

# **IIT Dataverse: infrastrutture per la conservazione e la condivisione dei dati di ricerca**

**Valentina Pasquale**

Research Data Management and Stewardship Office Manager,  
Research Organization Directorate

**Anna Maddalena**

Application Management Office Manager,  
ICT Directorate

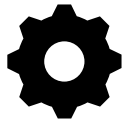
[rdm@iit.it](mailto:rdm@iit.it)

# Perché Dataverse?

**Dataverse** è un'applicazione web **open source** per condividere, conservare, citare ed esplorare i dati della ricerca, sviluppata a Harvard University.



Open-source, installato *on premises* e gestito *in-house*



Facilmente configurabile e integrabile con i nostri sistemi (ma non solo...)



Adottato a livello mondiale (e con una community molto attiva di sviluppatori ed utenti, vedi GDCC)



Offre supporto per schemi di metadati standard e customizzabili

**FAIR**

Offre supporto alla realizzazione dei principi FAIR





IIT DATAVERSE

[Terms of Use](#)



<https://dataverse.iit.it>

Metrics

1 Download

IIT Dataverse is a place for IIT scientists to preserve and share datasets supporting original research and used by Harvard University.

If you have IIT institutional credentials, click on "**Log in**" in the top right bar to start using IIT Dataverse.

To **get started**, have a look at the links below:

- [FAQ's](#) – more about depositing data in IIT Dataverse
- [Deposit checklist](#) – a short guide to dataset creation
- [IIT Dataverse Intranet page](#) – a more comprehensive reference for IIT Dataverse resources

Search this dataverse...

Find

[Advanced Search](#)

## FINDABLE

- ✓ Integrato con DataCite per l'assegnazione di DOIs.
- ✓ Censito in [www.re3data.org](http://www.re3data.org).
- ✓ Censito in [FAIRsharing.org](http://FAIRsharing.org).
- ✓ Integrato con OpenAIRE.
- ✓ Indicizzato sui principali motori di ricerca.

## ACCESSIBLE

- ✓ Accessibile attraverso protocollo standard HTTPS.
- ✓ APIs per upload / download dei dati.

## INTEROPERABLE

- ✓ Supporto per metadati disciplinari.
- ✓ Integrabile con vocabolari esterni.

## REUSABLE

- ✓ Supporto per diversi tipi di licenze.
- ✓ Possibilità di implementare *restricted datasets*.



# Di cosa parliamo oggi

- Quali dati in IIT Dataverse?
- Flusso di pubblicazione
- *Restricted data sharing*
- Supporto per *big dataset*
- Dataverse nella strategia istituzionale per i dati della ricerca

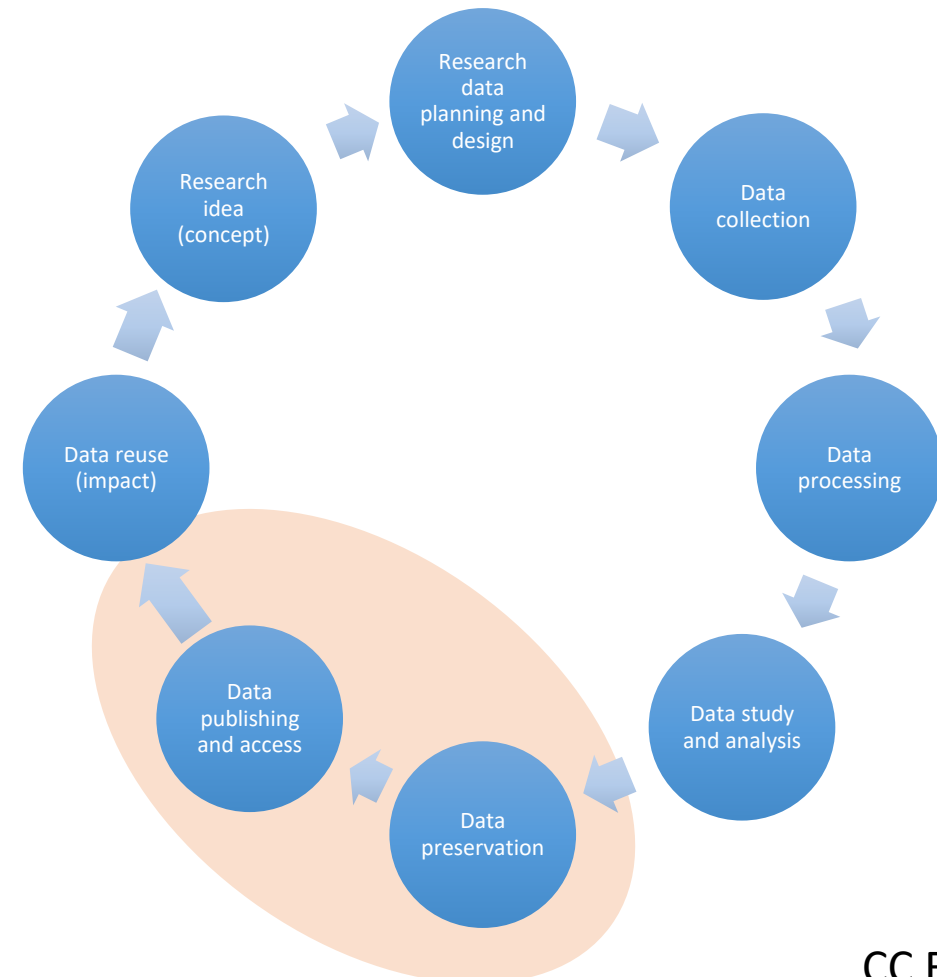
# Quali dati in IIT Dataverse?

- In IIT Dataverse accettiamo **dati di ricerca pronti per la conservazione a lungo termine e/o la pubblicazione.**
- Non si possono caricare in IIT Dataverse:

 **dati personali**

 **dati ad alto rischio**

 **pubblicazioni scientifiche**





# Data Risk Classification

The purpose of the Data Risk Classification is to establish the rules for classifying and securing information used by the IIT Foundation

## Low Risk



The information and systems are of a Low level of risk if:

- They are not of a High or Medium level of risk
- The information or ICT services are typically addressed to the public or external bodies
- The loss of Confidentiality, Integrity or Availability of the information would not have a significant impact on the Intellectual Property or on the mission or reputation of the Foundation, nor would it have a significant economic or legal impact.

## Medium Risk



The information and systems are of a Medium level of risk if:

- They are not of a High level of risk
- The information or ICT services are not typically addressed to the public or external bodies
- The loss of Confidentiality, Integrity or Availability of the information can have a mildly adverse impact on the Intellectual Property or on the mission or reputation of the Foundation, or a limited economic or legal impact.

**\* NOT personal data**



## High Risk



The information and systems are of a High level of risk if at least one of the following conditions are true:

- The loss of Confidentiality, Integrity or Availability of the information can have a significant adverse impact on Intellectual Property or on the mission or reputation of the Foundation, or a significant economic or legal impact
- The protection of the information is required by Law
- The protection of the information is required by contracts or certifications of national or international standards (such as the ISO standards)

# Bilanciare apertura e sicurezza

- E' **responsabilità** dei nostri **Principal Investigator** valutare quali dati condividere, anche in base alla classe di rischio associata.
- Forniamo **documentazione**, **linee guida** e **supporto** per operare queste valutazioni e decidere quali asset usare.

Data Risk  
Classification



Misure di sicurezza

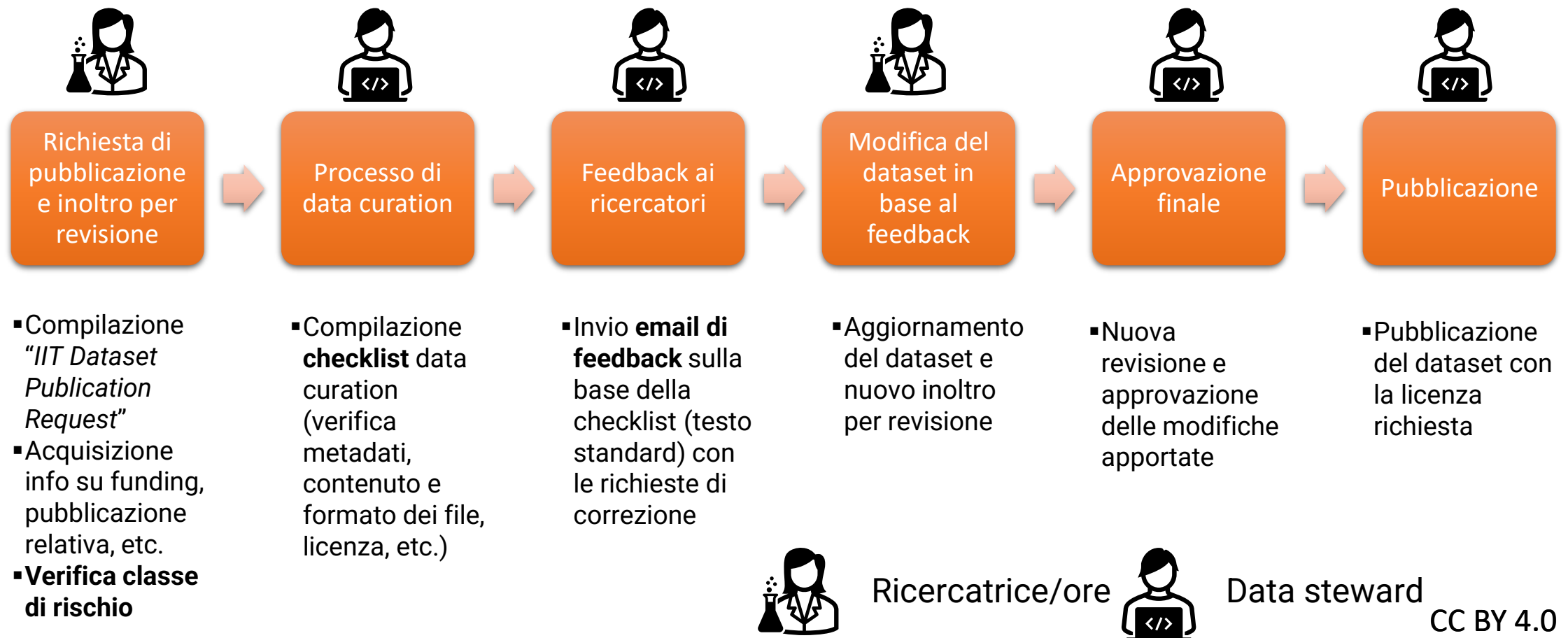


■ Website

■ Dataverse

■ Cloud ■ Internal storage

# Flusso di pubblicazione

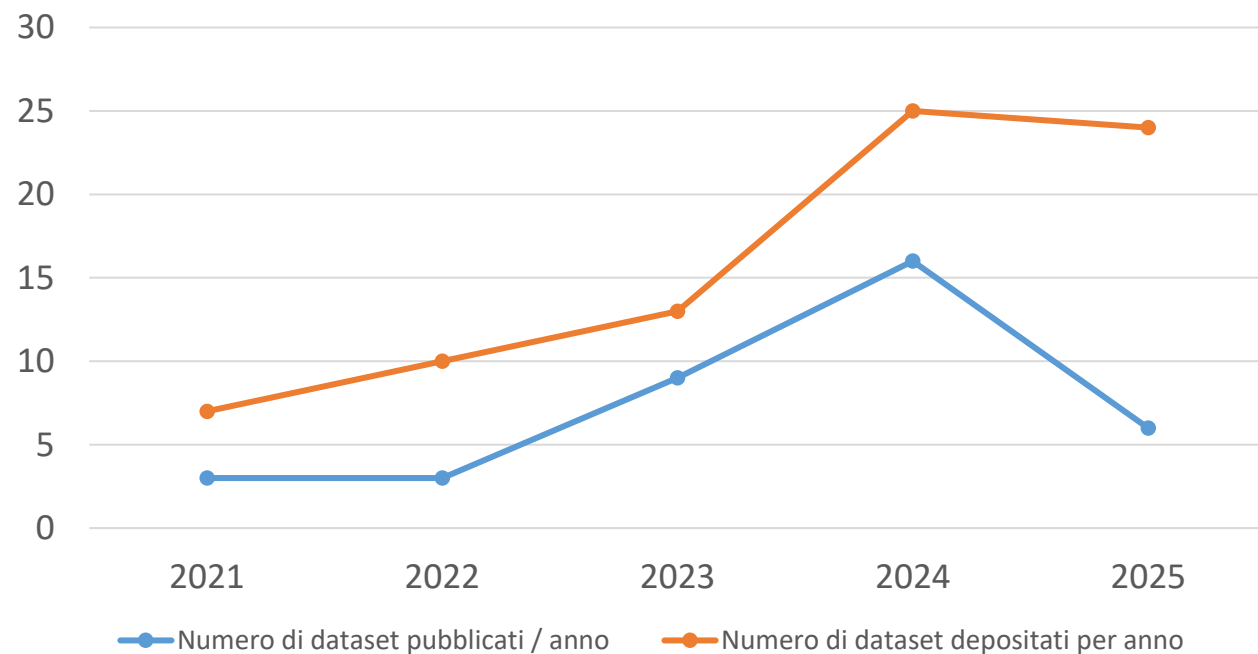




# Qualche numero su IIT Dataverse

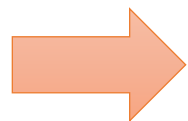
- L'uso per il deposito aumenta nel tempo (costante nel 2025)
- Nel 2025 c'è una flessione del numero di dataset pubblicati (alcuni sono attualmente in revisione per la pubblicazione).
- Tutti i dataset pubblicati sono *open*.
- Tutti i dataset pubblicati sono legati ad una pubblicazione scientifica e/o ad un progetto finanziato.

IIT Dataverse - numero di dataset per anno



# Nuovo flusso: “restricted data sharing”

- Introduzione della possibilità di pubblicare “**dati di ricerca**” **non completamente anonimizzabili** in modalità “**restricted**” (es., riprese video raffiguranti volti)
- In questo caso solo i metadati saranno resi pubblici (licenza CC0), mentre i dati saranno scaricabili da un utente esterno solo previa richiesta e firma di un **Data Access Agreement** fra IIT e l'Istituzione richiedente (in presenza di alcuni requisiti di base).



***Cosa è necessario prevedere?***

# Restricted data sharing (*ongoing*)



Aggiornamento dei *terms of use*, per abilitare il caricamento di dati personali (comunque limitato)



Attivazione login utenti esterni (via EduGAIN)



Gestione privacy dei dati personali degli utenti esterni (informative)



Predisposizione dei *custom dataset terms*



Predisposizione del *Data Access Agreement* (da customizzare per ogni dataset)



Diagramma del flusso di lavoro (attività, strumenti, ruoli, responsabilità)

\*in collaborazione con **Legal Affairs Directorate** e **ICT Directorate**

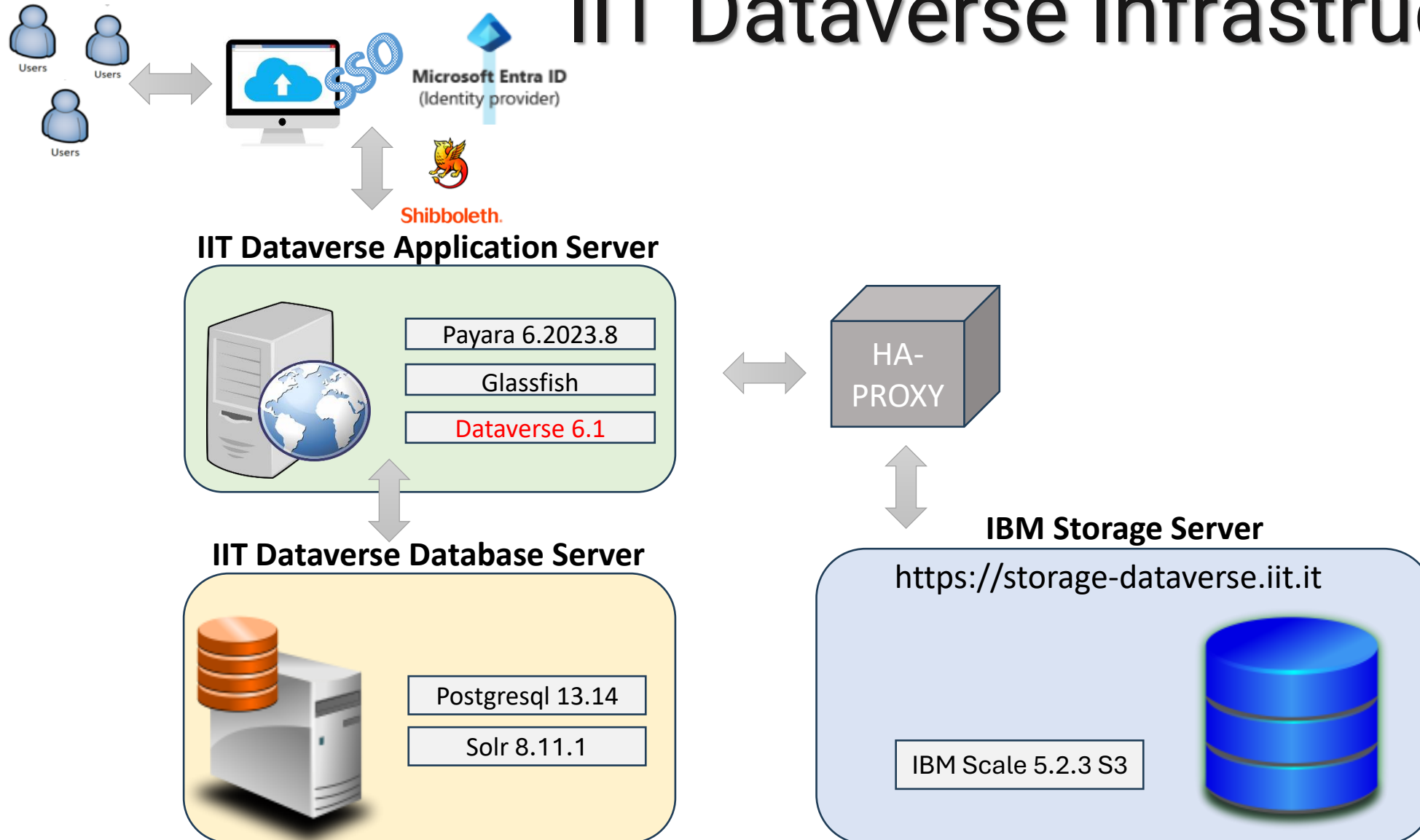
# Supporto per *big dataset* (ongoing)

- In IIT Dataverse accettiamo **dataset sino a 50 GB\***, ma riceviamo numerose richieste per dataset di dimensioni oltre i 100 GB e cerchiamo di soddisfarle quando possibile.
- Per questi dataset è necessario offrire un supporto extra:
  - Utilizziamo un S3-compatible storage (IBM Storage)
  - Upload / download tramite API (uso del DV Uploader in *upload*, fornitura di *script* per il *download*)
  - Sperimentazione in corso sull'uso del *direct upload / download*
  - Possibile uso di Globus per *big data transfer* integrato con Dataverse

\* Sulla base dei termini di utilizzo standard.

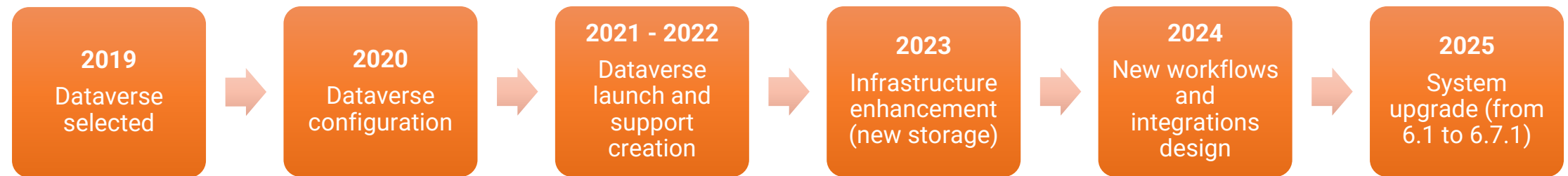


# IIT Dataverse Infrastructure

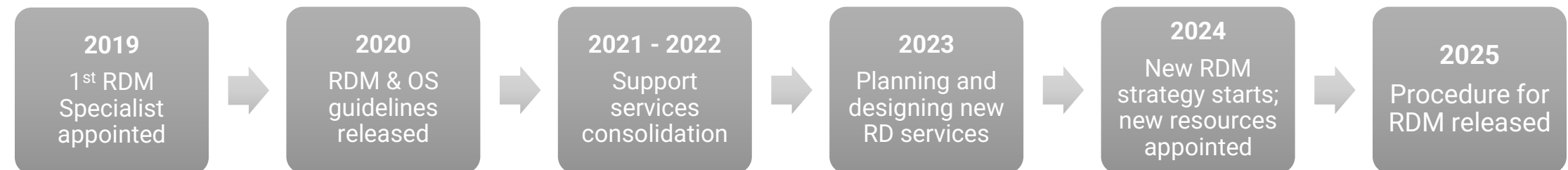


# RDM @ IIT: 2019 - 2024

## *IIT Dataverse*



## *Strategy & policies*



## Procedura per “Research Data Management”

- Rilasciata ad ottobre 2025
- Si affianca al Codice di Comportamento e di Condotta Scientifica, richiamando il mandato alla conservazione dei dati di ricerca e all’adozione dei principi FAIR
- E’ strutturata in base alle fasi del Research Data Lifecycle (pianificazione, raccolta, documentazione e verifica, conservazione, condivisione)
- Utile per puntare i ricercatori alle risorse messe a disposizione (strumenti, linee guida, manuali)
- Definisce ruoli, azioni e responsabilità

# Nuovo progetto RDM @ IIT: 2024 - 2028

Which tools can I use to manage research data?

## Infrastructure, tools, and services

Build an ecosystem of seamlessly integrated tools and services to each phase of the research data lifecycle.

- ❑ Infrastructure services
- ❑ Application services
- ❑ Platform services

How can I manage research data?

## Policies, processes and standards

Provide policies and guidance on all data-related processes.

- ❑ Data FAIRification
- ❑ Data protection (privacy, IP)
- ❑ Data sharing and preservation

Who is going to support me in managing research data?

## Support and training

Appoint data professionals (stewards, IT specialists, developers) to build a multi-layer support service.

- ❑ Strengthen institutional centralized support
- ❑ Build decentralized first-line & discipline-specific support

## Community-building and coordination

Engage researchers, coordinate collaboration and convergence

**Aim: Unlocking the full potential of Research Data as the main fuel to ML/AI-based innovative scientific approaches**

[Appendix 4 – Data Management of the Strategic Plan 2024-2029]

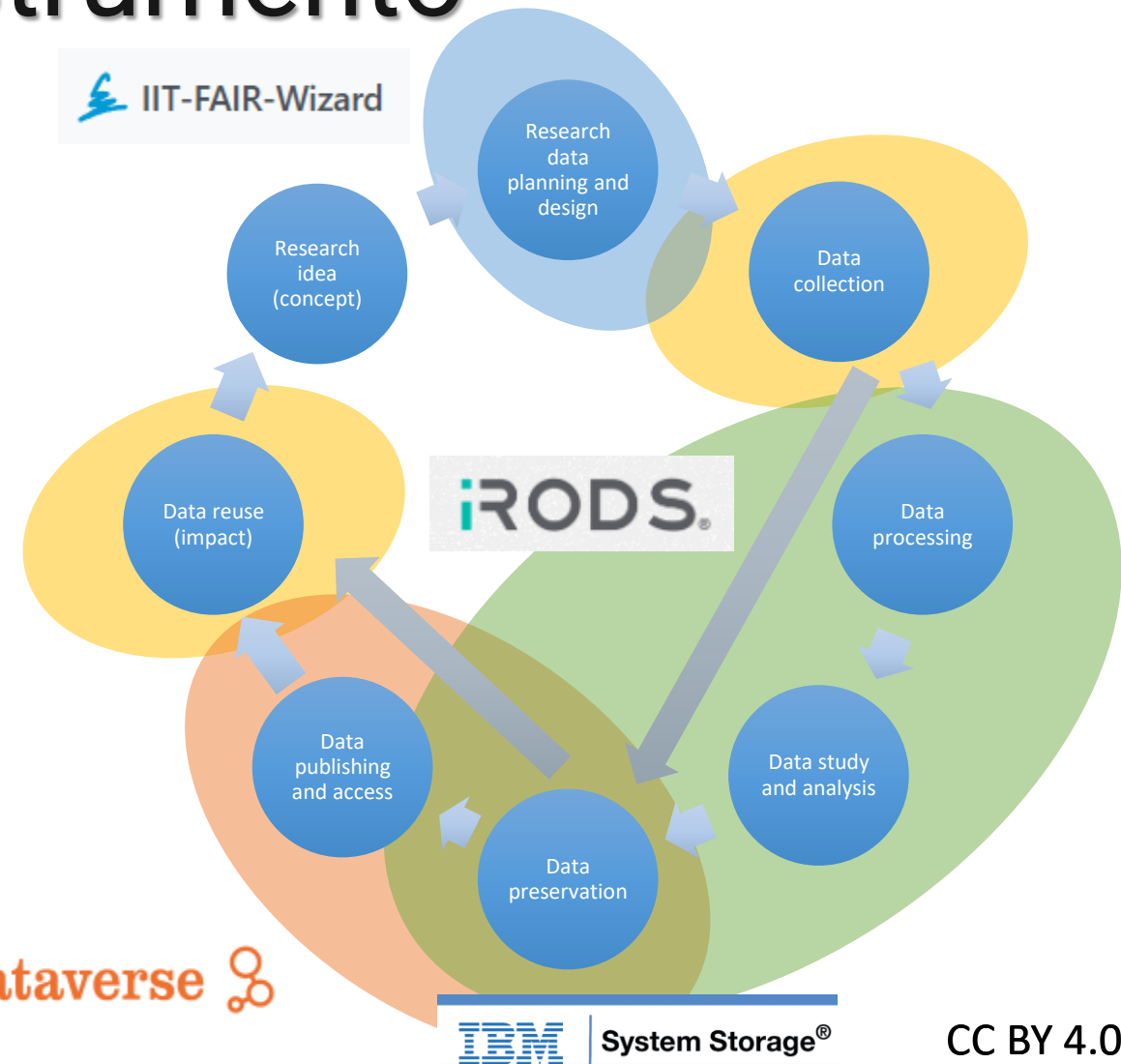
*Strategic overview and methodologies*



# Ogni fase ha il suo strumento

- FAIR Wizard: gestione Data Management Plans
- iRODS + Data Portal + API clients: piattaforma per l'organizzazione dei dati, la collaborazione interna, la conservazione e la tracciabilità
- IIT DATAVERSE: condivisione e pubblicazione dei dati
- IBM Storage: uso attivo dei dati

Lavoriamo sull'integrazione degli strumenti per creare un flusso di lavoro che sia il più possibile "seamless".



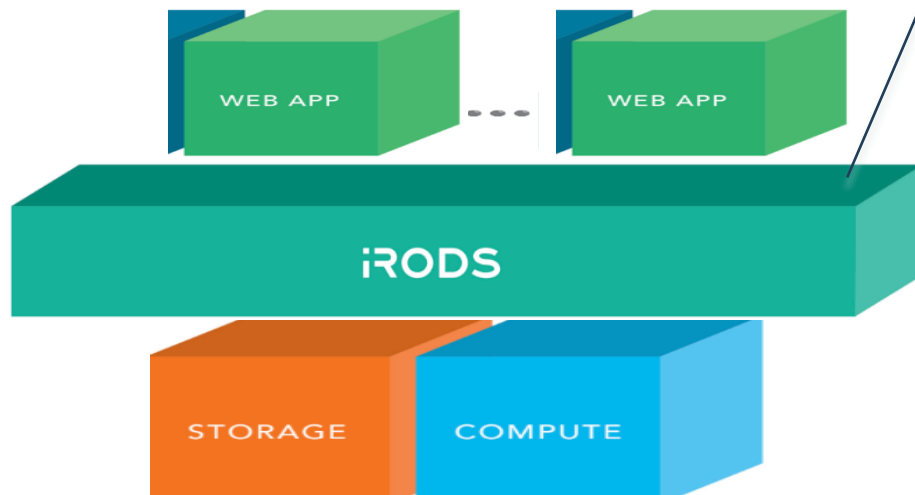
# Piattaforma per la gestione dei dati

iRODS provides a layer of abstraction which integrates with your pre-existing infrastructure.

This flexibility allows your infrastructure to continue to change over time.

**iRODS.**

 **Authentication via Institutional accounts**



**Virtualizzazione dei Dati:** filesystem virtuale e accesso a sistemi di archiviazione distribuiti

**Data Discovery:** catalogo dei metadati e ricerca su larga scala

**Automazione dei Workflow:** motore di regole per l'automazione dei flussi di lavoro

**Collaborazione Sicura:** garantita tramite ticket, permessi e distribuzione federata

Questa piattaforma deve garantire ai ricercatori la possibilità di **salvare, organizzare, documentare, analizzare** i dati con facilità, creando dei dataset **FAIR-by-design**, pronti per essere conservati e condivisi in modo aperto quando possibile.

# Sfide per il futuro (Dataverse)

- **Incrementare** il supporto per i ***big data*** (*direct upload*, Globus)
- **Incrementare** il supporto all'uso di *standard & custom metadata schema*
- **Integrare** la nuova piattaforma con Dataverse, per facilitare il deposito e la condivisione di dati FAIR-by-design
- **Tracciare** i dataset pubblicati da ricercatori IIT (non solo in IIT Dataverse), anche ai fini della valutazione



# Grazie per l'attenzione!



[rdm@iit.it](mailto:rdm@iit.it)