Look Before You Leap

Adventures in curating & preserving research data

Shahira Khair & Grant Hurley
October 20, 2020

Portage Global Water Futures Webinar

Hello!

Grant Hurley
Digital Preservation Librarian, Scholars Portal

Shahira Khair
Data Curation Librarian, University of Victoria

Previous slide image: City of Toronto Archives, "The diving horse at Hanlan's Point," fonds 1244, item 192.
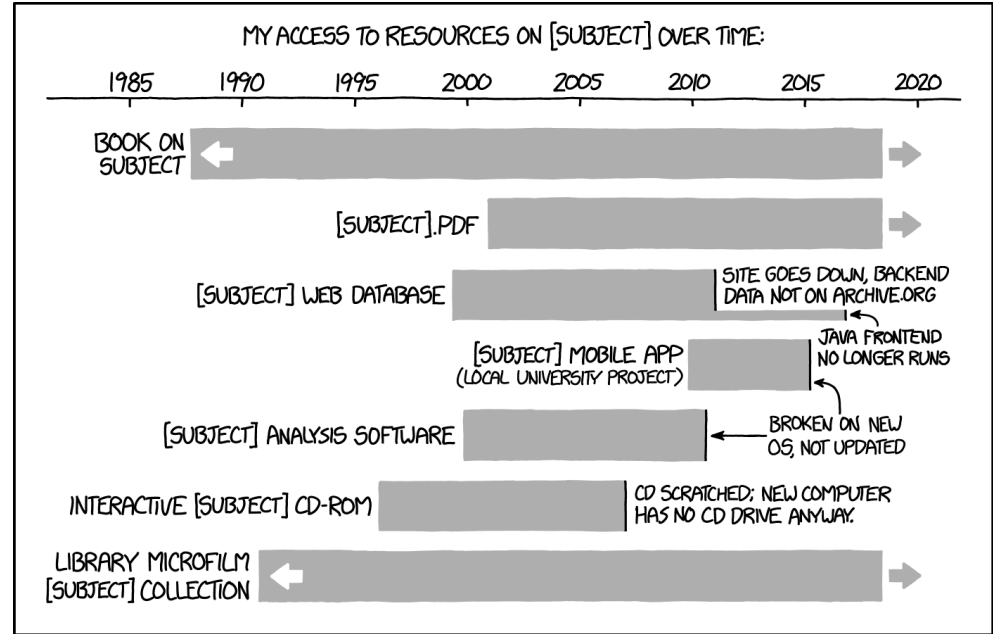
Land acknowledgements

Tkaronto (Toronto) is the traditional territory of many nations including the Mississaugas of the Credit, the Anishnabeg, the Chippewa, the Haudenosaunee and the Wendat peoples and is now home to many diverse First Nations, Inuit and Métis peoples. Toronto is covered by Treaty 13 signed with the Mississaugas of the Credit, and the Williams Treaties signed with multiple Mississaugas and Chippewa bands.

The City of Victoria is located on unceded Coast Salish territory, where I live and work on the lands of the Lekwungen (Songhees), Esquimalt and WSÁNEĆ peoples whose long standing relationships with this land continue to this day.

Learning Objectives

- Introduction to key concepts in digital curation and preservation

- Understand how choices made curating data for deposit can improve or harm the prospects of that data's preservation into the future

- Identify small but impactful ways to improve reusability of datasets now and in the future

Digital fragility

MY ACCESS TO RESOURCES ON [SUBJECT] OVER TIME:

BOOK ON SUBJECT

[SUBJECT].PDF

[SUBJECT] WEB DATABASE — SITE GOES DOWN, BACKEND DATA NOT ON ARCHIVE.ORG

[SUBJECT] MOBILE APP (LOCAL UNIVERSITY PROJECT) — JAVA FRONTEND NO LONGER RUNS

[SUBJECT] ANALYSIS SOFTWARE — BROKEN ON NEW OS, NOT UPDATED

INTERACTIVE [SUBJECT] CD-ROM — CD SCRATCHED; NEW COMPUTER HAS NO CD DRIVE ANYWAY.

LIBRARY MICROFILM [SUBJECT] COLLECTION

IT'S UNSETTLING TO REALIZE HOW QUICKLY DIGITAL RESOURCES CAN DISAPPEAR WITHOUT ONGOING WORK TO MAINTAIN THEM.

Photo: Jess Whyte

Xkcd comic #1909, 2017

Digital fragility

- Media used to store data degrade and age

- Many layers of mediation required between the physical data object and you:
  - Storage medium
  - Operating system and file system
  - File format and/or character encoding
  - Software application
  - Display

- Ease of deletion, corruption - but also replication, integrity checking

Digital fragility

- The need for stewards who make long-term commitments to keeping data accessible

- Broad definitions of the role of responsible stewards: FAIR/TRUST principles, certification standards (CoreTrustSeal, ISO 16363)

Q Popular Latest   *The Atlantic*   Sign In

TECHNOLOGY
The Irony of Writing Online About Digital Preservation
Last month, *The Atlantic* published a lengthy article about information that is lost on the web. That story itself is in jeopardy.

MEREDITH BROUSSARD NOVEMBER 20, 2015

Research Data Lifecycle

Plan Create Process Analyze Share Preserve Reuse
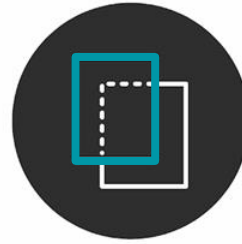
Data Curation

Share

- Broadly refers to the active management of research data over its lifecycle

- At the end of a lifecycle, the emphasis is towards improving reproducibility and reusability

- Data curators collaborate with researchers to store and responsibly share their data in ways that are Findable, Accessible, Interoperable and Reusable (FAIR)

A Data Curator's Day:

- Providing **consultations** with researchers

- **QA/QC audits** and dataset reviews

- Preparing datasets for **deposit** to a repository

- Augmenting datasets to improve FAIRness

  - **Organizing** files

  - Creating **documentation** or metadata

  - Implementing metadata **standards**

 - Applying **persistent IDs**

{"id":72614,"identifier":"SP2/OOVOQR","persistentUrl":"https://doi.org
{"id":6974,"datasetId":72614,"datasetPersistentId":"doi:10.5683/SP2/OO
01","UNF":"UNF:6:pN85EhT2dQgoglt4irvVHw==","lastUpdateTime":"2020-04-2
non-exclusive, non-transferable, royalty-free licence, for use in Cana
For the sake of clarity, the license being granted by Infoway to User
media, or technology now known or hereafter developed, on terms and co
respect to any and all components thereof. Except as expressly set for
granted, assigned, conveyed, transferred or provided to User by virtue
markings or notices placed upon or contained within any Dataset.","dis
representatives (the "Indemnified Parties"), from and against any and
expenses that they may incur in consequence of a breach by User of its
personal identifying information or health-related personal informatio
of Caregivers"},{"typeName":"otherId","multiple":true,"typeClass":"com
{"typeName":"otherIdValue","multiple":false,"typeClass":"primitive","v
{"typeName":"otherIdValue","multiple":false,"typeClass":"primitive","v
Health Infoway"}}]},{"typeName":"datasetContact","multiple":true,"typ
{"typeName":"datasetContactAffiliation","multiple":false,"typeClass":"
{"typeName":"dsDescription","multiple":true,"typeClass":"compound","va
R.A.Malatest & Associates Ltd., a third party research firm to conduct
care, coordination of care provided by healthcare professionals and va
{"typeName":"subject","multiple":true,"typeClass":"controlledVocabular
{"typeName":"keywordValue","multiple":false,"typeClass":"primitive","v
{"typeName":"topicClassification","multiple":true,"typeClass":"compoun
{"typeName":"publication","multiple":true,"typeClass":"compound","valu
evaluation/3226-2014-the-value-of-digital-health-for-caregivers-in-can
Health Infoway"},"producerAbbreviation":{"typeName":"producerAbbreviat
{"typeName":"grantNumber","multiple":true,"typeClass":"compound","valu
{"typeName":"distributor","multiple":true,"typeClass":"compound","valu
{"typeName":"distributorURL","multiple":false,"typeClass":"primitive",
{"typeName":"dateOfDeposit","multiple":false,"typeClass":"primitive","
[{"typeName":"geographicCoverage","multiple":true,"typeClass":"compoun
["Provinces"]}]},"socialscience":{"displayName":"Social Science and Hu
["National Panel. Panelists were recruited into the panel via telephon
for this survey was 27%. This compares to commercial panels that typic
{"typeName":"dataCollector","multiple":false,"typeClass":"primitive","
{"typeName":"collectionMode","multiple":false,"typeClass":"primitive",
(±4.4%)**Results may or may not be generalizable to Canadians as a who
{"id":77321,"persistentId":"","pidURL":"","filename":"2014 Infoway Can
sav","originalFormatLabel":"SPSS Binary","originalFileSize":283915,"UN
{"description":"Survey questions for National Survey of Caregivers","l
{"id":72616,"persistentId":"","pidURL":"","filename":"Caregiver and Di
d8b681b19033","rootDataFileId":-1,"md5":"802203f6d6aa08904c18611e37ae8
["Documentation"],"dataFile":{"id":104726,"persistentId":"","pidURL":"
{"type":"MD5","value":"e318547b8ca03b8718697f2558023479"},"creationDat
UNF:6:pN85EhT2dQgoglt4irvVHw== [fileUNF]"}}

Data Preservation
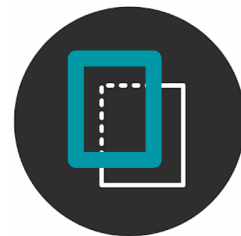
Preserve

- A definition: "The series of managed activities necessary to ensure continued access to digital materials for as long as necessary" (Digital Preservation Coalition, 2016)

- What does this mean in practice?

- The broad goal is to always to:

  - Improve the **preservation prospects** of data into the future using preservation strategies

  - Ensure users can access preserved materials in the form best suited to their needs

  - Monitor and mitigate preservation risks while doing no harm to the materials preserved
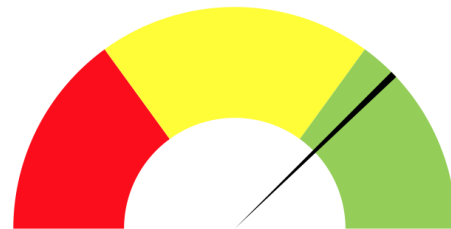
A Digital Preservationist's Day:

- Developing and documenting preservation **policies and workflows**

- Creating and verifying **checksums**

- Gathering and structuring **descriptive, rights, administrative, and technical metadata** to inform and support preservation

- **Migrating/normalizing** files where required in preservation-friendly formats and for access purposes

- **Storing** preservation package in friendly storage environments, with multiple, geographically separate copies

- Exploring and support approaches for **complex data objects** like emulation, software preservation, web archiving, and more!

PASSAGE.BK!

What is this?
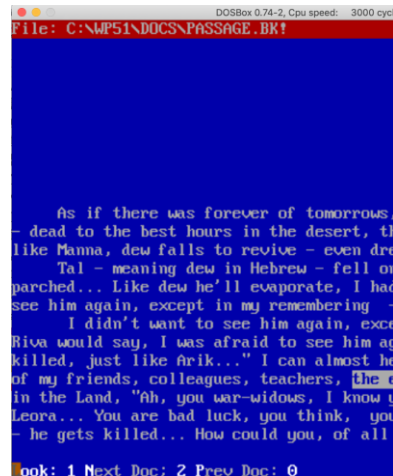
WordPerfect 5.1!

Preservation strategies & levels of commitment

Bit-level: periodic checksum validation of digital objects to ensure they have not been modified or corrupted, but no commitment to ensuring files are still accessible

Normalization/migration: converting files to preservation friendly formats, where possible, either upon receipt (normalization) or later on when files are at risk (migration)

Emulation: maintaining access to original files *and* their originating software/operating systems, and then running these together in a current computing system

# Black Mirror - Bandersnatch (2018, Netflix)
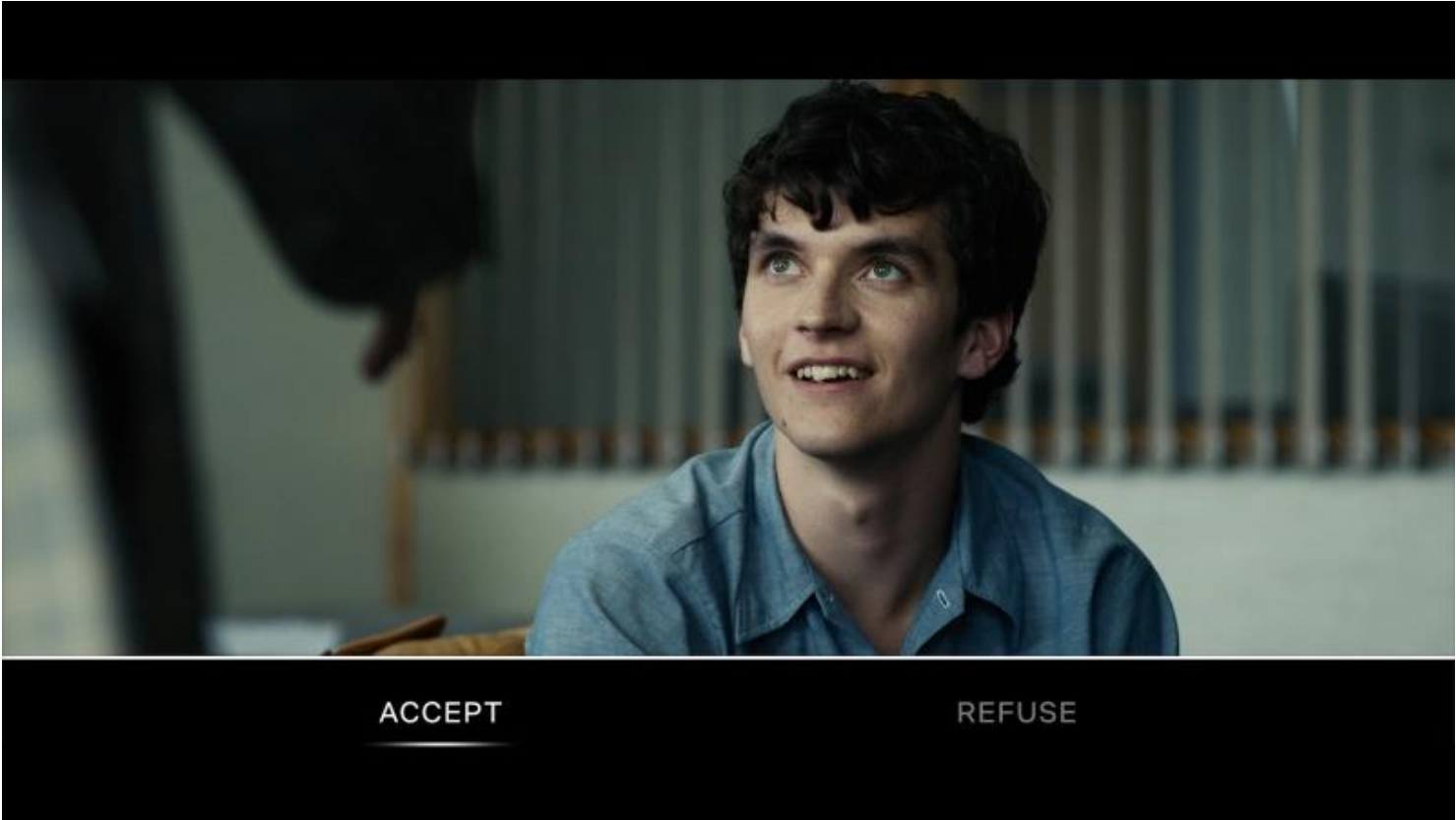
Scenario 1: File Formats

You receive an email from a grad student to deposit a dataset from their recently published mixed methods study on childhood bullying.
They have the green light to share from the university's REB.

They attach their dataset as a .nvp file, an NVivo workspace. NVIVO is proprietary software, but luckily your university provides licenses!

You proceed to open the file, and the workspace is really well curated.

- Files and folder structures and well named and organized
- Data are linked and interpretable within their original context
- Data are fully anonymized
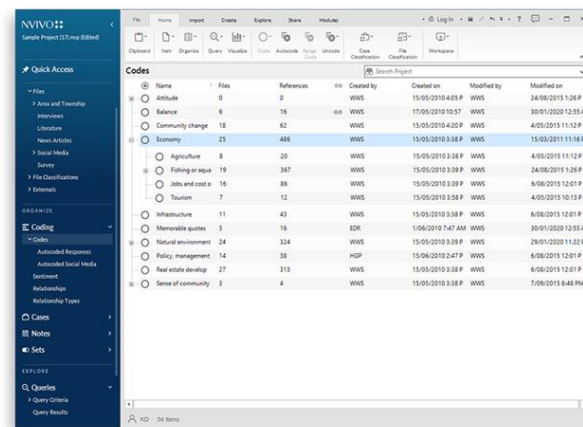
qsrinternational.com

Do you recommend: a) keep original file format

b) export into alternate file formats

Native file-formats

- Retains the original quality and context, which can support reproducibility & reusability:
  - Internal connectivity and interoperability
  - Internal documentation and metadata

- Reduced accessibility and interoperability for proprietary formats

- Accessibility of a file in a proprietary or non-documented format cannot be easily guaranteed; nevertheless original files usually kept even if not accessible

- Preserver could commit to maintaining original software (including specific version), if resources available and technically possible

- Is the format identified in PRONOM?
  - NVIVO is not currently documented!

Present → **Future**

Alternate file-formats

- Increased accessibility of open file formats

- Loss of information during transformation process

- Time consuming to recreate documentation and functionality

- Fewer challenges for preservation: known, documented formats using consistent standards are much easier to preserve (UTF-8 encoded CSVs or TAB files ftw!)

- Ideally, originating software is documented to assist in loss of context

Present → **Future**

Scenario 2: File selection / appraisal

During a consultation, a faculty member 2 weeks away from retirement brings you a flash drive containing a **20gb** zip file containing 20 years of data on nutrient measurements in nearby freshwater bodies.
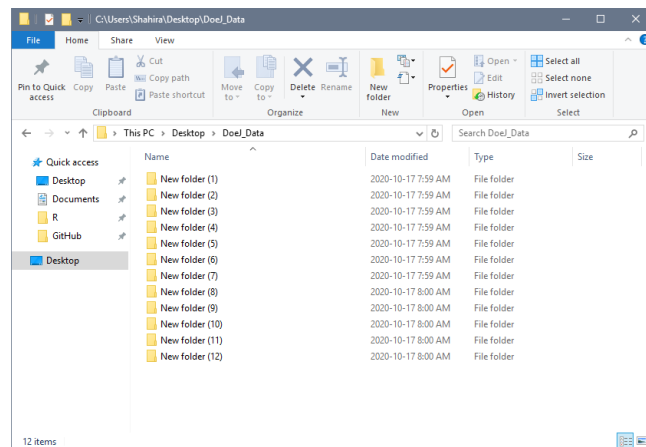
You extract the files to find:

- A dozen unnamed folders corresponding to funded projects
- Hundreds of files, many with multiple versions and drafts of the same files

Do you recommend: **a) keep it all, maintaining its current organization**

b) try to curate, keeping the most recent/complete versions of files

c) don't accept the data

Keep it all

- Impact of dataset structure on discoverability and reusability

- Impact of dataset size on storage and reusability

- Requirements of institutions, funders, publishers

- May be difficult for users to understand context unless effort is made to document structure and relationships between files

- Possibly complicated access/rights provisions to maintain for different components

- Greater quantity, size of files = more preservation resources needed (storage, processing, human)

Present ➝ **Future**

Selection and appraisal

- What constitutes a "dataset"?

- Considerations to prioritize:
  - Complete vs partial
  - Raw vs processed
  - Published vs unpublished

- Organization can inform as well:
  - Filing structure
  - Naming conventions

- Easier to apply modular preservation/access policies with selection:
  - Published data published, maintained in public repository with consistent cross-collection preservation policies
  - Unpublished, raw or incomplete files could be retained by university archives as evidence of researcher's process

- Consistent naming conventions, structure easier to parse/interpret

Present   **Future**

Scenario 3: Documentation

You receive a "dataset in review" notification from Dataverse:
"Replication Data for: Biodiversity assessment of Canadian wetlands", has been submitted for review.

You proceed to open the dataset and review its contents.

1. Four comma separated value (*.csv) files
   - Files are titled Figure_1 to Figure_4
   - Variables titles in files are moderately descriptive

2. In Dataverse only minimal citation metadata has been completed
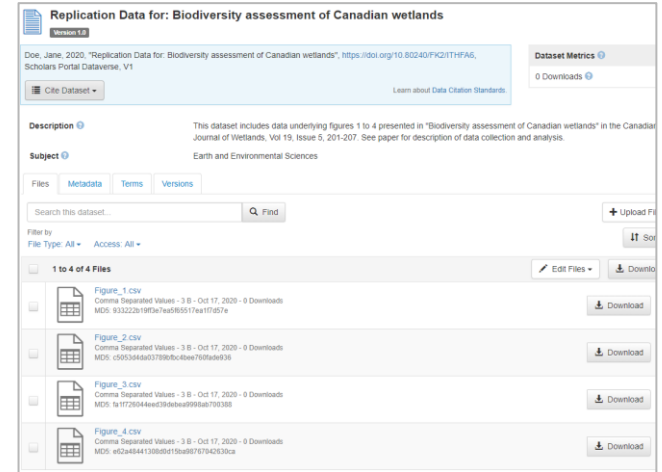
Do you recommend: **a) accept and publish**

                           b) return to author

Metadata

- Citation-block metadata to support discovery
  - Reference related outputs via PIDs

- File-level metadata for understandability and reusability
  - Descriptive
  - Administrative
  - Structural

- Descriptive, administrative, structural metadata can flow into preservation management systems, joining with technical metadata during preservation processing

- Relationships to other items (e.g. journal articles, reports) - also can be maintained if link is made

Present ➔ **Future**

Documentation

- Context supporting reproducibility and reusability, with goal of "independent understandability"

- Time consuming and challenging to (re)create after an investigation

- "Independent understandability" is key to preservation standards (e.g. OAIS) too!

- Documentation assists in understanding required dependencies, functionality, context, thereby informing migration or emulation strategies

Present → **Future**

Scenario 4: Licensing

During an audit of recent submissions to your data repository, you find a dataset proposing a new method of seismic response prediction for building structures.

The dataset includes:

1) Experimental software, which allows users to specify a range of inputs to model outcomes
2) A historical dataset of recent seismic events in the Pacific Northwest
3) A readme file that described how the dataset was created for academic purposes

The default repository setting of CC0 is enabled, but no licensing information is specified in any of the provided documentation
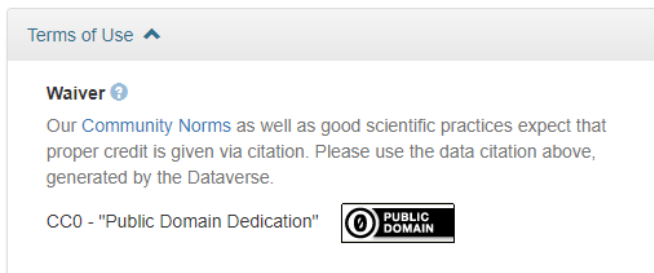
Do you recommend: **a) do nothing - leave as is**

b) contact the author for clarification

Rights for Creators & Users

- Supports attribution and furthers research impact for creators

- Limits reuse and distribution options for users

- Manages risk and liability for creators

- Access and other rights (e.g. copyright, embargoes) are ideally documented in preservation systems

- Tracking, documentation of rights key part of repository certification standards

Present ➡ **Future**

Rights for Host

- Permissions for metadata augmentation/improvement

- Permission for platform migration as service develops

- Permissions for future deselection decision making

- Permission for host/steward to migrate formats as needed, perform other preservation functions

- Is supplied software truly open? If so, it could be used as part of an emulation strategy

Present   **Future**

Takeaways

- Principles of reusability also broadly support preservation

- Putting the effort in at the time of deposit and curation ensures that technical debt is not deferred to a point when it cannot be easily repaid

- Consider the preservation prospects of a dataset that:
    - Is poorly documented and contextualized
    - Was not appraised/selected for long-term research value
    - Has files in closed or proprietary formats
    - Has unclear or undocumented rights

- Will the data steward prioritize it in terms of long-term resource commitments for preservation and access?

Thank you!

Shahira Khair

skhair@uvic.ca

Grant Hurley

grant@scholarsportal.info

City of Toronto Archives, "Dive from shore No.1 (slide)," fonds 200, Series 372, Subseries 41, Item 619