Challenges and Opportunities of Adding Non-Standard Data to an Existing Repository

"Driving Square Pegs Into Round Holes"

May 28, 2019
Canadian Research Software Conference, Montreal

Doug Mulholland, Paulo Alencar, Don Cowan

David R. Cheriton School of Computer Science, University of Waterloo

Les Stanfield – Ecohealth Solutions

canarie

GWF Global Water Futures

FLOWING WATERS
INFORMATION SYSTEM

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science

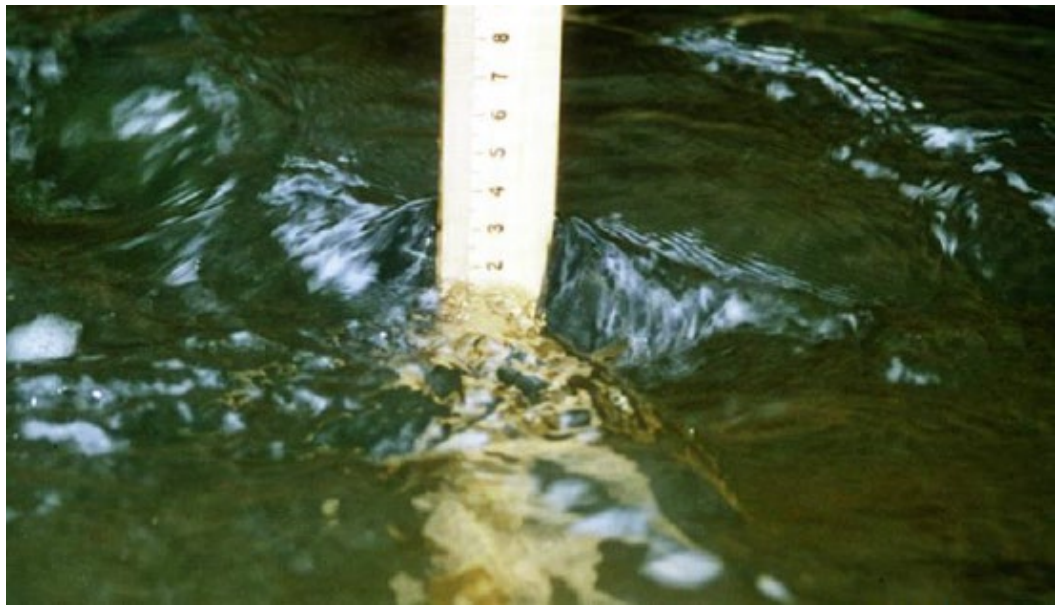
Context:

- Ontario Stream Assessment Protocol (“OSAP”) is a standard for collecting and recording stream data
  - About 25 years old, fairly widely used across Ontario and growing but not universally adopted by all practitioners
  - Solid science behind each methodology (reproducibility, rigour)
  - Regular reviews by steering committee (practitioners and proponents)
  - Minimal government support
- Flowing Waters Information System (“FWIS”) was created to hold OSAP data
  - Operational, fairly widely used and growing, partner with domain experts
  - One component of iEnvironment (CANARIE project)
  - Built with our “WIDE” toolkit (metadata-driven architecture)
  - Data summarized, shared (researchers, others)

UNIVERSITY OF WATERLOO

Challenges (1): Minor Variations

- Numerous (1000s) datasets are being discovered that should be “associated” with iEnvironment/FWIS either directly or partially
  - Most are historical using various “flavours” of OSAP
  - E.g., “Hydraulic Head” – measured in mm to closest 5 mm, usually recorded as an integer but stored as a text string; used for calculating “discharge” (aka “flow volume”)



UNIVERSITY OF WATERLOO

Challenges (2):

- Some technicians incorrectly record hydraulic head values as cm with decimal points → FWIS generates a warning diagnostic in the report
- One non-standard dataset (valuable historical data) was recently uploaded but included actual flow meter measurements that were converted to decimal hydraulic head values
  - Will add a “Computed decimal value” attribute (checkbox) to the table to suppress the diagnostic
- Most other variations are much more significant
  - E.g., Fish distribution summary query – “what fish (#, species, length, weight, etc.) were found where/when/by whom”
  - Query is generated by database “view” computed in response to request

UNIVERSITY OF WATERLOO

Challenges (3): "Data Independence"

- Government collects data from numerous sources, including reports from FWIS
  - Published as "Aquatic Resources Area" ("ARA" map layer/SHP file from OMNRF/"Land Information Ontario" – terms of licence permit republication), updated monthly
  - ARA layer includes "source" attribute (where did this record come from – FWIS, Conservation Authority name, consultant name, …)
- FWIS now includes ARA layer (except FWIS-sourced records) → FWIS also reports record source
  - No source code changes
  - Only the database view for the summary query was adjusted, along with support for ARA table storage and ARA data refreshes

UNIVERSITY OF WATERLOO

Challenges (4): "Exploit Some Structures"

- Recently asked about Terrestrial Data – plants, animals, invertebrates, …
  - Will lever existing aquatic locator and project collection strategies (location name, site code, date, project title/code) as well as logging, security facilities and other infrastructure
- Data is "related" to OSAP ("part of the story" – habitat suitability, degradation, …) but not at all like OSAP
- Environmental DNA ("eDNA") – growing popularity for aquatic species
  - Currently suitable for presence/absence but not abundance
- Will build on lessons learned in FWIS
  - Importance of a steering committee, government participation
  - Sustainability plan

UNIVERSITY OF WATERLOO

Challenges (5): Very Different Data, Minimal Commonality

- Marine Data – plants, fish in a salt water environment
- Essentially of no interest to freshwater researchers
- Makes use of WIDE facilities and similar features to FWIS but different user groups, data structures, etc.
- Also builds on lessons learned in FWIS, managed in a separate database
- Can be accessed from FWIS if need arises (“proxy” tables – SQL Anywhere database engine supports access to one database from within another)

UNIVERSITY OF WATERLOO

Approach (1):

- API and demonstration/utility program extended
  - curl/JSON access to API (R, PHP, VBA, Perl, Python, Java, JavaScript, Google Sheets Javascript, …)
  - Support for table/view create/alter/drop
    - Change data structures without changing any program code
  - Web forms and listings automatically generated
  - Permissions default appropriately (owner access only)
- Special security structure added for access to database structure
- Lever database/metadata-driven testing
  - Automated test generation

UNIVERSITY OF WATERLOO

Approach (2):

- Data comparison and synchronization tool
- Declarative software agents
- Manage descriptive metadata publication
  - Data providers routinely change their data, often in response to a user (i.e., researcher) discovering anomalies
- Global Water Futures (“GWF”) Connection: “Linking Stream Network Process Models to Robust Data Management...", Dr. Bruce MacVicar (U. of Waterloo, Civil Engineering)

UNIVERSITY OF WATERLOO

Special Concerns: "Vulnerable" Data

- Controversial or politically sensitive areas of study
- Researchers who may be considering retirement or a move to a different organization
- Contact Us

Questions?
Thank You!
Contact Us

Doug Mulholland, Paulo Alencar, Don Cowan

David R. Cheriton School of Computer Science, University of Waterloo

{dwm | palencar | dcowan} @csg.uwaterloo.ca

Les Stanfield, Ecohealth Solutions

Les.Stanfield@outlook.com