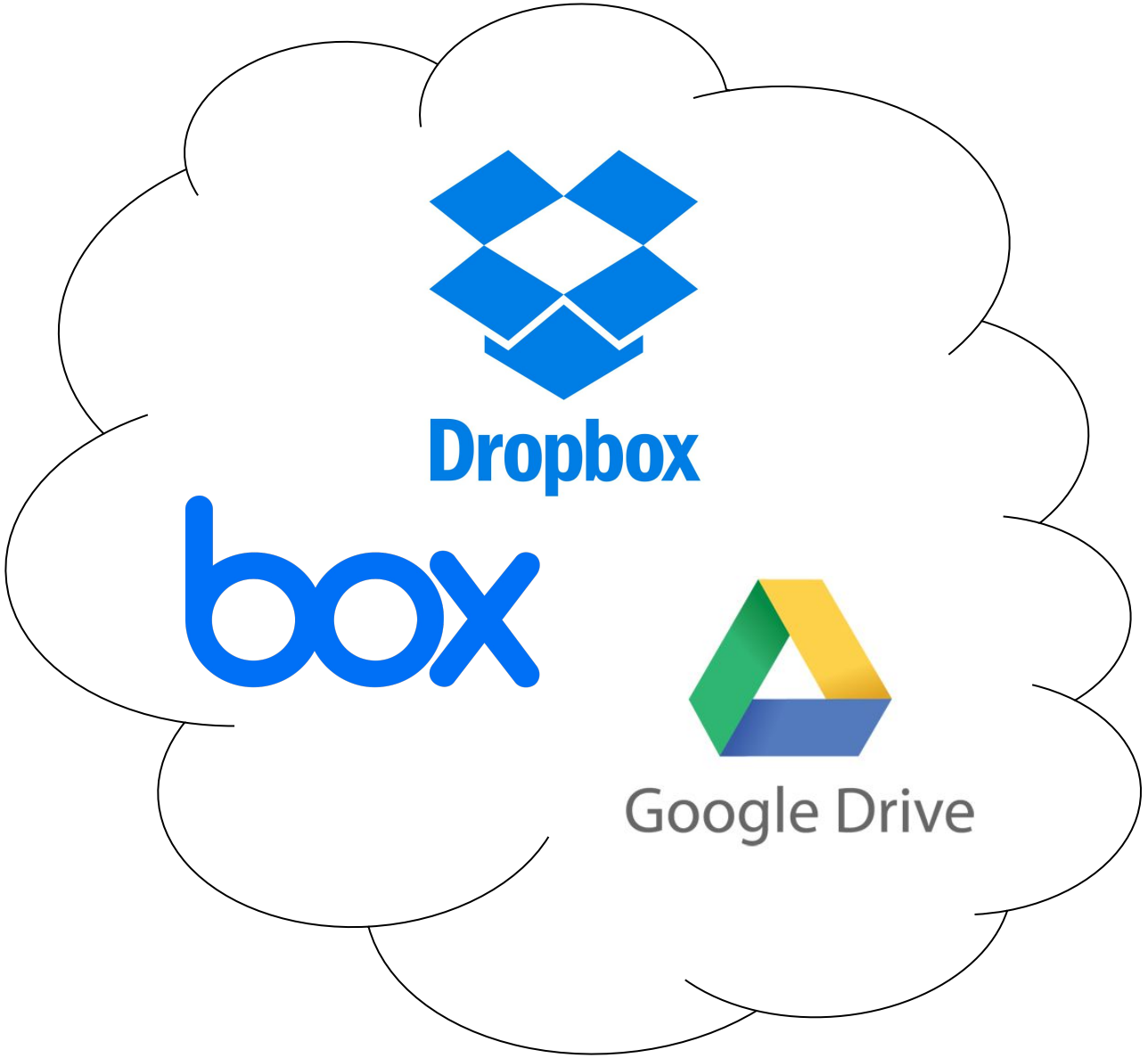
Plan to Preserve

- What will be preserved?
- Where will it be preserved?
- Back ups?
- Version control?
- Policies for access, sharing, and reuse
  - Obligations for sharing
  - Security and access control
  - Sensitive data
  - How long?
  - Intellectual property issues
  - Responsible parties

To the Cloud!

- Convenience
- Accessibility anywhere
- Cross platform
- Enhanced sharing
- Low cost

- But…
  - Privacy???????
  - Delay (slow or non-existent internet)
  - Storage, but not much else
  - File formats and semantics still matter



Better Opportunities for Data Sharing and Preservation

- Data archives/repositories
- Functionality for collaboration and archival/preservation
- Still very much discipline specific
- Impact is higher if you choose carefully!

DataONE

GitHub

PANGAEA.
Data Publisher for Earth & Environmental Science

DRYAD

knb

figshare
credit for all your research

HYDROSHARE

CUAHSI
HIS
Sharing hydrologic data

NSIDC
National Snow and Ice Data Center

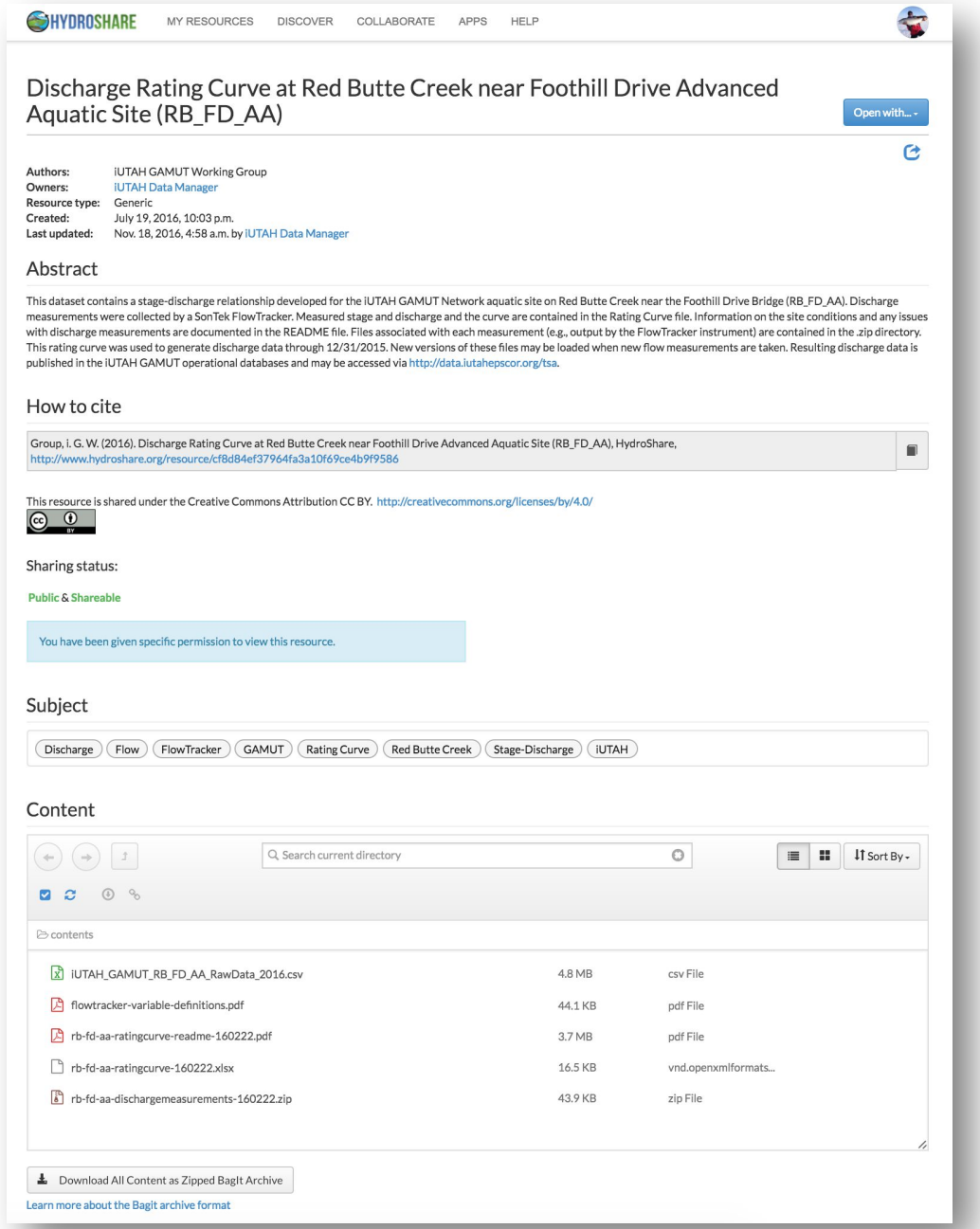
Which Repository to Use?

- For CZ Net – we are recommending some repositories

- For others:
  - Does your research sponsor stipulate a repository?
  - Does your scientific discipline have a repository?
    - Disciplinary data is best stored together where researchers are likely to find it
  - Code – many use GitHub (and there are ways to make your code citable)

Repository Challenges

- How to best use the repository?
- How to avoid data misinterpretation?
- Size limitations for individual users
- Level of required metadata – where to set the bar?
- Dataset review
- Ongoing curation



8. Maintain effective metadata

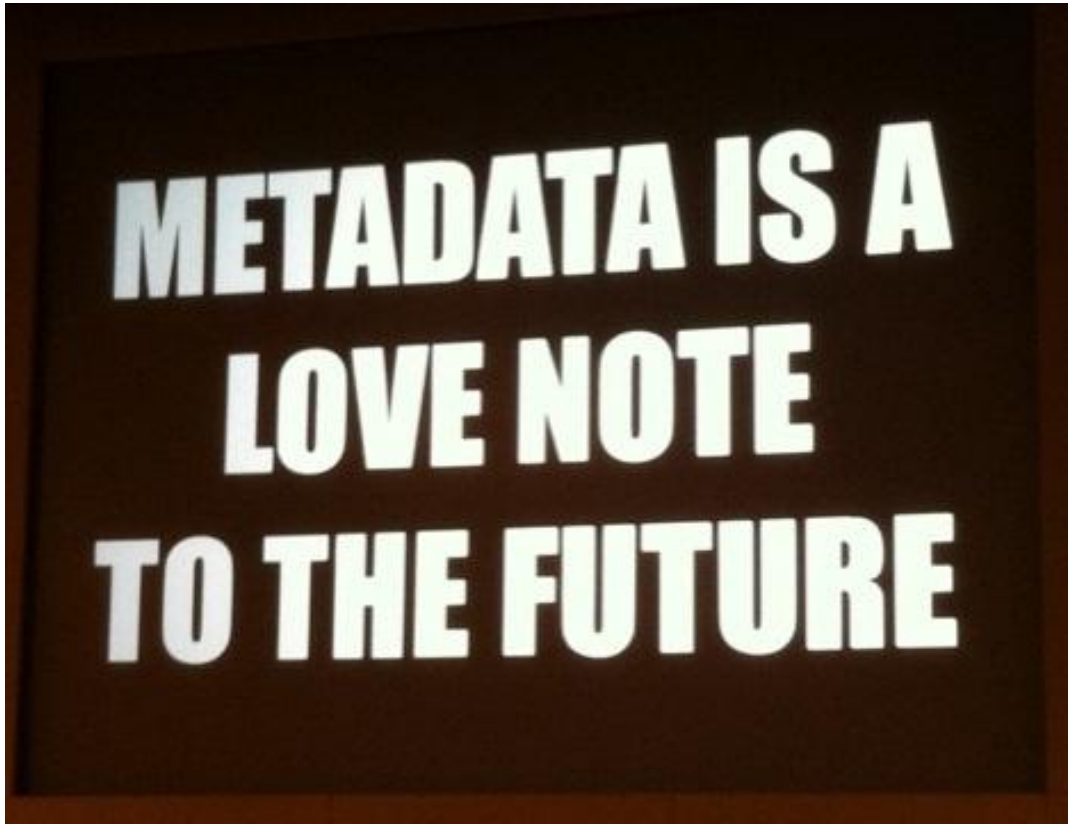
- Borer et al.: ***“Do not underestimate your ability to forget details about a study!”***

  - When did the tree that was stuck in my stream cross section get removed????
  - You may not analyze your data until years down the road
  - Exact details of methods, names, files, etc. will become fuzzy

What is Metadata?

- **Metadata is "Information about Data"**
  - **WHO** created the data?
  - **WHAT** is the content of the data?
  - **WHEN** were the data created?
  - **WHERE** is it geographically?
  - **WHY** were the data developed?
  - **HOW** were the data developed?

Content, quality, condition, and other characteristics



Metadata Content and Format

- What metadata are needed?
  - Details that make data findable
  - Details that make data meaningful
- How will metadata be created?
  - Lab notebooks?
  - Curation by you or a data manager?
  - Automatically generated by a sensor or instrument?
- What format will be used for the metadata?
  - Standards may be chosen by community or dictated by an agency
  - May be dictated by the repository into which you deposit the data

Metadata Extend to Samples for Reuse & Reproducible Science (upcoming webinars!)

- Use of globally unique and persistent identifiers for samples
  - to support unambiguous citation
  - to support linking of information in distributed data systems and with publications
- Standards for metadata to document the diverse range of samples and collections to make sample Findable & Accessible
- Best practices for sample and collection cataloguing, including a broad range of issues from interoperability to persistence of catalogues
- access policies for sample metadata & samples/specimens

Slide from Kerstin Lehnert

Sharing Data: The Golden Rule

- When you ***provide*** data to someone else, what types of information are they going to need to understand the data?

- When you ***receive*** a dataset from an external source, what types of details do you want to know about the data?

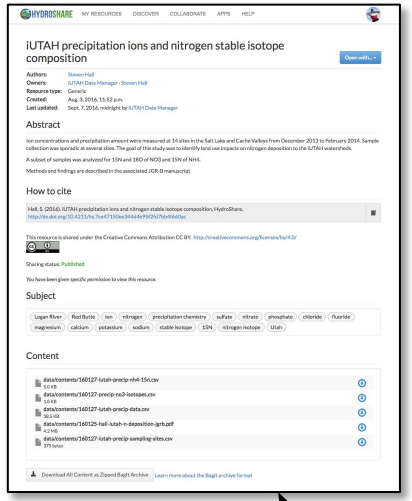
Sharing Data

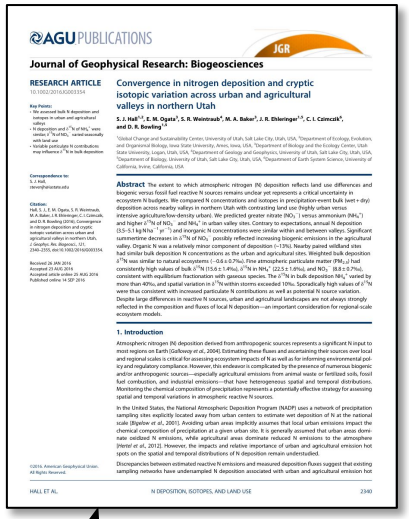
- **Providing data**:
  - Why were the data created?
  - What limitations do the data have?
  - What does the data mean?
  - How should the data be cited if it is re-used in a new study?

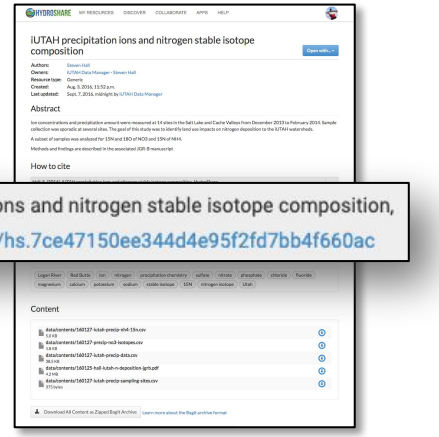
- **Receiving data**:
  - What are the data gaps?
  - What processes were used for creating the data?
  - Are there any fees associated with the data?
  - In what scale were the data created?
  - What do the values in the tables mean?
  - What software do I need in order to read the data?
  - What projection are the data in?
  - Can I give these data to someone else?

Data Sharing: Planning

- Address questions related to timing, organization, and authorship up front
- Considerations:
  - Which products will be generated and shared?
  - Are there sensitivities around any of the products?
  - How will data/files be organized?
  - Who is responsible for data/metadata creation and curation?







Hall, S. (2016). iUTAH precipitation ions and nitrogen stable isotope composition, HydroShare, https://doi.org/10.4211/hs.7ce47150ee344d4e95f2fd7bb4f660ac

Steven verified his data and metadata were correct but kept the data private

Steven submitted his paper for publication and responded to reviews

Steven published his data in HydroShare and received a DOI

With a little help, Steven deposited his dataset in the online HydroShare repository

Steven collected his data in the field and transformed into a sharable format

Steven published his paper and cited published data in HydroShare

Acknowledgments

Upon manuscript acceptance all data are publicly available online at the Hydroshare database: http://dx.doi.org/10.4211/hs.7ce47150ee344-d4e95f2fd7bb4f660ac. We thank four

The Steven Hall Story

Summary

1. Don’t mess with the raw data
2. Use descriptive file names
3. Use descriptive file headers
4. Archive data in non-proprietary data formats
5. Consider data entry
6. Automate analyses
7. Consider storage media
8. Maintain effective metadata

Following these steps help prepare datasets for deposit and sharing in a repository

Questions?

Contact us:

cznet@cuahsi.org