iUTAH Data Management Update

Amber Spackman Jones
Jeffery S. Horsburgh
iUTAH Summer Symposium
7-17-2015
InformaDon
Entropy

Example of the normal degradation in information content associated with data and metadata over time ("information entropy").
Information Entropy

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

“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.”

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.
What if instead?

Curated data published in a data repository

Data synthesized and leads to another publication

Data annotated by additional users

Paper using data is published

Time

Information Content of Data and Metadata
iUTAH Data Policy

- Applies to all datasets created or developed with support from iUTAH

- Recommended reading for all investigators. Available at: http://iutahepscor.org/resources/documents/iUTAH_Data_Policy.pdf

- In general, open data policy to maximize the impact and broad use of datasets collected by iUTAH research teams

- Researchers should have expectation of first rights to analyze and publish data

- Researchers expected to provide high quality datasets with sufficient metadata

- Data products should to be considered at outset of data collection
iUTAH Data Policy: Data Typology

A. Primary iUTAH datasets and research products (e.g., raw and QAQC sensor data, baseline sampling datasets, general iUTAH datasets for the iUTAH community).

B. Support from iUTAH provided, but created by a specific investigator or group to support particular research question/goal.

C. Types A and B but subject to IRB restrictions.

D. Proprietary data that may be subject to licensing, copyright, other restrictions.

---

Timeframes

- Published as soon as results are created
- Finalized data submitted within one year of completion of data collection activities
- Same timeframes as A and B, but may require anonymization
Publication:
- Assigns a citation and a URL
- Data and metadata are discoverable
- Datasets are archived and curated
iUTAH Data Policy: Data Collection Plans

• **ALL** data creation efforts with **ANY** funding from iUTAH (salary, travel, sampling, equipment, etc.) **MUST** submit a brief plan to the Data Policy Committee **PRIOR** to funding.

• Plan should include:
  1. Identification of types of data to be collected/created
  2. Brief description of methods, data formats, and data products
  3. Timeline for data generation and expected publication
  4. Identification of who will have access to preliminary data during collection
  5. Identification of limits to access
  6. Information on collaborators/co-authors of data products or publications
Data Publication System

http://repository.iutahepscor.org

- Web-based system for iUTAH researchers to submit and publish data and models.
- System supports curation of datasets.
- Integrates the submission and presentation of data and metadata.
- Supports discovery and access of datasets to a wide audience.
- Supports storage and archival.
- Datasets are private until approved by a moderator.
Data Publication System
http://repository.iutahepscor.org

- Organized into datasets. Each dataset consists of multiple resources.
- Supports submission of metadata-only record.
- General level metadata. More specific metadata may be submitted as a resource.
- Metadata records submitted to provide insight into what work iUTAH participants are conducting.
Data Publication Tutorials
http://iutahepscor.org/data_modeling.shtml
http://data.iutahepscor.org/mdf/About/Training_Materials/

• Developed short videos to provide training on aspects of data publication.
• Four videos:
  – Overview
  – Data Collection Plans
  – Data Publication System: Search and Discovery
  – Data Publication System: Submittal and Publication
Data Submission Best Practices

Use Descriptive File Names

- Use only plain ASCII characters
- Brief, but descriptive of content
- Include a “readme” file when using many files

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?
Data Submission Best Practices

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 both format and syntax (e.g., the structure within the file)
Energy Balance of and Isolated Urban Tree

This dataset contains energy-related measurements in and around a relatively isolated urban tree on the University of Utah campus. Several measurement techniques are used to describe the spatial distribution of surface temperature, and fluxes of sensible heat, latent heat, radiation, and moisture. The tree is a Freeman

Isolate Tree (isTree) Dataset Metadata

Contact:
Brian Bailey
bbaileyeng.uta.edu

Coordinate System:
y = North
x = East
z = upward
(a,y,z) = (0,0,0) (tree trunk, tree trunk, ground @ tree)

Data table files preprended 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 coordinate position (9.10 m, 6.65 m, 0).

Sampling Rate: 20 Hz

Data Table:

| Column Titles in the data file 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. w20 wind speed from C3AT3 at heights of 1.5m, 4.0m, 7m, 10m (in-Dir) in meters/second | 6. w20 wind speed from C3AT3 at heights of 1.5m, 4.0m, 7m, 10m (in-dir) in meters/second |
| 7. h20 water vapor concentration from RH20 (grams H2O per meter^3) air | 8. h20 water vapor concentration from RH20 (grams H2O per meter^3) air |
| 9. T20 air temperature measured by EC150 at a height of 4m (degrees Celsius) | 10. T20 air temperature measured by EC150 at a height of 4m (degrees Celsius) |
| 11. CO2 CO2 concentration measured by EC150 at a height of 4m (milligrams CO2 per meter^3) air | 12. CO2 CO2 concentration measured by EC150 at a height of 4m (milligrams CO2 per meter^3) air |
| 13. solar radiation measured by EC150 at a height of 4m (degrees Celsius) | 14. solar radiation measured by EC150 at a height of 4m (degrees Celsius) |
| 15. T20 air temperature measured by EC150 at a height of 4m (degrees Celsius) | 16. T20 air temperature measured by EC150 at a height of 4m (degrees Celsius) |
| 17. RH20 relative humidity measured by EC150 at a height of 4m (degrees Celsius) | 18. RH20 relative humidity measured by EC150 at a height of 4m (degrees Celsius) |
| 19. RH20 relative humidity measured by EC150 at a height of 4m (degrees Celsius) | 20. RH20 relative humidity measured by EC150 at a height of 4m (degrees Celsius) |

Citation:
Biofilm response to nutrient enrichment in urbanizing streams: nutrient-diffusing substrate bioassay

This dataset contains the results of a nutrient limitation study conducted in the Logan River, the Middle P. River, Red Butte Creek and the Jordan River. We measured the response of stream biofilms to nitrogen and phosphorus enrichment with nutrient-diffusing substrates. Biofilm response was measured as chlorophyll a and ash-free dry mass. The mean, standard deviation, and sample size of biofilm responses to nutrient enrichment treatments are included in this dataset. This dataset includes the concentration of total nitrogen, total phosphorus, NH4-N, NO3-N, and SRP in stream water samples collected from each site.

Data and Resources

Biofilm_nutrient_bioassay.csv
This file contains the results of an experiment conducted in the Logan...

ash-free dry mass bioassay chlorophyll nitrogen nutrient limitation nitric phosphorus stream biofilms water quality


Logan River discharge, temperature, and conductivity dataset

This dataset contains information on water quality infrastructure projects that were funded in the RI through the Blue River Restoration Program. The data contains information on water quality variables collected at multiple sites along the Logan River, including water temperature, conductivity, and pH. The data also includes information on the type of infrastructure project and the cost associated with each project.
Survey Data Viewer:  
http://data.iutahepscor.org/surveys/
<table>
<thead>
<tr>
<th>Concern Rating</th>
<th>Water Quality Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Drinking Water Supply</td>
<td>very good</td>
</tr>
<tr>
<td>Groundwater Beneath Community</td>
<td>neither good nor bad</td>
</tr>
<tr>
<td>Nearby Mountain Rivers and Lakes</td>
<td>very bad</td>
</tr>
<tr>
<td>Downstream Streams and Rivers</td>
<td>very good</td>
</tr>
</tbody>
</table>

Survey Data Viewer

Familiarity with How Much Household Spends on Water Each Month
Has Lawn Where They Live
Person Responsible for Watering Lawn
Participation in Water Activities
Originally from Utah
Satisfaction with Overall Quality of Life
Owns or Rents Home
Has Ties to Farming
Sex of Respondent
Age of Respondent
Highest Level of Formal Schooling
Zipcode
Questions?

http://data.iutahepscor.org

Amber Jones
amber.jones@usu.edu

Jeff Horsburgh
jeff.horsburgh@usu.edu