A Curated List of Appraisal Challenges

DPC Clinic - Reappraising Appraisal - September 04, 2024

Erin Clary and Amanda Tomé

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Introductions

- Erin Clary, Curation Coordinator
  - Along with a team of curators, review new data deposits submitted to the Federated Research Data Repository (FRDR) and provide support to researchers before, during and after deposit.
- Amanda Tomé, Preservation Coordinator
  - Responsible for all preservation activities related to deposits in FRDR, including appraisal and retention of datasets.

FRDR
Federated Research Data Repository

DFDR
Dépôt fédéré de données de recherche

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Agenda

01. What is FRDR?

02. High-level Overview of Curation & Preservation workflow

03. Current Appraisal Process

04. Challenges

05. Next Steps

What is FRDR?

The Federated Research Data Repository (FRDR) is a bilingual publishing platform for sharing and preserving Canadian research data.

- National service, launched in full production in Feb 2021.
  - Open to Canadian researchers in any discipline
  - Collections for research groups, labs, and cross-institutional research programs
  - Purpose-built for medium- to large-sized data
  - Default storage allocation of 1TB
- Initially a collaboration between the Canadian Association of Research Libraries and the Compute Canada Federation
- Funding now administered through the Digital Research Alliance of Canada

Currently ~ 502 published datasets, 247 TB

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

What is FRDR?

Our mission is to facilitate the findability, accessibility, and reuse of Canadian research data, and to act as a leader in the stewardship of Canadian research data

- General-purpose repository option
  - No restrictions on file type or format
  - Some deposits include code or software
- Open access repository (temporary embargo allowed)
  - Participating in national Controlled Access Management initiative
- Submitters and their licensors retain all ownership rights
  - Creative Commons license options (default), or custom license or terms of use, if necessary
- Curation and Preservation services

Curation & Preservation Services

Curation

- DCN CURATE(D) steps
- Check & Understand files
- Screen for sensitivities
- Licensing & attribution
- Assess FAIRness
- File ID, file tree, virus scan
- Curation log

Updates

- Only with approval of depositor(s)
- Curator: documentation, metadata, file names, structure, addition or removal of files, link to related outputs
- Depositor (preferred): changes to content, format transformations

Appraisal

- Evaluate data for long-term preservation
- Depositor contributes: Preservation Q
- Curator contributes: Appraisal checklist

Repository Storage

- Dissemination copy
- Medium-term, beyond duration of project (10+ years)
- Discovery and access
- Bit-level preservation
- Geographically distributed copies

Archival Storage

- Preservation copy
- Long-term
- Disaster recovery
- Active monitoring, management, risk assessment
- Reappraisal over time

What is Appraisal?

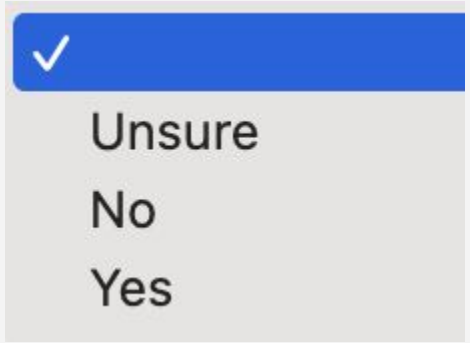
“Appraisal for preservation is the process of determining whether a dataset has sufficient long-term archival value to merit the work of monitoring, managing, storing and sustaining access to that data, as well as related systems and workflows, persistently over time.”

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

Long-term Preservation Question

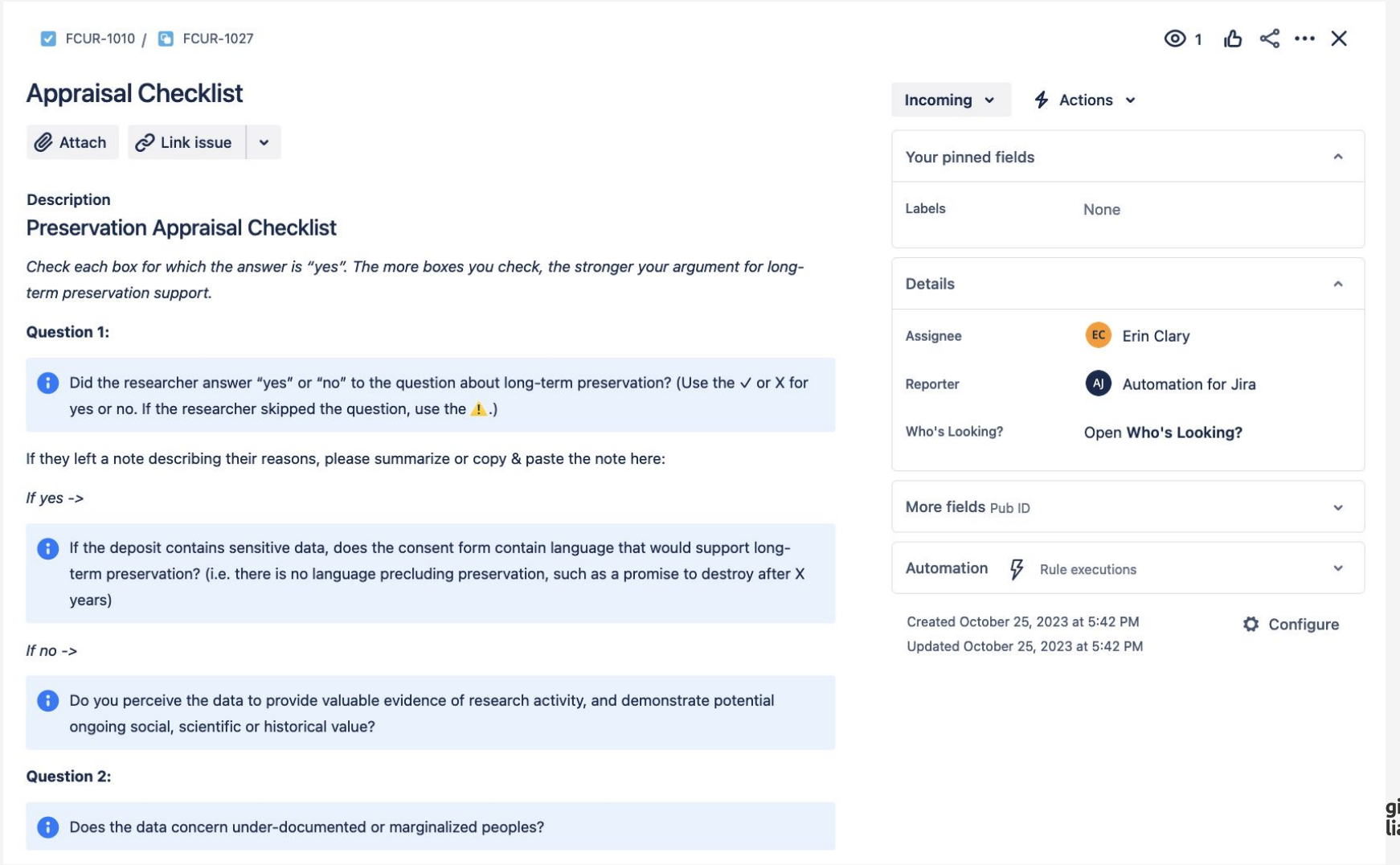
All datasets submitted to FRDR will be publicly available for at least 10 years. Some datasets with long-term value (more than 10 years) will be preserved for long-term access. If you think your dataset should be retained for the long-term, you are welcome to participate in the appraisal process. Please leave a comment here indicating, for example, potential ongoing social, scientific, or historical value.

Do you intend for this dataset to be preserved longer than 10 years?



Comment: _____

Appraisal Checklist for Curators



FCUR-1010 / FCUR-1027

1

Appraisal Checklist

Attach | Link issue

Description

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.

Question 1:

Did the researcher answer "yes" or "no" to the question about long-term preservation? (Use the ✓ or X for yes or no. If the researcher skipped the question, use the ⚠️.)

If they left a note describing their reasons, please summarize or copy & paste the note here:

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?

Question 2:

Does the data concern under-documented or marginalized peoples?

Incoming | Actions

Your pinned fields

Labels: None

Details

Assignee: EC Erin Clary

Reporter: AJ Automation for Jira

Who's Looking? Open **Who's Looking?**

More fields Pub ID

Automation | Rule executions

Created October 25, 2023 at 5:42 PM
Updated October 25, 2023 at 5:42 PM

Configure

Ease of Preservation

Preservation is not a one-size-fits-all activity and may be carried out differently depending on a variety of factors. This chart is meant to help you advise the researcher how well you can preserve the deposit, based on the choices they've made and your institution's preservation capabilities.

-All file formats are open and in formats already managed by the archive

-The data has been issued under an open license and is easily shared

Recommendation: Use existing preservation workflows

-Some file formats are open but there may be issues preserving all the content

-The license selected is somewhat restrictive

Recommendation: Preserve using existing workflows where possible; Re-appraise at a later date

-The data might be best preserved by the community it concerns

-The deposit contains sensitive data and the consent forms do not clearly support preservation actions

Recommendation: Seek guidance

Challenges

Research Domains / File Formats / Understanding Appraisal and Preservation / Capacity

Research Domains

- Agricultural biotechnology and food sciences
- Agriculture, forestry, and fisheries
- Basic medicine and life sciences
- **Biological sciences**
- Chemical sciences
- Civil engineering, maritime engineering, and mining engineering
- Clinic medicine
- Computer and information sciences
- Health sciences
- **Earth and related environmental sciences**
- Economics and business administration
- Electrical engineering, computer engineering, and information engineering
- Environmental biotechnology
- Environmental engineering and related engineering
- History, archaeology and related studies
- Humanities and the arts
- Materials engineering and resources engineering
- Mathematics and statistics
- Mechanical engineering
- Medical and biomedical engineering
- Nanotechnology
- Other natural sciences
- Other social sciences
- Political science and policy administration
- **Physical sciences**
- Psychological and cognitive sciences
- Social and economic geography

Based on Canadian Research and Development Classification (CRDC)

Research Domains - Class and Subclasses

- Space and solar physics
- Cryosphere process
- Cloud physics
- Combinatorics and discrete mathematics
- Paleobiology
- Biomolecular and medicinal chemistry
- Nanotechnology
- Computational mathematics

Fukś, H. (2024). Magic squares of subtraction of order 4. Federated Research Data Repository. https://doi.org/10.20383/103.0903

Research Domain - Dataset Examples

- SuperDARN 2022 RAWACF
- Magic squares of subtraction of order 4
- Data file for "Polycyclic aromatic hydrocarbons alter the hepatic expression of genes involved in Sanderling (Calidris alba) pre-migratory fuelling"
- Cranidial and pygidial measurements, and cranidial landmark data of Sahtuia carcajouensis and Mackenzieaspis parallelispinosa

Research Domain Challenge

If we don’t understand the work being carried out in the various research domains, how can we predict current value and future reuse of the data?

File Formats

- Preferred formats
- Migration can lead to significant loss
  - Example: Rendering of files with mathematical equations
- Don’t have a good handle on research data formats
- Some research relies on proprietary software
- Some researchers develop their own file formats


vno


oeb


fastq


qmd

File Formats Samples

Formats	Domain
.las - Lidar	Geometric and earth systems observation, Surface water hydrology
.nii - NIfTI	Central Nervous System
.rawacf - SuperDARN	Solar Physics
.tilt - Tilt Brush	Arts (arts, history of arts, performing arts, music), architecture and design
.bw - BigWig	Epigenetics and epigenomics
.mif - MRtrix Image Format	Other natural sciences

File Format Metrics Results

- Results from file format metric scan (March 2024)
  - 6 million files in FRDR
  - 1,186,213 where identified as ‘unknown’ – corresponds to 332 file formats
  - Misidentified files ~ 350,000 files (usually misidentified as plain text)
  - 400 file formats that are unknown

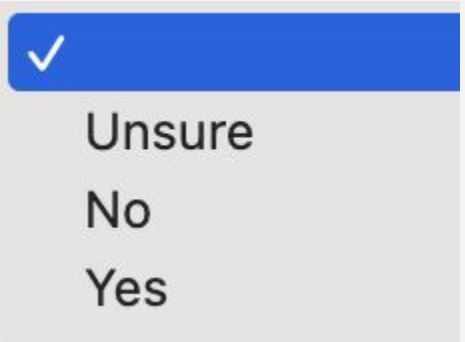
File Format Challenge

How can we successfully weigh long term reproducibility of the data in the appraisal process if there are so many file formats that are not understood, if migration pathways will not suffice, and emulation capabilities are limited whether from licensing and copyright challenges or because of the technical infrastructure and knowledge needed to perform this work?

Researcher Participation in Appraisal

All datasets submitted to FRDR will be publicly available for at least 10 years. Some datasets with long-term value (more than 10 years) will be preserved for long-term access. If you think your dataset should be retained for the long-term, you are welcome to participate in the appraisal process. Please leave a comment here indicating, for example, potential ongoing social, scientific, or historical value.

Do you intend for this dataset to be preserved longer than 10 years?

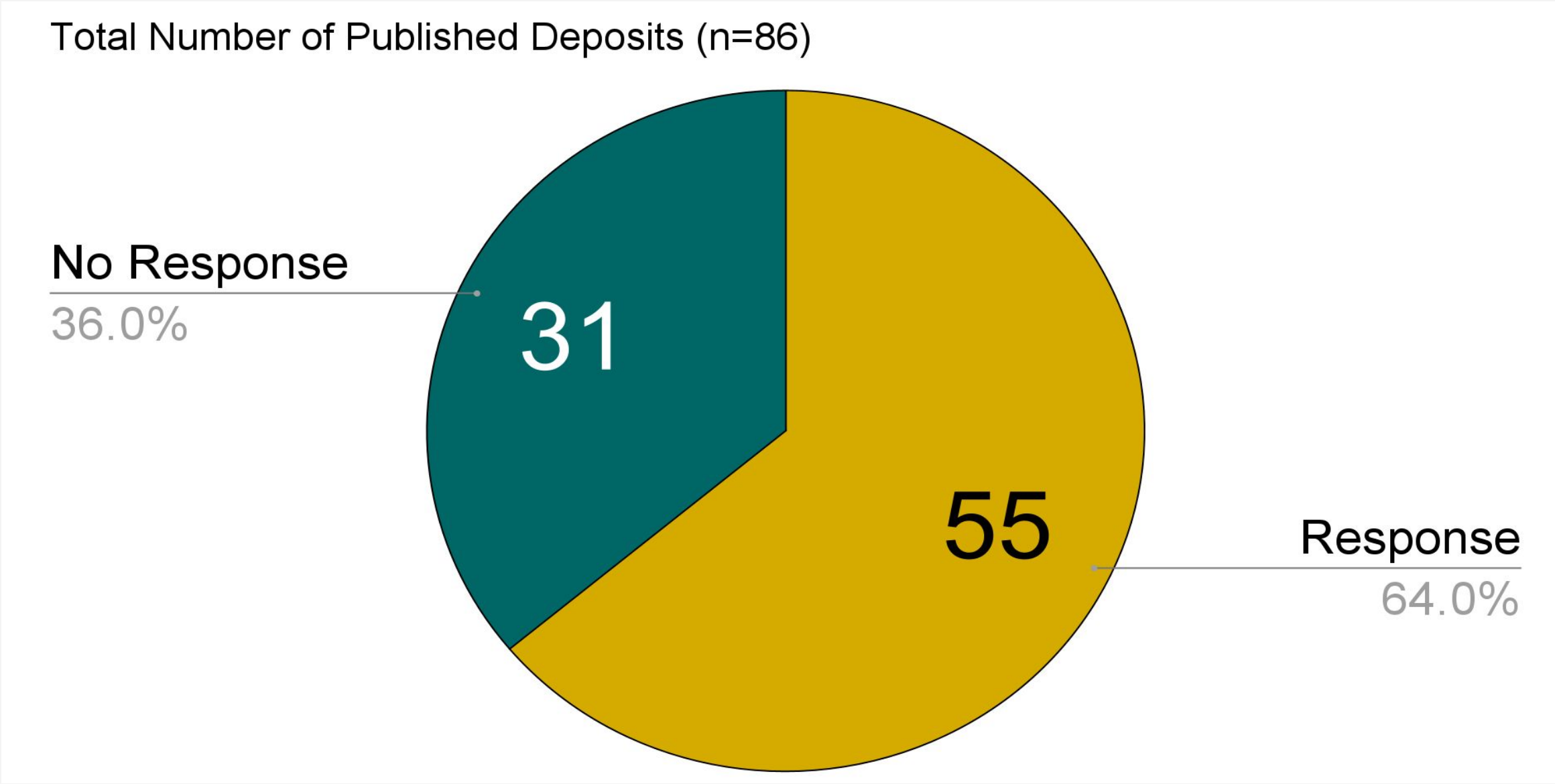


Comment: _____

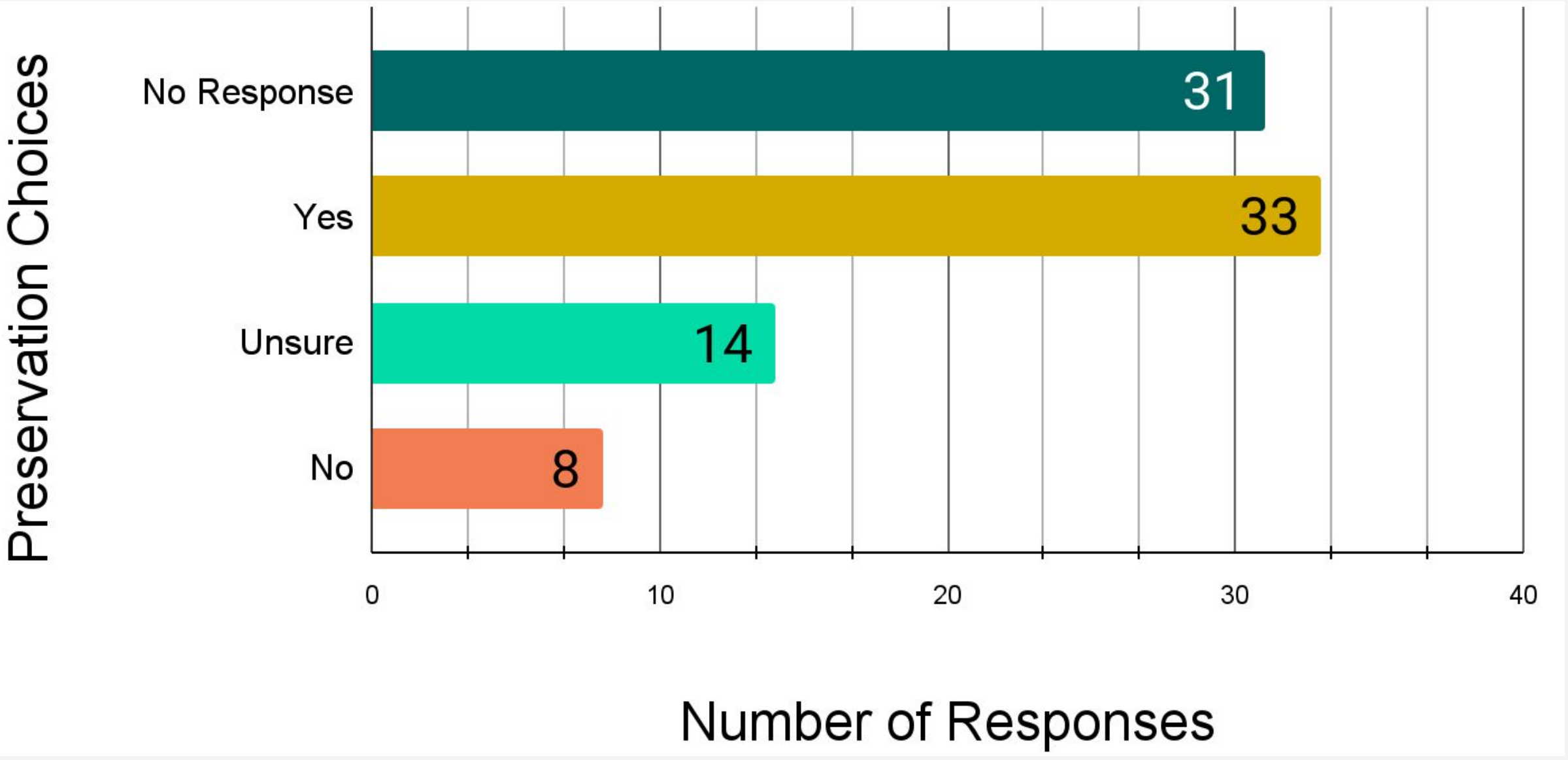
Initial Thoughts

1. Do researchers understand what preservation and appraisal is?
2. Do researchers believe that preservation only starts after 10 years?
3. Not all preservation capabilities are the same, should this information be provided?
4. Long-term isn't defined. Does it need to be?
5. What happens if we disagree with the researcher's choice?
6. All researchers would want their data to be kept and remain accessible in the future.

Researcher Engagement: 2023-05-04 through 2024-04-30



Breakdown of Responses



Researcher Sample Comments

YES	UNSURE	NO
Modelling output spanning 70 years could support various meteorological and hydrological studies	Ideally remain available for reproducibility, but would be possible to regenerate using code	10 years is enough to satisfy funder policy
Data acquisition required qualified personnel, expensive equipment, custom software, and a significant amount of time	Phenomenon not newly observed, but associated paper presents new theories to describe this phenomenon	May remain useful, but will likely have been modified or extended, and available in another form
Some minerals analyzed are very rare and may not be able to be analyzed again	Hope these results will establish guiding principles that will remain important for more than 10 years	In current form, only useful for verifying study results
Dataset is intended to serve as a benchmark to compare models in a standardized setup	Data does not document an experiment, but is intended to be used as reference material	Content unlikely to be relevant after 10 years
Used to support climate research which requires long-term data	It's difficult to forecast long-term value	Long-term access not required

Researcher Participation in Appraisal Challenge

- Are we engaging enough with our researchers about their data and providing enough educational resources and information about what happens to their data once deposited?
- Is the long-term preservation question itself causing more challenges or giving us important information?

Capacity

- All previous challenges have lead us to the capacity challenge
- Staff compliment
- Increase capacity by liaising with subject matter experts
- Can reproducing the data lead to better appraisal?
- Balance the needs between supporting researchers and publishing data

Capacity Challenge

How do we continue to build capacity and balance the need to upgrade skills and create networks that will aid with appraisal decisions with the needs of supporting researchers who are prioritizing publication and discovery of their datasets?

Other Challenges

- Model data - what do we keep?
- Datasets with multiple licenses or custom terms of use
- Datasets comprised of data and code/software
- Data collected, obtained, scraped or derived from third party sources

Next Steps

- Develop educational material around preservation and appraisal activities for researchers, curators and preservationists
- Embed preservation at the earliest stages of research - such as in Data Management Plans
- Develop data retention schedules for research data - where possible
- Utilize the appraisal checklist in a way to facilitate discussion about long term preservation that can aid with appraisal decisions

Thank you!

Please keep in touch! <support@frdr-dfdr.ca>

FRDR
Federated Research Data Repository

DFDR
Dépôt fédéré de données de recherche

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

 alliancecan.ca

 @Alliance_Can

 /AllianceCan

 support@frdr-dfdr.ca