Universiteit Utrecht

Using iRODS to manage, share and publish research data

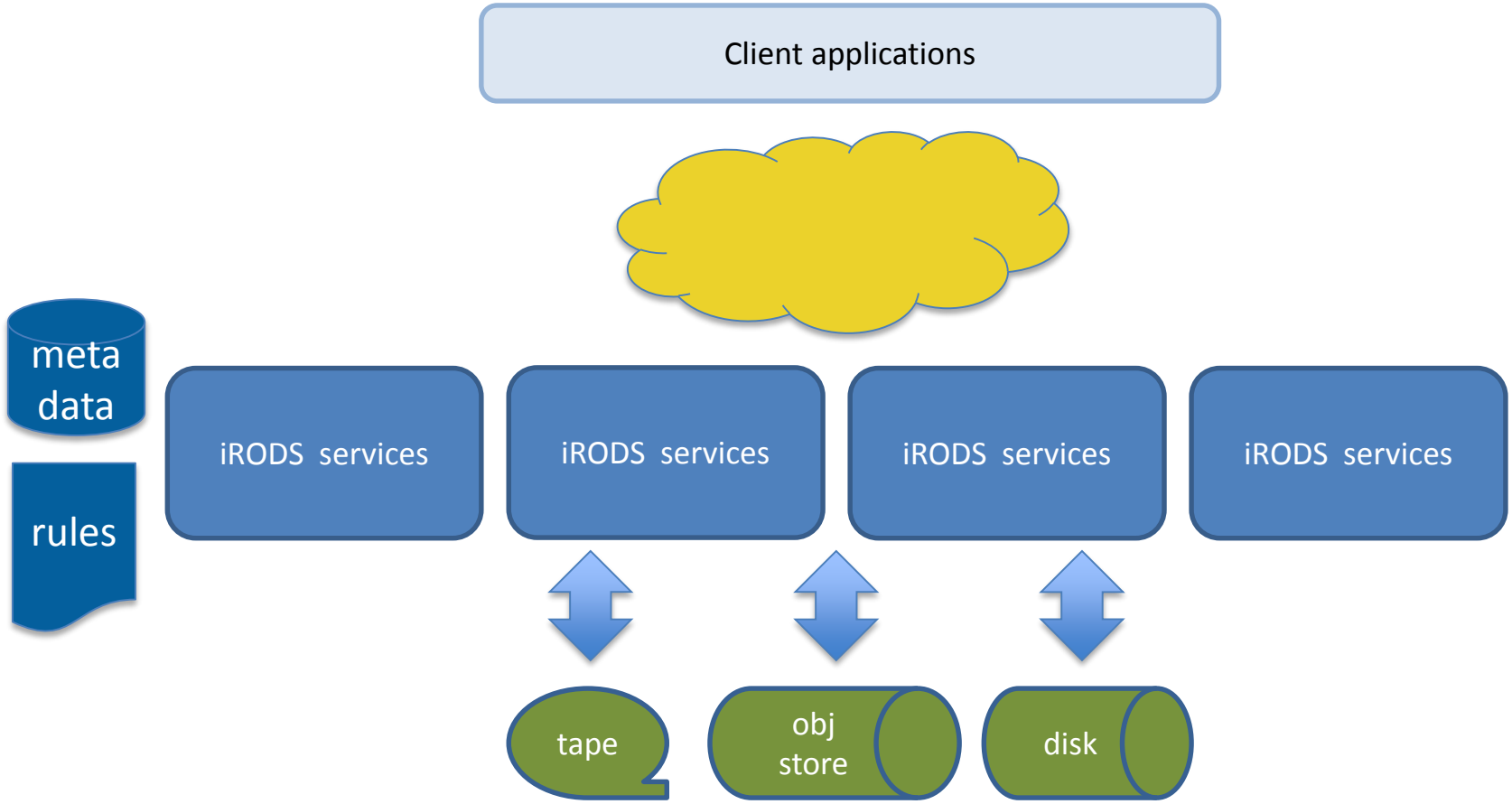
Ton Smeele
ITS/ResearchIT, Utrecht University

What is iRODS

- **data grid** software, to logically order and use large amounts of *data* along with their *metadata* from sources spread across multiple locations, computers and storage infrastructures
- An **integrating infrastructure** situated in the layer between user applications and operating system/storage facilities
- Includes services that allows users to access, manipulate and manage data in a **uniform way**, annotate the data with metadata and share data and metadata with other users.

Support for metadata is crucial when you need to manage your research data

iRODS unifies and manages access to persistent data



iRODS Rule Engine: operationalize policies

Policies:

- Ensure that *primary* research data is kept safe in at least two geographically spread locations managed by the university.
- Ensure data is properly described and accounted for.

Partial rule:

GIVEN (a new data file is being added)
WHEN (file has metadata "primary research data") , (file is of type image)
THEN
(***auto-extract*** the geo-location, date using the file's EXIF properties) **AND**
(**store the file** on our own Utrecht located bulk storage system) **AND**
(**store a replica** of the file on our Almere located storage system) **AND**
(**maintain** geo-location, actor and date as **metadata with the file**)

Why iRODS as Research Data Management platform

- **scalable platform**
  - can manage billions of files, petabytes of data
  - infrastructure/vendor neutral solution

  can be used to manage large/many data collections

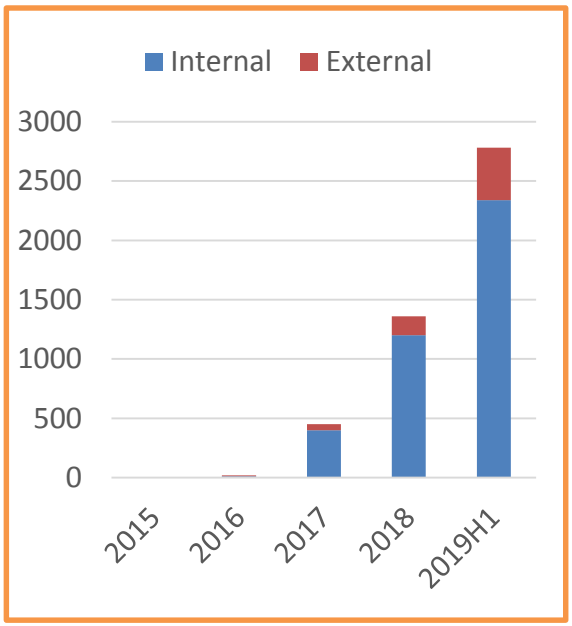
- **enforces data policies**
  - secures sensitive data
  - auditable controls

  supports demonstrable research integrity

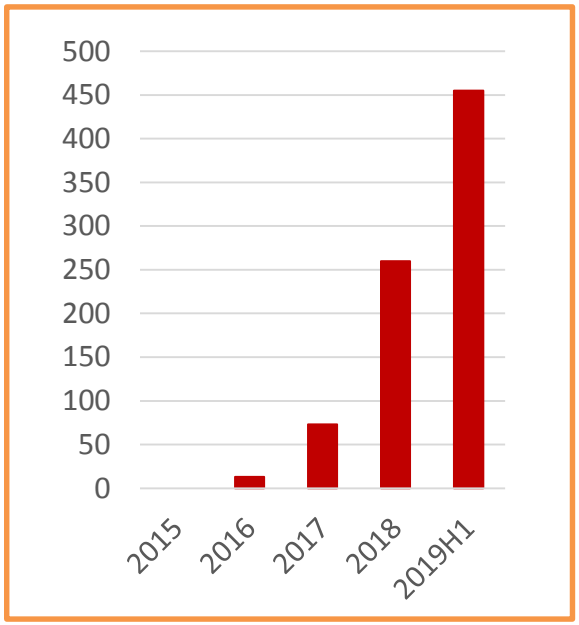
- **manages metadata alongside the data**
  - metadata based data policy execution decisions
  - data workflow automation

  facilitates research data workflows

Utrecht University iRODS managed research data



2800 Users (440 external)
8 production zones



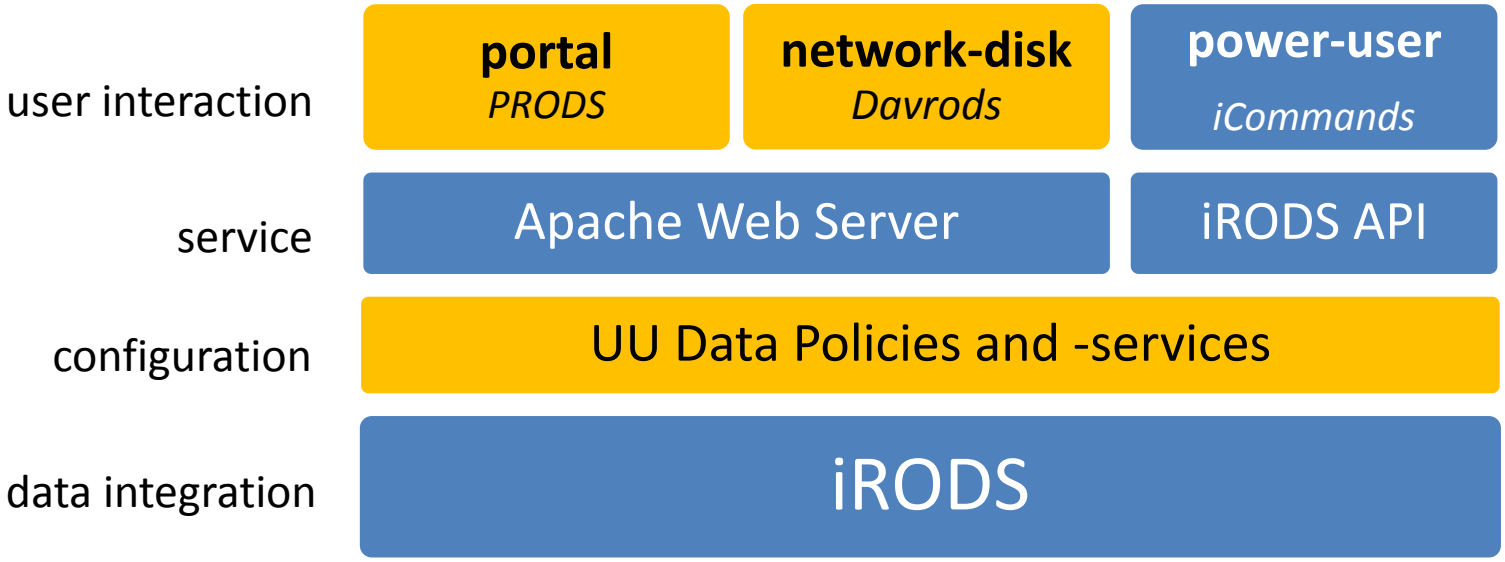
450 TB Data

production instances only, figures are indicative

Our iRODS implementation is called "Yoda":

preconfigured iRODS based system, delivered and supported as a service

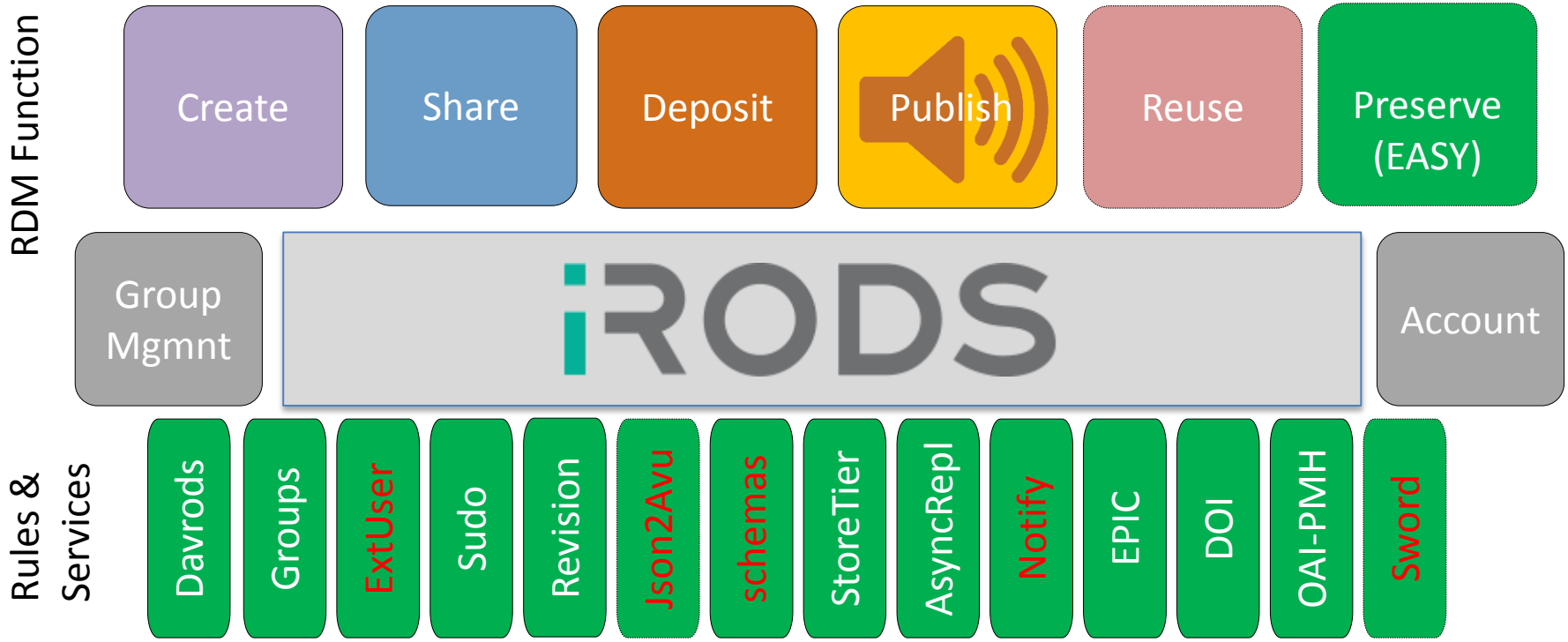
- enhanced with (graphical) user interfaces, policies and rules



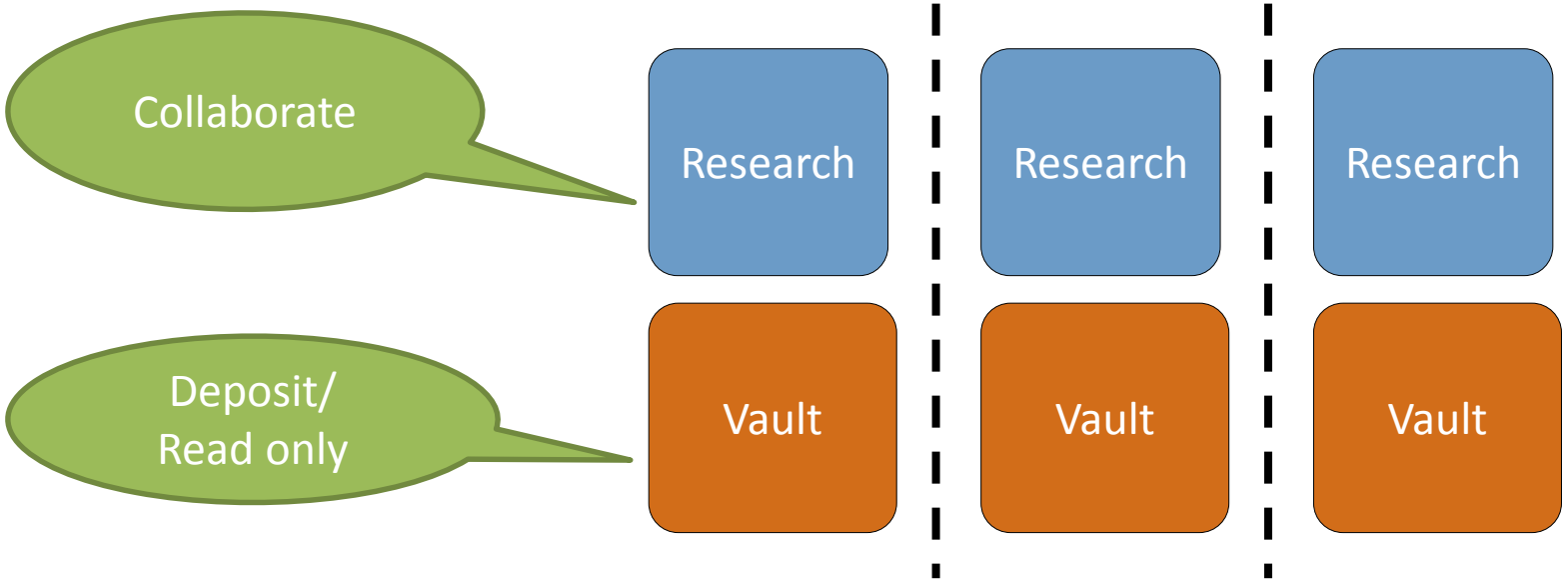
10,000 lines of rules *14 custom microservices*

YODA

iRODS implementation for Research Data Management



Yoda Data compartments

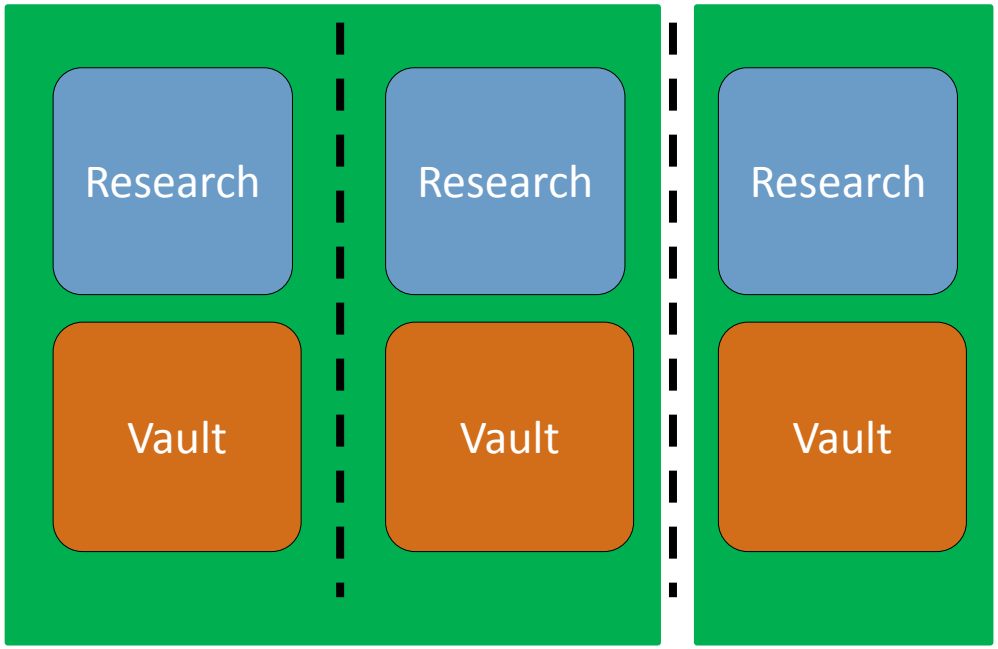


Each data compartment relates to an iRODS group

Yoda Communities ("category")

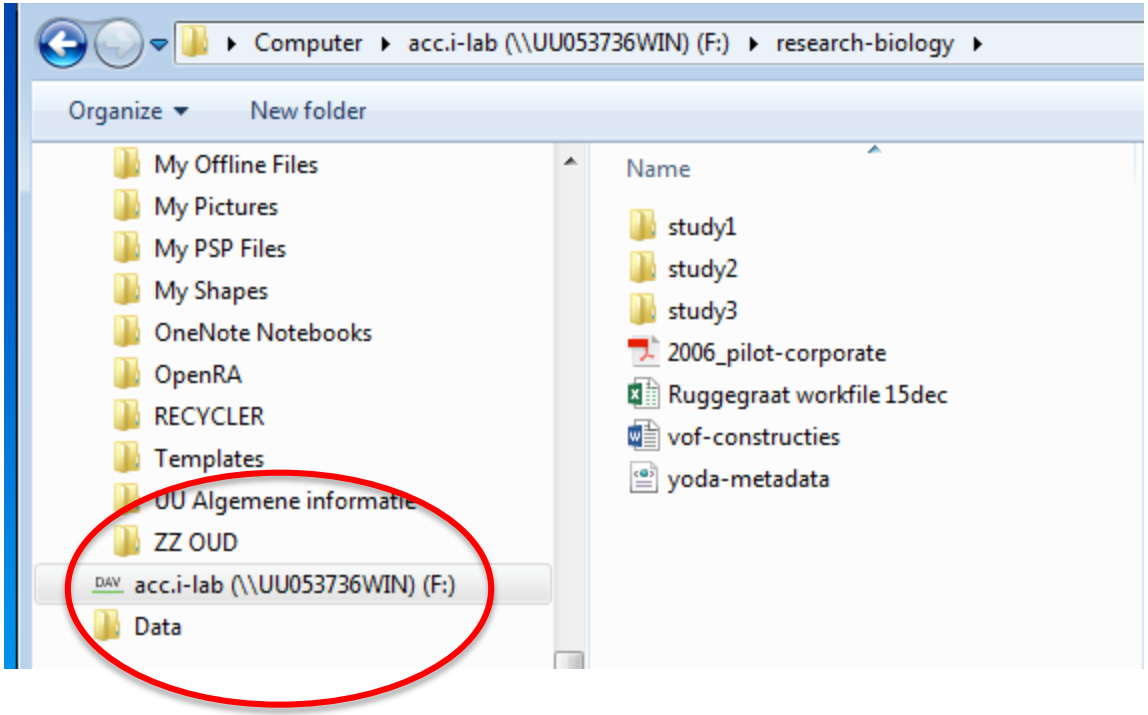
A community comprises of multiple data compartments

Per community:
- cost calculation/invoicing
- appointed datamanager(s)
- metadata schema



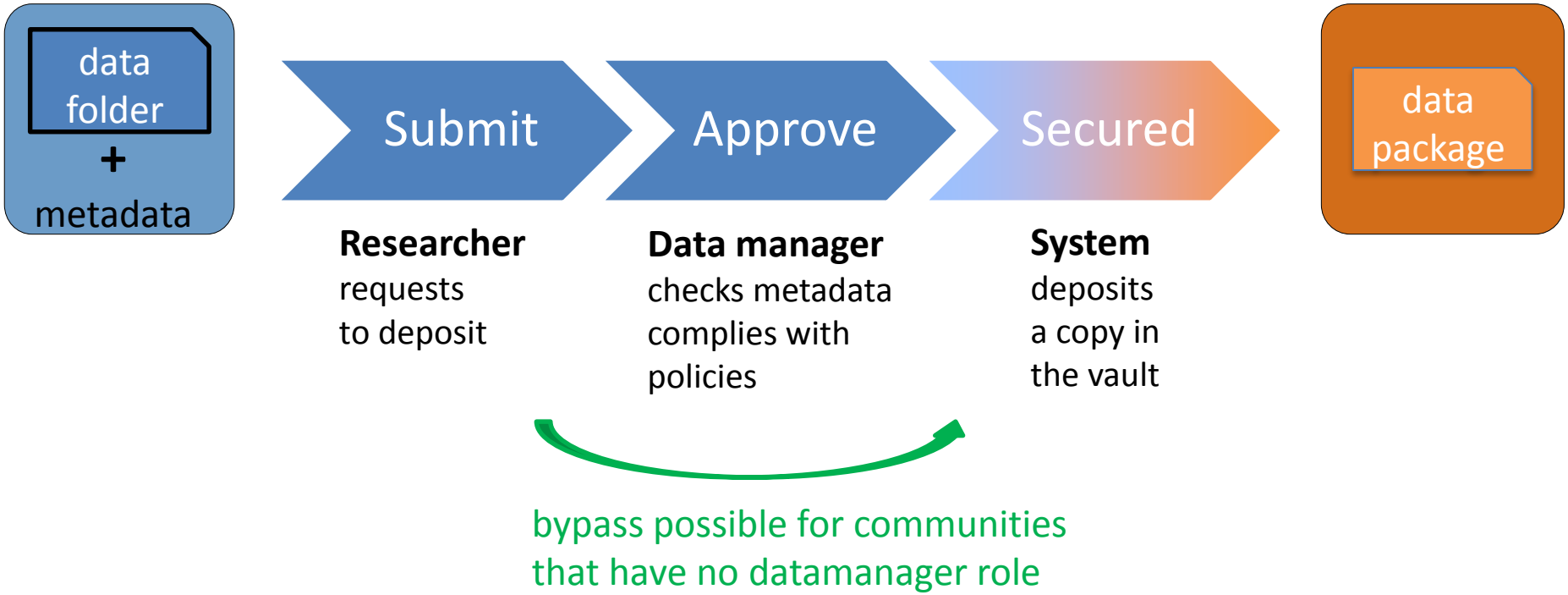
Community concept implemented as metadata on iRODS groups

Collaborate during research via the Yoda disk



WebDAV access from anywhere on any workstation using Davrods

Data Deposit workflow



FAIR Data Publication workflow

Researcher
requests to publish

Data manager
checks metadata complies with publication policies

System
publishes the metadata and provides internet access to data if classified as "Open"

support for sensitive data

'FAIR' Research Data Management using iRODS

Research

Collaborate safely as a group ("Research" folder)

Vault

Maintain integrity, deposit a folder in the vault

Allow FAIR reuse, publish a data package

demonstration

support for sensitive data

Yoda manages data during/after research

Research

Collaborate safely as a group ("Research" folder)
-> membership self-managed by researchers

Vault

Maintain integrity, deposit a folder in the vault
-> metadata can vary per community,
-> datamanager approves deposit

Allow FAIR reuse, publish a data package
-> datamanager approves publication, DOI citable data

Yoda is available under GPL license
at https://github.com/UtrechtUniversity

Thank you

More info:
Ton Smeele a.p.m.smeele@uu.nl