Computer Systems Group
University of Waterloo

Connecting Systems Together is No Small Feat

WATERLOO
CHERITON SCHOOL OF COMPUTER SCIENCE
cs.uwaterloo.ca

Doug Mulholland
Paulo Alencar
Don Cowan

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

canarie

Introduction/Overview

- A (probably incomplete) snapshot of stream, benthic and Groundwater data sources
- A researcher/biologist's perspective
  - Where to start?
  - Who's got what?
  - How to connect results?
- How can systems/repositories help?
- Recommendations/Conclusions

Ontario Stream, Benthic and Groundwater Data Snapshot

- Flowing Waters Information System
  - Ontario Stream Assessment Protocol
  - Biological data (fish), physical structures, flow, temperature with some benthic
  - Partnerships with Ontario Conservation Authorities, researchers, ad hoc organizations (“Stream Monitoring and Research Teams” –“SMART networks”)
  - Capture vulnerable data in sibling system (“iEnvironment” – more flexible, user-configurable)

Ontario Stream, Benthic and Groundwater Data Snapshot (2)

- Ontario Government - Ontario Data Catalog
  - Environment and Natural Resources – 475 datasets
    - Aquatic Resource Areas – data collected through Scientific Collection Permit Reports (what fish were found where, when)
    - Ontario Benthos Biomonitoring Network (Dr. Chris Jones)
    - Many others (Provincial Water Quality Monitoring Network, …) – just knowing that a dataset exists is no small accomplishment

Ontario Stream, Benthic and Groundwater Data Snapshot (3)

- Royal Ontario Museum (Ichthyology Collection) – "…over one million specimens of approximately 7,000 species from around the world…"; published at gbif.org (Global Biodiversity Information Facility)
- Ontario Conservation Authorities (CAs) – 36 organizations across Ontario, mostly in southern Ontario, set up by watershed, generally well-defined boundaries with no geographic overlap and local priorities
- Government of Canada – Fisheries and Oceans Canada – major project underway now to organize their data

Ontario Stream, Benthic and Groundwater Data Snapshot (4)

- Private Corporations – KISTERS North America (California based software company) “WISKI” relational database and water data analytics platform
- Municipalities – many used to collect biological data but have handed off stream monitoring/sampling to a CA
- Private engineering/consulting companies that specialize in various aspects of sampling and monitoring

Ontario Stream, Benthic and Groundwater Data Snapshot (5)

- Not-for-profit/charitable foundations
  - Oak Ridges Moraine Groundwater Program
  - Water Rangers (citizen science, test kits, water quality)
  - Gordon Foundation (DataStream Initiative – collect and distribute water quality, chemistry, flows, analyses)
  - The Land Between (Central Ontario from Frontenac Arch to Georgian Bay – First Nations partnerships, identify sensitive species, environmental challenges)
  - Great Lakes Observing System (Ann Arbor MI) – “Seagull”/Smart Great Lakes initiatives for data collection/dissemination – mostly lake data at present

Researchers - Connections

- Researchers – where does a researcher start???
- Our partners often just want to connect the dots – “a little of this and a little of that”
  - e.g., combine benthic and fish data, or fish, channel structure and temperature -> “Habitat Suitability Index”
- Impossible to predict what questions researchers will ask
- Data quality validation
- Biologist’s perspective – “link the systems”

- Whoa, just a minute – what does “link the systems” mean?

Connecting Systems

- Some commonality between the data
  - Usually location-based (i.e., “study sites”)
  - Co-ordinate systems vary (Lat/Long vs UTM vs …)
  - Local organizations identify sites with a site code (e.g., “DN001WM” – somewhere on the Don River (Toronto), often numbered roughly in km from the mouth of the river … but by no means always … “001” is one of 27 sites on various tributaries of the Mississippi River in Eastern Ontario)
  - Sometimes there’s common data but no guarantee that the same data will be presented from different sources (can be significant QA/QC concerns)

Connecting Systems (2)

- Simply link to another organization's home page or a dataset within the organization
  - Rarely any easily managed record-level common key, but there is data of interest to our researchers
  - A few systems provide a Digital Object Identifier (DOI)
    - doi.org – ISO 26324 (at the dataset level) but the DOI does not address frequently changing datasets
  - Want to validate periodically that the organization and target dataset continue to exist (automated link validation)
  - When reasonable, link to home page, provide navigation guidance to data of interest (e.g., gbif.org)

Connecting (3) … Home Page

- Advantages:
  - Site can provide guidance on DOI use, citation requirements, republishing
  - Data Sharing Agreement may be required
- Disadvantages:
  - Researcher still has work to do to associate records, assuming common key exists
- Challenges:
  - Starting URL may change
  - Navigation instructions change
  - Data often reported in inconsistent form (units of measure, observation conventions)

Connecting Systems (4)

- Record-level links
  - Enables researchers to immediately see supporting or contradictory evidence to a hypothesis
  - Build evidence for/against a correlation
  - Usually more difficult to maintain (systems may maintain an entirely opaque key scheme

Connecting Systems (5)

- API (Application Program Interface)
  - Requires some level of programming/scripting
  - MapServer(.org) – WMS/WFS (Web Map Server/Web Feature Server) – Open Geospatial Consortium standard API
  - WQX protocol for water quality data (US EPA and USGS)
  - Proprietary/custom (KiWIS – Kisters WISKI, FWIS Data Access API)
- Most (young!) researchers are pretty used to retrieving tables with a single R expression

Connecting (6) … Observations

- Publish metadata, data dictionary (with meaningful explanations) – what, where, when, ideally who
- Collect and report “project” metadata (why was the data collected at this site, how were study sites selected)
- Consider the decision that a researcher must make (should I use this data or not, is it valuable in my context?)
- CoreTrustSeal – forces repositories to document and formalize their processes (although probably overkill for small repositories today)

Conclusions/Recommendations

- As data repositories proliferate, data overlaps and redundancies more likely
- Different connection and linkage approaches have advantages and disadvantages
- Improve Findability, Accessibility, Interoperability and Reuse (FAIR principles)
- Are you focused on the researcher?
- Automated tools help a lot but don’t solve all the problems

Thank You!

- Comments/Questions?
- Any of this resonate with your sector?