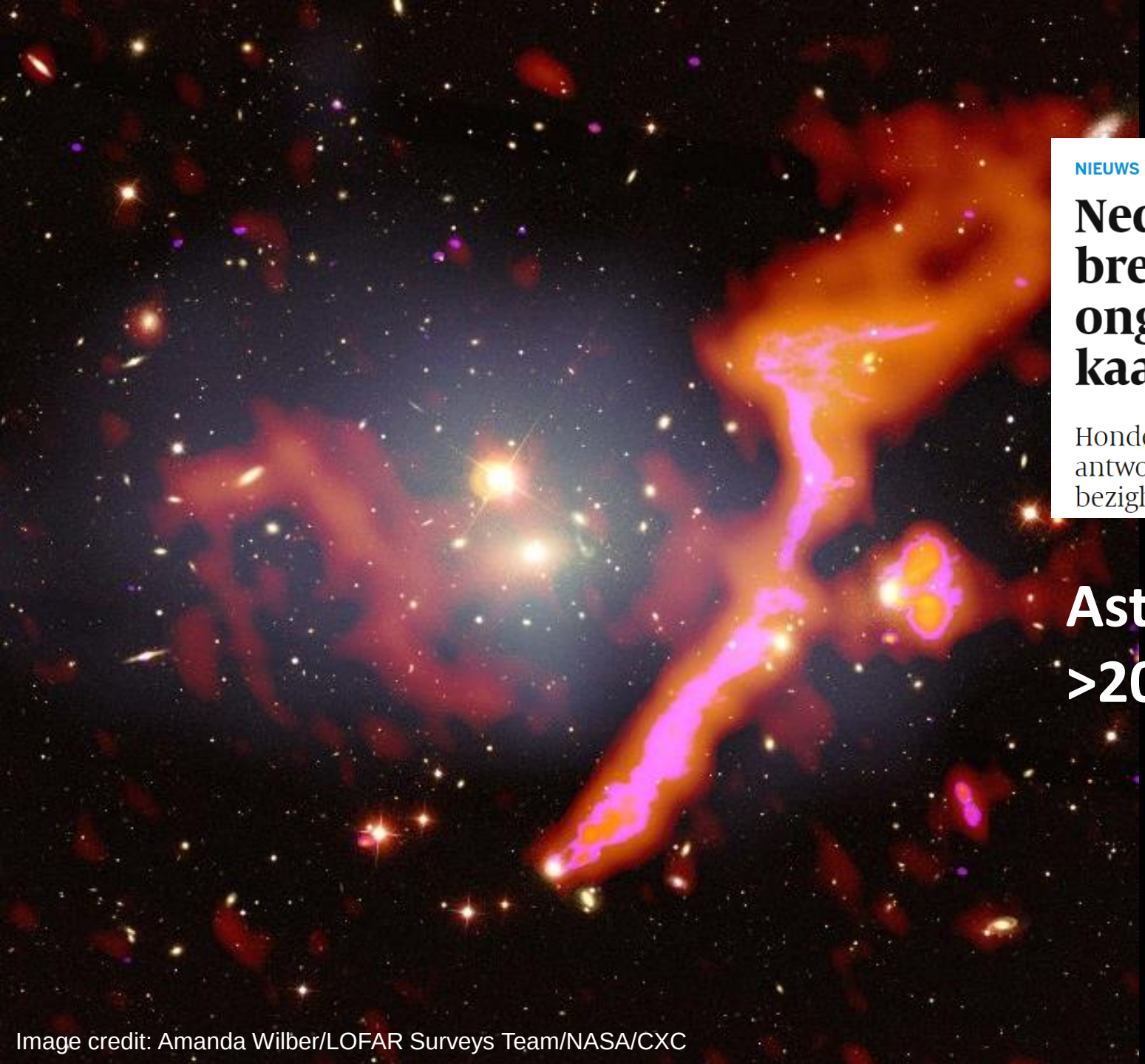
IRODS IN CONTEXT

EXPLORING INTEGRATIONS BETWEEN IRODS AND RESEARCHDRIVE / OWNCLOUD

HYLKE KOERS, GROUP LEADER DATA MANAGEMENT SERVICES

Introducing SURF

- SURF is the collaborative ICT organisation for Dutch education and research
- SURF offers students, lecturers and scientists in the Netherlands access to the best possible internet and ICT facilities
- SURF is a cooperation; its members are
  - Universities (14) & UMC's (8)
  - HBO (33) & MBO (43)
  - Other research organizations in the Netherlands



NIEUWS RADIOTELESCOOP LOFAR

Nederlandse radiotelescoop brengt noordelijke hemel ongekend nauwkeurig in kaart

Honderdduizenden nieuw ontdekte sterrenstelsels en antwoorden op raadsels die astronomen al jaren bezighouden. Dat is de eerste oogst van een grootschalig

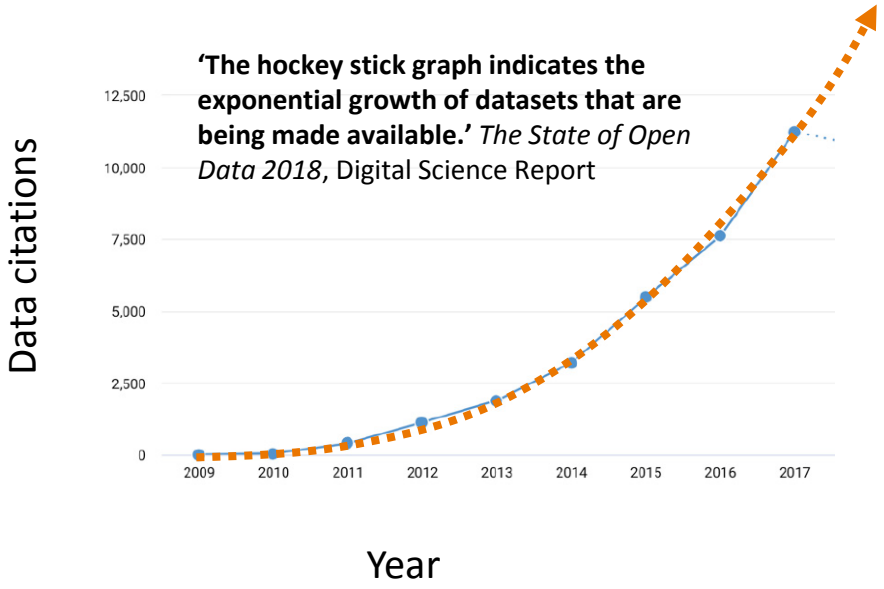
Astronomy meets Big Data: >20 Petabyte

Image credit: Amanda Wilber/LOFAR Surveys Team/NASA/CXC

Drivers for better RDM at Dutch research institutes

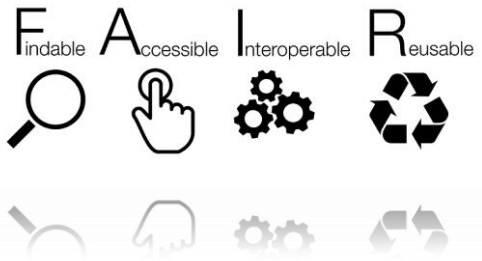
Lots of data

'The hockey stick graph indicates the exponential growth of datasets that are being made available.' *The State of Open Data 2018*, Digital Science Report



Research becomes more data-intensive and more interdisciplinary – and researchers need the right tools to do their job (in a way that complies with their institute's policies & guidelines)

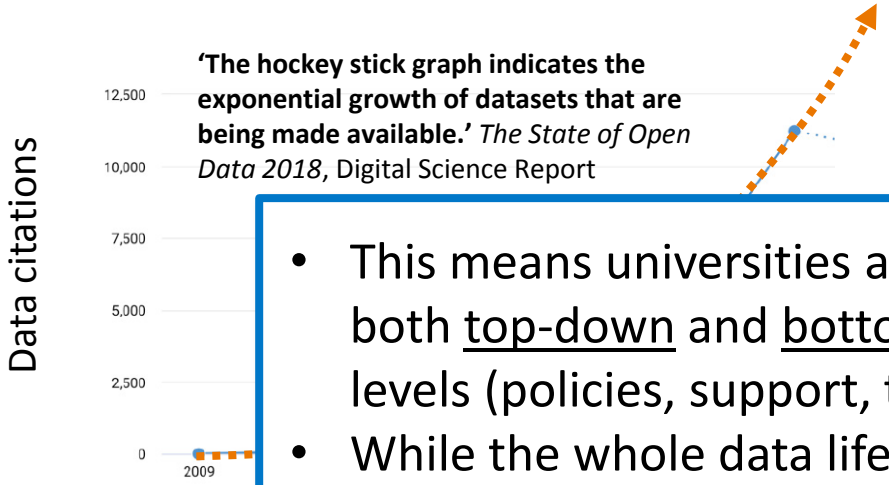
Lots of attention, lots of ambition



The FAIR Principles and Open Science are on the agenda of university boards, funders and the government.

Drivers for better RDM at Dutch research institutes

Lots of data



'The hockey stick graph indicates the exponential growth of datasets that are being made available.' *The State of Open Data 2018*, Digital Science Report

Lots of attention, lots of ambition

- This means universities and faculties experience a sense of urgency – both top-down and bottom-up – to offer better support for RDM on all levels (policies, support, technology, etc.)
- While the whole data life-cycle is relevant, long-term archival and publication of data are often seen as a priority.

Research becomes more data-intensive and more interdisciplinary – and researchers need the right tools to do their job (in a way that complies with their institute's policies & guidelines)

The FAIR Principles and Open Science are on the agenda of university boards, funders and the government.

The plot thickens… introducing our lead actors:

Stefan

Mara

Ayoub

This is **Stefan**. He's a bright and already accomplished postdoc in bio-informatics

- Used to working with large data
- Happy at the command line
- Used to writing her own data processing & analysis scripts
- Needs to adhere to University's policies regarding data archival.

This is **Mara**. She's a bright young PhD student in social sciences

- Data is usually small and in standard office formats
- Likes her GUI
- Uses standard analysis tools like SPSS
- Needs to adhere to University's policies regarding data archival.

This is **Ayoub**. He's a bright and driven data steward passionate about FAIR data.

- His job is to make sure that all data produced at the university is properly managed: archival, publication, right metadata standards.
- He wants to provide researchers the right tools and that fit into their daily workflow.
- Needs a consistent view on what data is produced

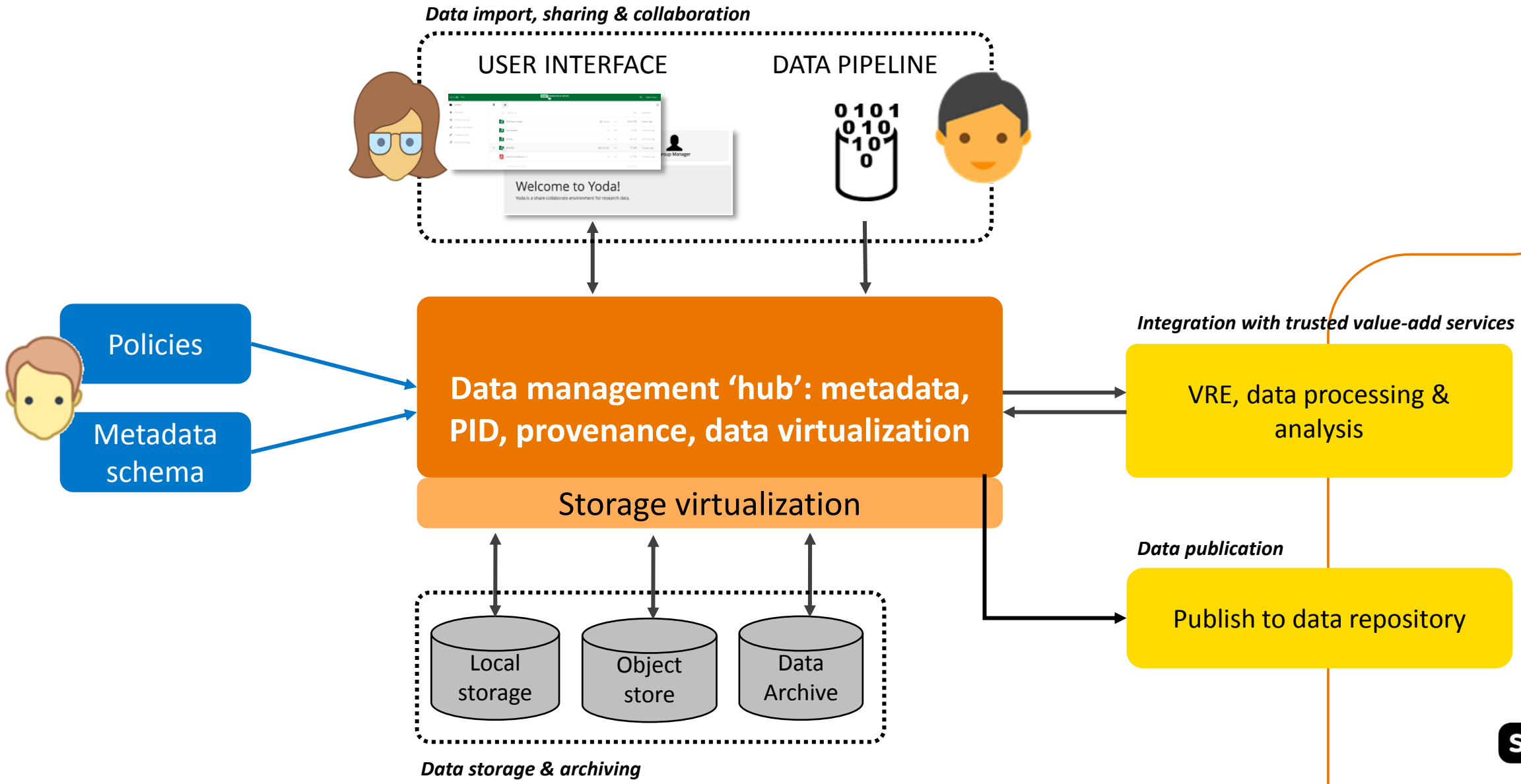
How to meet the needs of these different actors?

Especially with different institutes have common needs but different local contexts…

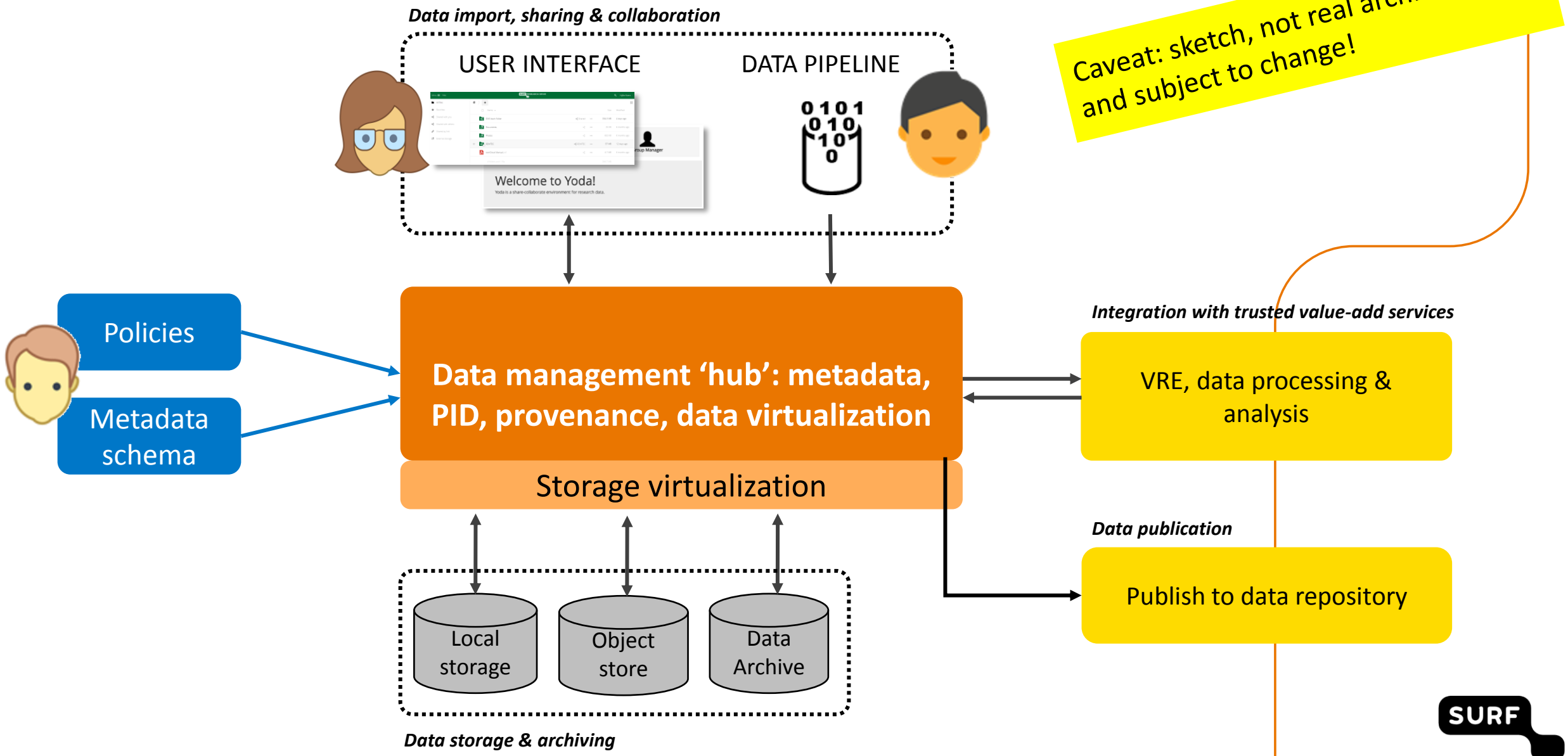
How to meet the needs of these different actors?

Especially with different institutes have common needs but different local contexts…

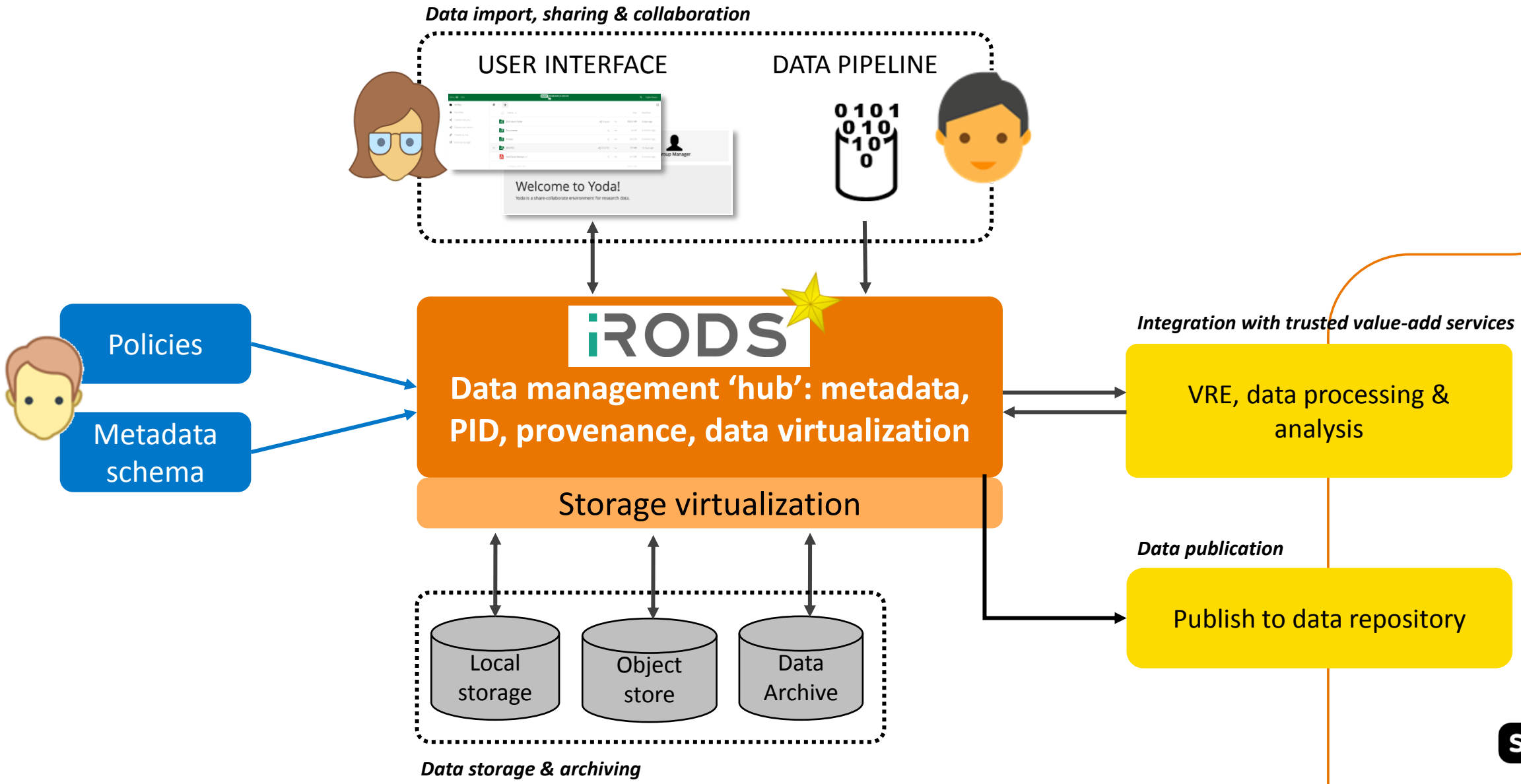
Our approach: a modular ‘framework’ for RDM



Our approach: a modular ‘framework’ for RDM



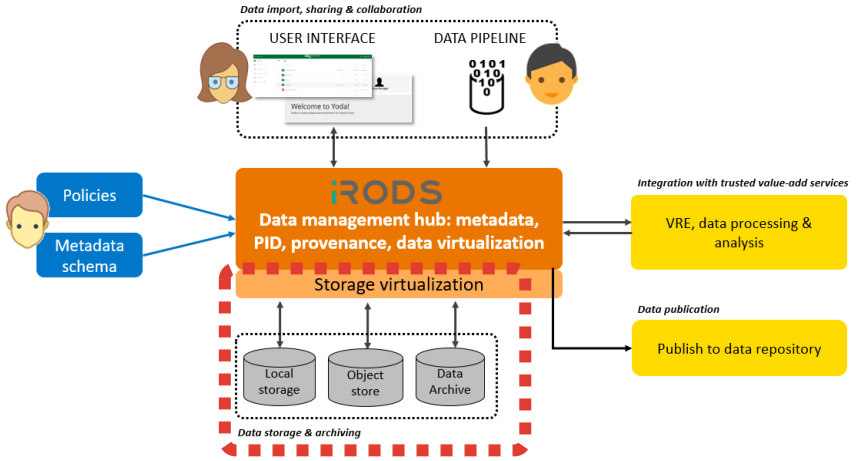
Our approach: a modular ‘framework’ for RDM



RDM Platform module (1): *Storage scale-out service*

- **SURF Data Archive** offers large-scale, cost-effective (and “green”) storage for long-term data preservation
- **The iRODS-to-Data Archive connector enables institutes to connect their iRODS-based RDM platform to the SURF Data Archive – with minimal installation and minimal overhead.**
- Provides layers of **storage abstraction** and **virtualisation**, iRODS rules attached to the services in order to **automate storage tiering and data movement tasks**.
- Can be **configured** and **tailored** to individual needs and policies re: long-term preservation
- A common use case is to deploy the Data Archive as a **scale-out solution** alongside the institutional repository.

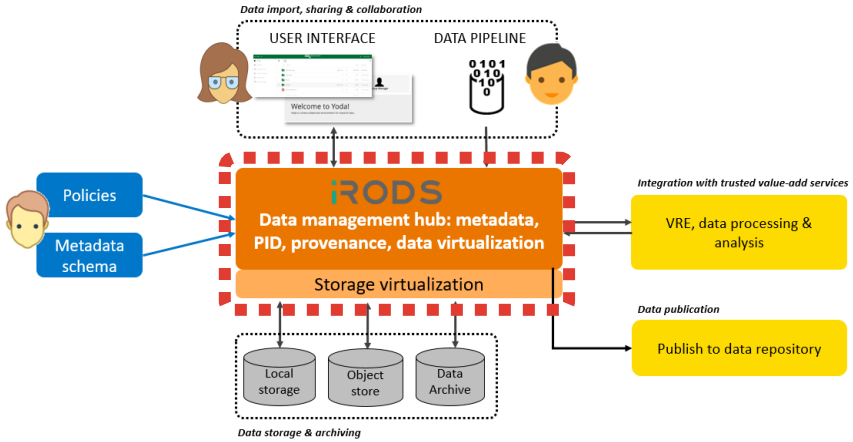
Developed and tested in POC’s and pilots with UU, ASTRON, MUMC, and others



RDM Platform module (2): *iRODS hosting*

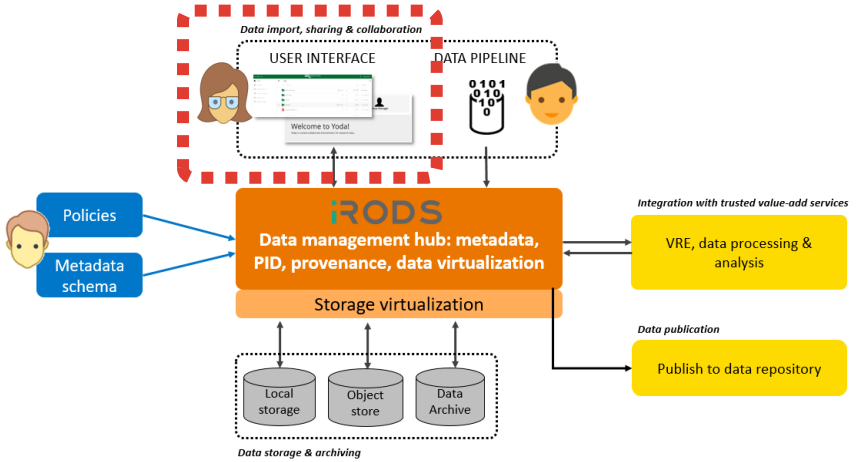
- iRODS is middle-ware: **powerful and versatile;** but also **requiring specific expertise** to set-up, configure, and integrate
- **The iRODS hosting service (PaaS / Iaas) allows institutes to benefit from the value that iRODS delivers - without having to develop detailed and specific expertise**
- Support available for **customization** and **integration** in local context
- Accelerating the development of iRODS-based RDM services at a reduced total cost of ownership.

Testing through POC's and pilots with UvA, WUR, and others



RDM Platform module (3): *User Interfaces*

- **iRODS does not come with a graphical UI out-of-the-box**, while many researchers (and data stewards) need a GUI to work effectively
- Fortunately, iRODS can be integrated with existing portals and/or with purpose-built front-ends.



RDM Platform module (3): *User Interfaces*

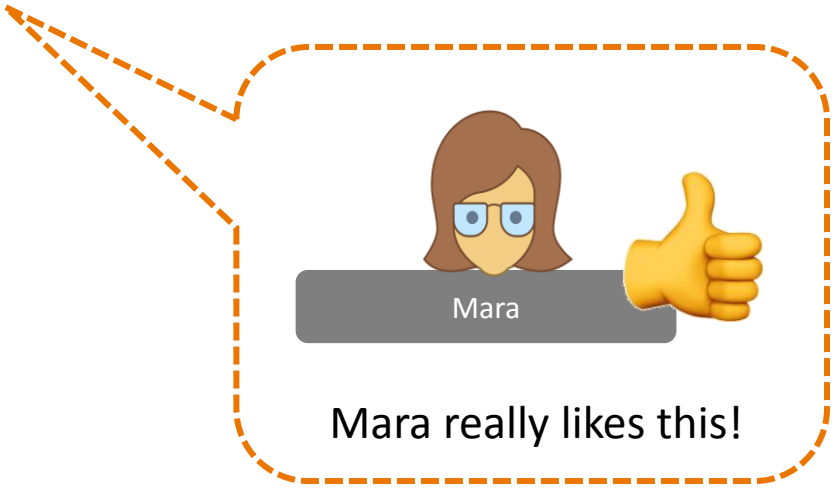
- **iRODS does not come with a graphical UI out-of-the-box**, while many researchers (and data stewards) need a GUI to work effectively
- Fortunately, iRODS can be integrated with existing portals and/or with purpose-built front-ends.

SURF Research Drive

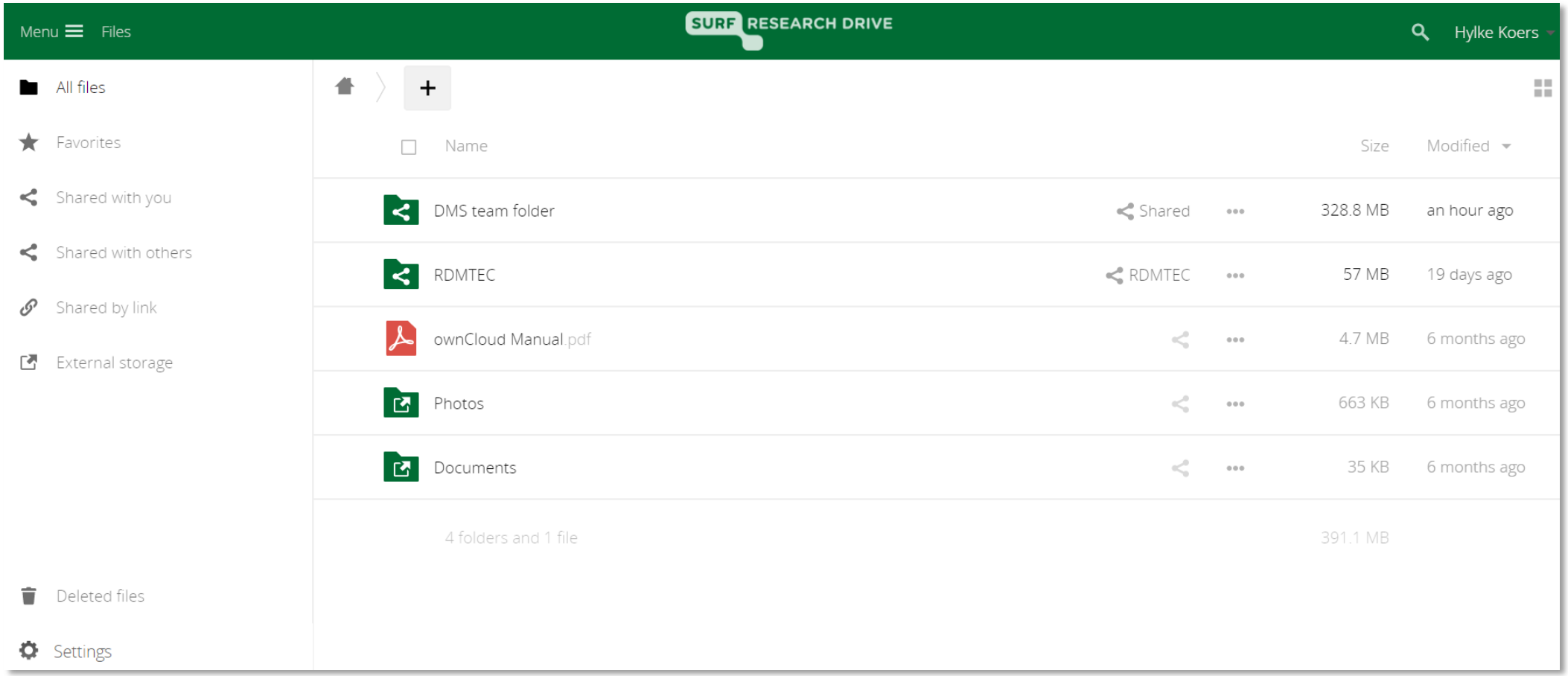
- Sync & share of research data
- One view for all research data
- Built on Owncloud technology: intuitive, easy-to user interface
- Large scale data collection for research teams
  - Limitless Storage
  - Secure
  - Integration with SURF HPC Services
- Supports Data Stewardship
  - Collaborative working with external parties
  - User and quota administration

SURF Research Drive

- Sync & share of research data
- One view for all research data
- Built on Owncloud technology: intuitive, easy-to user interface
- Large scale data collection for research teams
  - Limitless Storage
  - Secure
  - Integration with SURF HPC Services
- Supports Data Stewardship
  - Collaborative working with external parties
  - User and quota administration

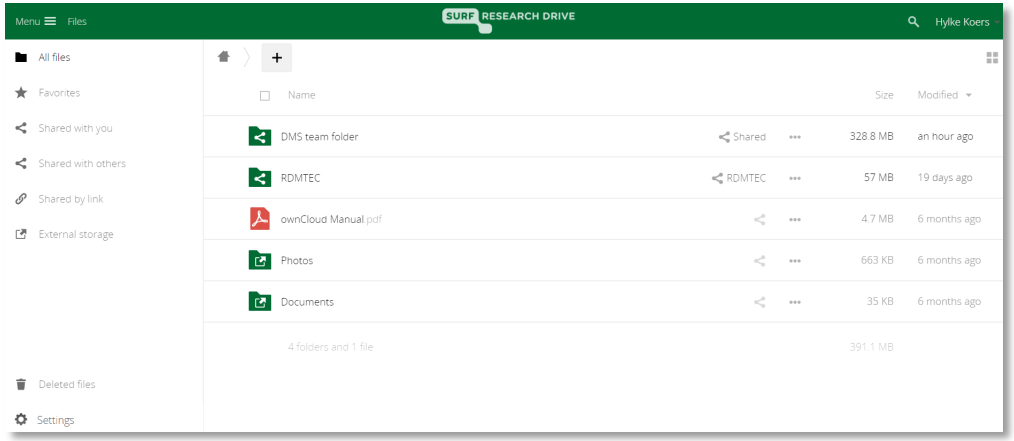


Here is what it looks like



SURF Research Drive

- Well suited to support the earlier phases of the data life-cycle:
  - Sync & share of research data
  - Easy UI
  - Collaboration facilities



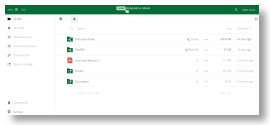
But…

- No metadata
- No integration with core RDM facilities later on in the data life-cycle – notably data archival or publication

SURF Research Drive & iRODS – combining the best of both worlds

So, we set out to extend ResearchDrive by integration with RODS:

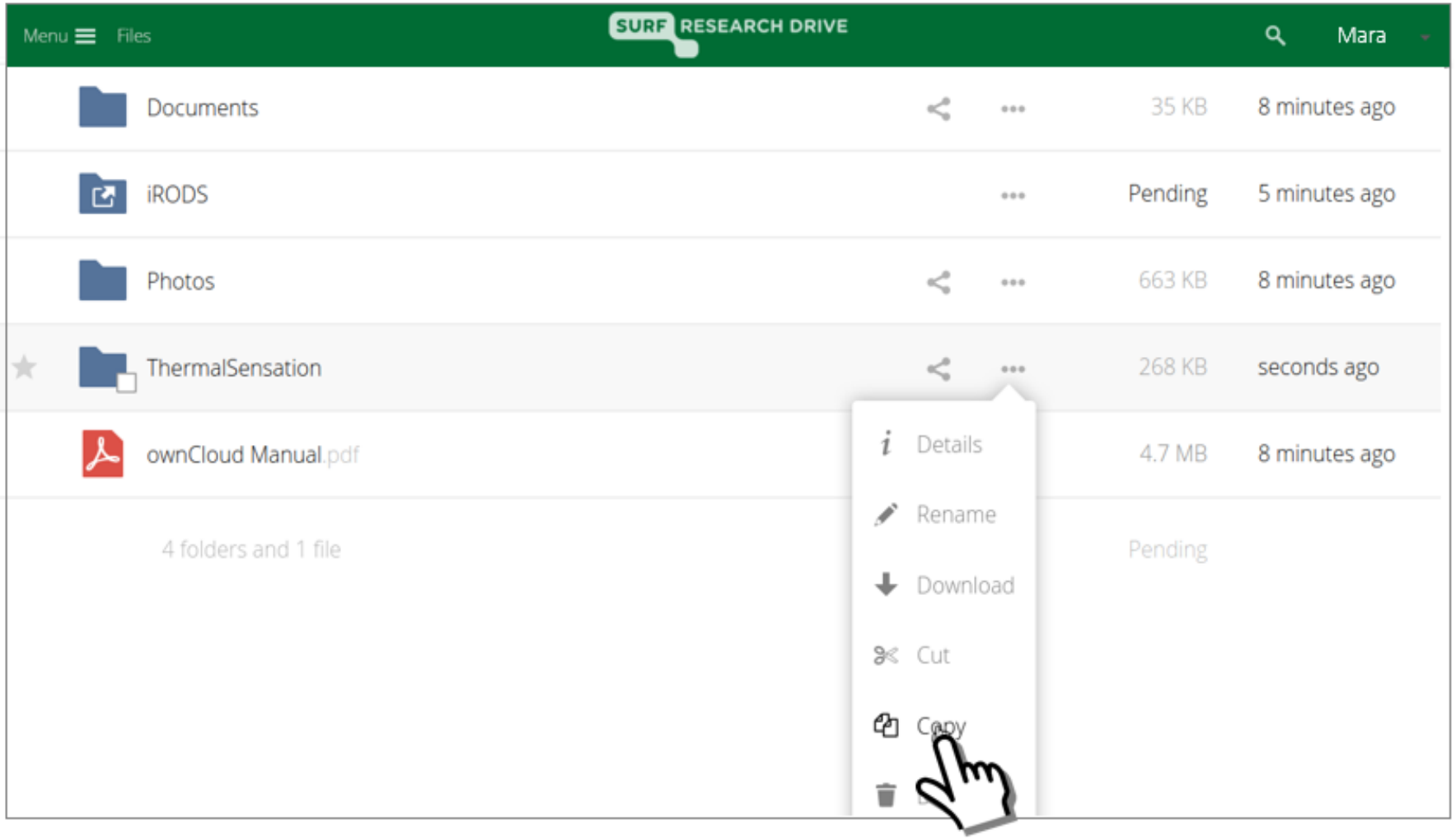
- **User Experience:**
  - User **can add metadata** from within the ResearchDrive environment.
  - Use can ‘archive’ or ‘publish’ from the ResearchDrive environment.
- **Behind the scenes, ResearchDrive is integrated with iRODS**
  - iRODS maintains the ‘source of truth’ metadata records
  - iRODS serves as point of integration to ensure consistent user experiences between Research Drive users (Marc), iRODS command-line users (Stefanie), and institutional data steward (Ayoub)
- ‘Archival’ and ‘Publication’ workflows codified in **Apache Airflow**, working in unison with iRODS rule engine



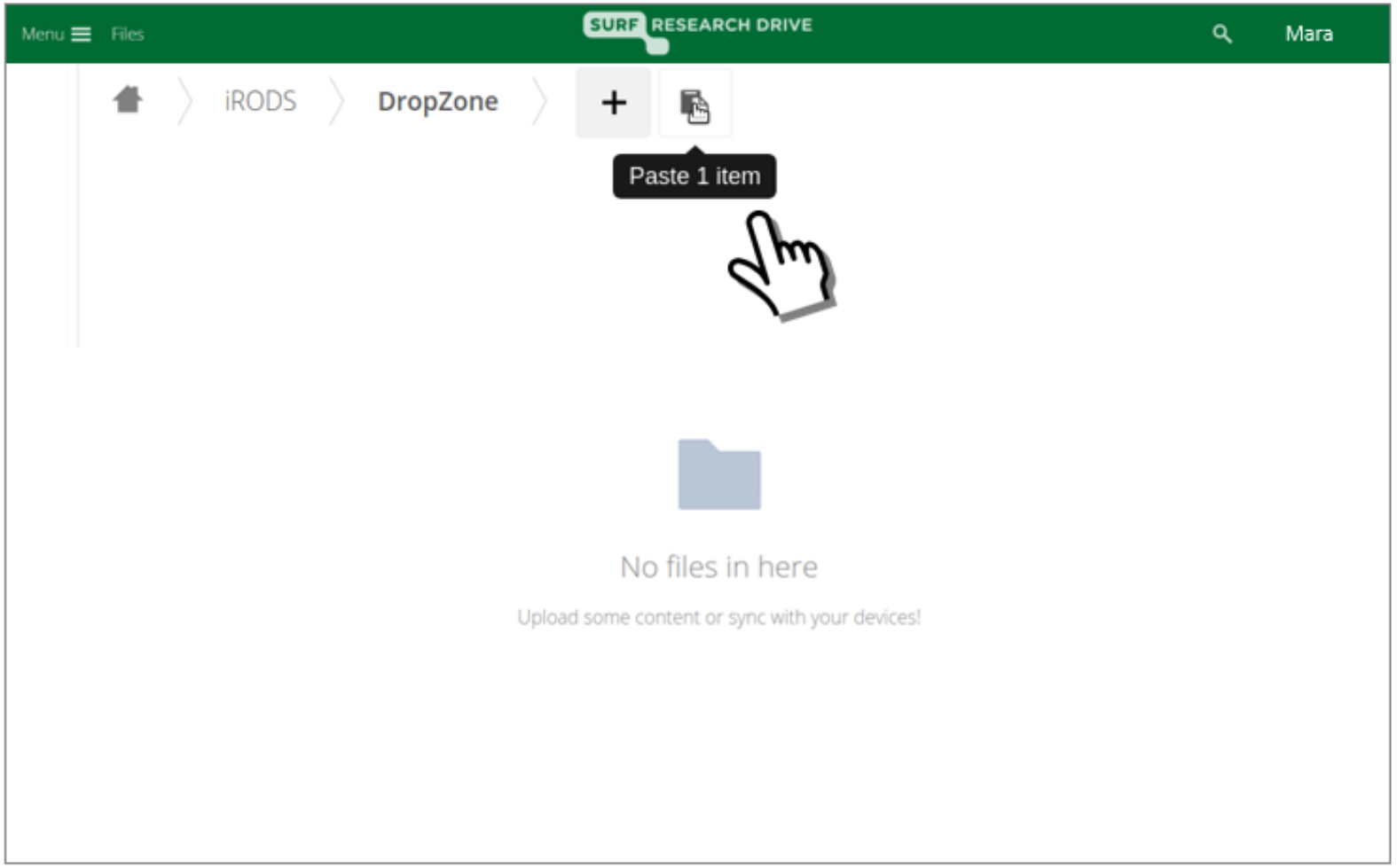
& iRODS = HAPPY (👩 👨 👱)

DEMO TIME

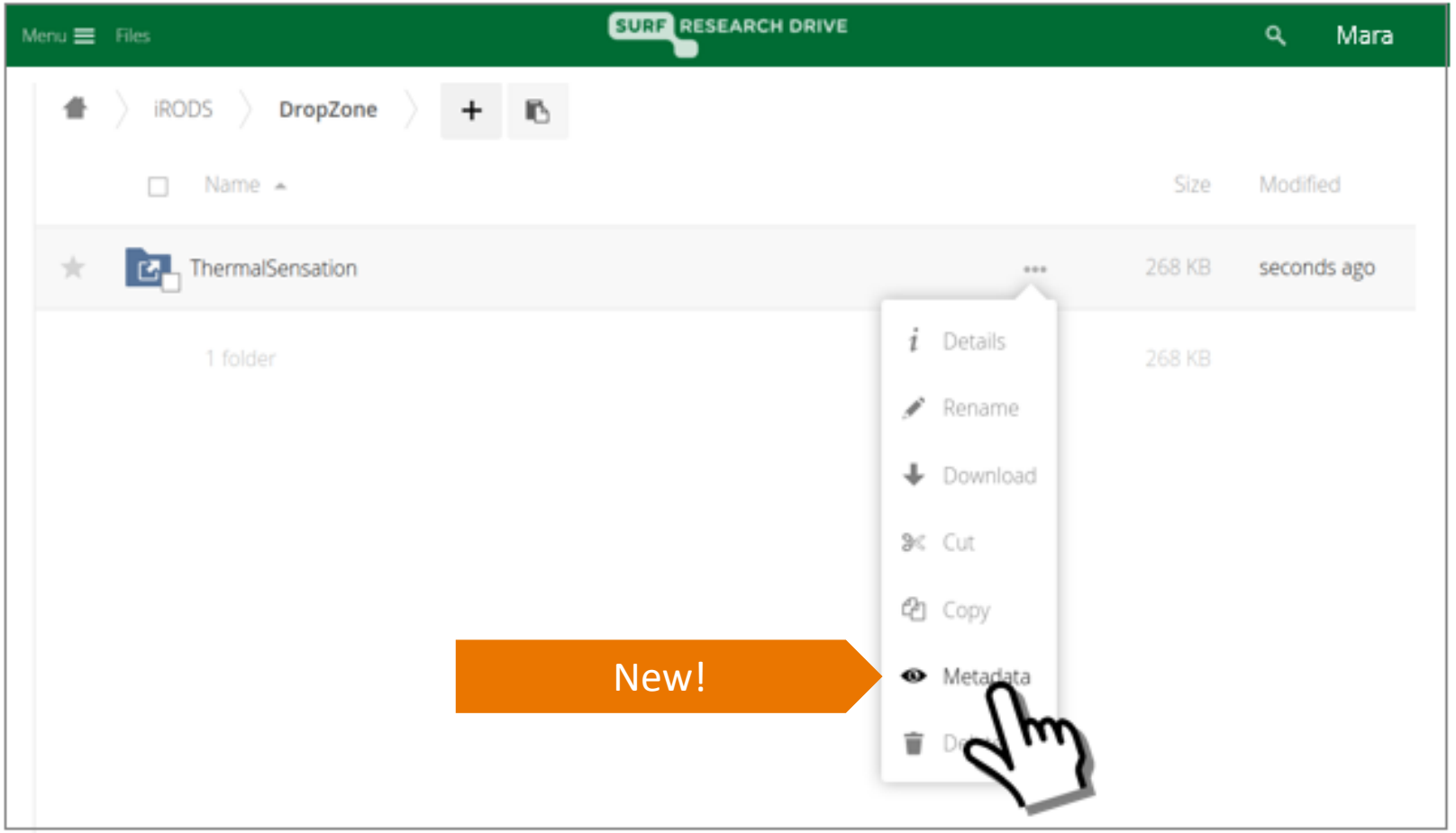
Mara (researcher) copies folder



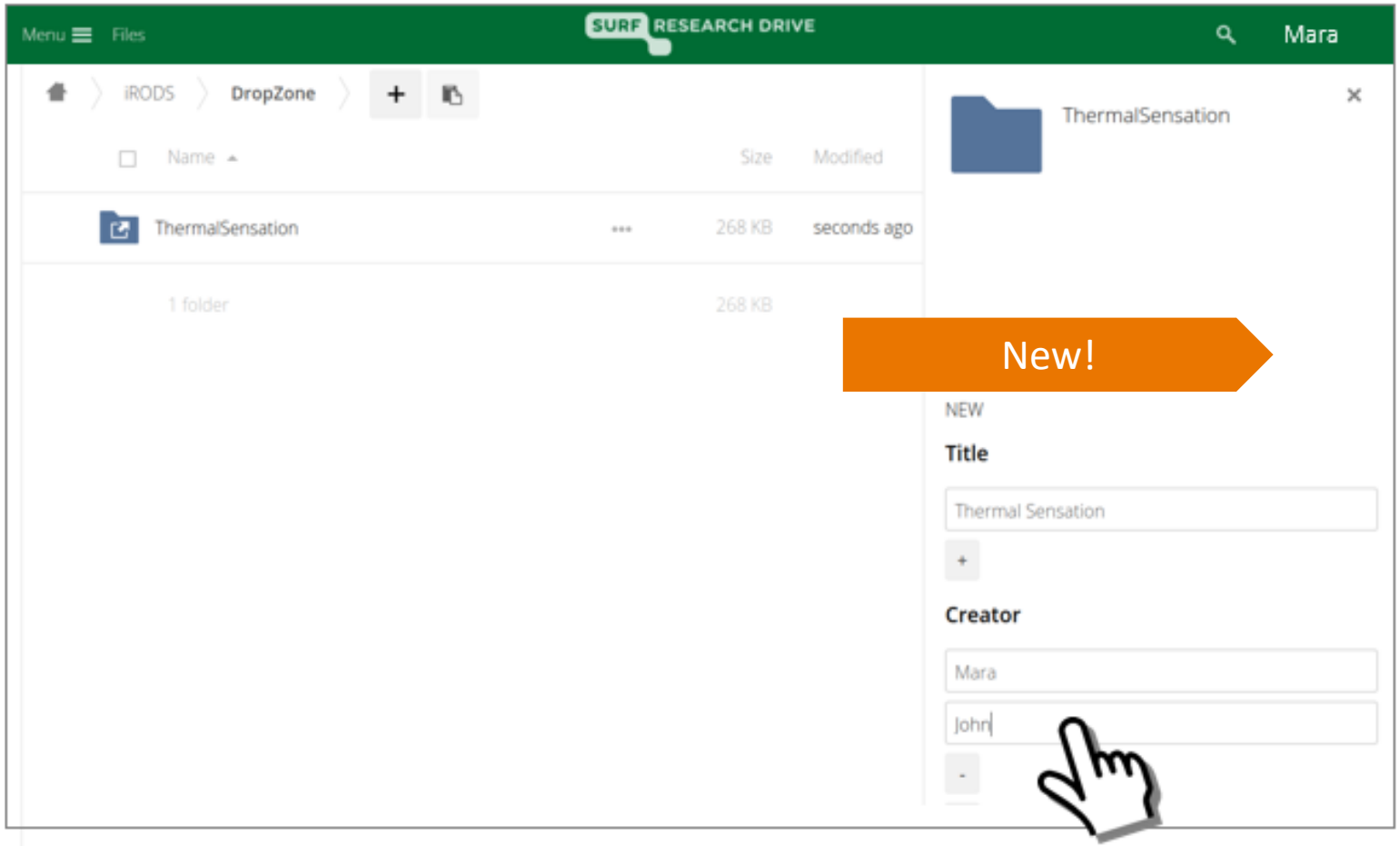
Mara (researcher) pastes folder into Archive dropzone



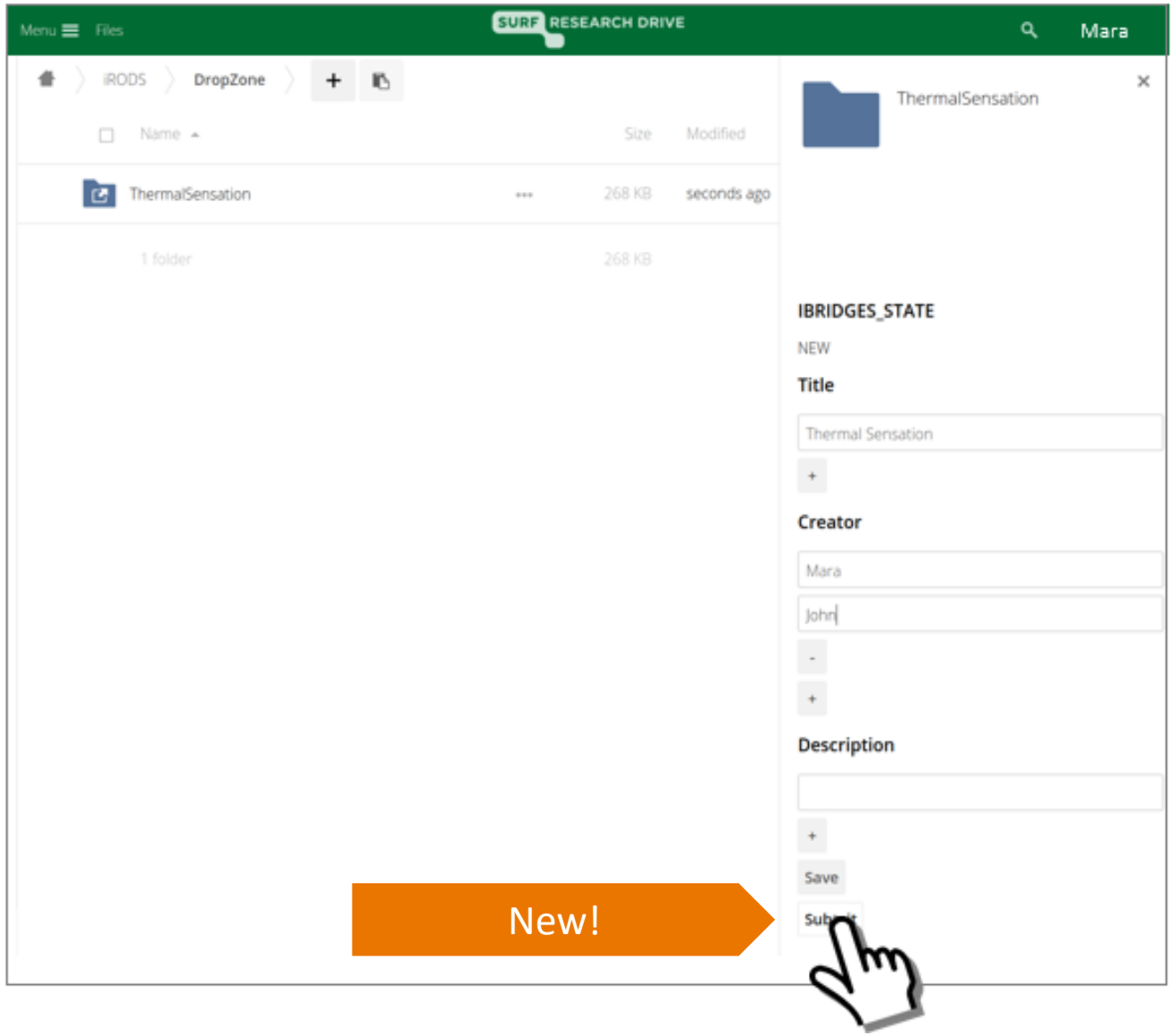
Mara (researcher) selects folder for submission and proceeds to add metadata



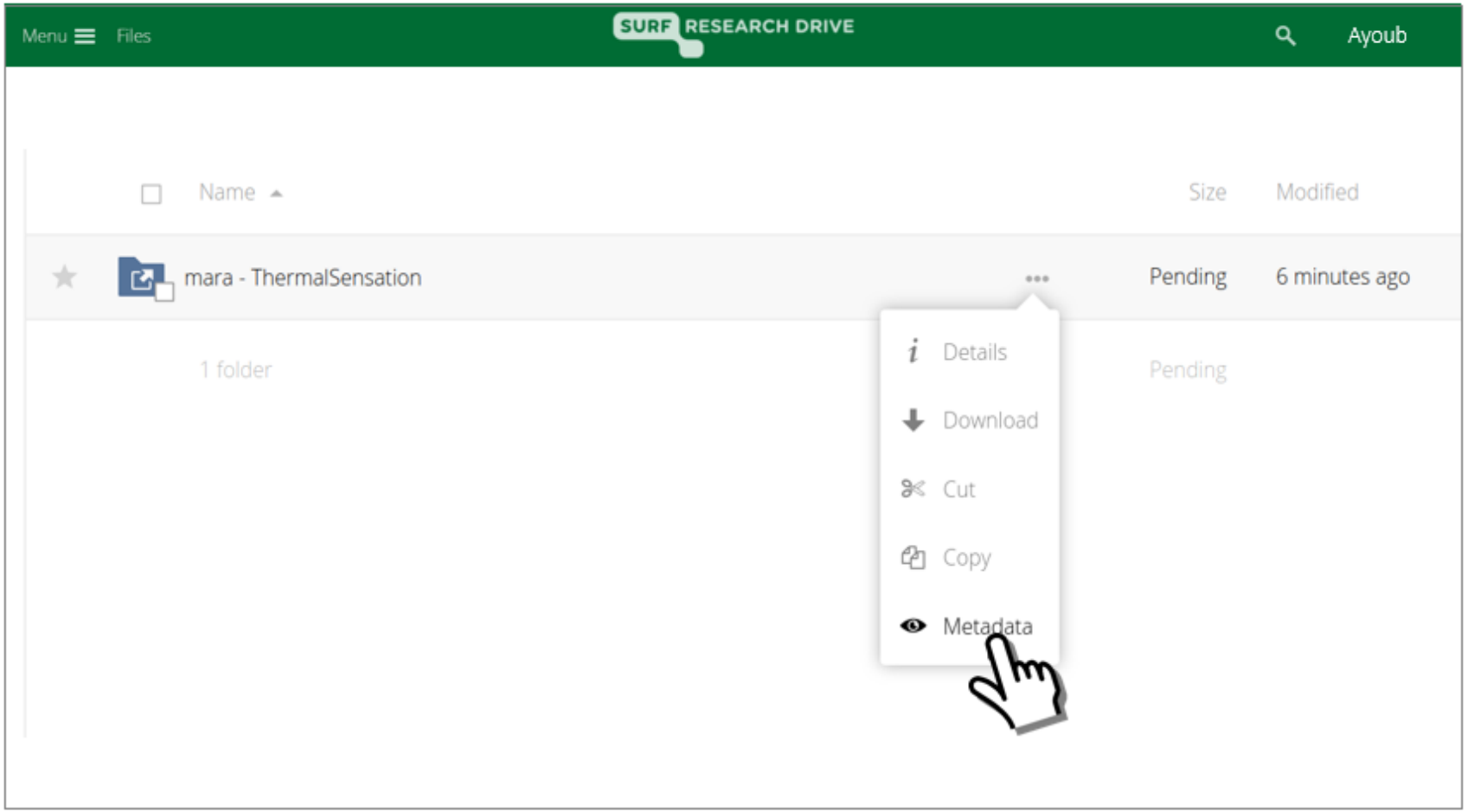
Mara (researcher) adds metadata



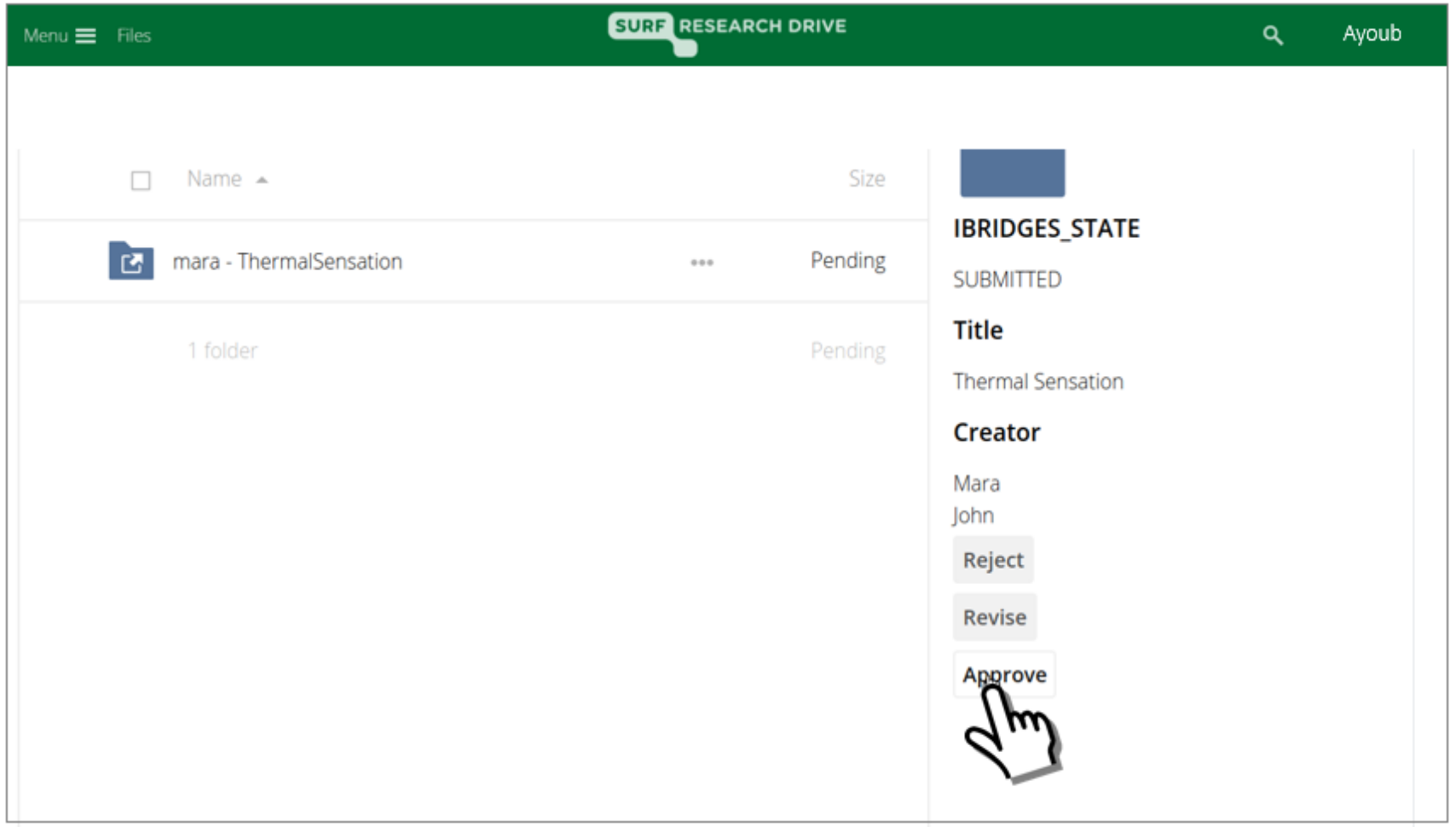
Mara (researcher) submits collection to Archive



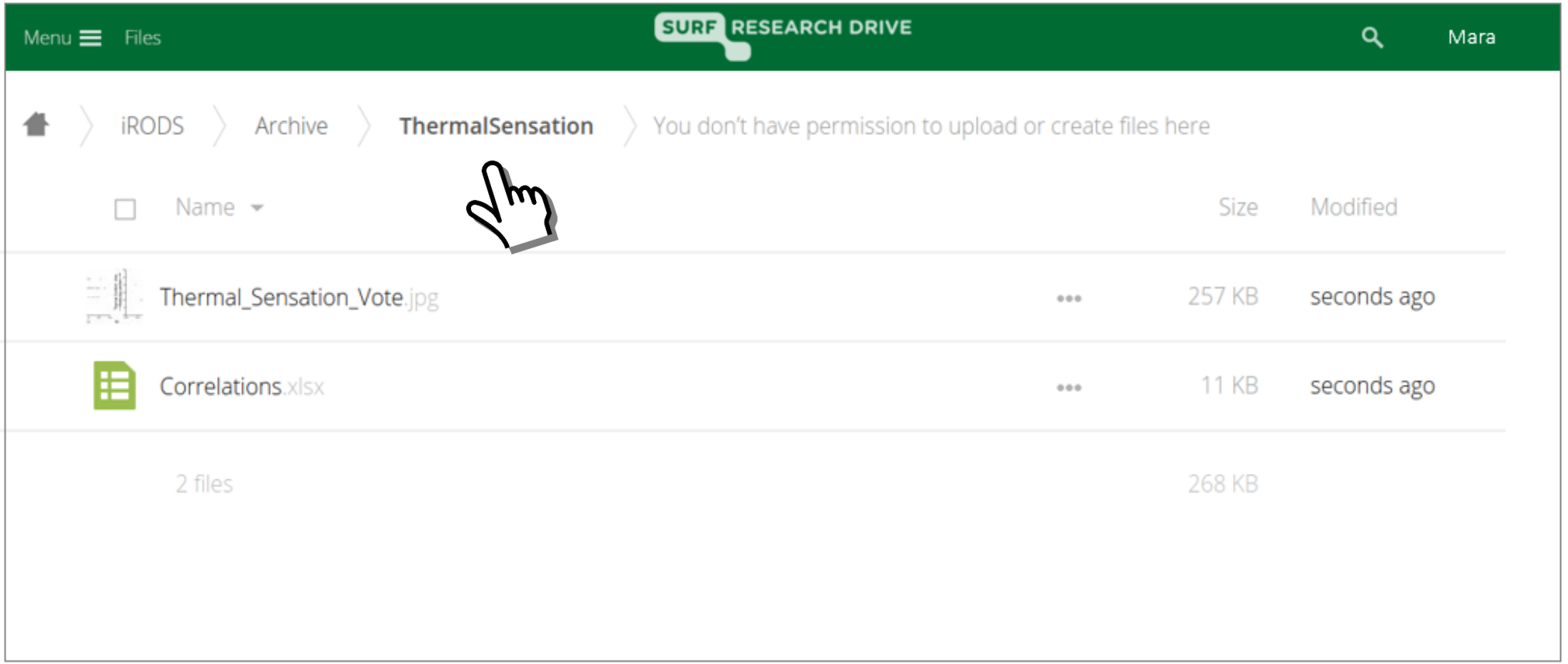
Ayoub (data steward) selects submitted collection



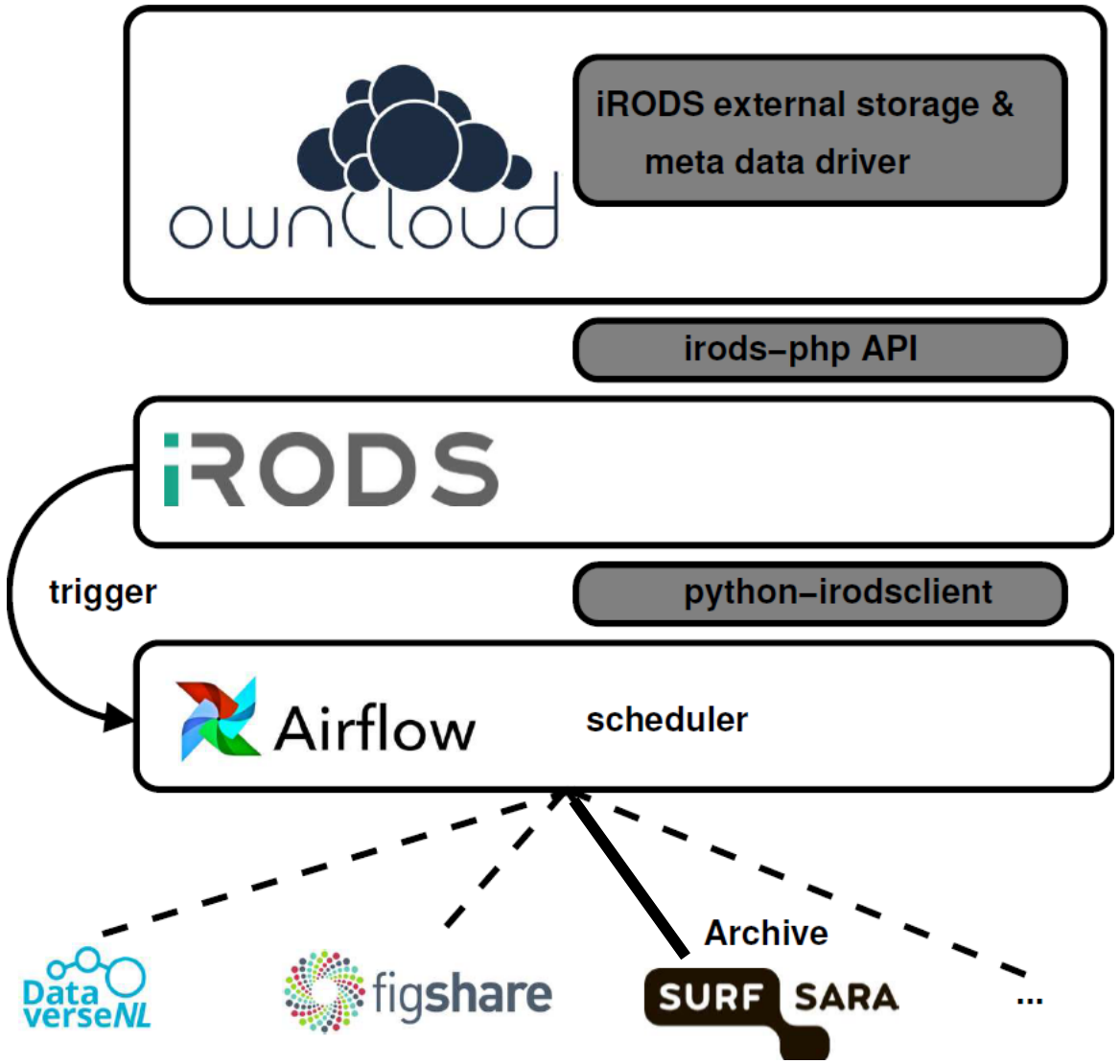
Ayoub (data steward) approves submission



Mara (researcher) checks that her data collection in now in the Archive

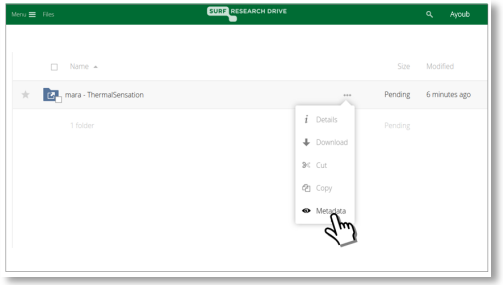


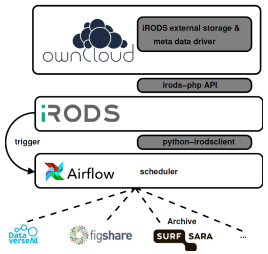
Technology stack



Summary

- We’re exploring an integration between ResearchDrive (Owncloud) and iRODS
- Benefits:
  - Support researchers who want to have an intuitive, easy-to-use GUI yet also have a need for RDM facilities like data archival and publication.
  - iRODS layer ensures consistency across the ecosystem and the different actors (prevent disconnected systems)





Kudo’s to Stefan Wolfsheimer and the rest of the SURF team for developing the PoC and gathering initial user feedback.

Next steps & future work

- User test the iRODS – ResearchDrive integration with current ResearchDrive users
- Firm up PoC code to ‘pilot grade’, looking in particular at scalable and robust user authentication and authorization
- Explore further extension to trigger data publication workflows – integrating with e.g. DataVerse, B2SHARE, 4TU.Datacenter, SURF Data Repository, Figshare, etc.

- *Still exploratory work – your feedback very welcome!*

ANNEX

User authentication & authorization

- Current POC:
  - Authentication and authorization through manually-entered usernames and passwords in Owncloud iRODS app
- Ambition
  - Single Sign-on
  - User identification and authentication through SURFconext and Science Collaboration Zone (existing SURF services)
  - Authorization through tokens from OAuth2 authorization server (via iRODS PAM modules and OwnCloud iRODS app)

Sequence diagram