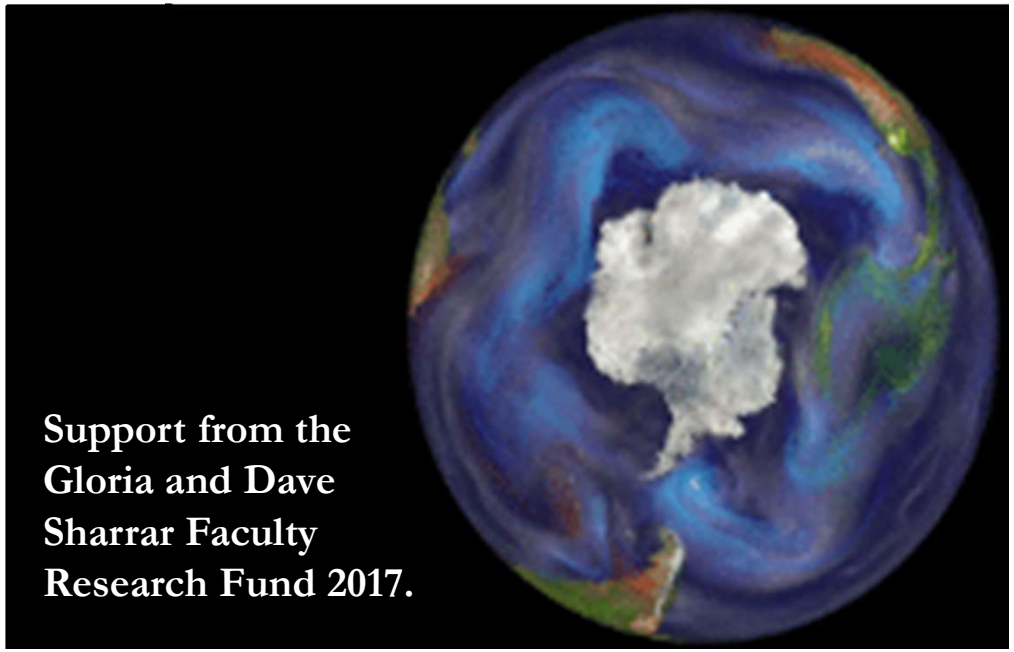
THE UNIVERSITY of TENNESSEE UT KNOXVILLE

BIG ORANGE BIG IDEAS

The Data Life Aquatic: Oceanographers' Experience with Interoperability and Re-usability

Wade Bishop (presenting),
Carolyn Hank & Joel Webster
School of Information Sciences
University of Tennessee

February 5, 2019
14th IDCC
University of Melbourne



Support from the Gloria and Dave Sharrar Faculty Research Fund 2017.

Assessing re-usability with FAIR

- Whether the re-user will be human or machine, the judge for data's re-usability will ultimately be that re-user conducting the analysis, making discoveries, and generating new data.
- A re-user's perspective could outline considerations for the functionality and design of data/metadata, and also the tools used to locate, access, and re-use.
- Operationalizing **re-use** is required for assessment and one way this can be done is speaking with actual re-users.
- The purpose of this study is to better understand how re-users discover and evaluate data.
- This perspective likely differs from other aspects that make data "curatable" and/or some machine-readable aspects.

Fairly sizeable list of derivative puns

- Several original FAIR Data Principle authors formed a FAIR Metric group to evaluate claims from many repositories and resources that they were already "FAIR."
- The group conducted focus groups to assess their metric guidelines addresses their principles, but found that not every metric, or even the Principles, were always understood how intended and published a response to clarify their principles (Wilkinson et al, 2018).
- Still, others in Europe, the United States, and beyond like GOFAIR, the Enabling FAIR Data Project, and others took the overall FAIR framework and translated the original principles to serve their own purposes.

Wilkinson, M. D. *et al.* (2018). A design framework and exemplar metrics for FAIRness. *Sci. Data* 5:180118 doi: 10.1038/sdata.2018.118

Oceanographic data

- Data collection is often in real-time, collected once in a snapshot or streaming continuously through sensors, and across broad geographic areas.
- When one gathers seafloor sediment, it is only later when researchers inland are connected to high and dry performance computing that the data are analysed.
- There is the implicit value of the scientific enterprise itself (e.g., contribute to a better understanding of the oceans), but also the value of knowing the contents of the exclusive economic zone (EEZ).
- The EEZ is each country's jurisdiction of the seafloor and ownership of the natural resources beneath the oceans in which to manage, conserve, explore, and exploit (NOAA, 2017).

Creating FAIR questions

- The interview questions were derived from the FAIR Data Principles and more details on that are in this paper:

Bishop, B. W. & Hank, C. F. (2018). Measuring FAIR principles to inform fitness for use. *International Journal of Digital Curation*, *13*(1). DOI: https://doi.org/10.2218/ijdc.v13i1.630

Recruitment

- Participants were recruited by contacting re-users of data in the Coastal and Marine Geology Program at two U.S. Geological Survey (USGS) Coastal and Marine Science Centers:
  - Pacific and Woods Hole
- Using a critical incident technique, ten oceanographers were asked to describe their most recent search for data.
  - NOAA-Marine (10)

Occupation and Education

1. What is your current job title?
2. How many years in total have you been working in your current job?
3. How many years in total have you been working with earth science data?
4. Describe your work setting.
5. Please indicate your credentials and degrees.
6. Please provide any other educational or training you have received that is applicable to performing your job.

What is your current job title?

- Half (n=5) identified their job title as Oceanographer or Research Oceanographer, and two as Geologists—with a specialization of the seafloor.
- For the remaining participants, one title reflects managerial responsibilities (Deputy Regional Manager), while two appear to indicate data management specific roles: Scientific Programmer and a Metadata Management Architect.

Years in current job and working with earth science data

- The average years spent working with science data, including all time in higher education, was almost 22 years.
- Participants' time in their current positions varied from 2.5 years to 30 years, with about 13 years being the average.
- These participants' expertise in locating science data through changes in data formats and information systems was apparent in their responses and detailed descriptions of how they locate and evaluate data.

Work Setting

- A few participants referred to field work, such as boats, cameras, and scuba gear.
- Still, the majority referred to the hardware and software used to analyse science data.
- Participants referred to their hardware as “heavy duty processing machines” and all types of computers, from laptops to clusters that access high performance computing to the cloud (e.g., Amazon Web Services), to conduct simulations and run models. The most mentioned “tools” were MATLAB, Python, and ArcGIS.

Education and Training

- Six of the ten hold PhDs, with the remainder having master's degrees.
- All but one were in the sciences; the outlier held a Master's in Art.
- Although participants were all asked about additional training they received to do their jobs, only two gave specific examples: MATLAB and ESRI workshops.
- Most participants indicated they were self-taught, and nearly all mentioned gaining new skills and knowledge from self-directed searches online (e.g., YouTube videos).
- The lack of formal training in data science and data curation for these scientists whose primary responsibilities are related to those tasks is a challenge found in many domains.

Method

- Phone interviews were conducted.
- The interviews were recorded and transcribed.
- The transcriptions were analyzed using NVivo.
- Grounded theory application of open, axial, and selective coding generated the following categories and broad themes across responses to the questions.

Critical incident prompt

- **Think of a recent search for data (or more). The following questions will determine how you discovered and evaluated that data for *fitness for use*.**

Interoperability

7. Was the data in a useable format?
8. How was the data encoded and was it using encoding common to other data used in your research (i.e., same format))?
9. Was the data using shared controlled vocabularies, data dictionaries, and/or other common ontologies?
10. Was the data machine-actionable (e.g., to be processed without humans)?

To be interoperable	**I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.** **I2. (meta)data use vocabularies that follow FAIR principles.** **I3. (meta)data include qualified references to other (meta)data.**

Usable format?

- Eight participants indicated that the data they located was in a useable format.
- Still, two were not usable without transposing what they found.
- "We've obtained is an NetCDF 3, let's say. So, we have to use an internal software, which is free, and it mods it to NetCDF 4 and that's about it."
- These additional steps to make data interoperable are logical models that could be built into machines to make data in similar transposable formats machine-actionable.
- The other participant with useable format issues was working with data presented in PDF, a bathymetry (the ocean floor) sheet map that was not actually encoded in something that could be easily made interoperable.
- Thus, much of the legacy data in oceanography will still require humans to transpose this invaluable dated data.

Common encoding?

- Nine of the participants indicated that the data was in a common encoding standard.
- Five indicated the encoding was NetCDF, two others citing text files, one used .mat, one GRIB 2 (i.e., gridded bathymetry), and one indicating a Shapefile.
- One participant said the data portal served up the data in any format they might need, serving up automatic translations, and was not specific to any particular format.
- This customizable system that serves up data in multiple common encodings solves many of the potential interoperability issues faced in re-use of data. “You have similar options for when you download these, you can bring it as text, as a list, CSV, or Excel, or whatever you want.”

Controlled vocabularies?

- Seven participants indicated controlled vocabularies, data dictionaries, and/or common ontologies were used,
- Still, three others did not know how values in their data were categorized.
- The Global Change Master Directory (GCMD) keywords and Southern California Coastal Water Research Project (SCCWRP) were named specifically, but marine sciences with fewer political boundaries and fewer variables does have well-established metadata standards to the point of invisibility to end users.
- For the participants that did not know if they were using a controlled vocabulary, it seemed as if their interpretation of using keywords from a thesauri was in fact a controlled vocabulary and the terminology of the question should be revised for each discipline.

Machine-actionable?

- The responses varied also for machine-actionable, as some data are not ready for processing without some human intervention.
- Seven were confident the data they re-used was machine-actionable, but three had issues that would be a barrier for machine-to-machine re-use.

Re-usability

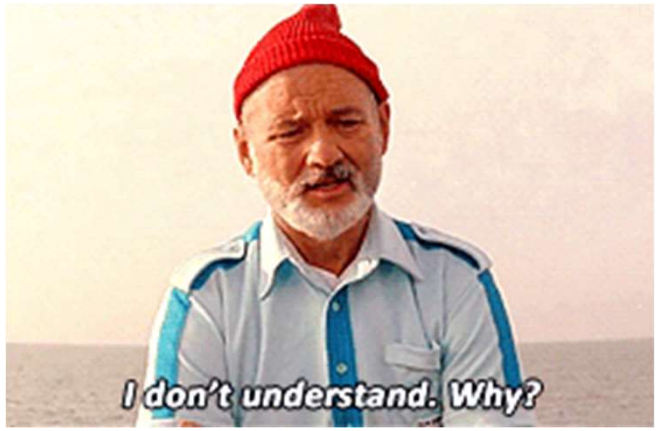
11. Were there any issues with the data that impacted reuse of the data (e.g., resolution)?
12. Did the data geographic scale or resolutions impact reuse of the data?
13. Did the coordinate systems used impact reuse of the data?
14. Did the metadata provide sufficient information for data reuse?

To be re-usable	**R1. meta(data) have a plurality of accurate and relevant attributes.** **R1.1. (meta)data are released with a clear and accessible data usage license.** **R1.2. (meta)data are associated with their provenance.** **R1.3. (meta)data meet domain-relevant community standards.**

Re-usabilty

- No data are perfect.
- When using data for purposes beyond its original collection and intended primary use, the imprecision, imperfect granularity, and versioning proliferation forced participants to accept and work with these challenges to re-use because the data are unique, valuable, and alternatives do not abound.

Re-use issues



- Although three participants indicated no issues with re-use, seven had challenges.
- The most common data issue is the lack of version controls. If errors occur in data collection (e.g., buoy drift), the data is only issued in a corrected version overwriting the old data.
- "Sometimes the data does change from month to month, in other words the data you extract one month might be different the next month and we don't always know when that remote dataset has been changed or why."
- This re-use issue of lack of documentation in large datasets is known, but no easy solution exists.

Known issues with precision and licensing

- Prior to GPS much of the legacy data has limitations because a reduced accuracy in navigation results in a lack of precision of that data.
- Humans may understand this better than machines without the context to introduce doubt into oceanographic data prior to the mid-90s.
- Humans face license agreements differently than is possible for machines.
- A human may contact a vendor and clarify licensing issues that may be unclear, where a machine would not process data with licensing limitations at all.

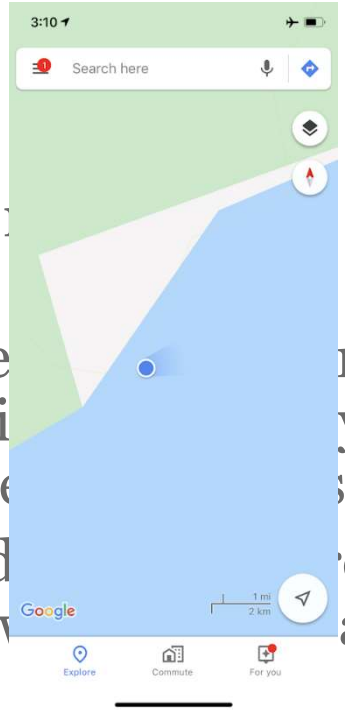
Remember the data are worth something

- In the US, the various agreements between federal agencies, diffusion of proprietary formats, and grey areas in data sharing present challenges for any re-use but one particularly rough for machines to sail through.
- "Like I look at shorelines, so I can share those shorelines that I generated from it, but actually sharing the imagery, I can't do that."

Geographic scale?

- Seven of the participants indicated that geographic scale did not impact re-use of the data.
- This is in part due to a need for points of data at specific locations where larger scale work with more data points across broader regions would.
- "You know gauge data is very site specific, so it doesn't really, you can't really extrapolate over a larger spatial area, so it is limited to a point in space."
- The three participants that indicated scale mattered were working with larger extents of the ocean.

Coordinate system?



- Similarly, eight participants also did
  coordinate system influenced re-use.
- Like scale, coordinate systems can be rted to others—meaning that data collected i ys can all be re-projected into the same syste ses.
- At least one participant said "we kind ed that. Because many times we couldn't tell ate systems was used."
- Scale and coordinate systems do influence findings, human and machine re-users may overlook the importance of these details and unknowingly alter findings, and a few hundred meters in the wrong direction along shorelines could be a big mistake.

Sufficient metadata?

- Finally, nine of the ten participants [illegible] metadata provided sufficient inforn [illegible] e-use.
- "I mean metadata is harder than da [illegible] nes."
- One participant was more elaborati [illegible] for the last 20-25% of the assessment…the [illegible] not be enough, you might have to actually get it and try it out yourself or send a query and ask somebody."
- This point is key because a machine might not readily know to ask someone or how to assess data once analysed and queried and building in those aspect of quality control is an additional step to making data machine-actionable.
- So, the answer to sufficient metadata might be yes for a human, but indeed no for a machine.

Interoperability & Re-usabilty discussion

- Many other data-curation tools, services, and guidance documentation inspired by the FAIR tsunami exists.
- Typically, reporting on re-users and re-use scenarios is hypothetical and abstract, rather than resulting from actual, direct data collected from real-world re-users or re-use cases.
- There are practical reasons for this discrepancy, data creation and curation practices and implications need to be well understood in order for discovery and access to get underway.

Conclusion

- Ultimately, the data is for re-use and re-users.
- This study is intended to contribute to this vital though less frequently investigated group of data curation stakeholders and in the domain of oceanography.
- In conclusion, despite a big push to enable machine-actionable data and many successes with artificial intelligence in data analyses, certain aspects of preparing, processing, and legally sharing data require human input for the foreseeable horizon.

Wilkinson, M. D., Verborgh, R., Bonino da Silva Santos, L. O., Clark, Tim, Swertz, Morris A., Kelpin, Fleur D. L., . . . Dumontier, Michel. (2017). Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science, 3*, e110. doi: 10.7717/peerj-cs.110

GI: Organization, Access, and Use

- Bishop, B.W. & Grubesic, T. H. (2016). *Geographic Information: Organization, Access, and Use of Geographic Information*. Springer.