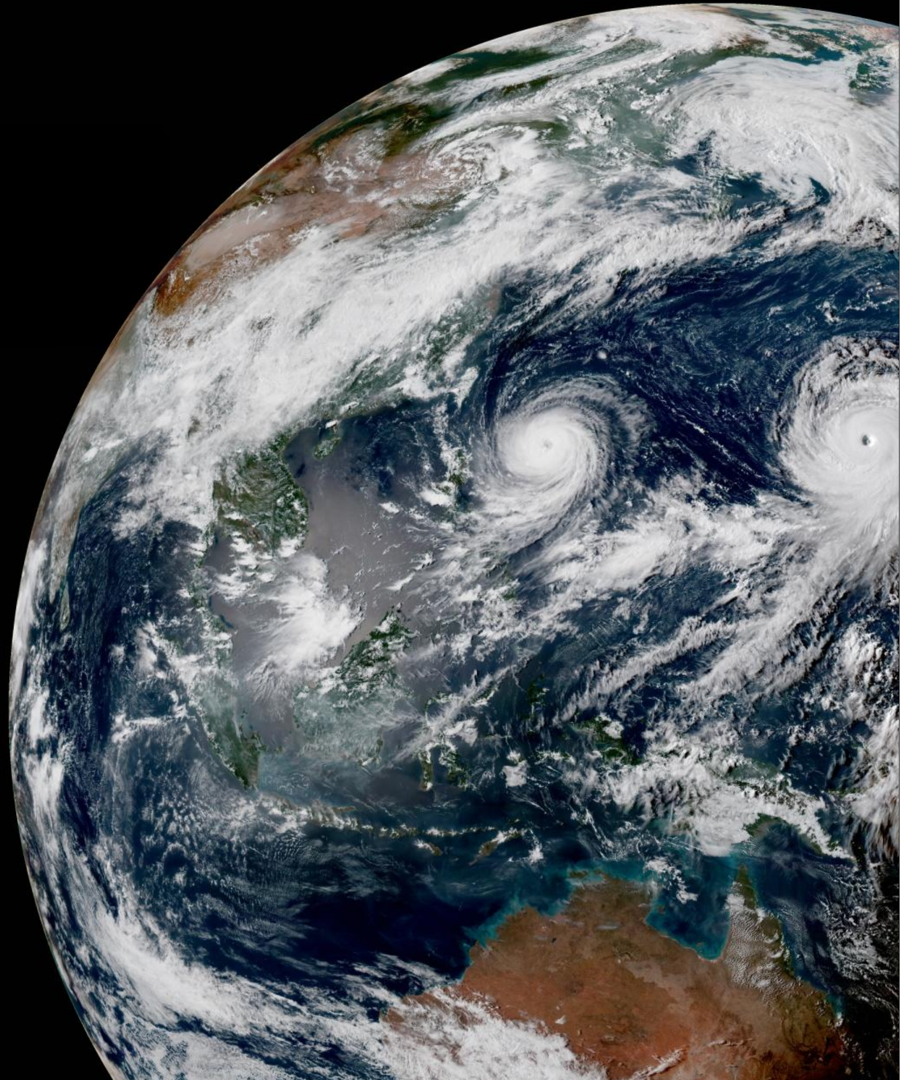
JEDI R2D2 and IODA for FAIR data management

Eric Lingerfelt (R2D2) and Stephen Herbener (IODA)

Evan Parker, Ashley Griffin, Fábio Diniz, Clémentine Gas, Benjamin Ruston, Christian Sampson, Travis Sluka, and Yannick Trémolet

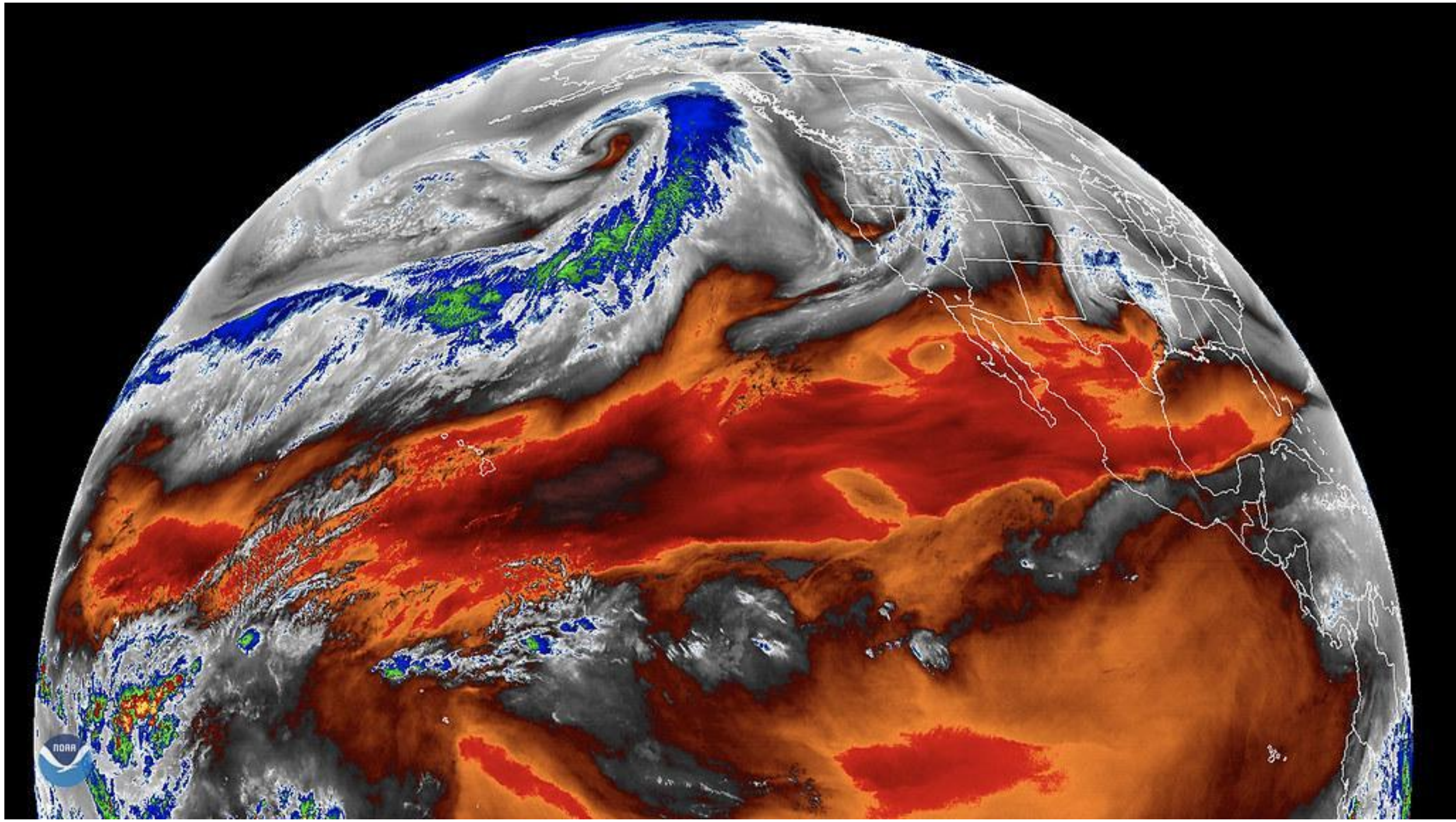
UCAR JCSDA



Community Space Weather Modeling and Data Assimilation Workshop, September 10, 2025

Talk Topics

- Big Data in NWP
- Data Management Challenges
- What is FAIR?
- FAIR is everywhere!
- JEDI-Skylab
- F & A with R2D2
- I & R with IODA
- What is R2D2?
- NRT Obs Processing
- AI in FAIR



Big Data in NWP

- As the state-of-the-art of NWP advances, there exist several factors that impact the need to handle big data effectively
- The amount of incoming observational data is increasing
  - New instruments coming on-line
  - Increasing number of channels on a given instrument
- An explosion of the number of points as model grid spacing shrinks
- The frequency of forecast cycling is increasing
  - Global 6hr cycle → Regional 15min cycle → …
  - Pushing toward the mode of “continuous DA” where new observations are assimilated as they become available within the same DA run

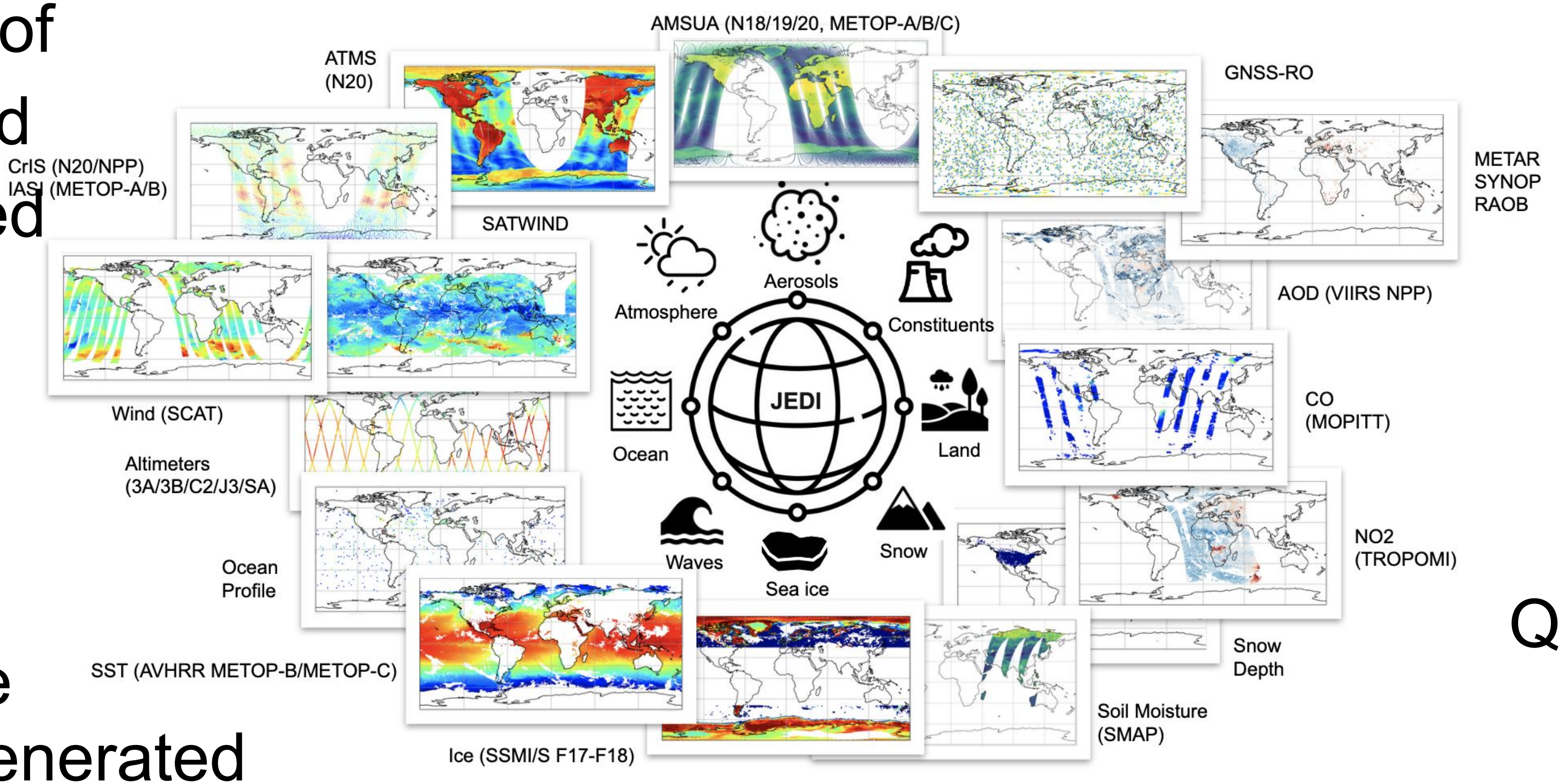
Big Data at JCSDA

Volume

Quantity of generated and stored data

Variety

Type and nature of the data



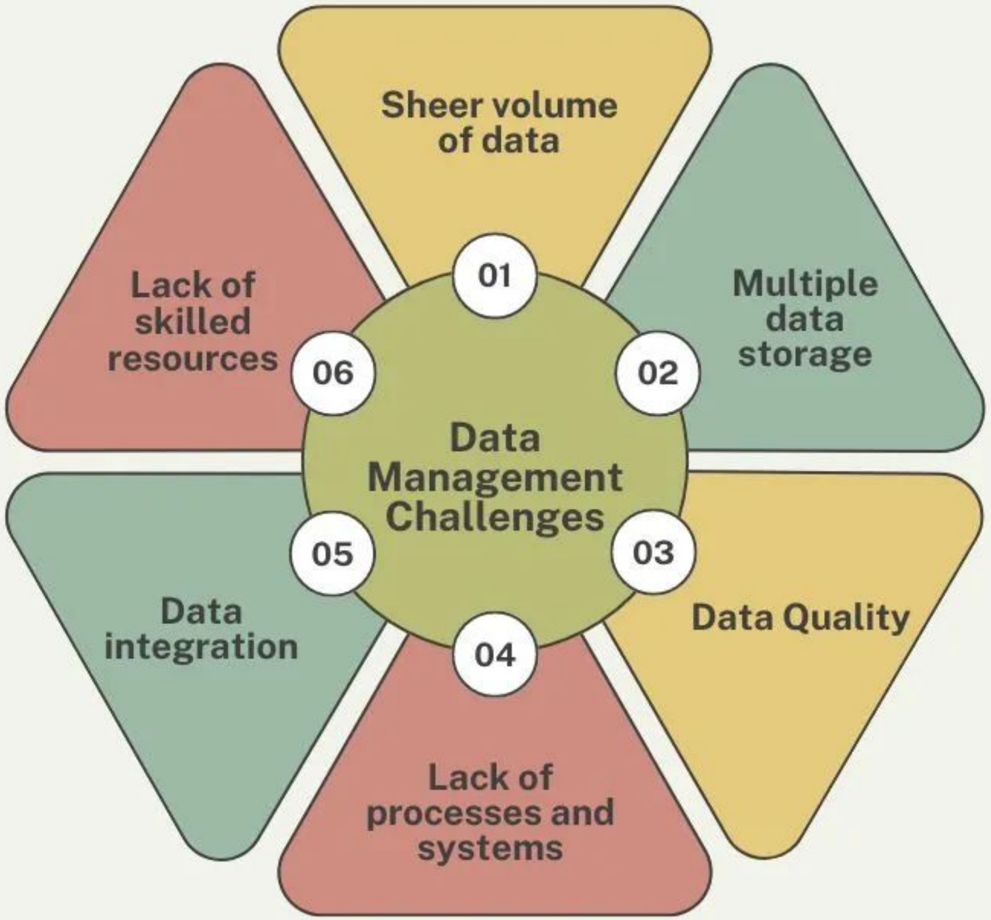
Velocity

Speed at which the data is generated and processed

Veracity

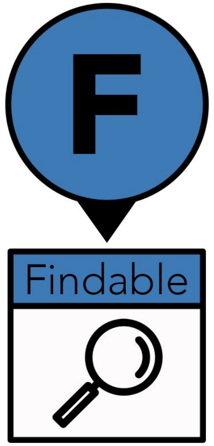
Quality and value of the data

Data Management Challenges

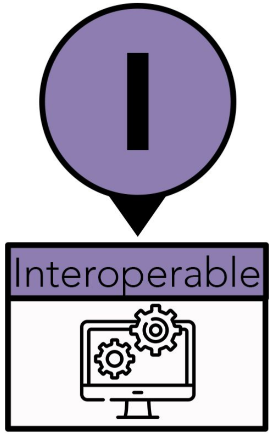


1. Sheer volume of data
2. Multiple data storage
3. Data Quality
4. Lack of processes and systems
5. Data integration
6. Lack of skilled resources
7. Data governance
8. Data security
9. Data automation
10. Data analysis
11. Going from unstructured to structured data

What is FAIR?

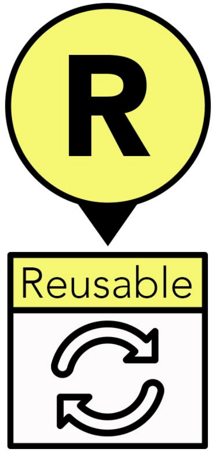


Data should be discoverable through rich metadata and *registered or indexed in a searchable resource*, often using globally unique and persistent identifiers.



Data needs to be in a common format, *use recognized terminologies*, and have metadata with formal syntax to allow for easy exchange and understanding between different systems.

Data and its metadata should be *readable by humans and machines*, and available through a trusted repository.



Data must have clear usage licenses and provenance along with *sufficient descriptive metadata* to allow for future validation and reuse.

FAIR is everywhere!

Make all scientific data FAIR

Follow the geosciences and demand best practice in publishing and sharing data, urge **Shelley Stall** and colleagues.

Scientific data are burgeoning — thousands of petabytes were collected in 2018 alone. But these data are not being used widely enough to realize their potential. Most researchers come up against obstacles when they try to get their hands on data sets. Only one-fifth of published papers typically post the supporting data in scientific repositories — as has been shown by *PLoS ONE*[1].

Too much valuable, hard-won information is gathering dust on computers, disks and tapes.

Scientists don't share data for many reasons. Those who create data rarely receive credit, and when they do, recognition is often limited to citations (see page 30). Scant support is available for curating data. These issues span all disciplines, but conversations are disconnected.

That's why more than 100 repositories, communities, societies, institutions, infrastructures, individuals and publishers (including the Springer Nature journals *Nature* and *Scientific Data*, see go.nature.com/2wbn4kj) have signed up since last November to the Enabling FAIR Data Project's Commitment Statement in the Earth, Space, and Environmental Sciences for depositing and sharing ▸

CORRECTED 5 JUNE 2019 | 6 JUNE 2019 | VOL 570 | NATURE | 27

Findable · Accessible · Interoperable · Reusable — Make Data FAIR — AGU100 ADVANCING EARTH AND SPACE SCIENCE

GO FAIR

AMS American Meteorological Society

Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats

Robert Crystal-Ornelas[1,12], Charuleka Varadharajan[1]✉, Dylan O'Ryan[1,2], Kathleen Beilsmith[3], Benjamin Bond-Lamberty[4], Kristin Boye[5], Madison Burrus[1], Shreyas Cholia[6], Danielle S. Christianson[6], Michael Crow[7], Joan Damerow[1], Kim S. Ely[8], Amy E. Goldman[9], Susan L. Heinz[7], Valerie C. Hendrix[6], Zarine Kakalia[1], Kayla Mathes[10], Fianna O'Brien[6], Stephanie C. Pennington[4], Emily Robles[1], Alistair Rogers[8], Maegen Simmonds[1,11], Terri Velliquette[7], Pamela Weisenhorn[3], Jessica Nicole Welch[7], Karen Whitenack[1] & Deborah A. Agarwal[6]

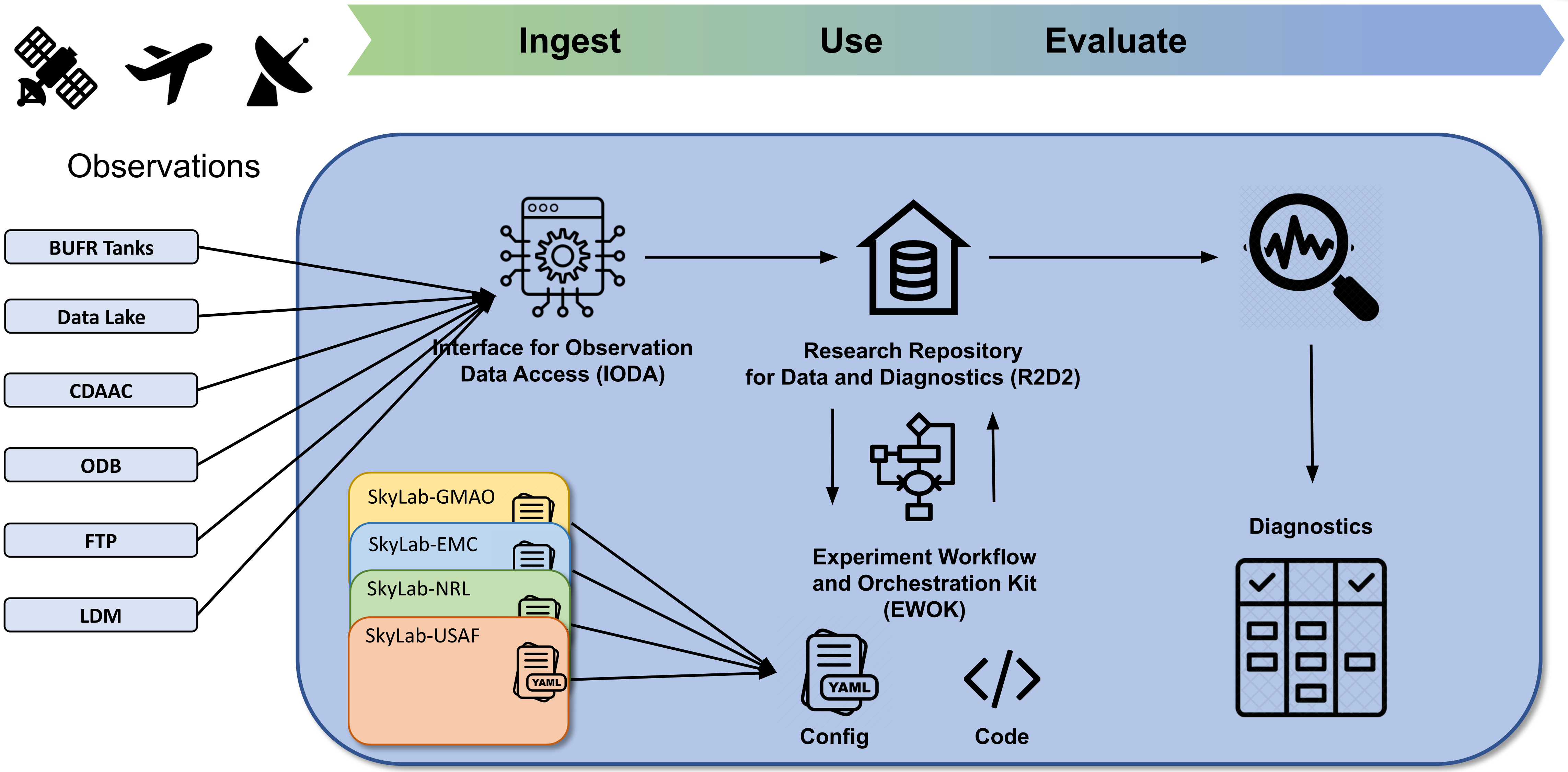
Data and Software Policy Guidelines for AMS Publications (updated Dec 2022)

The guidance presented here is designed to help authors make the data, software, and documentation supporting the research presented in AMS journals as open and accessible as possible to readers and users, in accordance with the FAIR (Findable, Accessible, Interoperable, and Reusable) Principles (Wilkinson et al. 2016). The guidance stems from the

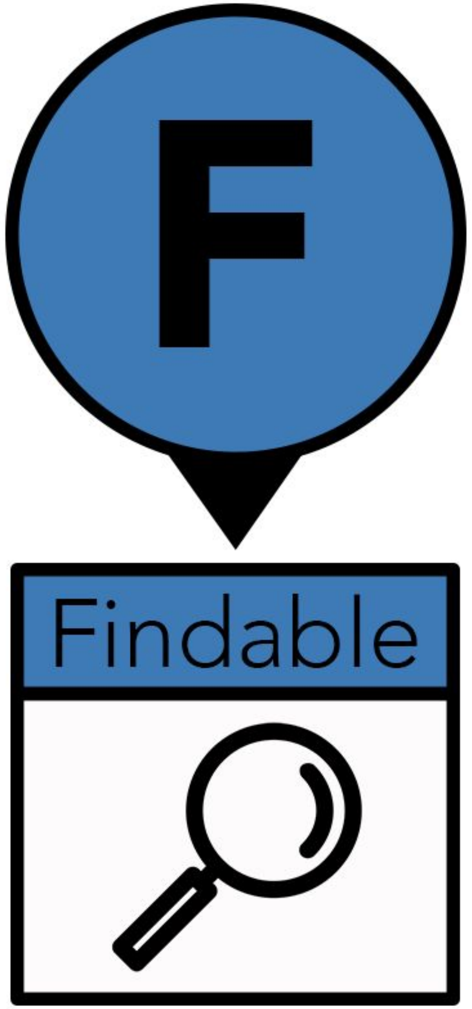
The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... Barend Mons ✉ + Show authors

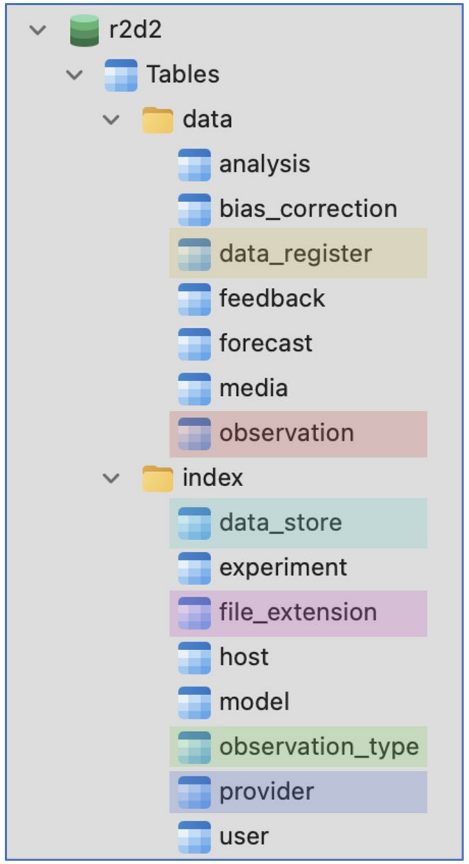
JEDI-Skylab: End-to-end DA Testbed System



Findability with R2D2



- Metadata and data are assigned a globally unique and persistent identifier
- Data are described with rich attributes
- Metadata and data are registered or indexed in a searchable resource



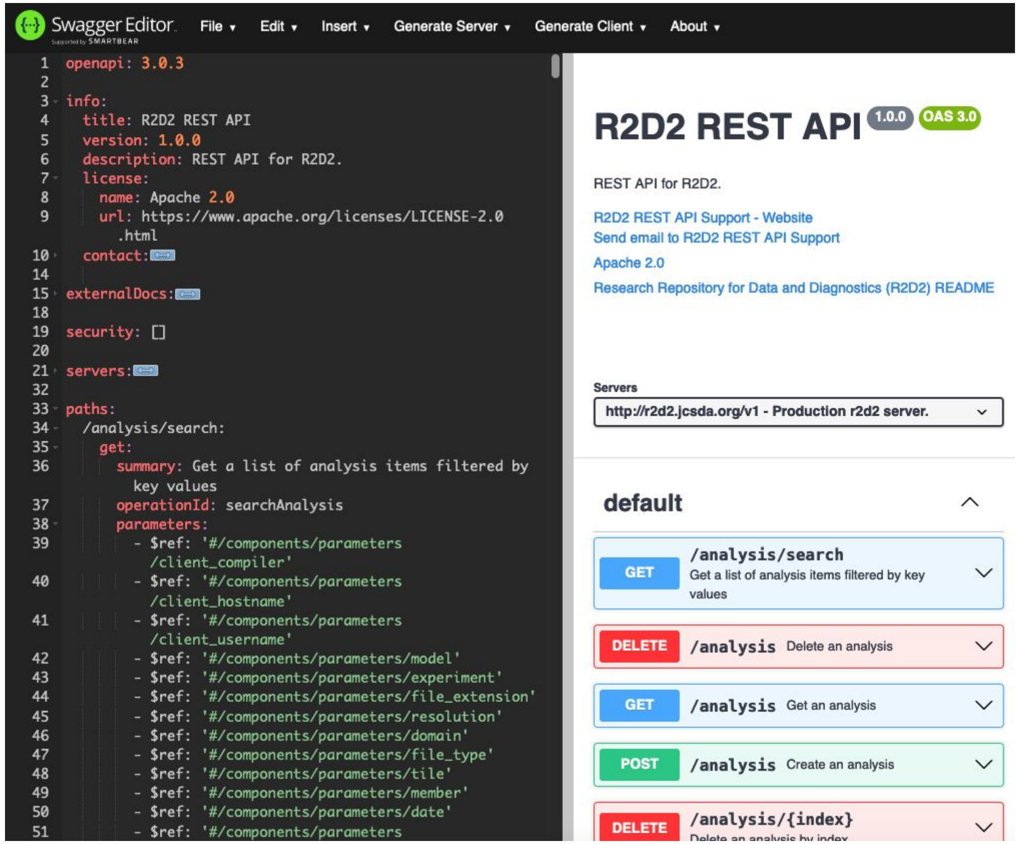
observation_index	15578
provider_index	89
observation_type_index	1235
file_extension_index	161
window_start	2022-02-15 21:00:00
window_length	PT6H
create_date	2023-05-12 01:50:56

data_register_index	4839463
data_store_index	32
item_index	15578
item	observation
create_date	2023-05-12 01:50:57

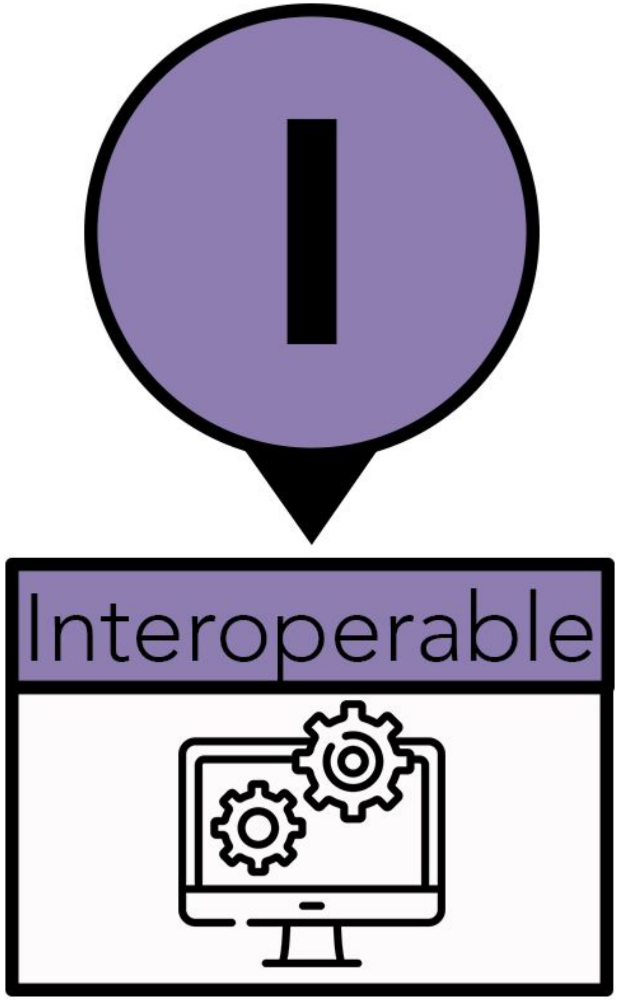
observation_type_index	1235
name	iasi_metop-c

Accessibility with R2D2

- Metadata and data are retrievable by their attributes using a standardized communications protocol
- The protocol is open, free, and universally implementable and extensible
- The protocol allows for an authentication and authorization procedure, where necessary

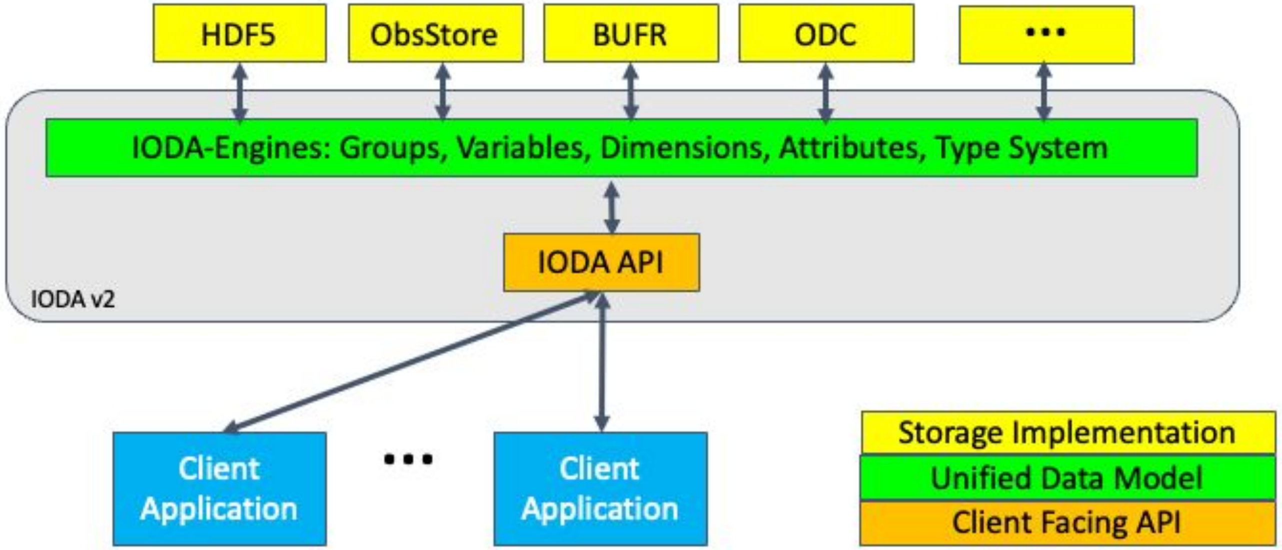


Interoperability with IODA

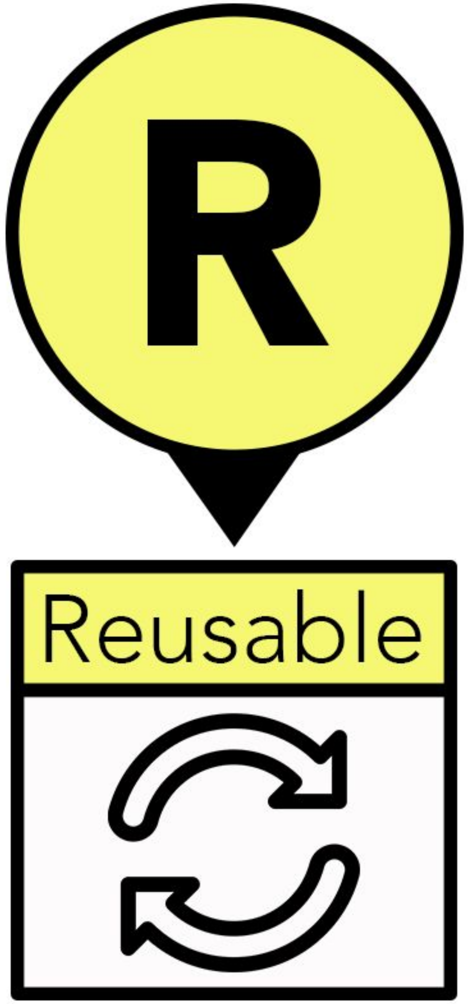


- IODA presents data contained in disparate storage implementations through a single unified data model
- The different storage implementations are “encapsulated”, effectively hiding their details from the clients of IODA
- This enables the clients of IODA to access and operate on data in a consistent manner regardless of how that data is stored

IODA Architecture



Reusability with IODA



- An important enabler of reuse is provided by the JEDI Observation and Model Data Conventions
- Variable names (airTemperature, brightnessTemperature, etc.)
- Metadata associated with the variables (latitude, longitude, sensorZenithAngle, etc.)

Convention Tables

Tables:

- Standalone web page
- JCSDA-internal link for OBS team comments

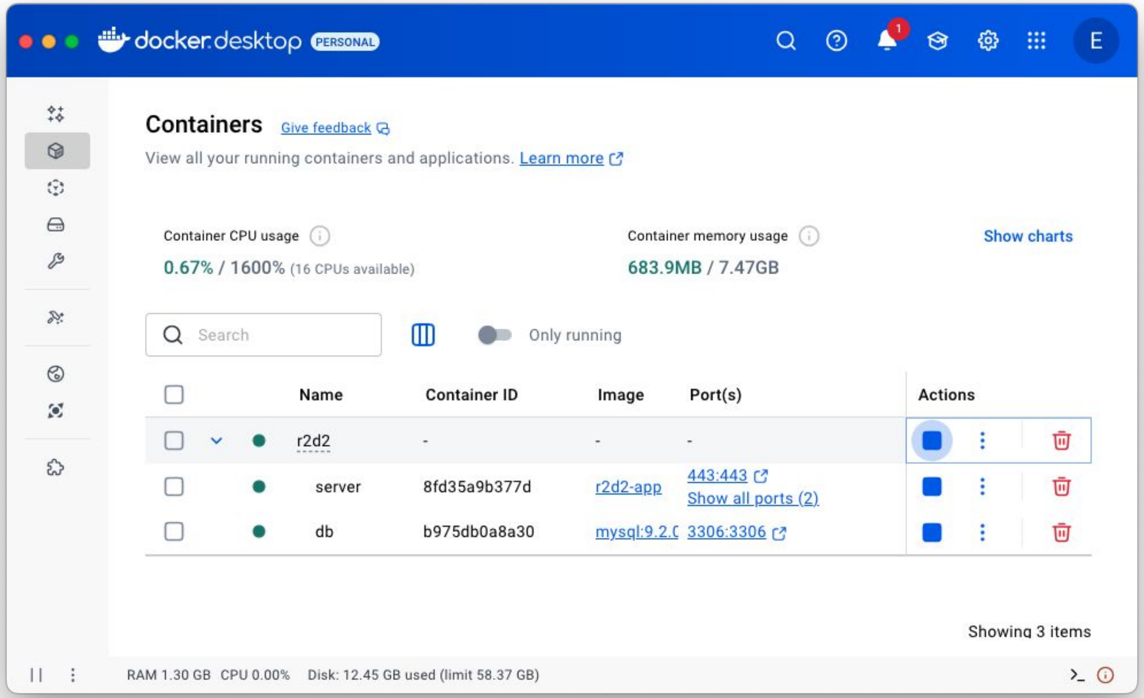
Conventions for JEDI Data - Tables

Name	Dimension 1	Dimension 2	Dimension 3	Recommended Dimensions	Data Storage Ty
radiance	Location	Channel		Location, Channel	float
spectralRadiance	Location	Channel		Location, Channel	float
scaledSpectralRadiance	Location	Channel		Location, Channel	float
brightnessTemperature	Location	Channel		Location, Channel	float
brightnessTemperatureStandardDeviati	Location	Channel		Location, Channel	float
equivalentBlackBodyTemperature	Location	Channel		Location, Channel	float
thickness	Location	Channel		Location, Channel	float
emissivityError	Location	Channel		Location, Channel	float
bendingAngle	Location			Location	float
zenithTotalDelay	Location			Location	float
slantPathDelay	Location			Location	float
atmosphericRefractivity	Location			Location	float
albedo	Location	Channel		Location, Channel	float
reflectivity	Location	Layer		Location, Layer	float
horizontalReflectivity	Location	Layer		Location, Layer	float
verticalReflectivity	Location	Layer		Location, Layer	float
differentialReflectivity	Location	Layer		Location, Layer	float
equivalentReflectivityFactor	Location	Layer		Location, Layer	float
reflectivityMaxInColumn	Location			Location	float
reflectivityLowestScanLevel	Location			Location	float
radialVelocity	Location	Layer		Location, Layer	float
cosAzimuthCosTilt	Location	Layer		Location, Layer	float

What is R2D2?

R2D2 is

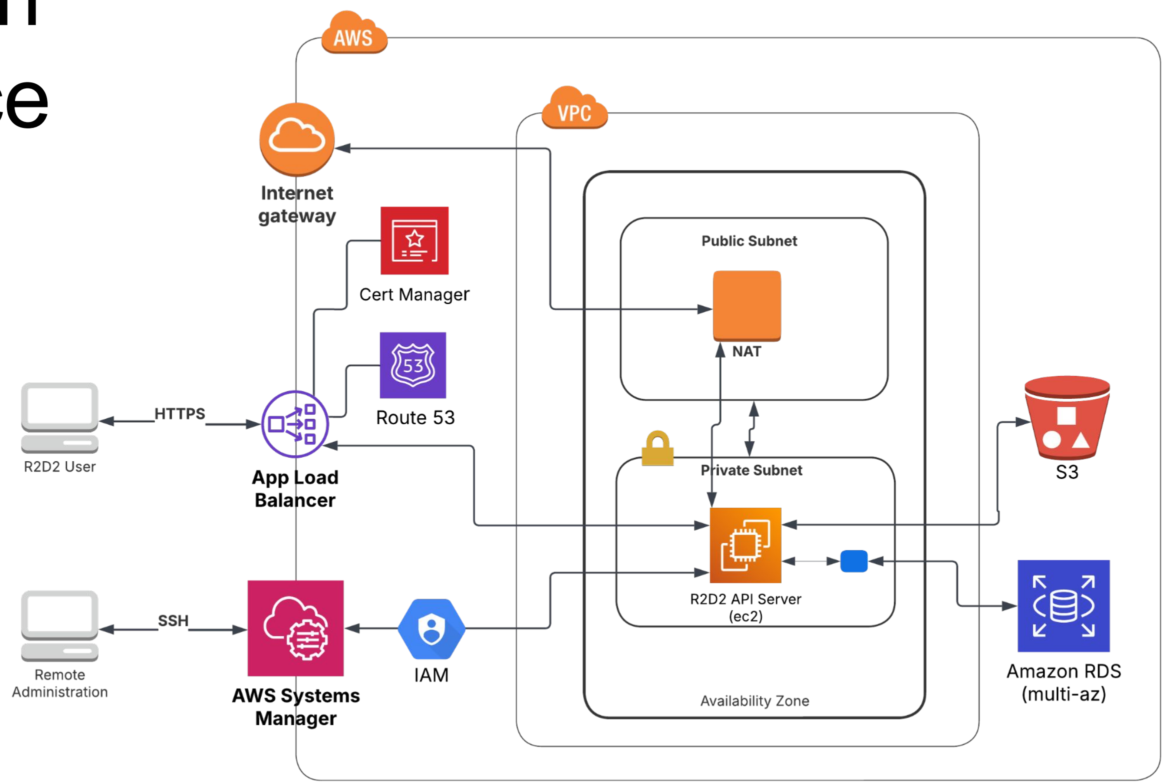
- an online data, artifact & configuration management client / server system
- for ***all*** JCSDA data assimilation workflows and analysis tools
- works on laptops, HPCs, and the cloud
- server is Docker-ready
- enables data protection and authenticated data access



What is R2D2?

R2D2 is

- a fully supported 24/7 production JCSDA cloud-based data service
- “black box” appliance
- data aggregator, data indexer, metadata register
- easy! `import r2d2`
  `r2d2.store()`
- portable, easy-to-maintain, scalable, lightweight, extensible



What is R2D2 not?

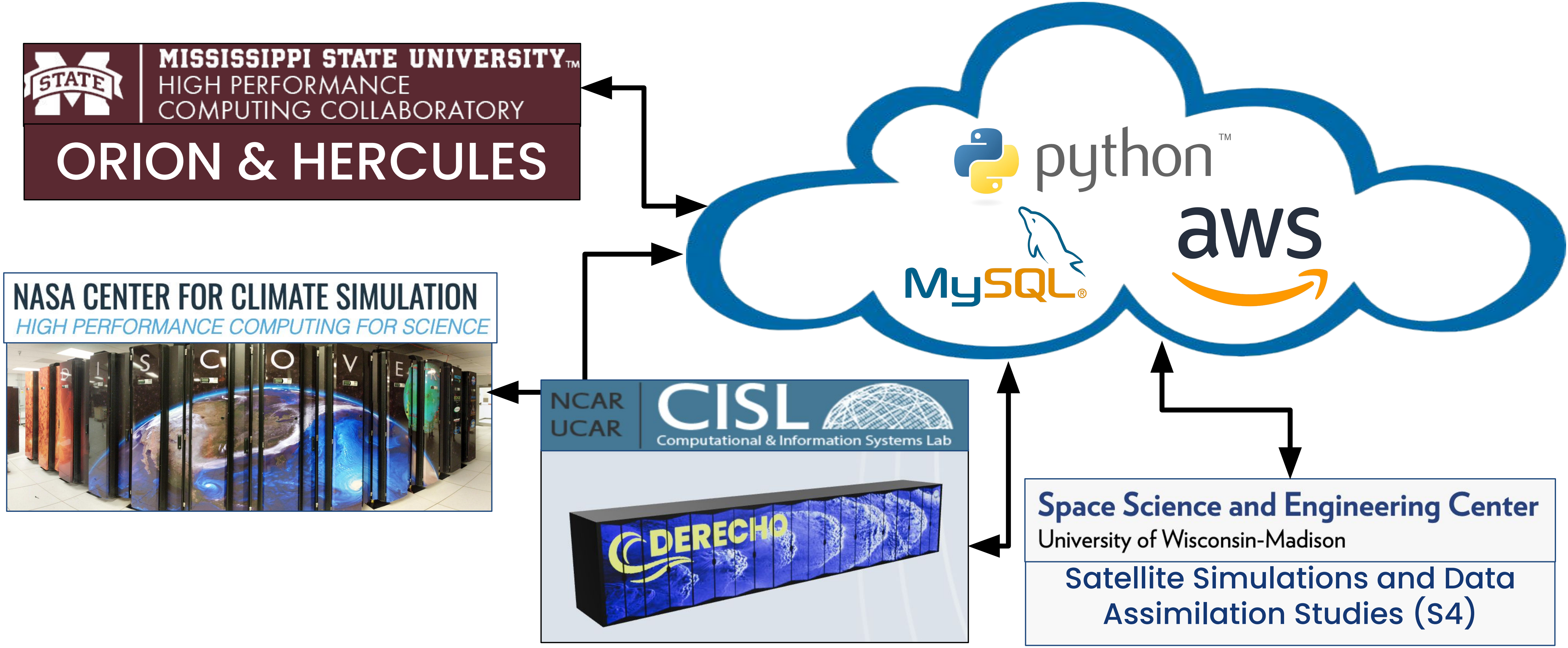
R2D2 is ***not*** a data provider.

R2D2 does ***not*** publish data.

R2D2 does ***not*** mint DOIs.

R2D2 Integrates Edge Data with Compute Resources

R2D2 enables data + compute proximity by providing ***seamless*** access to data stored at HPC centers and in the cloud. R2D2 stores data ***plus*** experiment and compute configurations.



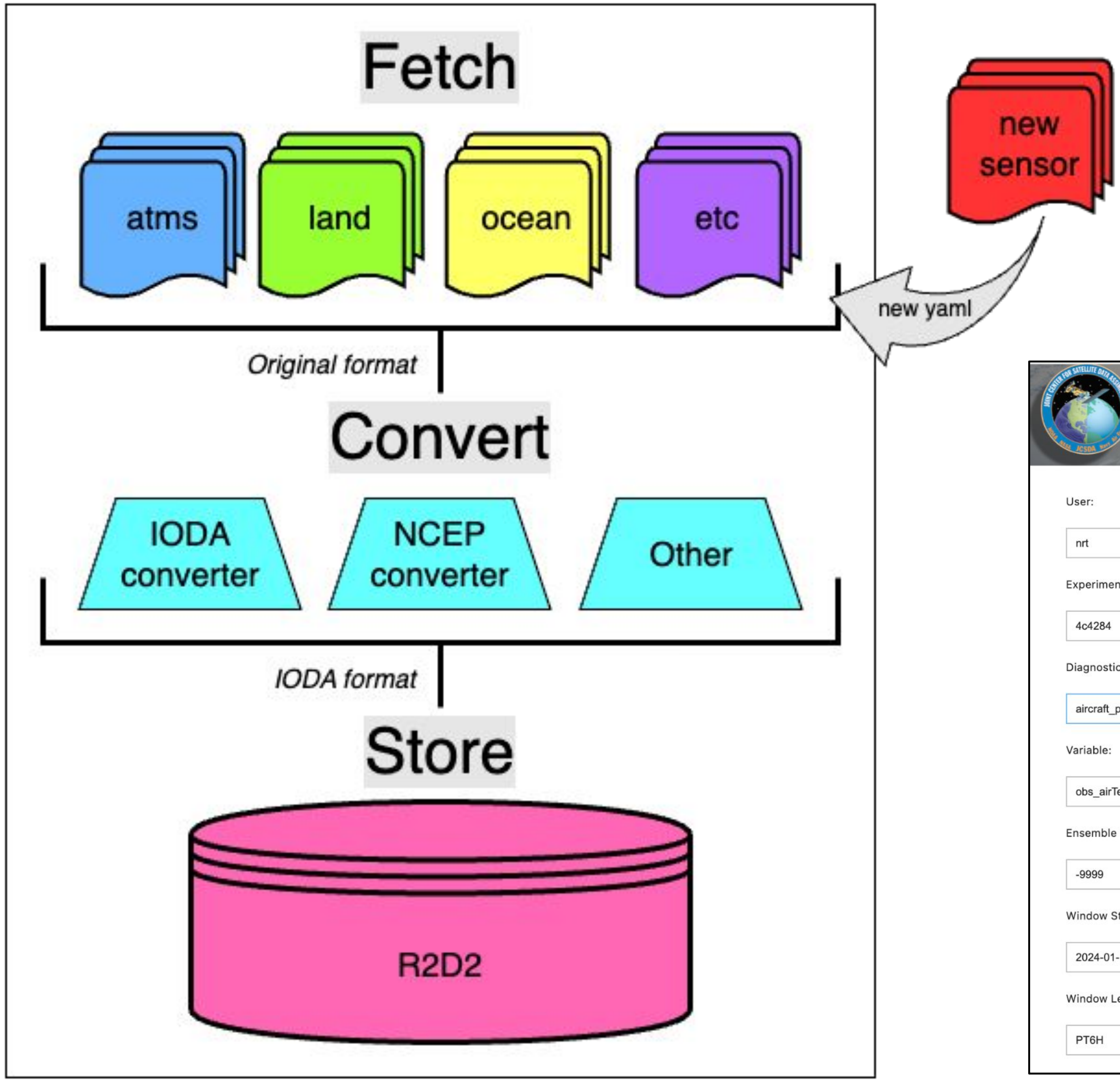
How do I use R2D2's client API?

```
import r2d2

r2d2.search(item, optional attrs,
limit_to_compute_host=False,
include_data_stores=False)

r2d2.store(item, attrs, source_file,
data_store (optional),
store_as_symlink=False)

r2d2.fetch(item, attrs, target_file,
data_store (optional))

r2d2.update(item, attrs, key, value)

r2d2.delete(item, attrs, data_store)
```

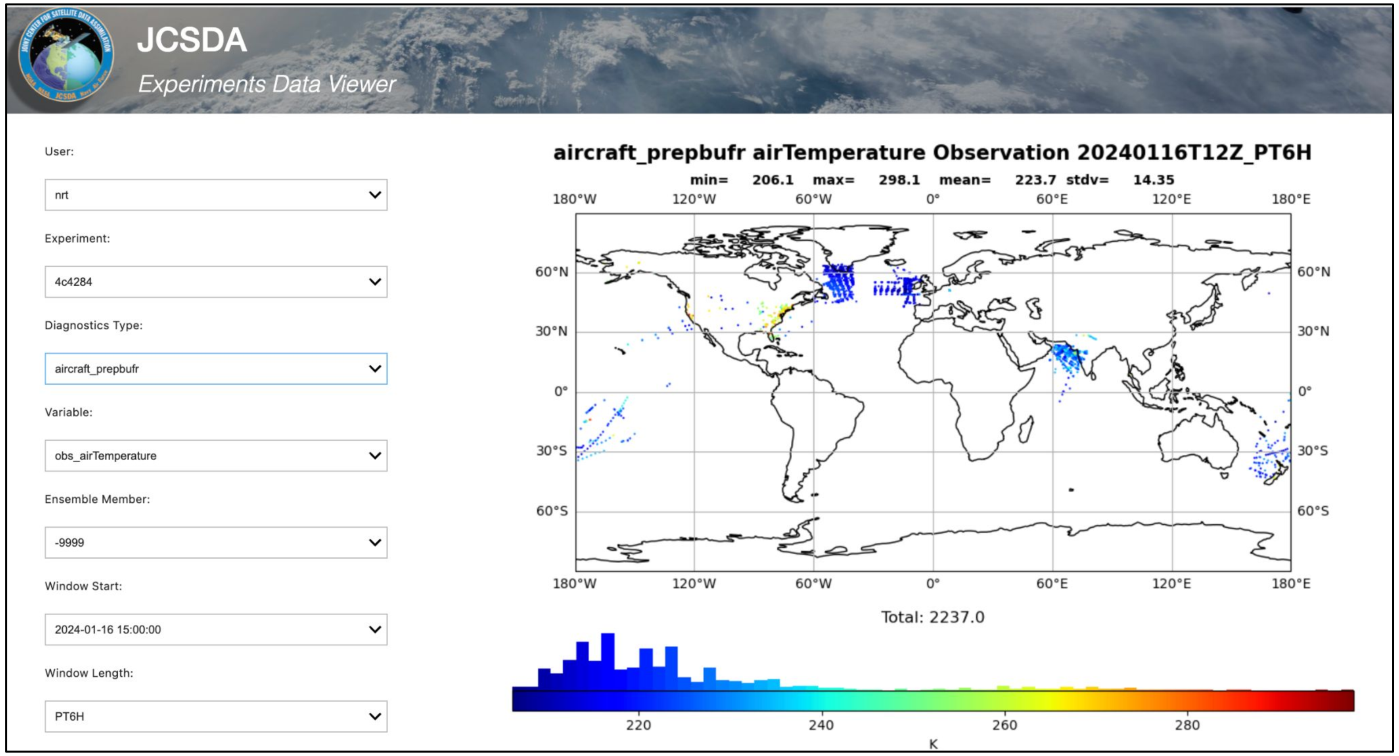
Required Data Item Keys

Item	Keys (^{R} indicates required)
analysis	model^R, experiment^R, file_extension^R, date^R, domain, file_type, tile, member
bias_correction	model^R, experiment^R, provider^R, observation_type^R, file_extension^R, file_type^R, date^R
diagnostic	experiment^R, file_extension^R, diagnostic_type^R, date^R
feedback	experiment^R, observation_type^R, file_extension^R, window_start^R, window_length^R, member
forecast	model^R, experiment^R, file_extension^R, resolution^R, step^R, date^R, domain, file_type, tile, member
media	experiment^R, observation_type^R, file_extension^R, plot_type^R, variable^R, window_start^R, window_length^R, member
observation	provider^R, observation_type^R, file_extension^R, window_start^R, window_length^R

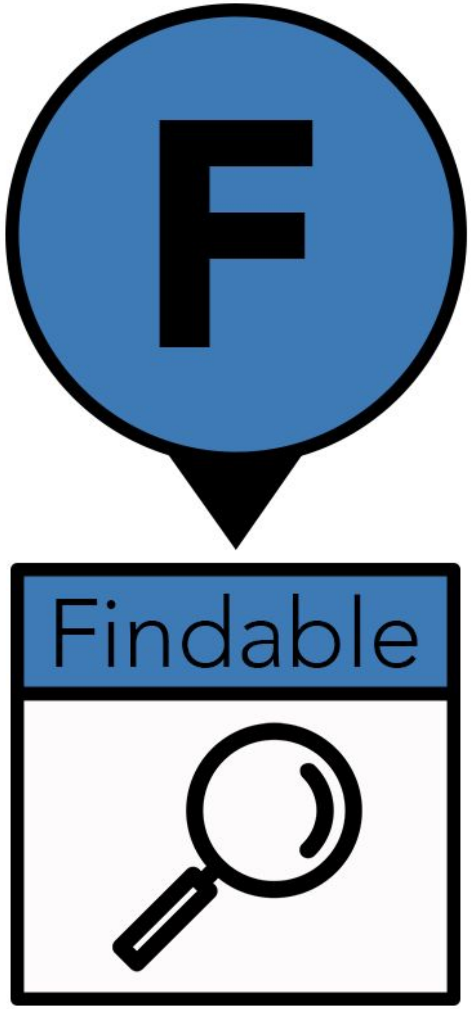
NRT FAIR Observational Processing Pipeline



FAIR enables progress towards "Continuous DA" in JEDI with near-real-time observational data pipelines



Findability with AI



Automated metadata generation

AI can scan large datasets to extract and standardize metadata, such as provenance, licensing, and methodological details. This reduces the burden on data creators and ensures consistency.

Smarter search

AI-powered search tools can use the comprehensive, standardized metadata to quickly find and recommend relevant datasets, even for complex scientific queries.

Knowledge graphs

AI can create knowledge graphs that semantically link disparate datasets across different sources, allowing for more powerful and contextualized searches.

Accessibility with AI

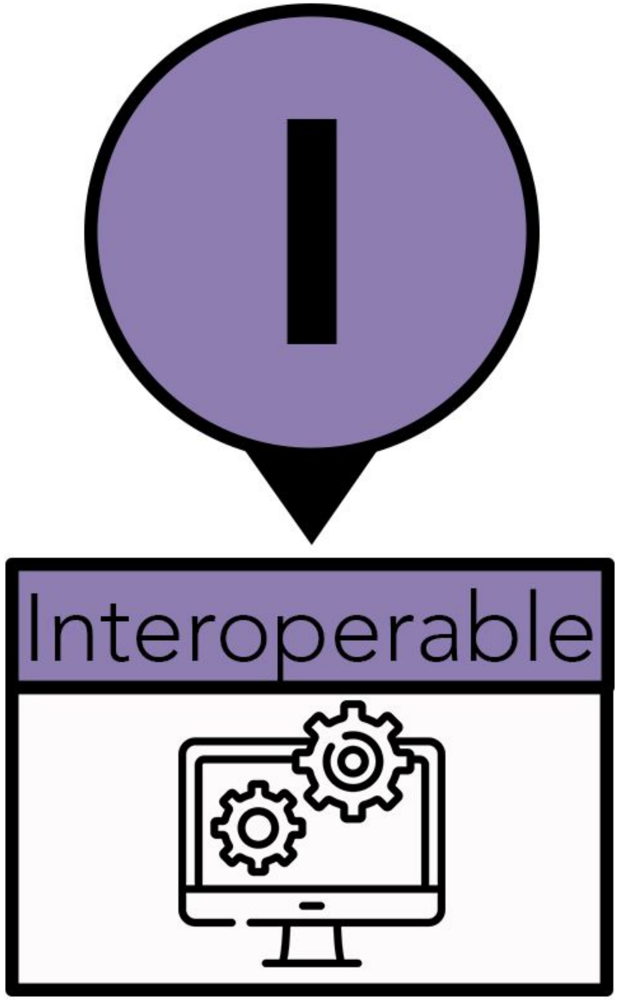
Seamless access across repositories

AI can be used to navigate complex data access protocols and train models across multiple, geographically dispersed data repositories, overcoming traditional barriers.

Optimized workflows

Advanced AI can streamline the process of transferring and retrieving large-scale datasets from storage to computing environments, such as supercomputers or cloud platforms.

Interoperability with AI



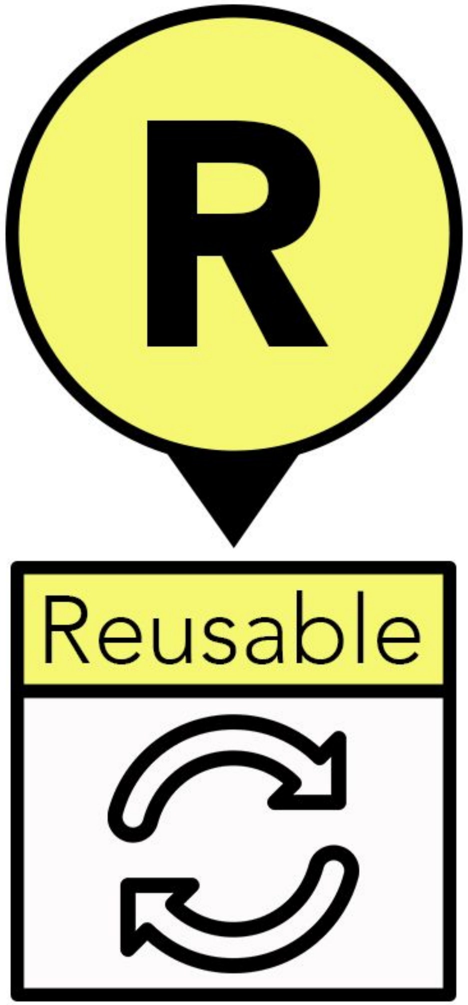
Data standardization and harmonization

AI and machine learning algorithms can analyze and reconcile data from hundreds of different sources, correcting errors and standardizing formats to ensure consistency.

Semantic mapping

AI uses semantic vocabularies and ontologies to add machine-readable context to data, enabling seamless mapping and integration between different systems.

Reusability with AI



Enhanced data quality and accuracy

AI algorithms can systematically detect and fix inconsistencies and errors in data, ensuring higher quality and reliability for any future reuse.

Predictive insights

AI-driven data enrichment can add valuable context to datasets enabling the data to be repurposed for new types of analysis.

Bias detection

AI can identify potential biases within datasets that could affect outcomes, allowing researchers to create more fair and trustworthy models.

Thank you!

Questions?

https://www.go-fair.org/