CANADA FOUNDATION FOR INNOVATION

MSIF Workshop Panel 4

Research facilities in a data-intensive and open-science world / Situer les installations de recherche dans un monde à forte concentration de données et à la science et aux données ouvertes

Moderator: Fred Whoriskey, Executive Director, OTN
Suzanne Beth, Principal Coordinator, Coalition Publica
Natalie Harrower, Executive Director, CRDCN / RCCDR
Kevin Krieger, Head Engineer, SuperDARN Canada

November 3, 2023

INN❂VATION
Canada Foundation for Innovation   Fondation canadienne pour l'innovation

CRDCN: pathways in an Open Science World

Dr. Natalie Harrower
Executive Director

CFI 2023 Major Science Initiatives Fund Workshop
University of Saskatchewan | 3 November 2023

CRDCN
RCCDR
Canadian Research Data Centre Network
Réseau canadien des Centres de données de recherche

The Canadian Research Data Centre Network is hosted by McMaster University, which is located on the traditional territories of the Mississauga and Haudenosaunee nations, and within the lands protected by the Dish with One Spoon wampum agreement.
I would like to give honour and thanks to the Haudenosaunee and Anishanaabe nations and to recognize our responsibility, as guests, to respect and honour the intimate relationship Indigenous peoples have to this land

Canadian Research Data Centre Network

A premier research and training platform for the quantitative social and health sciences in Canada. Est. 2000

https://crdcn.ca

1 of 19	Of Canada's Major Science Initiatives
2,400	Researchers
33	Secure data facilities
200+	Datasets
5,000+	Publications, articles, policy reports, etc.

Our Vision

To be Canada's national platform for leading edge research and training in the quantitative social and population health sciences, creating social, economic, health and environmental benefits for Canadians

Our Mission

Enable quantitative researchers to address vital questions

Train the next generation of highly qualified personnel in social and population health sciences

Contribute to evidence-based policy and decision-making by facilitating access to relevant, timely and well-documented data supported by state of the art computing facilities

THE CRDCN PARTNERSHIP

CRDCN serves as a bridge between:

- Statistics Canada
- 33 university partners
- Prospective partners in the government, non-profit and private sectors

We provide researchers with secure access to the most sensitive microdata files, facilitate analysis, provide training, and create partnership opportunities.

Statistics Canada microdata subject areas

Agriculture and food

Business and consumer services and culture

Business performance and ownership

Children and youth

Crime and justice

Digital economy and society

Economic accounts

Education, training and learning

Environment

Families, households and marital status

Government

Health

Housing

Immigration and ethnocultural diversity

Income, pensions, spending and wealth

Indigenous Peoples

Labour

Population and demography

Science and technology

Society and community

Transportation

Travel and tourism

Example data sets

Aboriginal People's Survey

Canadian Community Health Survey

Canadian Housing Statistics Program

Canadian Survey on Disability

Census (including linkage to hospital data)

COVID-19 Survey Data (vaccinations, benefits payments, Portraits of Canadians)

Educational and Labour Market Longitudinal Platform

General Social Survey

Housing and Environment Survey

Labour Force Survey

Life After Service Survey

Longitudinal Administrative Databank

Longitudinal Immigration Database

Programme for International Student Assessment

Public Service Employment Survey

Social Assistance data (Ontario and BC)

More than 200 data sets are available

CRDCN Canadian Research Data Centre Network

CRDCN
Research Policy Snapshot

Effect of heat and air quality on elderly mortalities and emergency visits

House Price Effects from Vancouver's SkyTrain Expansion

How well is cannabis legalization curtailing the Illegal market?

Impact of COVID-19 school closures on test scores

CRDCN Canadian Research Data Centre Network

Current Area of Focus: : Strategic Plan 2024-2029

Current Area of Focus: vRDC

The Virtual Research Data Centre (vRDC) is a new platform for that both centralizes the network, and provides remote access from authorized home/office work stations.

Overhaul of storage, compute and networking capabilities to optimize network capability, facilitate analysis and storage of large datasets

Designed to meet the requirements of the evolving cybersecurity landscape

CRDCN Canadian Research Data Centre Network
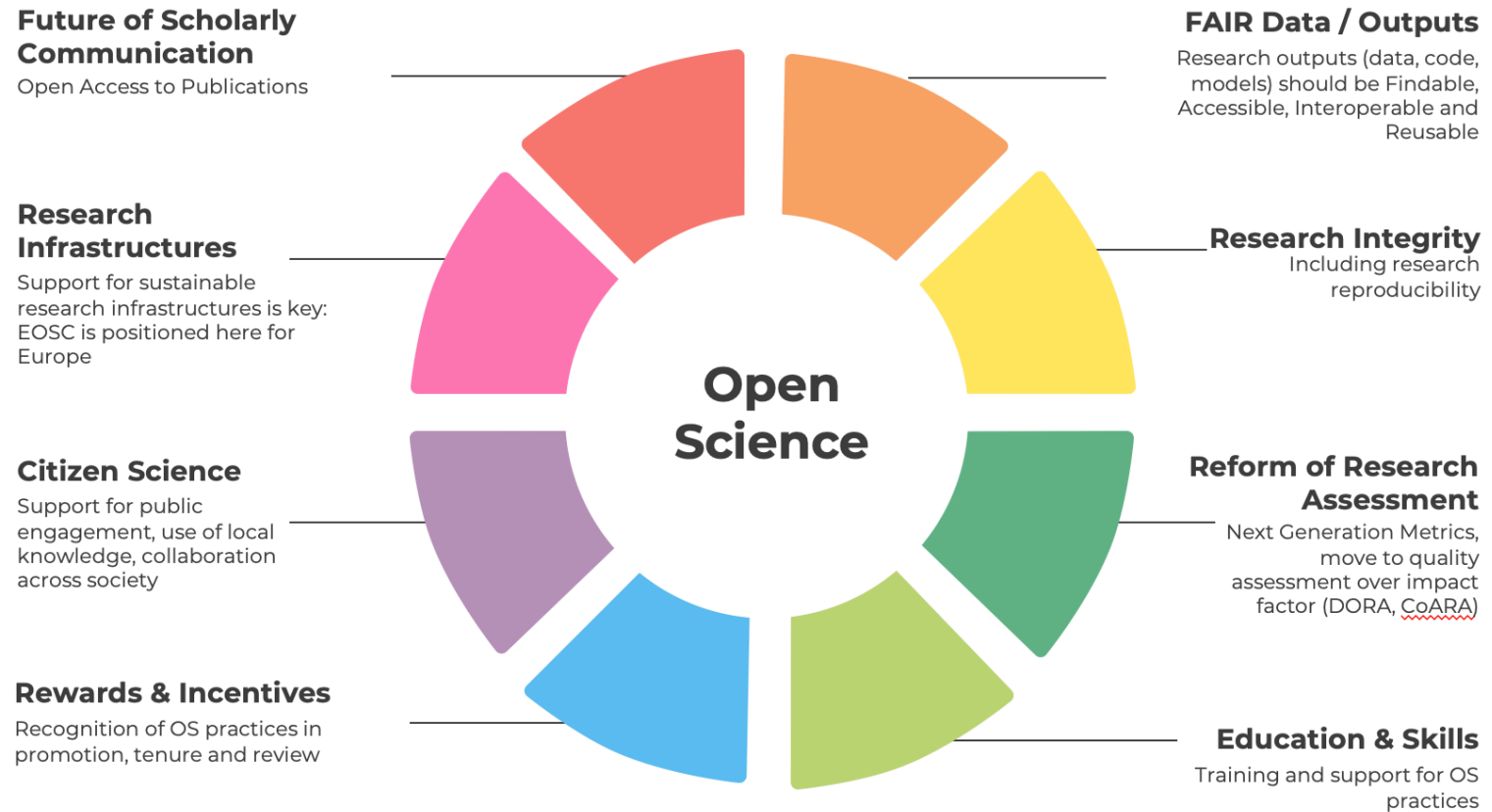
RCCDR Réseau canadien des Centres de données de recherche

CRDCN and Open Science

Adapted from the European Commission's
8 Priorities/Pillars of Open Science

FAIR Data / Outputs

Research outputs (data, code, models) should be Findable, Accessible, Interoperable and Reusable

Open Science

Facilitating data sharing is our core activity

however…

Sharing of sensitive personal data comes with significant considerations and restrictions to ensure privacy

What role can we play in the Research Data Lifecycle?

Standardisation is key --

FAIR for sensitive data

FAIR data does not equal Open Data

PIDs for sensitive data

Citation practices

Code sharing to enable collaboration, efficiencies, transparency…even to improve access process?

Support for DMPs, collaborative spaces, communities of practice

Relationship building and knowledge exchange re OCAP principles, CARE principles

Increase Access to data through innovative methods (e.g. Data visitation?)

International Collaborations

CRDCN Canadian Research Data Centre Network

What role can we play nationally?

CRDCN makes sensitive StatCan microdata available to researchers

Which also means we have a rich and varied community of researchers

What other elements of Open Science can we support?

Are there ways the MSIs can work together to collectively progress OS?

Policy, Advocacy, Support

Open Science is a complex jigsaw that requires all parts to move forward

Future of Scholarly Communication
Open Access to Publications

Research Infrastructures
Support for sustainable research infrastructures is key: EOSC is positioned here for Europe

Citizen Science
Support for public engagement, use of local knowledge, collaboration across society

Rewards & Incentives
Recognition of OS practices in promotion, tenure and review

FAIR Data / Outputs
Research outputs (data, code, models) should be Findable, Accessible, Interoperable and Reusable

Research Integrity
Including research reproducibility

Reform of Research Assessment
Next Generation Metrics, move to quality assessment over impact factor (DORA, CoARA)

Education & Skills
Training and support for OS practices

Open Science

Adapted from the European Commission's
8 Priorities/Pillars of Open Science

Leveraging Data, Research, and Policy Collaborations

Partnerships with purpose to benefit Canadians

November 21-23
Hamilton, ON

CRDCN 2023
Canadian Research Data Centre Network

crdcn.ca

Contact:

Natalie Harrower

natalie.harrower@crdcn.ca

CRDCN Canadian Research Data Centre Network
RCCDR Réseau canadien des Centres de données de recherche

Coalition Publica

→ Coalition Publica is a **partnership** between **Érudit** and the **Public Knowledge Project** (PKP) that supports the social sciences and humanities journal community in the transition towards **sustainable open access**

→ Coalition Publica is developing a non-commercial, open source **national infrastructure** for digital scholarly publishing & dissemination combining **Open Journal Systems** (OJS) and the **erudit.org** platform

1. The scholarly publication system & open access

Scholarly publishing:
a strategic sector at a tipping point

A key link in the research cycle

→ Publication of **research results** & starting point of new inquiry
→ Basis of **evaluation practices** in contemporary academia
(citations, impact factor)

Open access movement

→ Emergence in the 2000s in reaction to the "**Serials crisis**"
→ As OA's uptake is accelerating worldwide & **public policy** is
aligning with it, it is reshaping the **conditions & practices of
knowledge dissemination**

Open access & data management

Open access implies a **free circulation of research results** in order to maximize it, but it's not enough to ensure its **discoverability**


Findable
Metadata and data should be findable for both humans and computers

Interoperable
Data needs to work with applications or workflows for analysis, storage and processing

F A I R

Accessible
Once found, users need to know how the data can be accessed

Reusable
The goal of FAIR is to optimise data reuse via comprehensive well-described metadata

→ Coalition Publica offers a set of **services**, as well as **expertise,** to ensure and enhance the **discoverability and indexing** of Canadian scholarly journals in social sciences and the humanities

2.   Coalition Publica:
      a technological infrastructure

National open infrastructure

+300,000 articles
98% OA
5.7 millions unique visitors (2022)
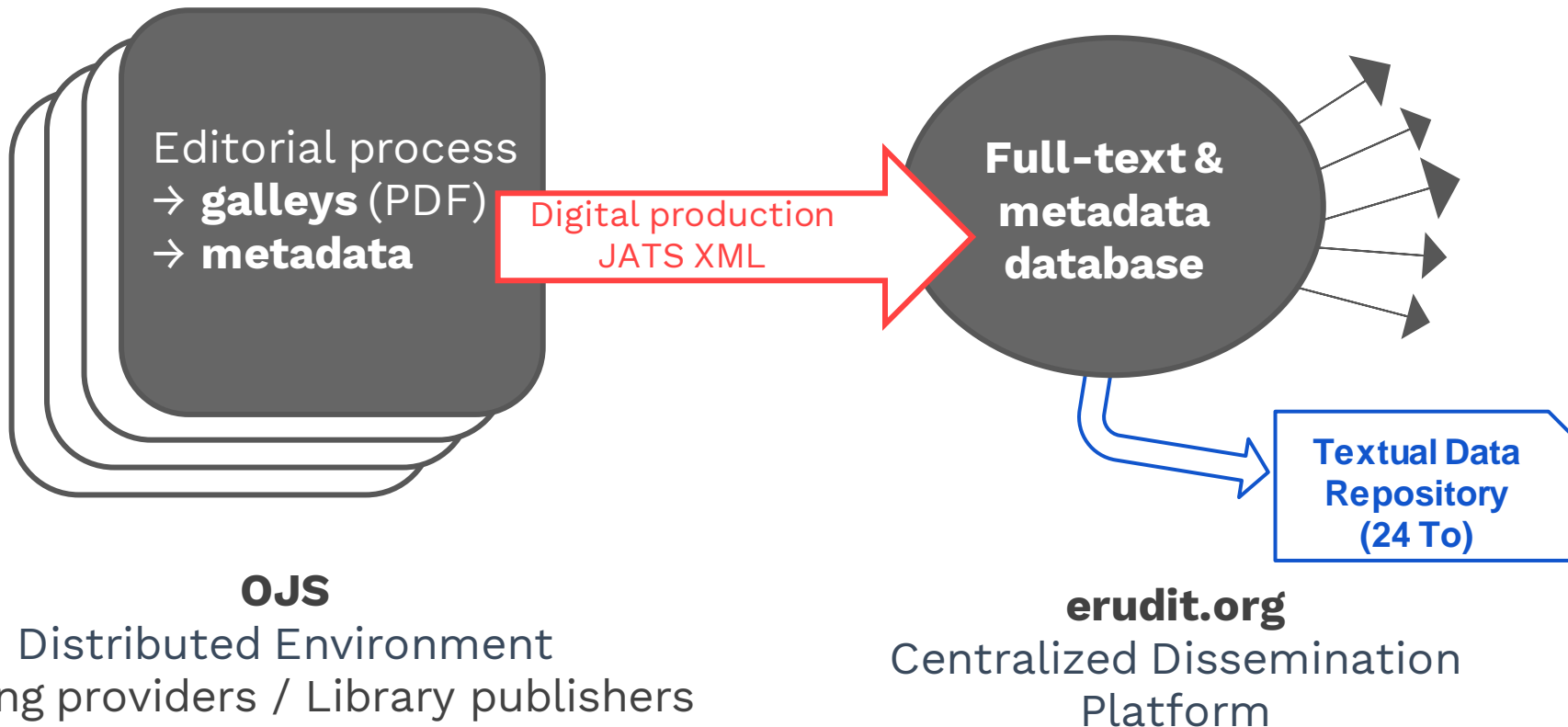
Editorial process
→ **galleys** (PDF)
→ **metadata**

Digital production
JATS XML

**Full-text &
metadata
database**

**Textual Data
Repository
(24 To)**

OJS
Distributed Environment
Hosting providers / Library publishers

erudit.org
Centralized Dissemination
Platform

Metadata curation

the bibliographic record, the ancestor of digital article metadata

Metadata link **data describing the article** (e.g. author, title, language, keywords, abstract, ISSN, volume) to the text itself, so that it can be **found, located & shared**

Metadata are standardized & can be **read both by machines (search engines) & by humans**

Metadata curation & discoverability

Metadata curation is key to widespread **indexing**, which determines articles' circulation, visibility and **discoverability**

Metadata are **harvested from OJS to erudit.org**'s tools, and **converted to metadata formats** used by other databases (**interoperability**)

→ as a centralized dissemination platform, erudit.org is a **showcase for journals**

→ it is also a **relay toward other discovery databases**, e.g. Google, Google scholar, library catalogues, Scopus, WoS, PubMed, **DOAJ** etc.

Coalition Publicas's textual data repository: a treasure trove

Natively **digital & digitized documents** from 4 main sources:

- articles disseminated on erudit.org
- Canadiana/RCDR
- Library & Archives Canada
- Bibliothèque & Archies nationales du Québec

24 To: one of the **largest French corpus** in the world

Research potential:

- **data mining**
- **large language models** training

3.   Expertise & funding:
       two other key elements

Coalition Publica: a human infrastructure

Coalition Publica's **top expertise** in digital scholarly publishing in Canada is enhanced by close **collaboration with academic libraries**

Support, documentation, training

- OJS **technical support**
- OA **best practices** (metadata, persistent identifiers, licenses)

Concertation, leadership & advocacy

→ Scholarly publication in Canada as a collaborative, **fair & sustainable ecosystem**

Canadian Research Knowledge Network
Réseau canadien de documentation pour la recherche

Fonds de recherche Société et culture
Québec

CARL ABRC

SSHRC = CRSH

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

Funding: Partnership for open access (POA)

- Elaborated with the **Canadian Research Knowledge Network** (CRKN) in 2014 (Canadian academic library consortium)

- Transformation of traditional commercial relationships (subscription sales) into an **ecosystemic collaboration**

- Historic five-year agreement 2018-2022: included **revenues for OA journals**, successfully extended for 2023-2024

- **International** expansion: POA with Couperin Consortium (France) & BICfB (Belgium)

Diamond OA, the model of Canadian journals

A 2021 study found that Canada is home to over **480 active scholarly journals** published by non-commercial entities (e.g., libraries, universities, associations etc.):

- primarily (76%) in the **humanities and social sciences**
- with **small editorial teams** of 1-6 people
- and for those publishing immediate open access, 97% (n=283) operate on a **diamond OA model**

Lange, J & Severson, S. (2021, 10 12). What Are the Characteristics of Canadian Independent, Scholarly Journals? Results from a Website Analysis. *The Journal of Electronic Publishing* 24(1) doi: 10.3998/jep.153
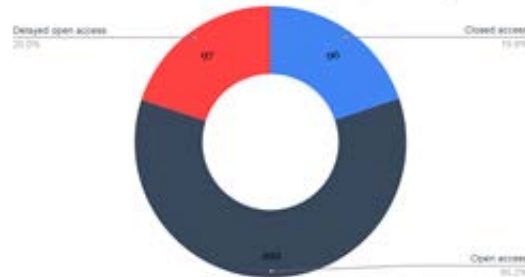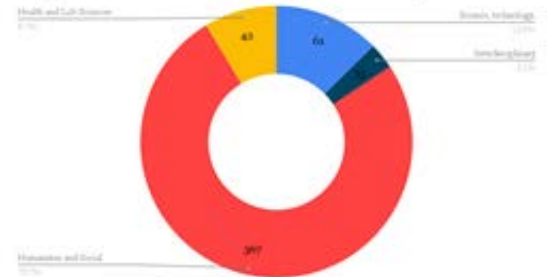
4.   A concluding note on Indigenous Knowledge dissemination

Indigenous studies & Indigenous Knowledge dissemination

→ **Indigenous studies journals** participating in Coalition Publica

Working areas include:

- Publication languages (possible but very rare)
- Metadata categories (CP Metadata Working Group's recommendations)
- Dissemination modes/media (audio/video vs article)

→ Indigenous Knowledge dissemination: issue that will require long & thoughtful **consultation with communities** – we're really at the beginning

Thank you!

Suzanne Beth
Senior Coordinator, Research & Innovation
Érudit Consortium

Data Management
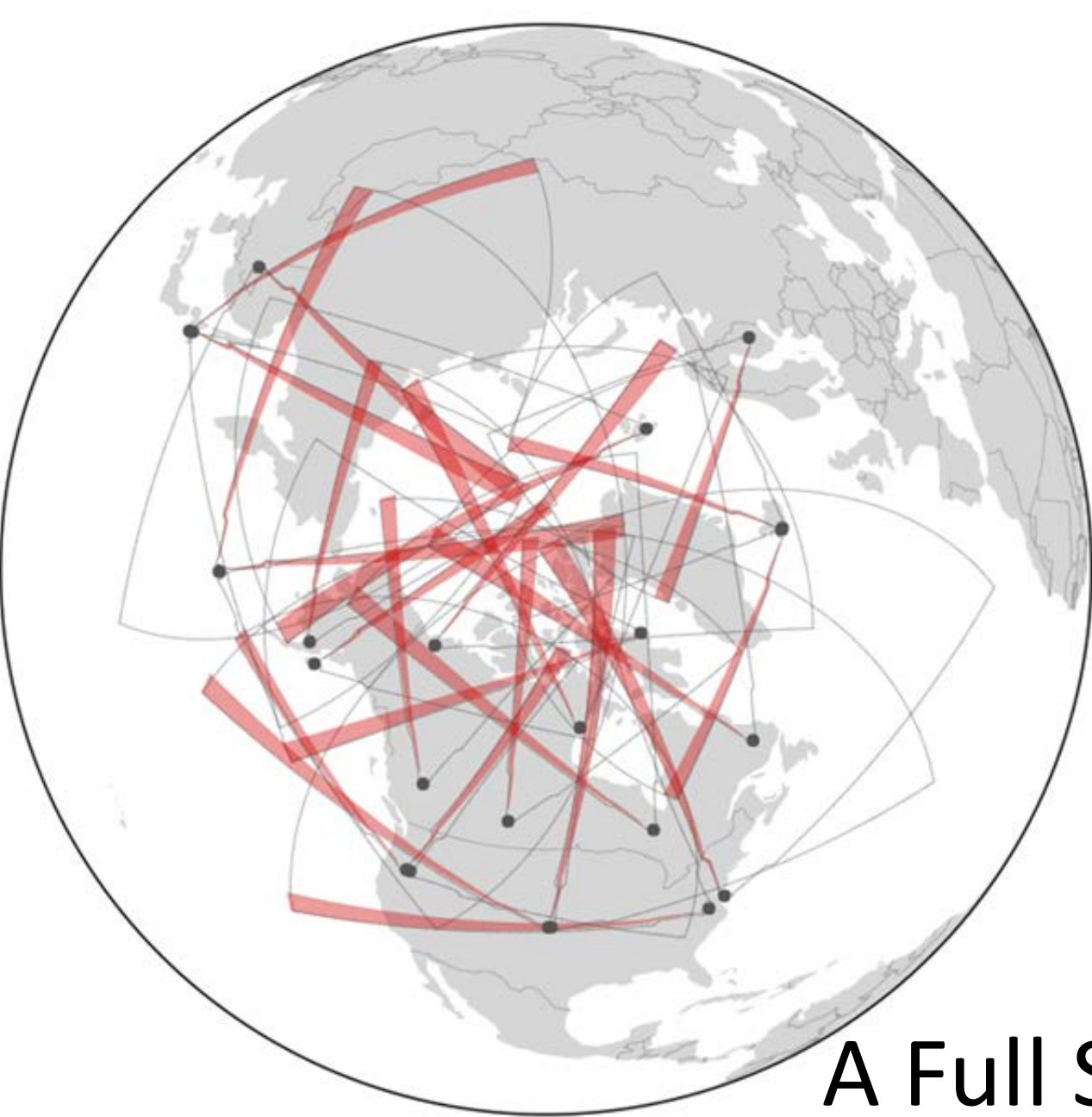## Strategies and evolution over the past decade

Kevin Krieger
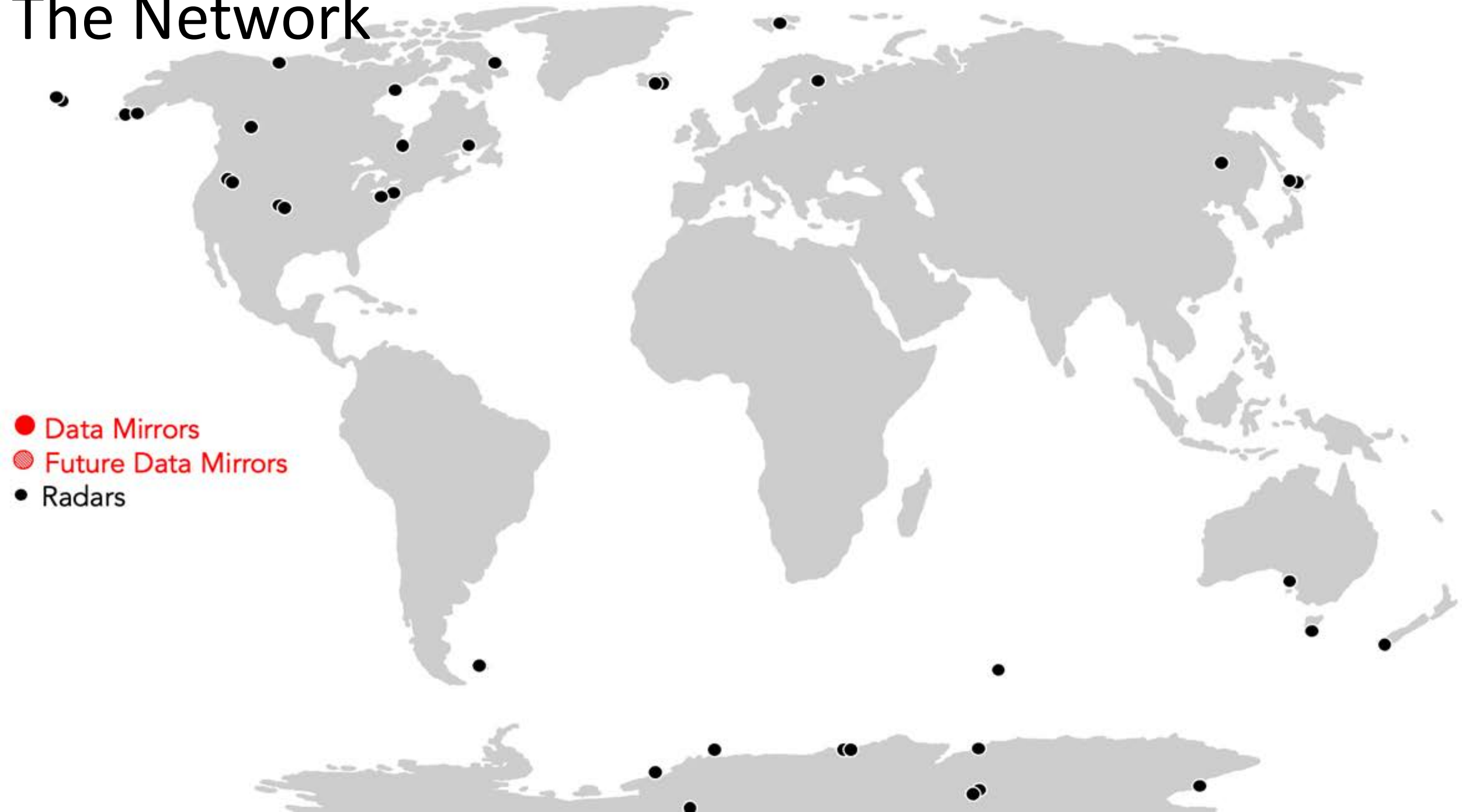November 3, 2023
MSIF Workshop, Saskatoon SK

The Facility

- 37 HF radars covering North/South
- 5 Canadian radars
- 16 Institutions
- 10 Countries
- Continuous operation since 1993
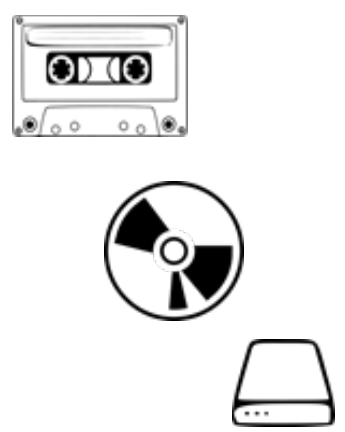- 10 GB of data per radar-month
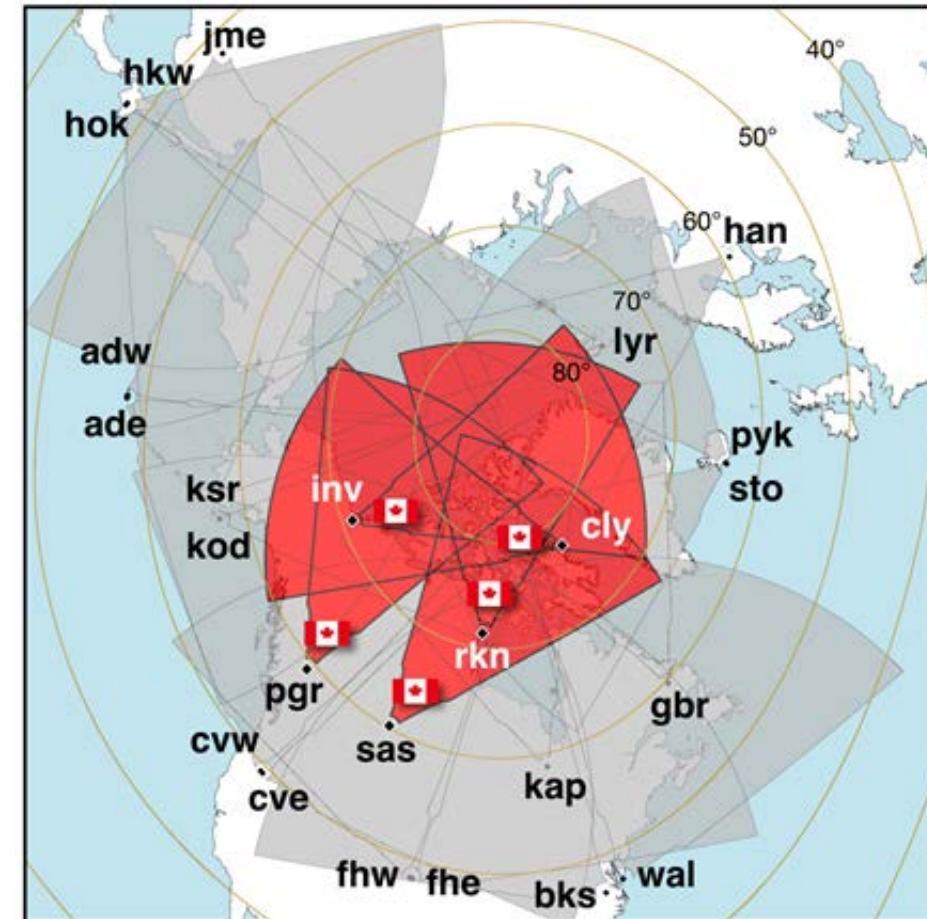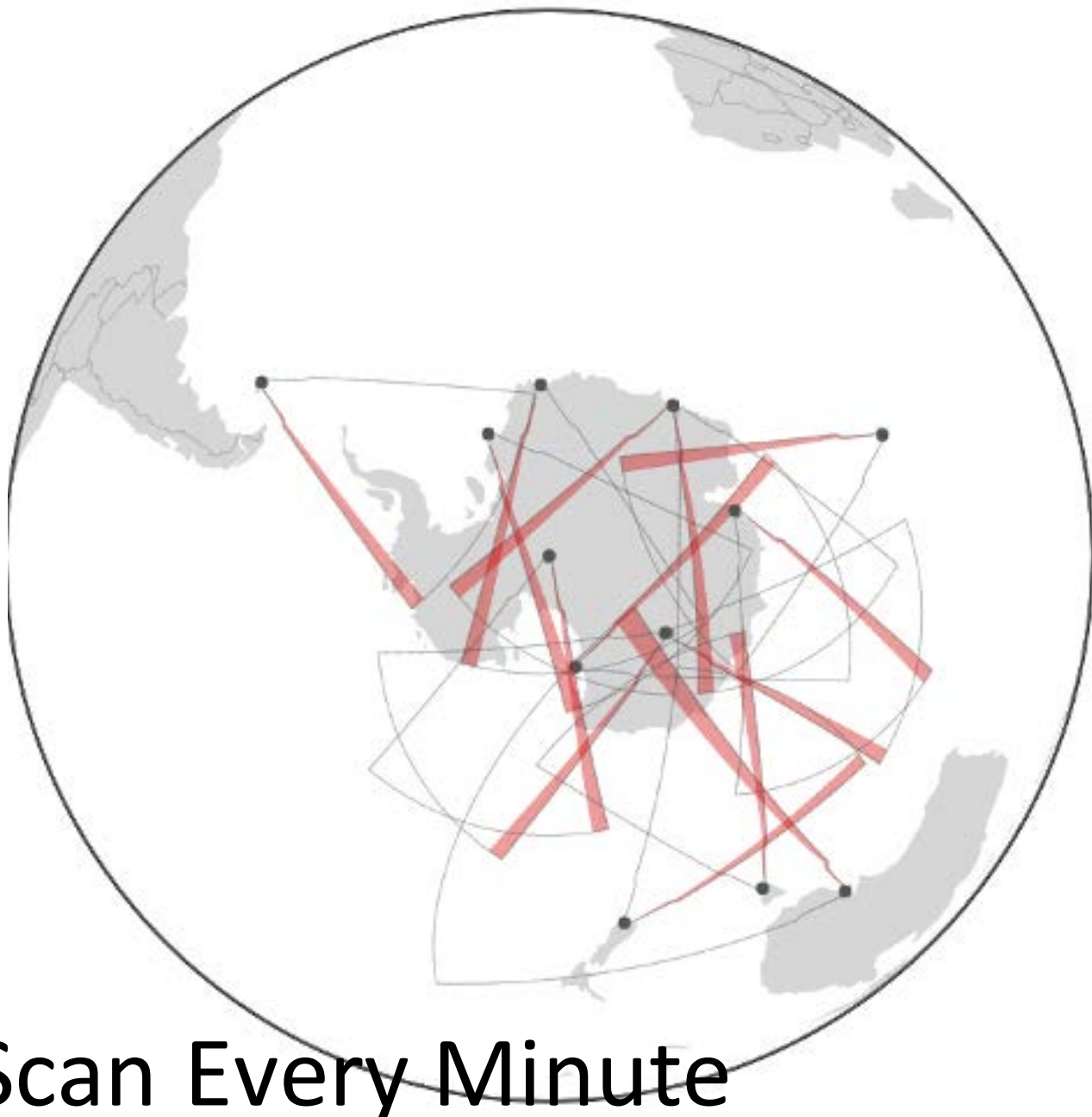
A Full Scan Every Minute

The Network

Data Mirrors
Future Data Mirrors
Radars

1993-2023: Three Decades of Data

1993092914t.dat.bz2

1993-2012	2013	2014	2015-2017	2018	2019-2023	2023-

Storage media

Working Group Online distribution

Local servers

'The Cloud'

Data loss and recovery event

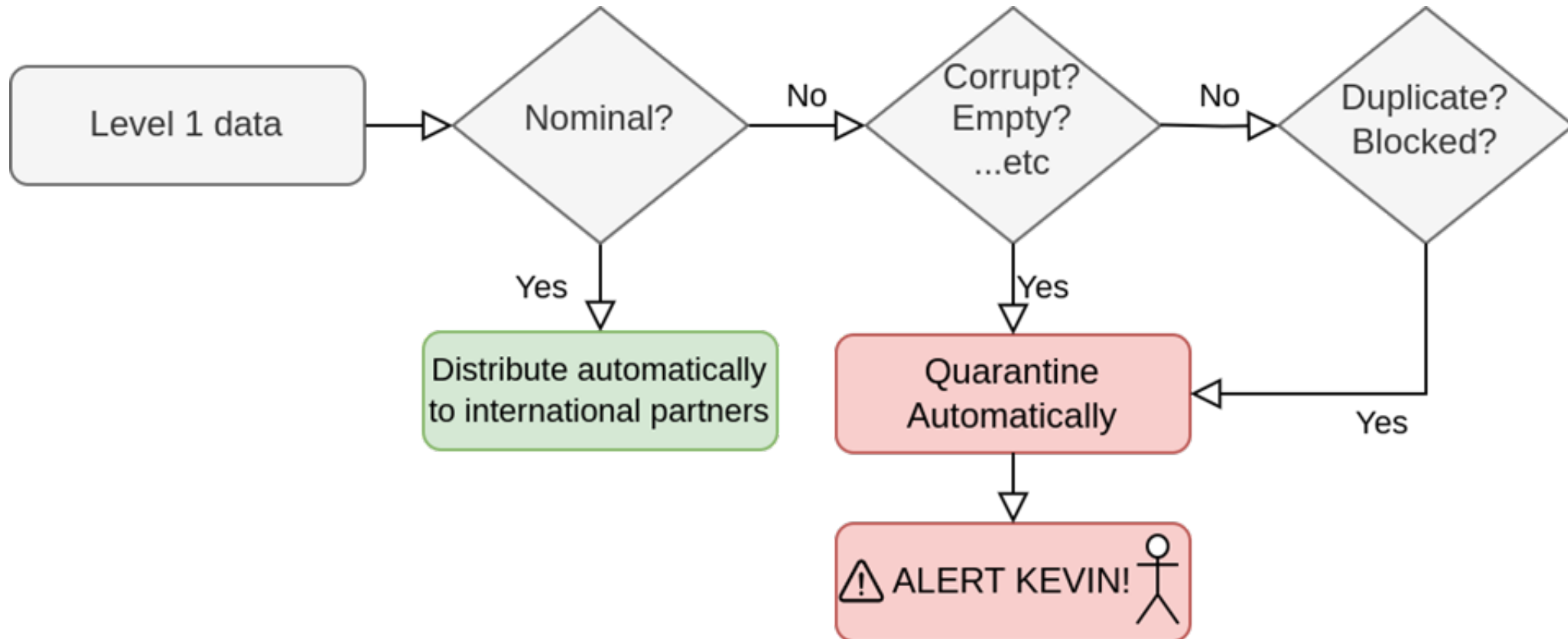**New Capabilities**

Ongoing challenges

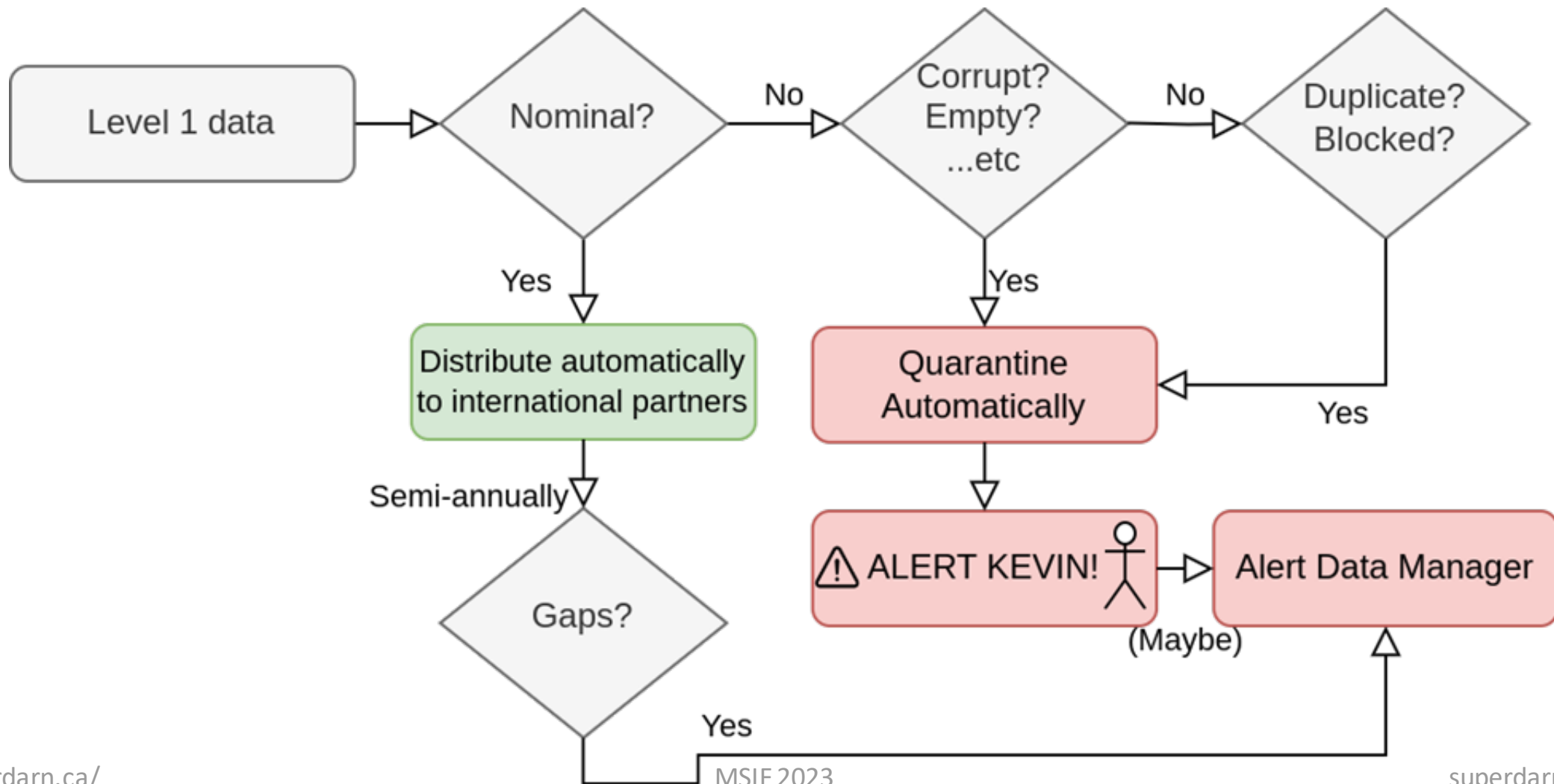# 2013/2014: Organization, Online Distribution

Challenges
- Storage media failures
- International groups
  - Different policies
  - Slow transfers
  - 'packet boating'
  - Sporadic firewalls
  - Consistency of funding
  - Limited expertise

Goal: High Quality, Open, Canonical Dataset

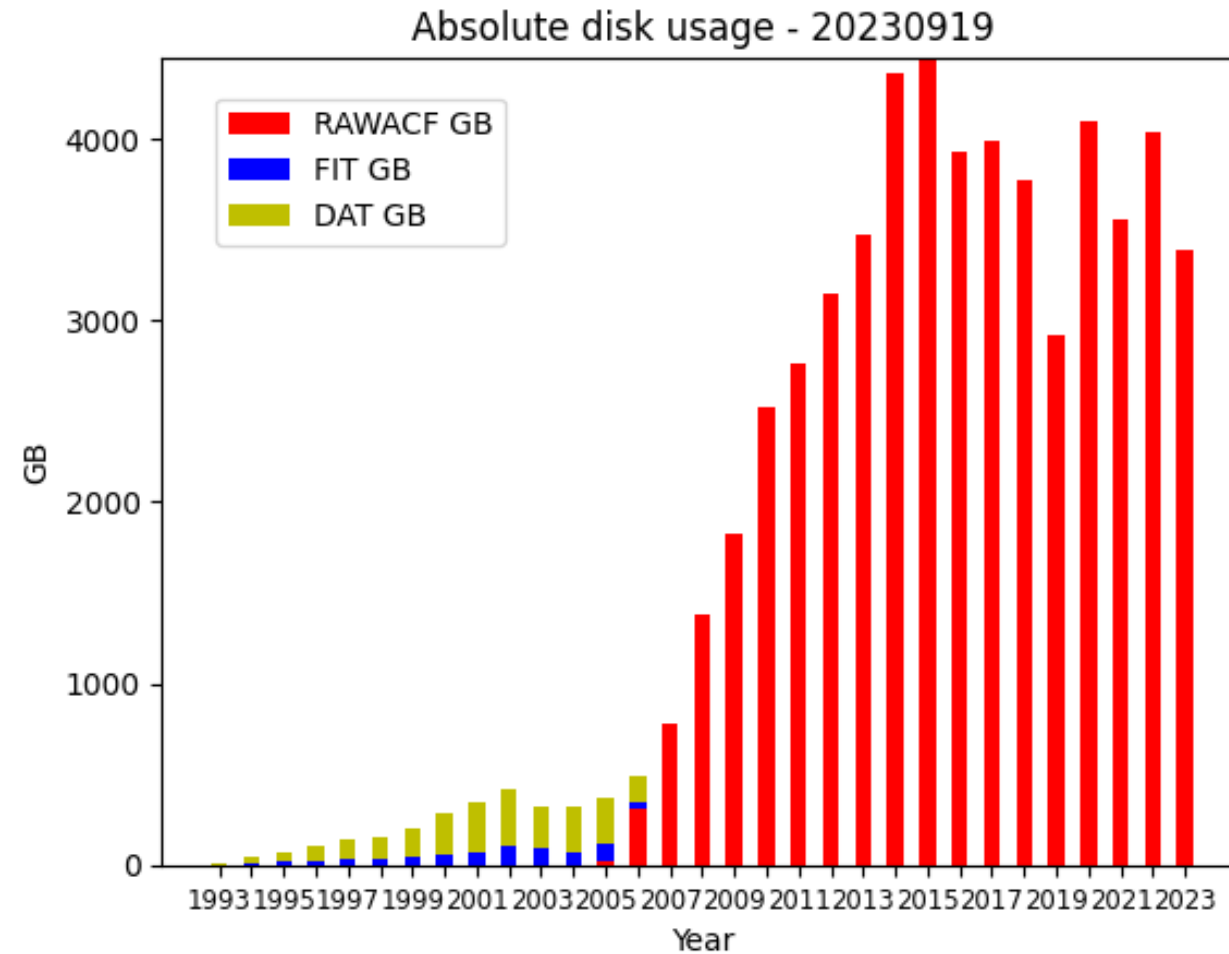# Goal: High Quality, Open, Canonical Dataset

2015-2017: "The Cloud"

Opportunities
- Scaling up space and users
  - 2013: 10TB, 10 users
  - 2023: 105TB, 50 users

Challenges
- Less administrative control
- Proprietary transfer protocol
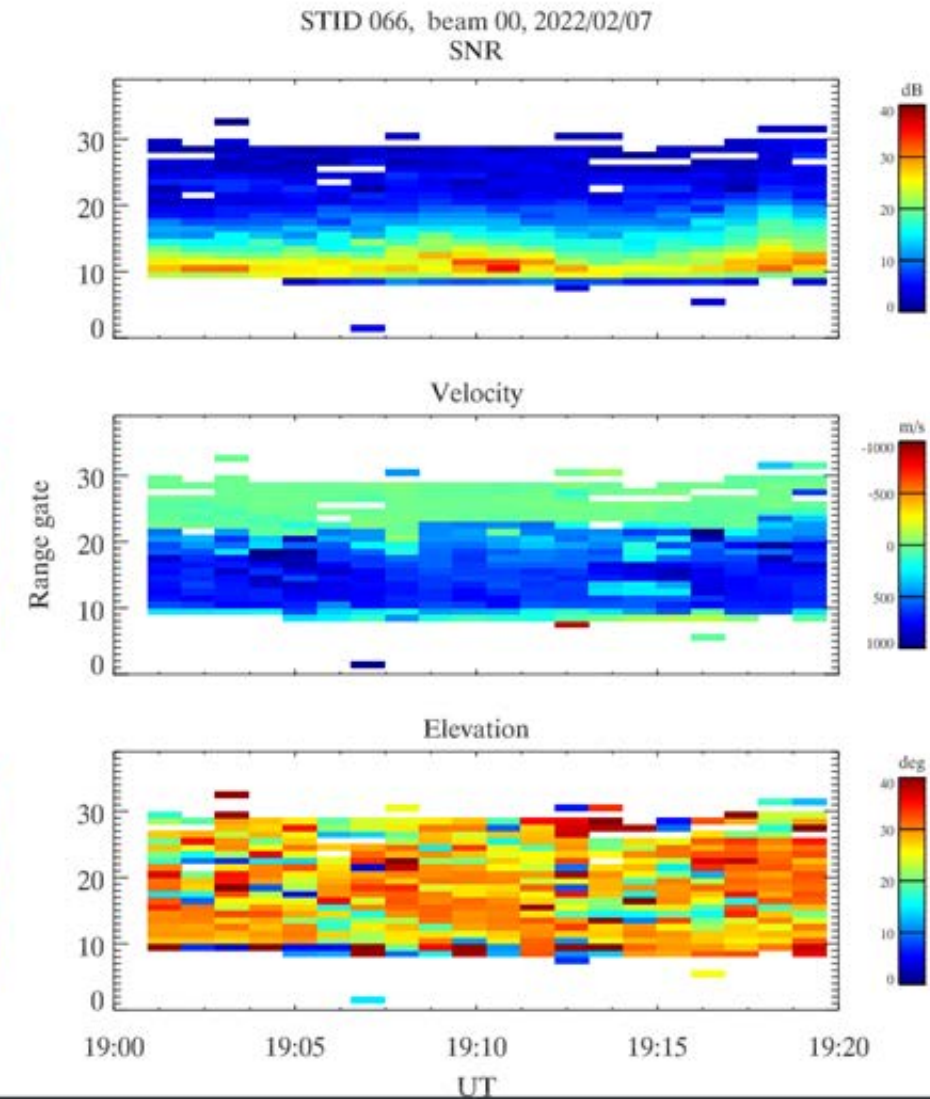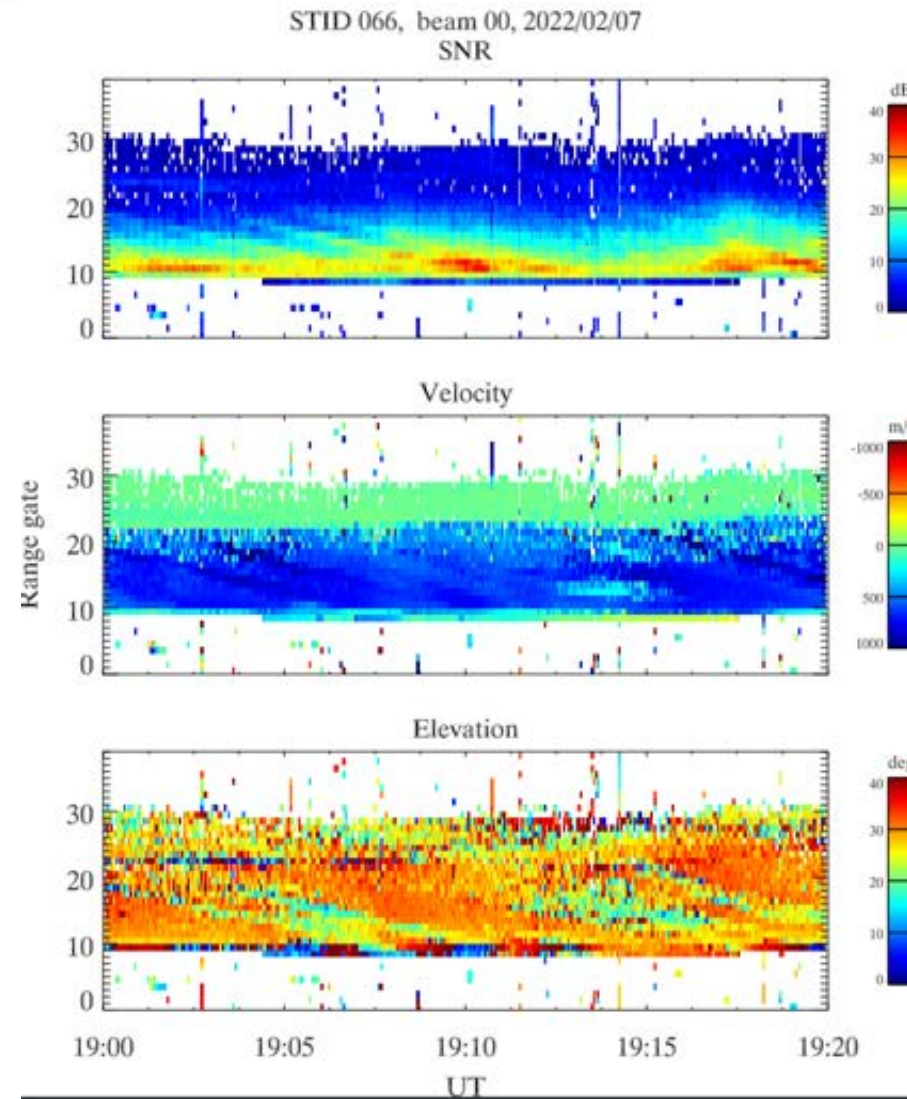 - More clicks!
  - Change is difficult

Absolute disk usage - 20230919

2019-2023: New Capabilities

Increased
Time resolution

means...

More data

2019-2023: Improved Resources

Improved Internet Access in Canada's North

```
------------------------------------------------------
Network Access (to 2012-12-14)                 400.00
GST                                             20.00
------------------------------------------------------
------------------------------------------------------
Network Access (to 2013-05-13)                 369.95
GST                                             18.50
------------------------------------------------------
```

Trans Date	Description	Amount
07/16/2022		$130.00

Trans Date	Description	Amount
08/16/2023		$115.00

Ongoing Challenges: 2023 and Beyond

- Varied funding agreements among international partners
- Diverging instrument hardware and software
- New instrument capabilities:
  - Increasing data quantity
  - increased data complexity
- Historical, proprietary data formats
- Lack of labour, expertise, funding among international partners

Three Decades of Data Management

Drivers of change
- Disruptive events
- Opportunities
- Defining goals
- Policies
- Procedures

Legend:
- 🔴 Data Mirrors
- 🔴 Future Data Mirrors
- • Radars