Trusted Data Repositories as an Essential Element of the Research Enterprise

Reyna Broadhurst, Associate Director, ORCID: 0000-0001-6975-6816
World Data System - International Technology Office, ROR: 01pwmqe95
2025-Oct-21, eResearch Australasia

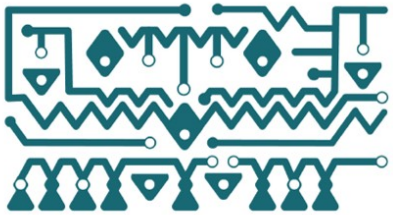
International Science Council

WORLD DATA SYSTEM

Acknowledgement of Country

We acknowledge the traditional owners of the land on which we meet today - the lands of Turrbal and Yuggera peoples, and pay our respects to their Elders past, present and emerging.

As representatives of the data repository community, we recognize the importance of revising our practices with data to meet the CARE Principles, being mindful that depends on relationships with Indigenous communities and recognizing unique concerns and priorities.



CARE Principles for Indigenous Data Governance

WORLD DATA SYSTEM

World Data System (WDS)

The **mission of the WDS** is to enhance the capabilities, impact, and sustainability of our member data repositories and data services by:

Creating trusted communities of scientific data repositories

Strengthening the scientific enterprise throughout the entire lifecycle of data and all related components creating first-class data that feeds first-class research output

Advocating for accessible data and transparent and reproducible science

International Science Council

The global voice for science

WDS was initiated by the the predecessor of ISC (ICSU) in 2008, and continues that relationship as an ISC member.

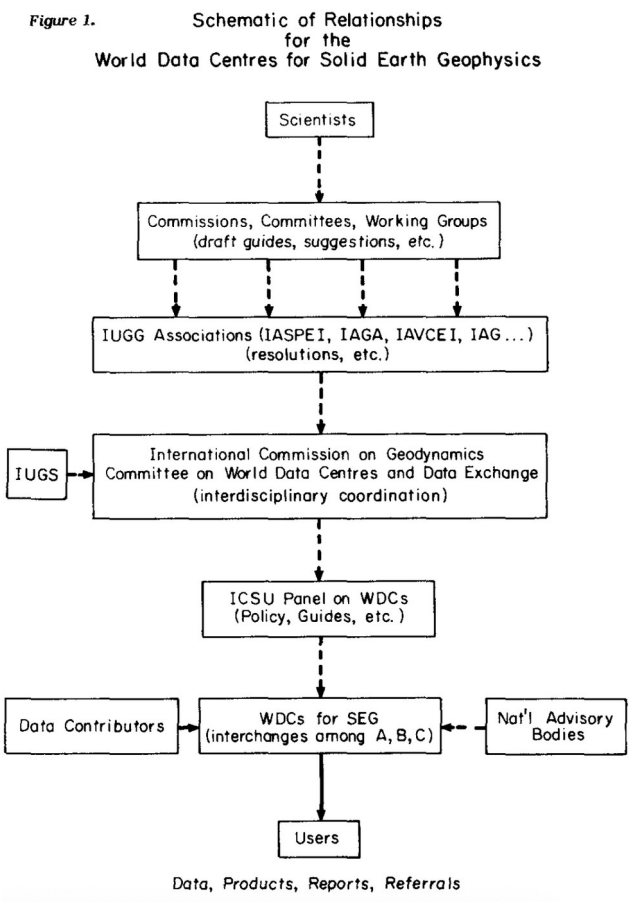
WORLD DATA SYSTEM

World Data System Origins

Data are the lifeblood of science *and the sharing of observations and measurements has always been essential to the formulation of new concepts and the evaluation of existing ones.*

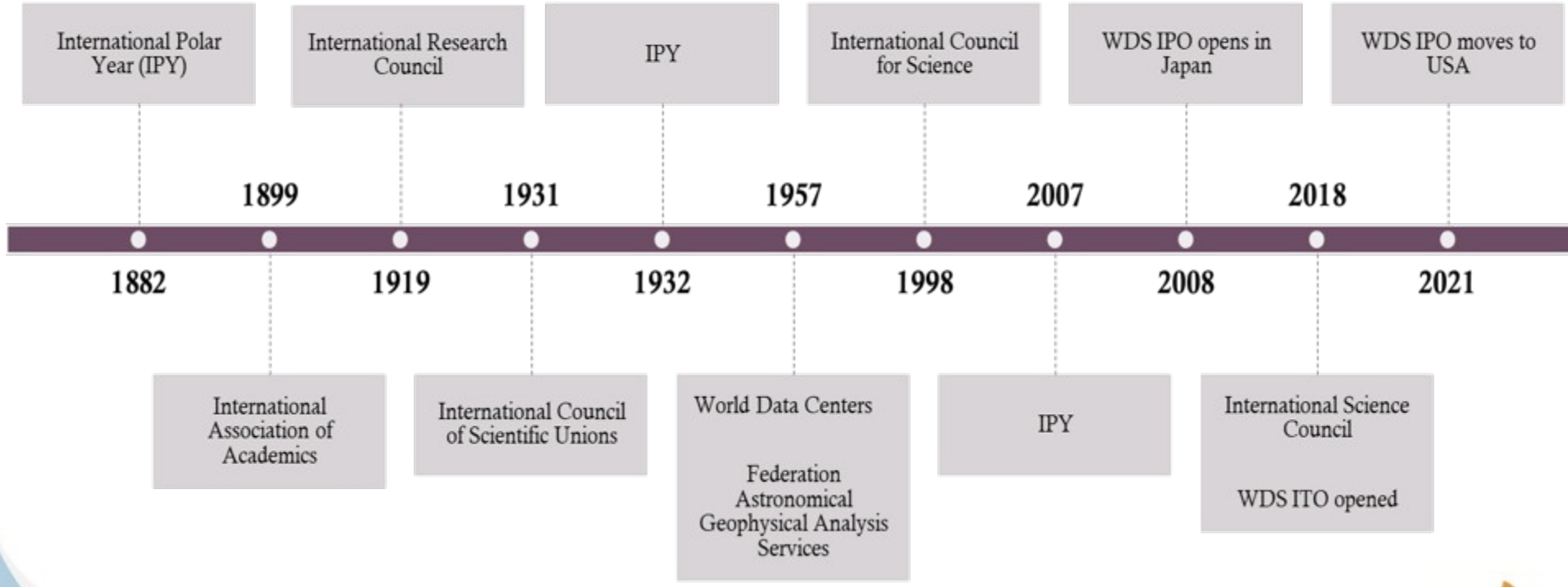
… it was recognized that ***a more formal mechanism would be needed to provide for the archiving and dissemination of the relatively large quantities of data*** *that would accumulate. To meet this need, the IGY founders* ***established a system of World Data Centers****.*

Modern data centers require a wide range of capabilities including duplicating, microfilming,digitizing, computer graphics, plotting and publishing. ***They must be able to meet the needs of users*** *for analog or digital data and for summaries, searches, maps and other data products.*



Lander JF. THE WORLD DATA CENTER SYSTEM. Episodes 1979;2:26-28.
https://doi.org/10.18814/epiiugs/1979/v2i2/005

World Data System History



WORLD DATA SYSTEM

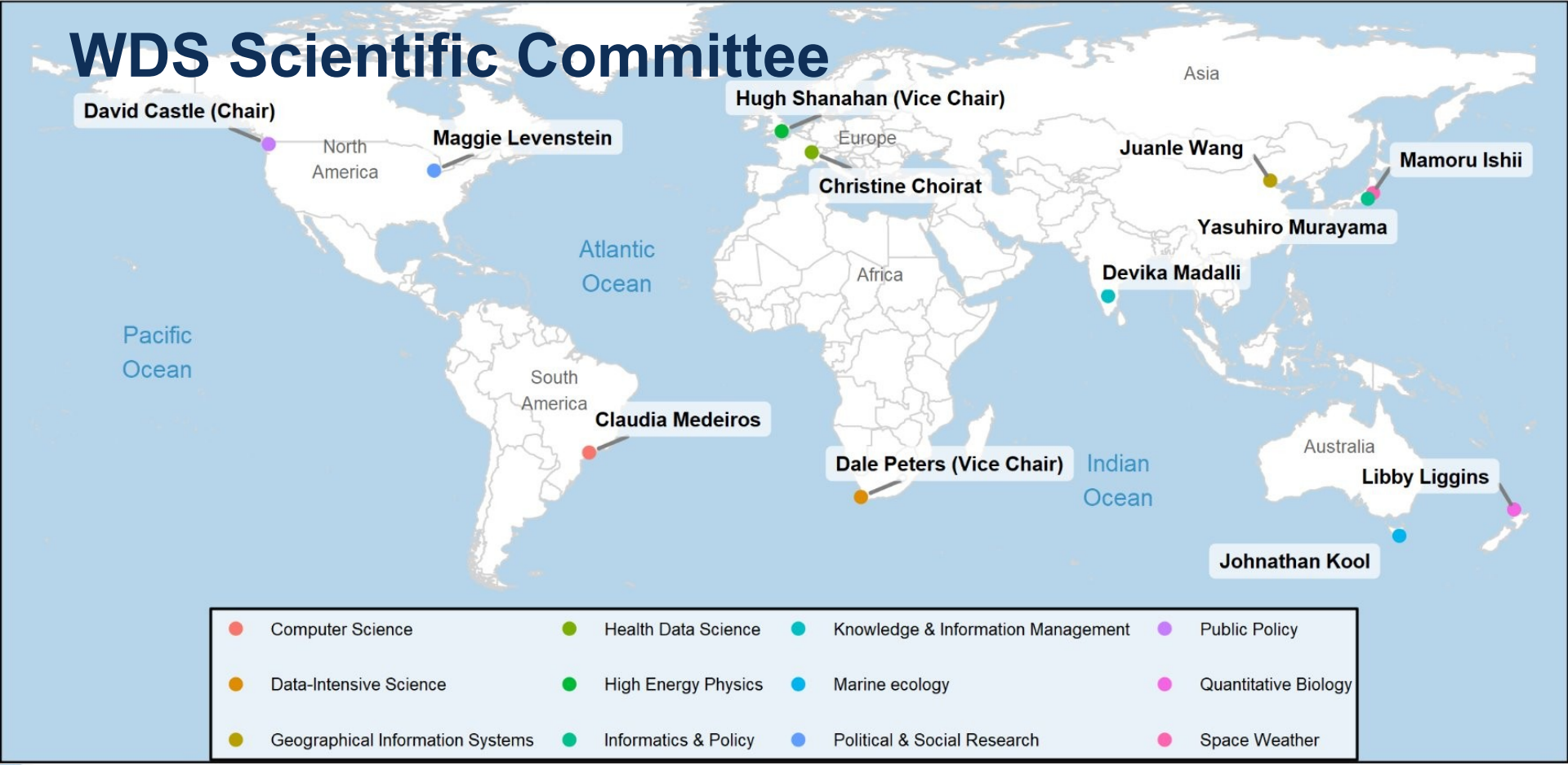
WDS Membership

Over 150 members representing 5 categories:
regular, network, partner, associate, and candidate

Become a member:
https://worlddatasystem.org/members/application_membership/

WORLD DATA SYSTEM

WDS Scientific Committee



WORLD DATA SYSTEM

WDS Staff

International Program Office (WDS-IPO)

- PI: Dr. Suzie Allard
- Executive Director: Meredith Goins
- Program Manager: Daniela Santos Oliveira
- Administrative Assistant: Katherine Read
- Informaton Specialist: Samantha Campbell

International Technology Office (WDS-ITO)

- PI: Dr. David Castle
- Director: Reyna Broadhurst
- Data Science Developers: Andrea Budac , Emilie Altman
- Polar Data Research Associate: Chantelle Verhey

WORLD DATA SYSTEM

Garbage In, Garbage Out

Data inputs are crucial to avoid erroneous, biased and harmful outputs.

Key considerations:

- Gaps & biases due to limitations in coverage (e.g., geographical, demographic, biodiversity, temporal)
- Sensitive/confidential data considerations
- Utilize & report trusted sources of inputs

Prompt: Garbage in, garbage out for AI algorithms (leonardo.ai)

Brooks Hanson, Shelley Stall, Joel Cutcher-Gershenfeld, Kristina Vrouwenvelder, Christopher Wirz, Yuhan (Douglas) Rao & Ge Peng. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. *Nature* 623, 28-31 (2023). doi: https://doi.org/10.1038/d41586-023-03316-8

WORLD DATA SYSTEM

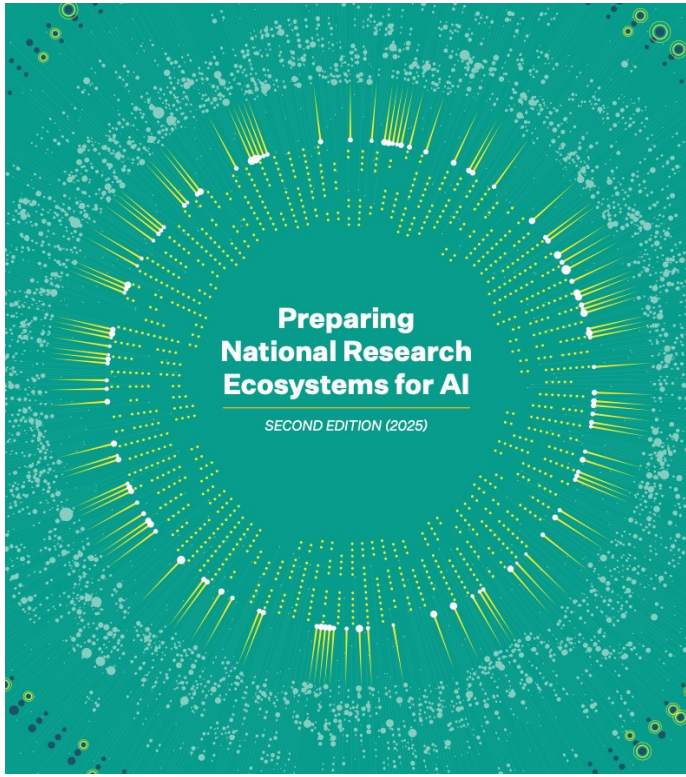
Garbage In, Garbage Out: Value of Curation at Discipline-Based Repositories

"Recognized, quality-assured data sets are particularly needed for generating trust in AI and ML, including through the development of standard training and benchmarking data sets…current data requirements set by funders and journals have inadvertently incentivized researchers to adopt free, quick and easy solutions for preserving their data sets. **Generalist repositories** that instantly register the data set with a digital object identifier (DOI) and generate a supporting web page (landing page) **are increasingly being used**….**This means that many of the deposited research data and metadata meet only two of the FAIR criteria**: they are findable and accessible. **Interoperability and reusability** require sufficient information about data provenance, calibration, standardization, uncertainties and biases to allow data sets to be combined reliably — which is e**specially important for AI-based studies**….**Disciplinary repositories**, as well as a few generalist ones, **provide this service** — but it takes trained staff and time, usually several weeks at least…..We also **urge funders to require that researchers use suitable repositories** as part of their data sharing and management plan. Institutions should support and partner with those, instead of expanding their own generalist repositories."

WORLD DATA SYSTEM

ISC Centre for Science Futures

- AI is a thematic focus area: https://council.science/theme/artificial-intelligence/
- **Regional workshops** held in Kuala Lumpur, Malaysia (Oct 2023), Santiago de Chile (April 2024) and Oman (Feb 2025)
- Project Lead is David Castle, WDS SC Chair
- Recent report summarizing **literature review and country case studies**: *Preparing National Research Ecosystems for AI*, International Science Council, 2025, https://council.science/publications/ai-science-systems,
    - While some commendable progress, there is a general lack of funding, digital infrastructure, regulations, ethics



WORLD DATA SYSTEM

Insights from International Science Council Report

Preparing National Research Ecosystems for AI: Strategies and progress in 2024. (n.d.). International Science Council. Retrieved July 23, 2024, from https://council.science/publications/ai-science-systems/

4 of the 45 issues that emerged from the literature review were for data quality:

1. **Accuracy**
2. **Bias and exclusion**
3. **Subject orientation of data vs. the interdisciplinary nature of AI research:** Most scientific knowledge comes from a specific subject. We need to encode and use it, while enabling communication between domains and allowing for the growing generation of interdisciplinary knowledge.
4. **Data coding and annotation:** AIs, and large language models in particular, require humans to code and annotate the data they use. These individuals must be aware of the risk of embedding cultural differences in the data during the annotation process.

WORLD DATA SYSTEM

ESIP Partnership

EARTH SCIENCE INFORMATION PARTNERS

Documentation enables users to understand the contents of a data set and should provide information and tools to ***increase data usability***.

Quality* information** determines the dataset's "fit-for-purpose***" for AI & affects the trustworthiness of AI applications.

Access affects the ***efficiency & reproducibility*** for AI R&D process.

Preparation help identify common ***data services and tools*** to reduce preprocessing burden for users including AI practitioners.

ESIP Data Readiness Cluster is a forum for community members to:

- Understand users' data needs for AI/ML Research and Development with environmental data
- Develop community standards, leading practices & tools for AI-ready data

WORLD DATA SYSTEM

ESIP AI Data Readiness Checklist

Quality

- Completeness
- Consistency
- ***Unbiasedness***
- Timeliness
- ***Uncertainty information***
- ***Provenance***

Documentation

- Standard/common metadata
- ***Tools & examples (codes, softwares, notebooks)***
- Data dictionary
- Identifiers
- ***License information***

Access

- ***Data format*** (prefer variety)
- Delivery options (prefer variety, including cloud)
- ***Security & privacy***
- ***Usage rights***

Preparation

- ***Labels/targets identified***
- *Recommended data splits*
- ***Missing values & outliers***
- *Recommended (or not-suitable) for (which techniques would be suitable and which would not)*
- *Past usage (links to documents and/or tutorials)*

ESIP Data Readiness Cluster (2023). Checklist to Examine AI-readiness for Open Environmental Datasets. Version 1.0. Earth Science Information Partners. https://github.com/ESIPFed/data-readiness [2023-11].

AI-Readiness Checklist Assessment Tool

ESIP AI-readiness Checklist

Dataset Name			Version 1.0, last updated 7/13/2023
Dataset Version			
Dataset Location/Link			
Assessor Name			
Assessor Email			
		Responses	**Definitions**	**Usability Score**
Is this raw data or a derived/processed data product?		(please select option)		
Is this observational data, simulation/model output, or synthetic data?		(please select option)		
Is the data single-source or aggregated from several sources?		(please select option)		
Data Quality				
Will the dataset be updated?		(please select option)		
	If the data will be updated, how often will it be updated?	(please select option)		
	If the data will be updated, choose the update frequency that best describes the dataset.	(please select option)		
	Will there be different stages of the update (e.g., updated with preliminary data first and replaced by a later update of the full record)?	(please select option)		
		If yes, what is the delay between different stages?	(short answer)	
	Should the new version of the dataset supersede the current version?	(please select option)		
		Please provide an explanation for "Other" response.	(short answer)	
Is there any documentation about the completeness of the dataset?		(please select option)	Completeness definition	
	If there is documentation regarding the completeness of the dataset, please provide a link to report/document.	(link to report/document)		
How complete is the dataset compared to the expected spatial coverage?		(please select option)		

Dropdown options

General dataset information

Proposed usability score

WORLD DATA SYSTEM

AI Data Readiness Workshop & Pilot Cohort

Target Audience: data repository representatives who are involved with curating metadata and data (or provisioning of tools for that curation).

Workshop

- 47 attendees from 12 countries (Australia, Austria, Canada, Chile, Columbia, Finland, Hungary, India, Portugal, South Africa, United Kingdom, USA)
- Disciplines: Agriculture, Earth and Ocean Sciences, Health, Neuroscience, Social Sciences, Generalist, and more

Pilot

- 12 data repositories included
- Intro, mid-term, and concluding meetings
- Dataset selection, worksheet completion,1-on-1 sessions
- 2 guest speakers (health data repository experience with checklist, earth sciences expert on bias characterization)

Cohort Feedback

- Expand **glossary**, especially to define that have multiple interpretations (e.g., raw data, processed data, derived data
- Provide **rationale** for each question
- Expand drop-down choices (beyond yes, no, not applicable), from **binary to maturity level**
- Include **scoring** and examples
- Add considerations such as **CARE Principles** and whether the dataset has **prior usage by an AI model**
- Consider nuances for different use cases, such as **static data versus dynamic data**

Further guidance needed:

- Bias
- Quality
- Data versioning
- Dynamic data
- Data curation
- Annotations/labels

Automation of checklist is also in consideration, potentially building on open source tools like the F-UJI FAIR dataset checker.

WORLD DATA SYSTEM

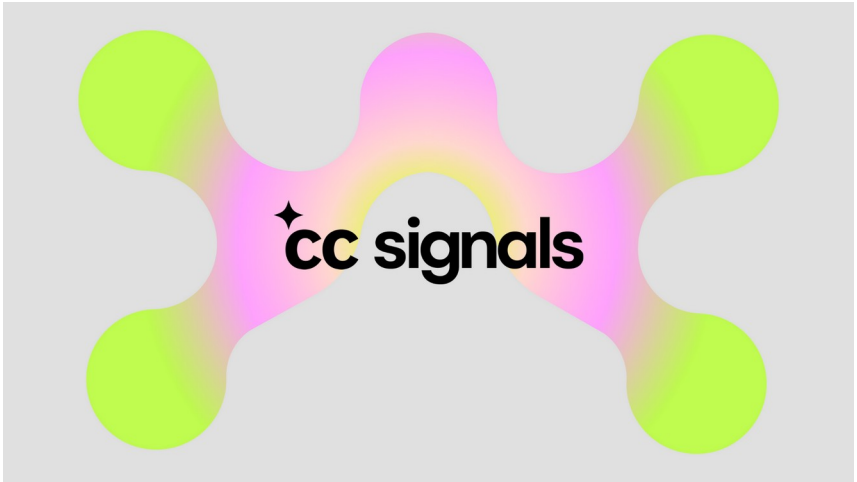
License Information & Usage Limitations

What is the license for the data?

Is the license standardized and machine-readable?

CC Signals - An emerging social contract for AI. The Declaring Party applies CC signals to a set of standard categories: Automated Processing, AI Training, Generative AI Training, AI Use, and Search.

```
Content-Usage: train-genai=n;exceptions=cc-cr
```



Indigenous Data Governance

Local Contexts

Traditional Knowledge (TK) & Biocultural (BC) labels enable customized expression of provenance, protocols, & permissions

Biological Diversity COP16 - Article 8(j)

- recognizes Indigenous peoples *"efforts in applying, preserving & maintaining their traditional knowledge, innovation & practices in relation to promoting the conservation & sustainable use of biodiversity."*
- Includes sharing of benefits from the utilization of genetic resources & digital sequence information on genetic resources, as well as traditional knowledge associated with genetic resources.

WORLD DATA SYSTEM

FORCE11 Data Usage Typologies WG

https://force11.org/group/data-usage-typologies/

Intent: to develop a **common typology of data uses**, along with their description and associated characteristics, so that repositories, publishers, institutions, funders, infrastructure providers and meta-researchers can consistently capture, compare and evaluate the ways in which data is used.

Deliverables

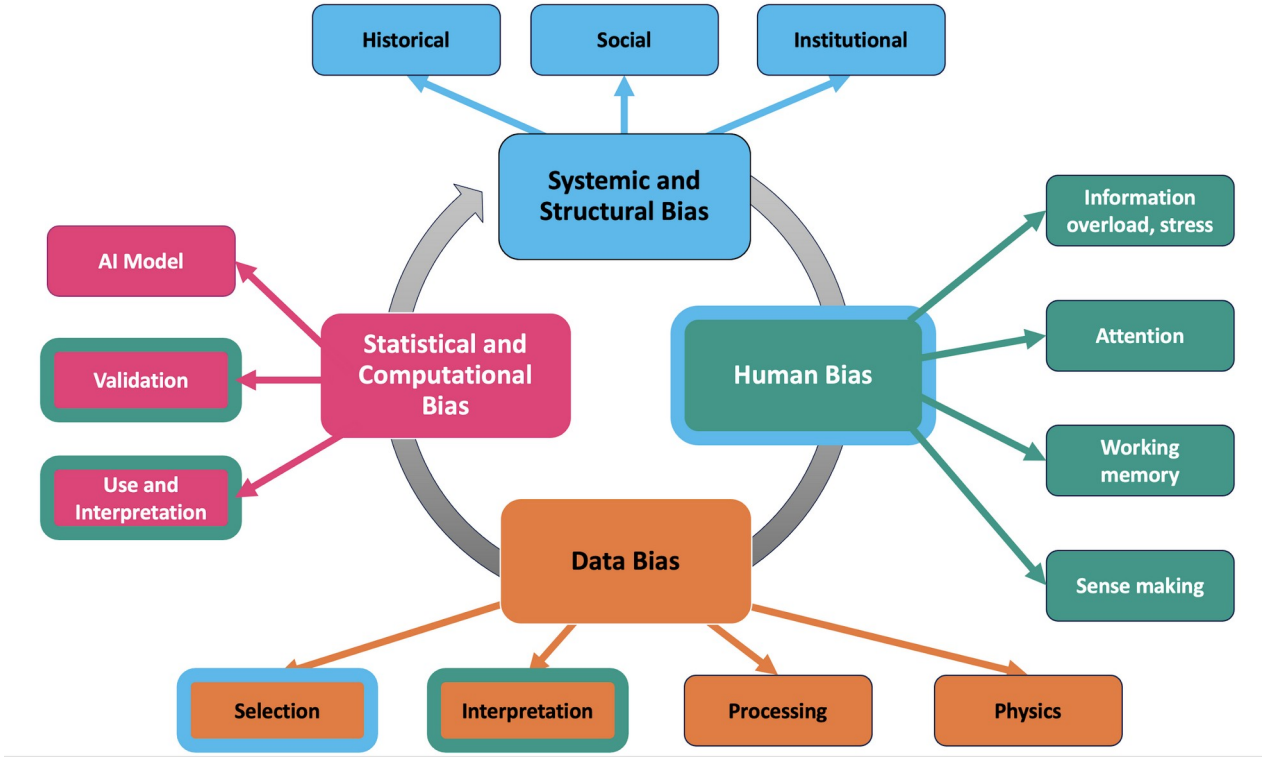
- Codified list of data use types
- Common set of characteristics to identify each of the use types
- Recommendations for how to capture data use types in metadata

While citation in research publications is important, many uses are not captured today:

- Disaster management systems (e.g., early earthquake warning systems, tsunami monitoring systems)
- Weather and climate model data assimilation
- AI models training and benchmarks
- Policy development
- Education
- Technology evaluation (e.g., new instrument testing)
- …

Categorization of Bias in AI for Earth Sciences

- Four main categories
  - Systemic & Structural Bias
  - Human Bias
  - Data Bias
  - Statistical and Computational Bias
- First step to measuring and mitigating these biases



McGovern, A., A. Bostrom, M. McGraw, R. J. Chase, D. J. Gagne, I. Ebert-Uphoff, K. D. Musgrave, and A. Schumacher, 2024: Identifying and Categorizing Bias in AI/ML for Earth Sciences. *Bull. Amer. Meteor. Soc.*, **105**, E567–E583, https://doi.org/10.1175/BAMS-D-23-0196.1.

WORLD DATA SYSTEM

Schema.org Metadata

Schema.org Cluster

- Develops domain-specific practices for publishing and harvesting structured Earth science data
- Dataset metadata guide

Ocean Data and Information System (ODIS)

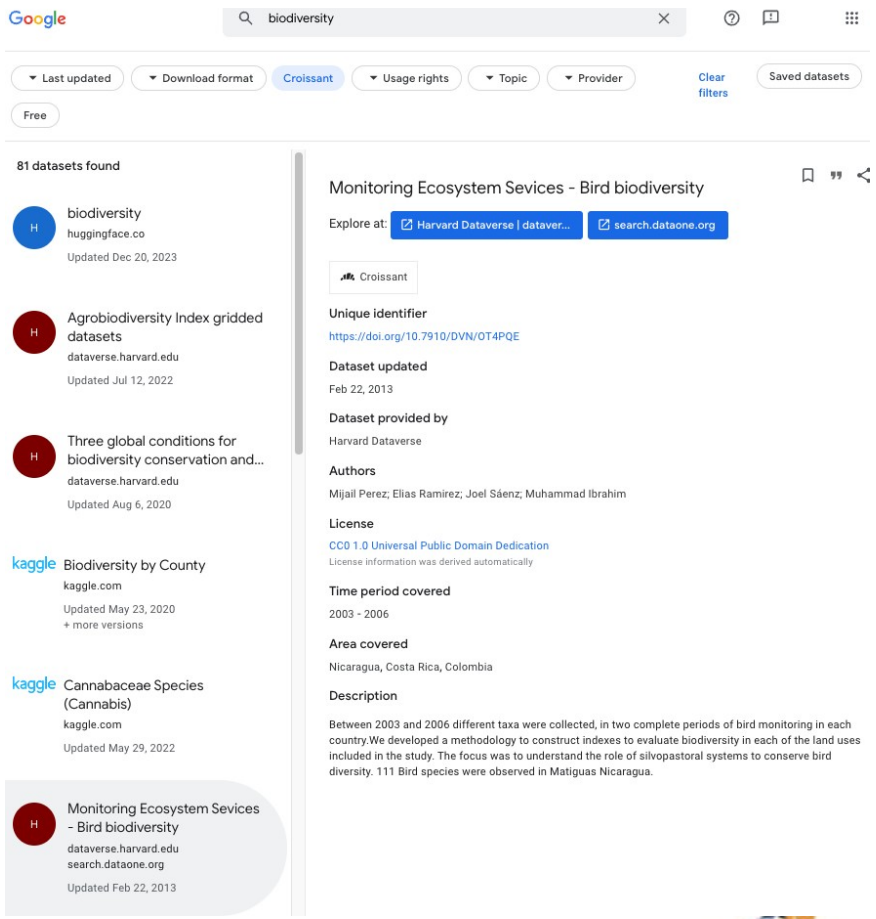
- Building an interoperable federated digital ecosystem for all ocean stakeholders, part of the Ocean Decade vision for a digital representation of the ocean
- Based upon web architectural patterns and use of schema.org/JSON-LD
- Profile documentation

Polar Data Search

- Created to aid researchers, stakeholders, & rights holders seeking polar data through a single interface
- Leveraging schema.org metadata for harvesting records from 20 polar data repositories
- Best practice guide for implementing schema.org

ML Croissant Extension

- Croissant is a format with 4 layers: metadata, resource file descriptions, data structure, and default ML semantics
- Specification: https://docs.mlcommons.org/croissant/docs/croissant-spec.html
- Major outcome of the ML Commons Croissant Working Group
- Used by Kaggle, Hugging Face, OpenML, etc.
- Croissant filter within https://datasetsearch.research.google.com/



Data Spaces, Federated Systems, Data Repositories & Interconnections

Continue to support **data sharing mechanisms and integrations** into data spaces, digital twins and federated systems, regional and disciplinary-specific.]

ODIS

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

GBIF

INTERNATIONAL DATA SPACES ASSOCIATION

DITTO Digital Twins of the Ocean

Copernicus Data Space Ecosystem

Build **relationship metadata and knowledge graphs** between repositories and related entities like harvesters and data spaces with FAIRsharing and re3data

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

OpenAIRE

FAIRsharing.org standards, databases, policies

WORLD DATA SYSTEM

WDS Data Stewardship Award

This award celebrates **early career individuals** (within 10 years of most recent academic degree) who have significantly improved the quality, integrity and accessibility of research data.

We are now open for applications for the **2026 Data Stewardship Award - due January 16**.

Scan the QR code to learn more criteria, nomination and selection process.

Past award winners: https://worlddatasystem.org/dsa-previous-winners/

World Data System International Program Office (WDS-IPO) is hosted by the University of Tennessee Oak Ridge Innovation Institute. This work is supported by a cooperative agreement (DE- SC0021915) with the U.S. Department of Energy Office of Science.
Contact WDS-IPO at wds-ipo@utk.edu.

World Data System International Technology Office (WDS-ITO) is hosted by the University of Victoria. This work is supported by the Digital Research Alliance of Canada and Ocean Networks Canada.
Contact WDS-ITO at ito-webadmin@oceannetworks.ca.

Thank You!

Questions?

Visit our website at worlddatasystem.org and wds-ito.org

Instagram @wds_ipo
Facebook @WorldDataSystem.International
LinkedIn @company/world-data-system
Bluesky @wds-ito.org
Vimeo @worlddatasystem

Reyna Broadhurst
ito-director@oceannetworks.ca

International Science Council

WORLD DATA SYSTEM