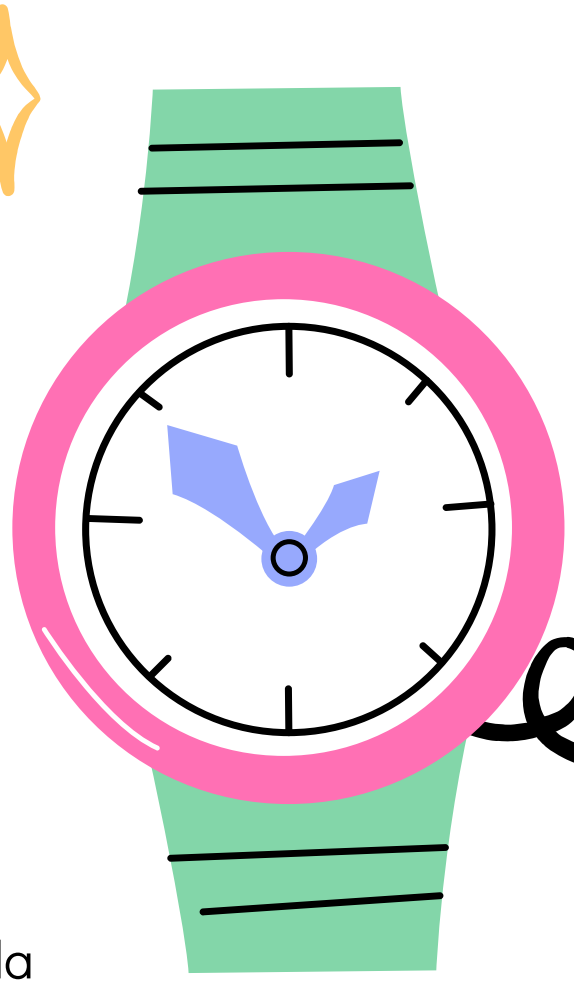
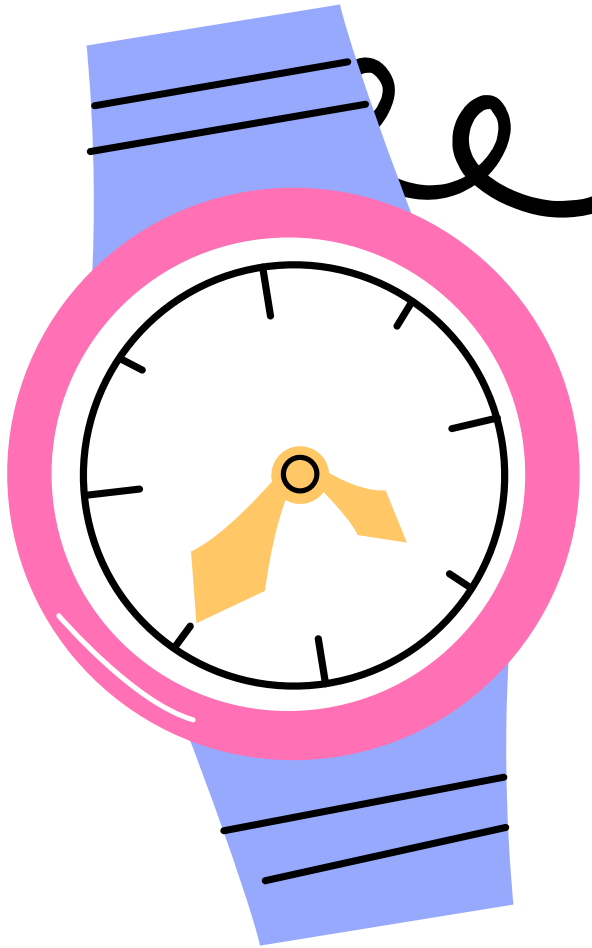
# Tick, Tock, Your Data's on the Clock



Erin Clary, Digital Research Alliance of Canada
Beth Knazook, Digital Repository of Ireland
Mikala Narlock, Data Curation Network / University of Minnesota

September 4 2024



# Background

# DCN Research

University of Minnesota
NYU
DRYAD
University of Michigan
Illinois
The Michael J. Fox Foundation for Parkinson's Research
Johns Hopkins University
University of Colorado Boulder
Cornell University
University Libraries Virginia Tech
Libraries University of Wisconsin - Madison
Princeton University
PennState
University of Virginia Library
Penn Libraries University of Pennsylvania
Duke University
Washington University in St. Louis University Libraries
University of Nebraska Lincoln

## Preservation

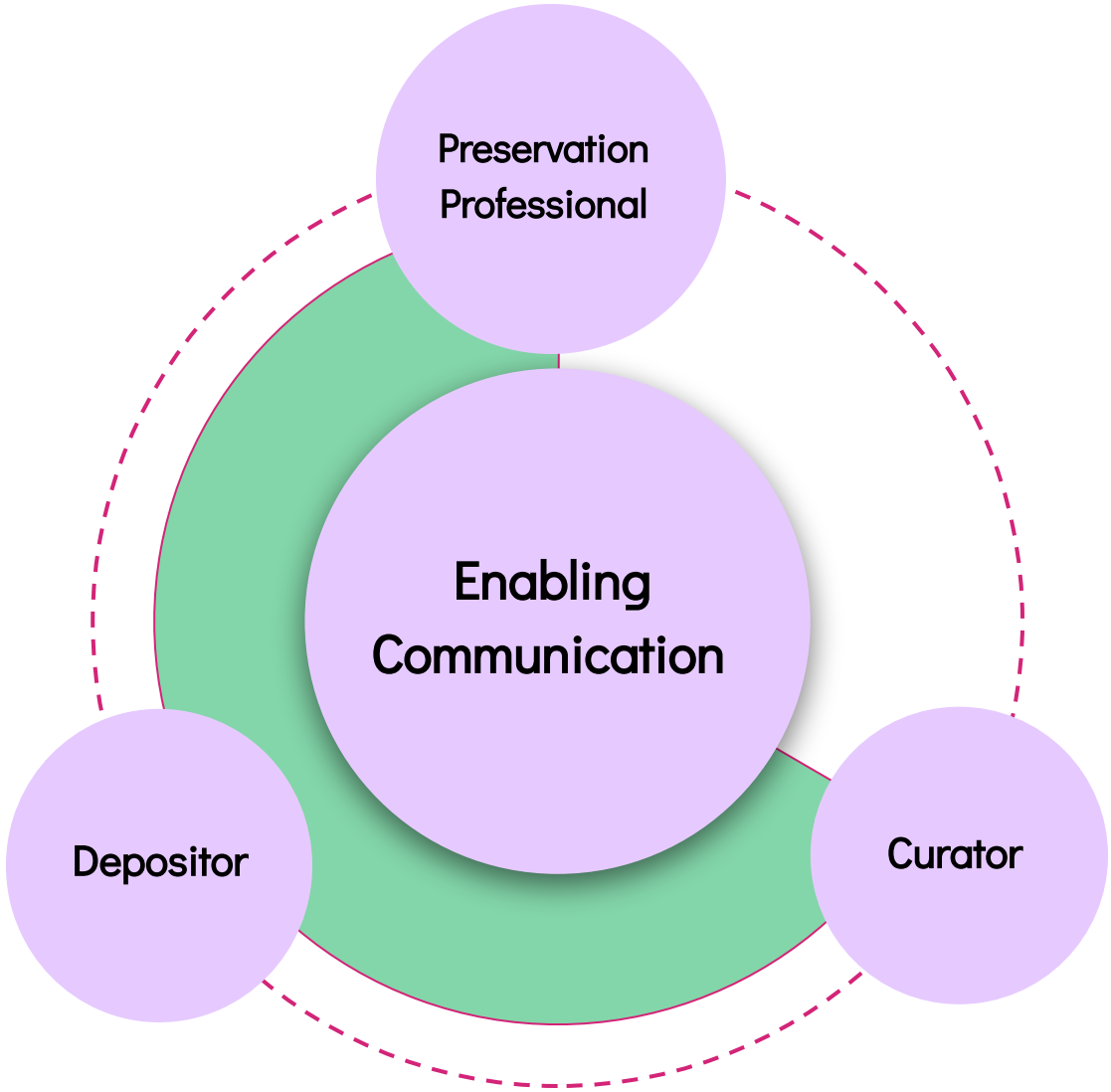
Data

Repository

Published paper describing research:
https://doi.org/10.17605/OSF.IO/HMZER

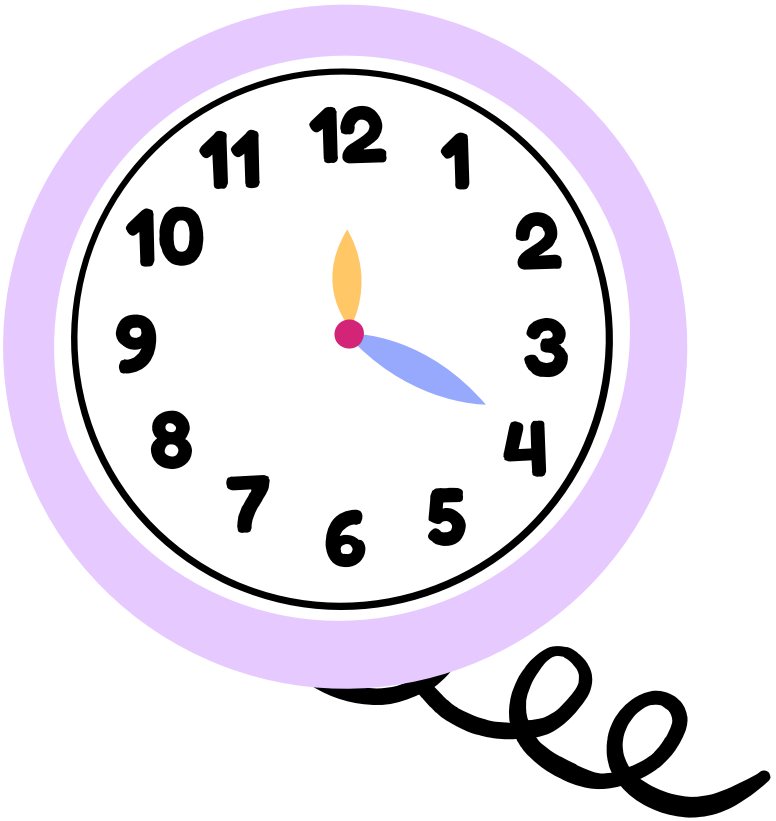
# What are the gaps in repository appraisal?

- Appraisal functions are split between acquisition and preservation
- Acquisition prioritizes the immediate usability of files
- The actual work of preservation does not influence acceptance decisions



# Research Questions

- Preservation activities currently being taken on research data?
- What is the average size of the dataset being submitted to the repository?
- Does the repository charge to deposit or preserve the data?
- Is there a stated data retention period or deaccession policy?
- Is software preserved alongside the data?



# Challenges


Increasingly complex datasets (e.g., blended software, code, and other data formats)

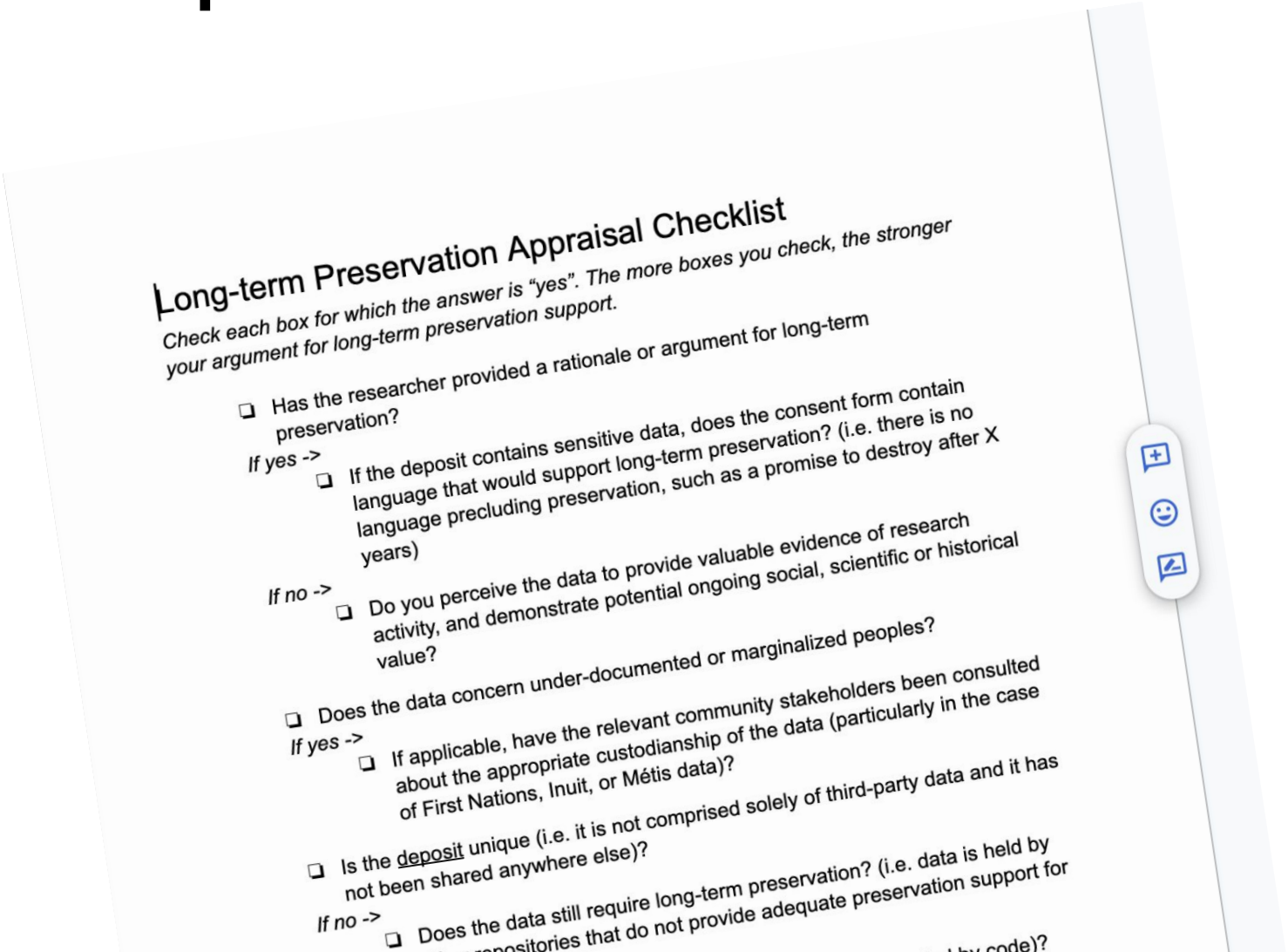
Standard metadata for long-term, large-scale preservation efforts

Review and retention policies

# Enter the appraisal checklist

# The Curatorial Preservation Review: Answering existential questions with a checklist

- How much effort are these data worth?
- Am I deleting a significant historical record if I choose not to migrate?
- Is anybody else keeping this stuff?
- What was the depositor thinking?! This could have been a spreadsheet.
- etc.!



Long-term Preservation Appraisal Checklist

*Check each box for which the answer is "yes". The more boxes you check, the stronger your argument for long-term preservation support.*

❑ Has the researcher provided a rationale or argument for long-term preservation?

*If yes ->*

❑ If the deposit contains sensitive data, does the consent form contain language that would support long-term preservation? (i.e. there is no language precluding preservation, such as a promise to destroy after X years)

*If no ->*

❑ Do you perceive the data to provide valuable evidence of research activity, and demonstrate potential ongoing social, scientific or historical value?

❑ Does the data concern under-documented or marginalized peoples?

*If yes ->*

❑ If applicable, have the relevant community stakeholders been consulted about the appropriate custodianship of the data (particularly in the case of First Nations, Inuit, or Métis data)?

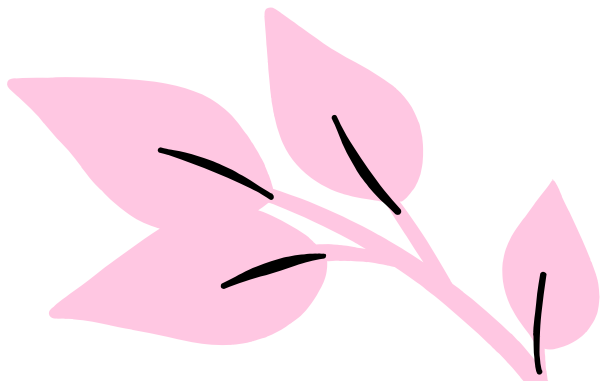
❑ Is the deposit unique (i.e. it is not comprised solely of third-party data and it has not been shared anywhere else)?

*If no ->*

❑ Does the data still require long-term preservation? (i.e. data is held by repositories that do not provide adequate preservation support for

# Modifiable checklist

http://doi.org/10.5281/zenodo.11371489



## Accompanying guidance:

Dorey, Jonathan, Hurley, Grant, and Knazook, Beth. (2022). **Appraisal Guidance for the Preservation of Research Data.** Zenodo. https://doi.org/10.5281/zenodo.5942236

# Structure of the checklist

## SECTION 1

Questions about what the researcher intends, and what the repository and its designated community values.

.

- These questions share a lot in common with existing curation guidelines and are intended to be answered ‘in the moment’: Are these data valuable *now*? Should we collect and steward them?
- It builds a rationale for any work that may be needed to preserve these datasets in the future.

## SECTION 2

Questions about the ease of preservation for the repository.  Is your organization prepared for the ongoing human, technological, and environmental costs required to preserve the data?

- These questions ask the curator to assess whether a researcher has complied with guidance on providing open or preferred datasets.
- It asks the curator to consider whether the costs of the dataset (in terms of storage, effort, and reappraisal) are justified by the value proposition presented in section 1.

# Structure of the checklist

RESULT

## Overall Preservation Recommendation

*Count the number of 'yes' answers to help inform your recommendation.*

| ❑ YES | ❑ NO |
| --- | --- |

- Needs to be calculated into long-term storage costs
- Should be included in long-term software and technical development forecasts
- Social investment made to open reappraisal to community consultation

- Store as long as it is useful
- Advise researchers of recommendation and timeframe for deletion
- **If uncertain, reappraise in the future**

# Questions

| Criteria | Questions |
| --- | --- |
| Preservation intent | • Is long-term preservation an important factor in the researcher's decision to deposit with your repository? |
| Relevance to mission | • Do the materials (continue to) meet your institution or repository's acquisition mandate or priorities, collection policy, domain specialty, or other priorities? |
| Value | • Does the data concern under-documented or marginalized peoples?<br>• Do you perceive the data to provide valuable evidence of research activity, and demonstrate potential ongoing social, scientific or historical value? |
| Uniqueness | • Is the deposit unique (i.e. it is not comprised solely of third-party data and it has not been shared anywhere else)<br>• Is the data unique (i.e. it is not the result of a model or generated by code)? |

# Questions

| Criteria | Questions |
| --- | --- |
| Cost / Economic case | • Is the dataset of a reasonable size (not very large or complex nor organized into many folders, sub-folders, and files)?<br>• Are the data easy to access via most personal computers and/or are platform-independent? |
| Rights & restrictions / Potential for redistribution | • Has the researcher assigned an open license to the files? |
| Preservability of content and context / Full documentation | • Are the data files well-documented (i.e. there is sufficient information to ensure that the files will be correctly interpreted over time, including a README, clear description of methodology and variables, and/or links to scholarly publication)?<br>• Have the data been provided in your repository's preferred or accepted file formats, supporting both preservation and access |

# Zen and the Art of Informed Reappraisal

The checklist is **NOT** a means of funnelling some portion of your datasets to the dumpster. (ugh, this one on the trash heap for sure)

The checklist is **NOT** an excuse for poorly curated datasets. (well, we're not keeping them anyway!)

—

You **CAN** reevaluate in light of changing community needs because you've recorded your rationale.

You **CAN** make the argument to preserve the really valuable stuff (because you know what that stuff is).

The checklist is a tool for mindful decision-making. It acknowledges data deaccession as a necessary and responsible tool for managing access to high-quality data.

# Questions?

Thank you!
Mikala, Beth & Erin