

The 3 "Arrrs" of Vulnerable Data – Retirement, Reorganization and Restructuring

Canadian Research
Software Conference

Doug Mulholland, Don Cowan, Paulo Alencar
University of Waterloo, Computer Systems Group

Les Stanfield, Ecohealth Solutions

canarie



Global Water Futures

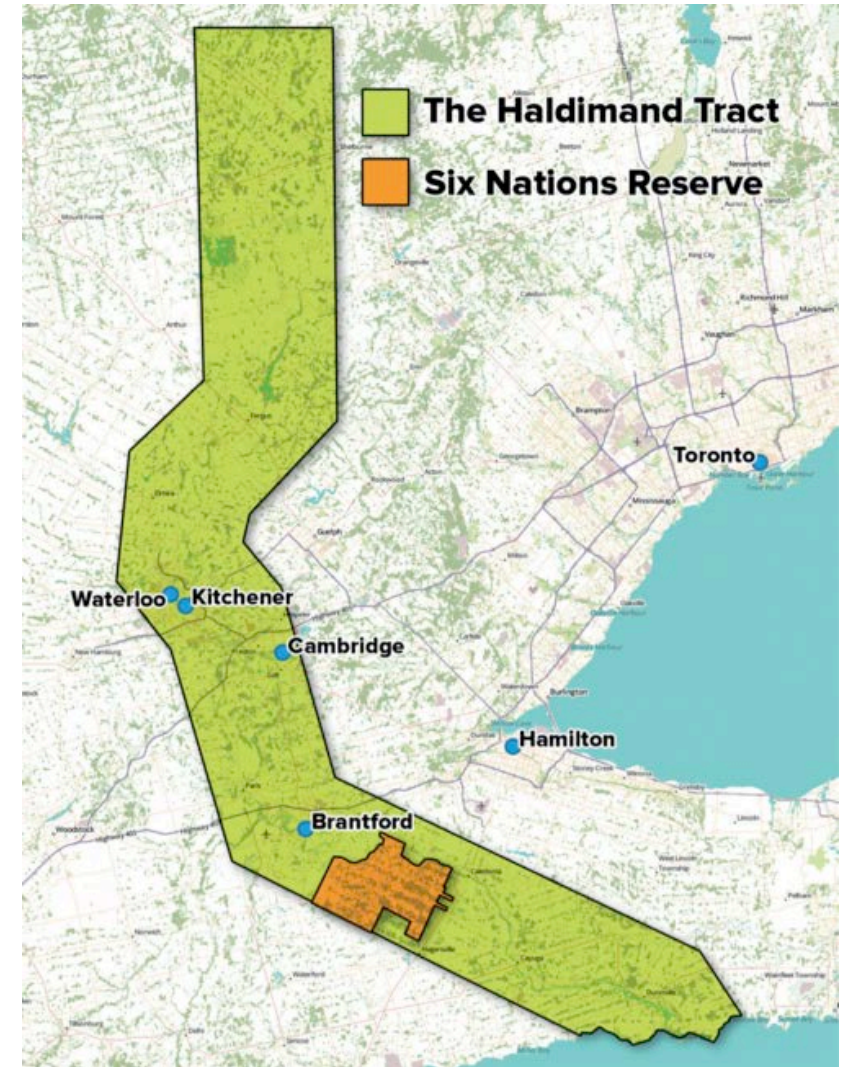
FLOWING WATERS
INFORMATION SYSTEM



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

Territorial Acknowledgement

The University of Waterloo acknowledges that much of our work takes place on the traditional territory of the Neutral, Anishinaabeg and Haudenosaunee peoples. Our main campus is situated on the Haldimand Tract, the land granted to the Six Nations that includes six miles on each side of the Grand River. Our active work toward reconciliation takes place across our campuses through research, learning, teaching, and community building, and is centralized within our Indigenous Initiatives Office - <https://uwaterloo.ca/human-rights-equity-inclusion/indigenousoffice>



Vulnerable Data

Objective:

- Raise awareness of an urgent, ubiquitous problem that's not often adequately addressed
 - Acute in the environmental sector where senior researchers have been working for years collecting increasingly large data sets with rapidly changing storage platforms, data collection standards and practices
 - Historical data is extremely valuable in the context of climate change, landscape change (urbanization, habitat loss), loss of biodiversity, invasive species

Clarification:

- “Arrr” – “Yes”, “I agree” in various online “Pirate” dictionaries! [oops, definitely not my position]
- “Argh” – “exclamation of annoyance, exasperation, or other negative factor”
...urbandictionary.com

Observation:

- I have more experience at losing data (systems software actually) than I'd like to admit. (shame factor)

Examples of Common Vulnerabilities

- Obsolete/failed media
 - Rapidly changing storage capacities, formats, volatile media



Left to Right: 8" floppy (1.2 Mb), 5 1/4" floppy (500 KB), 3.5" floppy (1.44 Mb), IBM 3420 reel tape (140 Mb @ 6250 bpi), IBM 3480 tape cartridge (200 Mb), Digital Equipment Corp. RK05 (2.5 Mb) and many more
Images obtained from the "Museum of Obsolete Media" – obsoletemedia.org and Wikipedia



Examples of Common Vulnerabilities (con'd)

Anything sound familiar?

- Encrypted media - “The password was written on a piece of paper wrapped around it.” or “...on a yellow sticky note”
- “I can well imagine some IT safety tech saying this is not safe and taking it to store it somewhere ‘safer’.”
- “...in my experience, it missed all the archived photos which were on another drive... Thousands of photos were lost.”
- “...data that was stored on a shared drive was not archived with this dataset...”
- “...in theory, all the files were on the shared drive, but god knows whether and where they are stored/archived...”
- “No one knows where or how to access them.”
- “The steering committee has disbanded.”
- “xxx is retiring soon... Very sad.”

Examples of Common Vulnerabilities – Key Human Factors

- Retirement
 - Senior scientists/practitioners today have lived through many different data storage formats, from perfect bound paper lab notebooks and paper publications to smart phones, USB memory sticks and cloud storage. How well did they record metadata or even the data itself? Who else is familiar with the various data sets, userids and passwords?
 - Publishers now routinely require supporting data and data processing code to be provided with the publication, but how complete is the data set?
 - What about work that wasn't published?
- Reorganization
 - Scientists and technicians in the environment sector are notoriously mobile (move from department to department or organization to organization)
 - Government departments, corporations, non-governmental organizations (NGOs) all routinely reorganize as the science changes
- Restructuring
 - If an individual or department is unable to routinely justify its work, they may be “restructured” (out of a job/budget allocation)
 - Management/strategic priorities are never static

Metadata Captured in iEnvironment/FWIS (“Flowing Waters Information System”)

Add rich, context-based metadata both at the project and sample event level:

- What was done – protocol, type of sampling
- When – date, time as needed
- Where – co-ordinates of study site (within roughly 50m is adequate for this type of work)
- Who – what organization and practitioner did the work – balance privacy and storage of personal information with “need to know” (what were the technician’s qualifications, experience, etc.?)
- Why? Too often overlooked or incomplete but often the most important – is it worth the effort to recover the data set?
 - Why was the study being done at all?
 - Why/how are sites chosen for this work?
 - Is this site/sampling typical or a reference site
 - 52 specific yes/no questions, including “Other” and “Comments” fields

Environment/FWIS Generic Data Upload/Archive Facility

Tabular Data

- Remote access, including MS Excel macros and Google Apps Script processes to enable table management
- Permissions management – facilitates keeping the data private, managed sharing or unrestricted publication
- Users can lever existing infrastructure (naming and mapping of study sites, content searching, project organization, photos/sketches, ...)
- Metadata recording/publication

Unstructured Content

- File metadata processing (e.g., photo location, camera type, lens, ...)
- Content and metadata indexing/searching with a broad range of supported file types

Sharing Data More Broadly

“Library of Congress” type of registry

- DOI? (Searchable?)
- One “start” shopping for data
- Encourage richer, context-based metadata – standard who/what/where/when often misses important details to enable answer to how valuable the data is for an unforeseen application
- Handling vulnerable data is consistent with FAIR guidelines for scientific data management and stewardship (Findable, Accessible, Interoperable, Reusable)
- Application of FAIR principles to workflows involving the preservation of vulnerable data and metadata

References

1. Lavoie, B., Gartner, R.: Preservation Metadata, 2nd edition. DPC Technology Watch Report, DPC Technology Watch Series, May 2013.
2. Library of Congress: PREMIS—Preservation Metadata: Implementation Strategies, v. 3.0. <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>
3. Wilkinson, M. et al., The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016.
4. European Commission, Turning FAIR into reality, Directorate General for Research and Innovation, 2018.

Thank You!

What are you doing for data that doesn't quite fit anywhere else?

Questions?

**Doug Mulholland
University of Waterloo, Computer Systems Group
and
Centre for Community Mapping (comap.ca)**

dwm@csg.uwaterloo.ca