Cyberinfrastructure to Support Large Scale, Collaborative Water Research in Utah: Critical Outcomes from the iUTAH Project

Amber Spackman Jones

Jeffery S. Horsburgh

and the iUTAH Cyberinfrastructure Team
• Cyberinfrastructure needed to support:
  - Interdisciplinary and cross-institutional teams
  - Diverse data collection and modeling efforts
  - The full data life cycle
• Includes storage, software, networking, computational, and human resources.

iUTAH: innovative Urban Transitions and Aridregion HydroSustainability
Cyberinfrastructure needed to support:
- Interdisciplinary and cross-institutional teams
- Diverse data collection and modeling efforts
- The full data life cycle

Includes storage, software, networking, computational, and human resources.

Figure courtesy DataOne
Cyberinfrastructure needed to support:

- Interdisciplinary and cross-institutional teams
- Diverse data collection and modeling efforts
- The full data life cycle

Includes storage, software, networking, computational, and human resources.
“All of the primary datasets collected as part of this project will be made freely and publicly available...”

- iUTAH Proposal Data Management Plan
Information Entropy

Figure taken from Michener (2006)

- Time of publication
- Specific details about problems with individual items or specific dates of collection are lost relatively rapidly
- General details about data collection are lost through time
- Retirement or career change makes access by scientists to "mental storage" difficult or unlikely
- Accident may destroy data and documentation
- Death of investigator and subsequent loss of remaining records

Information Content of Data and Metadata vs. Time
Information Entropy: What if instead…

- Curated data published in a data repository
- Data annotated by additional users
- Paper using data is published
- Data synthesized and leads to another publication

Time

Information Content of Data and Metadata
Data Policy and Data Management Training

• Developed project data policy to codify guidelines and timeframes for data sharing
• Conducted training
• Developed data publication tutorials
• Review and curation of submitted datasets

Changing the culture from “My Data” to “Our Data”
GAMUT: Gradients Along Mountain to Urban Transitions

40 sites

2000 data series

200 million+ values over 4 years

GAMUT: Accessing Sensor Data

**HydroShare:**
Data available in flat files (csv) and updated daily

**Time Series Analyst:**
Web interface, map and plot tools, and data export

**Web Services:**
Programmatic data access
GAMUT: Quality Control of Sensor Data

- Continuous, high frequency data require post processing
- Modifications to correct for common data errors
- Sensor Drift & Calibration
- Fouling
- Power Failure
- Icing
- Anomalies

GAMUT: Grab Sample Results

- Two efforts with different variable sets:
  - Monthly sampling: 2013-2014
  - Biweekly sampling: 2014-2016
- Variables: TSS, VSS, species of Nitrogen and Phosphorus, Total Coliform, E.coli, DOC, Fluorescence, Chlorophyll-a, Isotopes, Ions
Social Water Science Data: Considerations

- Developed methods for categorizing social science data
- Dimensions help us understand mechanisms and restrictions for how social science data can be shared

Social Water Science Data: Visualization

- Visualization of public intercept survey results
- Generic and reusable survey template
- Open source code

http://data.iutahepscor.org/surveys/

Sharing Diverse Data

HYDROSHARE [http://www.hydroshare.org]

- An online, collaborative system for sharing and publishing various data types, models, and code
- Support for collaboration-sharing with individuals, groups, or publically
- Ability to formally publish with a DOI
The Steven Hall Story

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

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

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.

Steven published his paper in HydroShare and cited published data.

Acknowledgments
Upon manuscript acceptance all data are publicly available online at the HydroShare database: http://dx.doi.org/10.4211/hs.7ce47150ee344d4e95f2f7bb4f660ac. We thank four...
Hydroinformatics Course

- 3 to 5 partner universities
- 30 - 45 students total across the campuses
- We focus on:
  - Data and the data life cycle
  - Databases and data models
  - Data visualization, transformation, analysis, and modeling
- Technologies we use:
  - Relational database management systems
  - Structured query language
  - Python Programming
  - R Statistical Computing

“My team used basic concepts from almost every class period and topic section in our term project. It was cool to see how all the individual skills added up to help us create and maintain hydrologic information.”
Undergraduate Training

- Creating the next generation of “Cyber-savvy” engineers and scientists
- Prototyping and developing new software applications
- Collaborating with iUTAH scientists
- Co-authoring CI-related papers
- Gaining practical experience and improving job prospects
Questions?

Jeff Horsburgh
jeff.horsburgh@usu.edu
435-797-946

Amber Jones
amber.jones@usu.edu
435-797-7147

Open Source Code Repositories

- WEBTSA – GAMUT time series data visualization
  - https://github.com/UCHIC/WEBTSA
- ODM Streaming Data Loader
  - https://github.com/ODM2/ODM2StreamingDataLoader
- ODM Tools Python – Sensor data management and QC
  - https://github.com/UCHIC/ODMToolsPython
- ODM2 Sensor – Sensor equipment management
  - https://github.com/UCHIC/ODM2Sensor
- iUTAH Utilities – Automated alerts, etc.
  - https://github.com/UCHIC/iUtahUtilities
- iUTAH Survey Data Viewer – Visualization of survey data
  - https://github.com/UCHIC/SurveyDataViewer
- iUTAH Data – Modeling and Data Federation Website
  - https://github.com/UCHIC/iUTAHData