

# Kando: An iRODS Compatible Data Organizer for CKAN

Tanmay Dewangan

# What is Data Commons? What is CKAN?

## Data Commons

- Collaborative platform where datasets are stored, organized, and shared
- Following FAIR (Findable, Accessible, Interoperable, and Reusable) principles

## CKAN (Comprehensive Knowledge Archive Network)

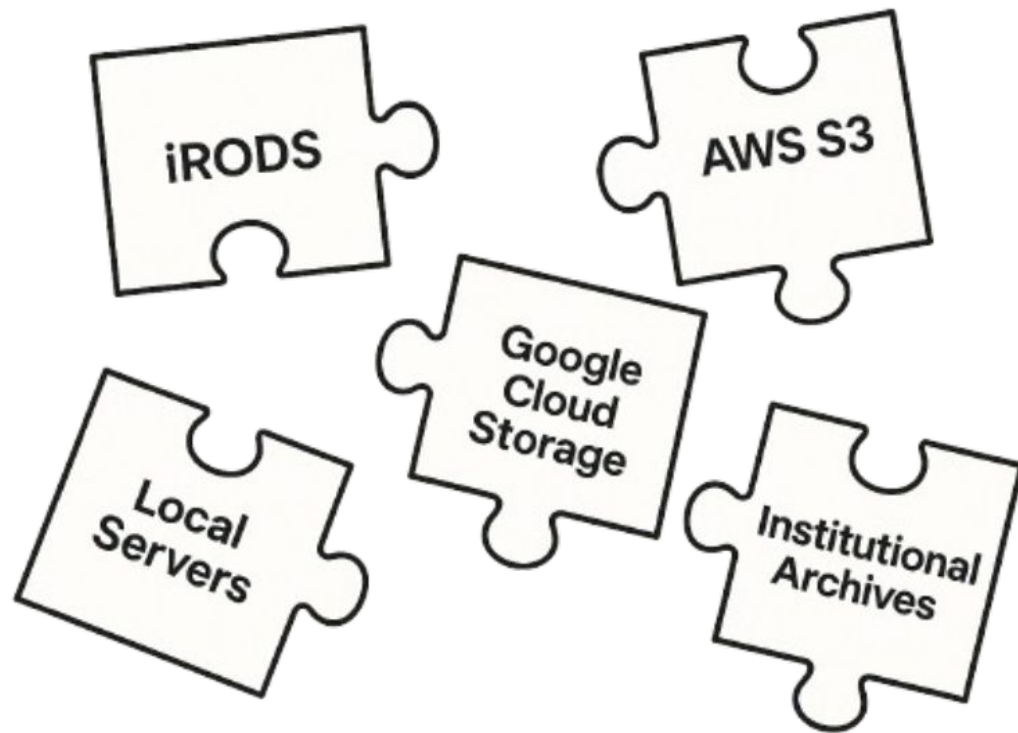
- open-source data commons system for publishing data, used by governments, research groups, institutions.



iRODS

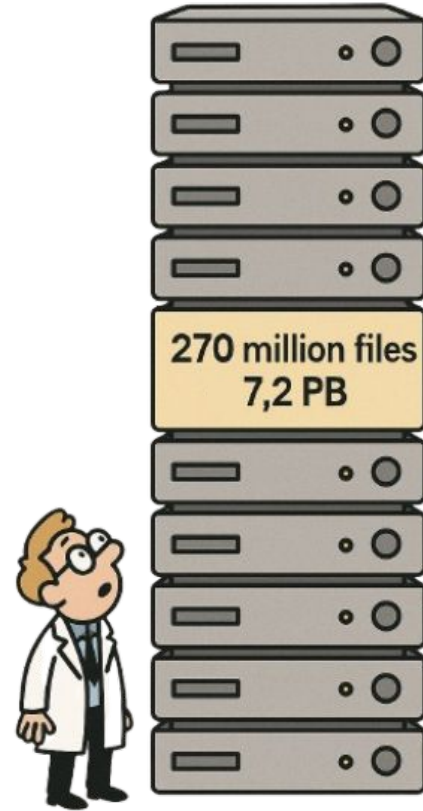
CKAN

# The Challenge — Fragmented Research Data

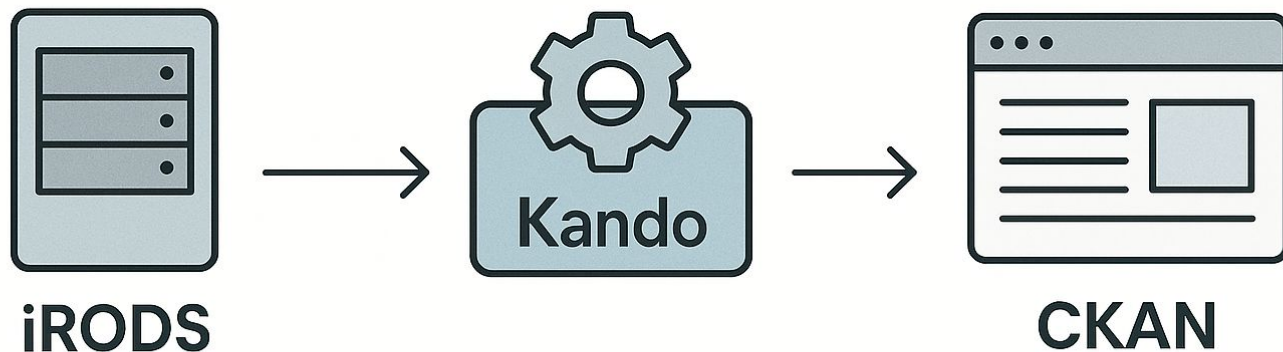


# CyVerse Scale & Impact

- 270 million files (7.2 PB)
- 78 million project files (3.2 PB)
- 28 million public files (470 TB)
- 270 curated datasets
- 15,000 TB downloaded last year



# Kando Architecture – iRODS to CKAN



## Steps:



# Metadata Standards



## Croissant

- Designed for machine learning datasets
- Ensures datasets can be easily understood and used by different machine learning frameworks

## DCAT (Data Catalog)

- A W3C standard for describing datasets in a catalog
- Enhances the discoverability of datasets by providing a common structure for metadata.

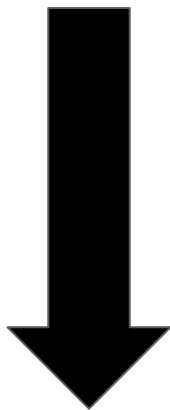


# CSV Files

String	Int	Date
a	1	2020-01
b	2	2020-02
c	3	2020-03



a	1	2020-01	b	2	2020-02	c	3	2020-03
---	---	---------	---	---	---------	---	---	---------



String	Int	Date
a	1	2020-01
b	2	2020-02
c	3	2020-03



a	b	c	1	2	3	2020-01	2020-02	2020-03
---	---	---	---	---	---	---------	---------	---------

# Parquet Files

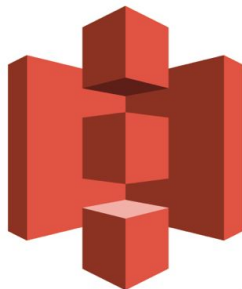
# Expanding to Cloud Integration: Why It Matters

- Many research datasets are now born in the cloud
- Public cloud storage is fast, cheap, and scalable
- Integrating cloud data into institutional data commons is key to supporting modern research workflows



# Cloud Integration of Kando

- Now supports **AWS S3** and **Google Cloud Storage buckets**
- Enables cloud-hosted data to be linked in CKAN data commons
- **Safe user-friendly workflow** as iRODS integration
- Next steps: full Parquet + metadata standardization for cloud workflows

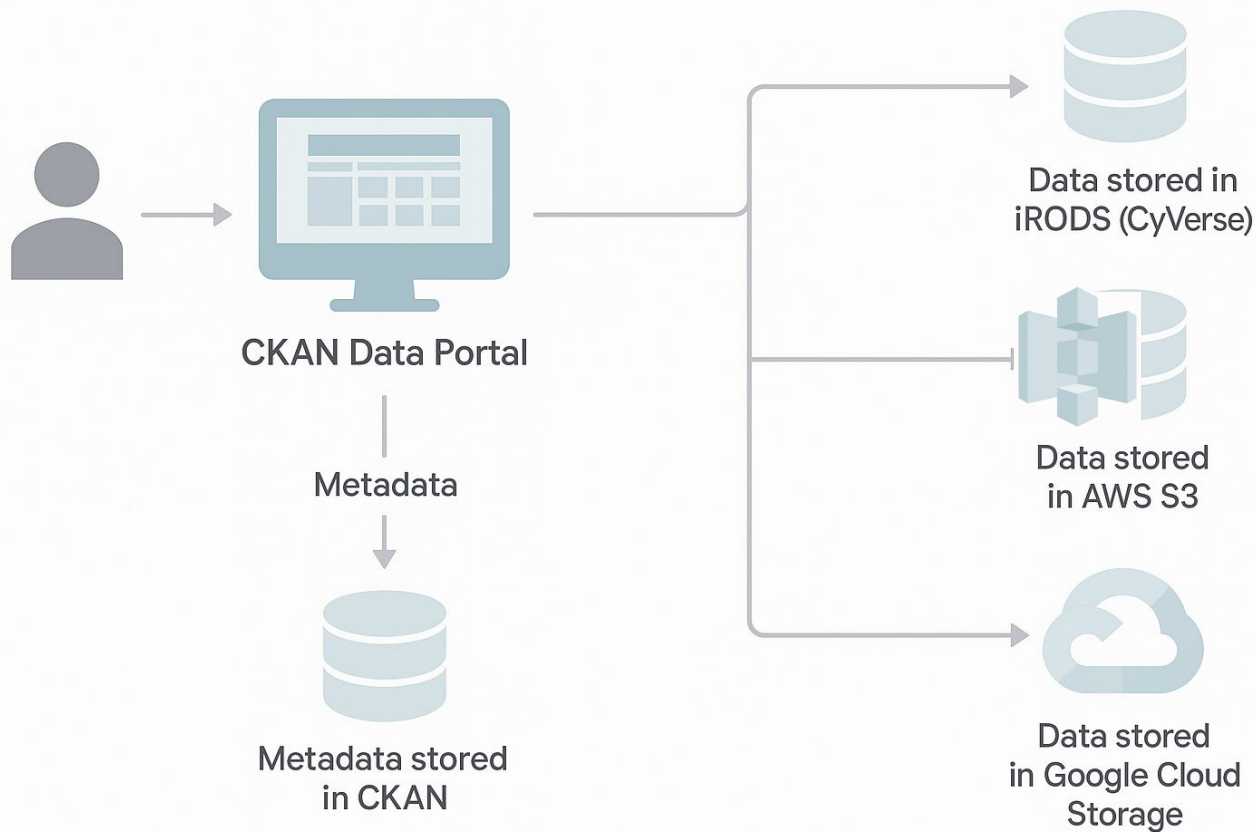


Amazon S3



Google Cloud Storage

# Where is the Data Actually Stored?



# Kando User Interface + Demo

[Replicate From CyVerse](#) [Generate Croissant JSON](#) [Generate DCAT JSON](#) [Upload Croissant JSON to CKAN](#) [Upload DCAT JSON to CKAN](#) [Replicate AWS Bucket](#) [Replicate GCS Bucket](#)

Username

Password


Path

☐ Convert CSV to Parquet

Submit

Output

# Why FAIR Data Matters for Researchers

- Reusable data → saves time and effort for future work
  - Interoperable data → supports cross-domain research
  - Compliant data → meets agency and publisher requirements
  - Discoverable data → increases research impact and visibility
- 
- A decorative graphic in the bottom right corner consisting of two overlapping curved shapes. The outer shape is light gray, and the inner shape is a darker blue, creating a layered, wave-like effect.

# Closing & Future Directions

- Integrated cloud + institutional data commons
- FAIR data alignment
- Simplified workflows
- Future Steps:
  - Private buckets / signed URL support
  - Parquet + metadata conversion for cloud
  - LLM-based metadata generation
  - Enhanced CKAN search & visualization





Thank you!  
Questions?