

# SURF super day 2019

*The data frontline in academia*

---

**Marc Galland, [m.galland@uva.nl](mailto:m.galland@uva.nl)**

Support scientist (data analysis and management)

University of Amsterdam (FWNI, SILS, Plant Physiology Department)

December 17, 2019

# Outline

1. **Current Research Data Management in academia**
2. **The “Green Data Vault” project**
  - a. History: Seed Valley
  - b. Vault workflow and architecture.
  - c. Milestone: decision to adopt iRODS & YODA (May 2019)
  - d. Test of YODA at the Green Life Sciences cluster (UvA)
3. **Future plans (also with SURFsara):**
  - a. Building use cases: compiling genomic datasets (iRODS)
  - b. Academic writing 2.0

# Outline

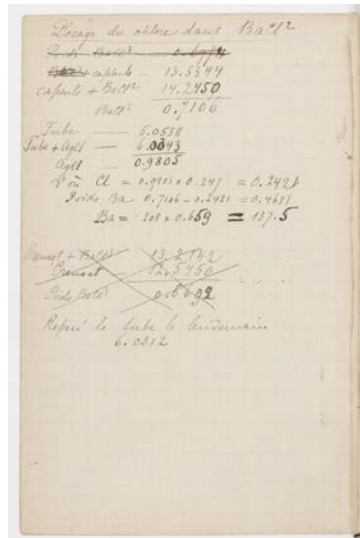
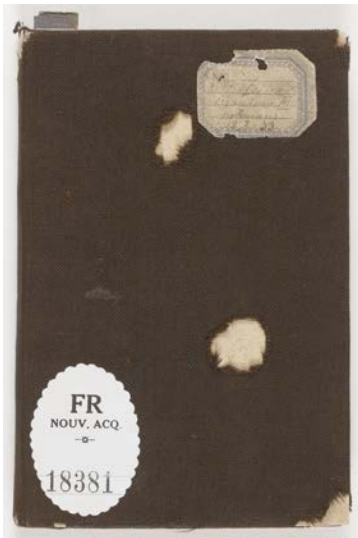
1. **Current Research Data Management in academia**
2. **The “Green Data Vault” project**
  - a. History: Seed Valley
  - b. Vault workflow and architecture.
  - c. Milestone: decision to adopt iRODS & YODA (May 2019)
  - d. Test of YODA at the Green Life Sciences cluster (UvA)
3. **Future plans (also with SURFsara):**
  - a. Building use cases: compiling genomic datasets (iRODS)
  - b. Academic writing 2.0

# Data are recorded in paper laboratory notebooks



Marie Curie Skłodowska  
(in 1921)

Laboratory notebooks

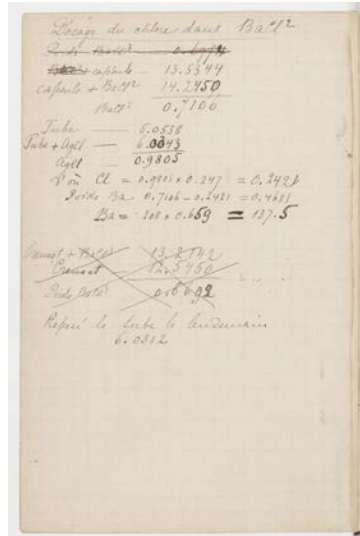
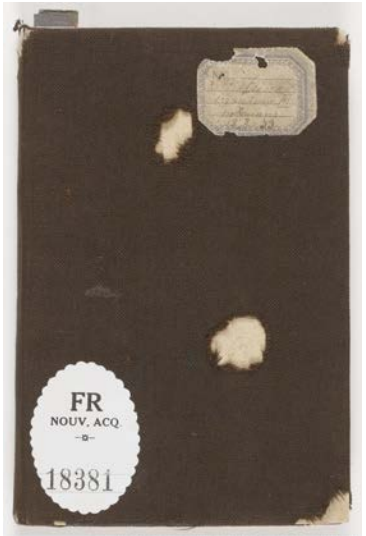


# Data are recorded in paper laboratory notebooks

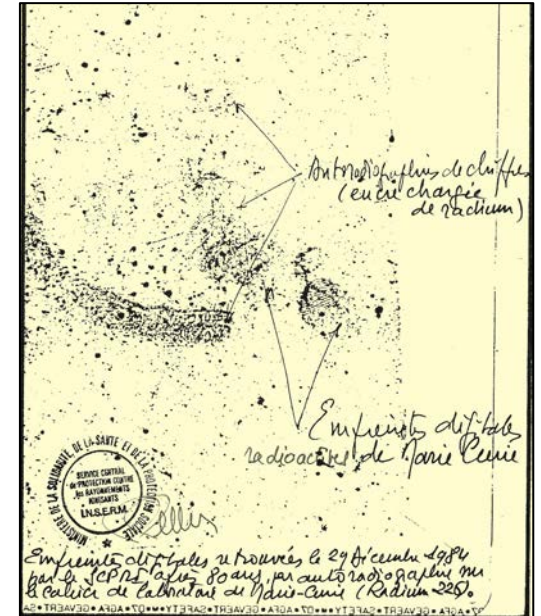


Marie Curie Skłodowska (1921)

Laboratory notebooks



Radioactive lab notebook



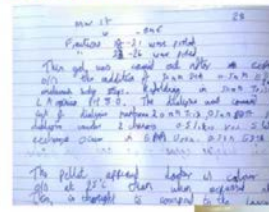
# Laboratory notebooks have no search button

Experimental notes today...

Notebook no.	Date 13 January 2017	
Title <b>The reality of today</b>	Continue	



Notebook no.	Date 13 January 2017
Title <b>Work for Sherlock Holmes</b>	



Presentation by Dr Marko Hyvonen

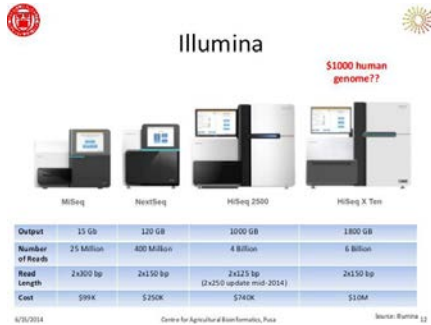
<https://doi.org/10.17863/CAM.7217>

# Data are (often not) backed up.



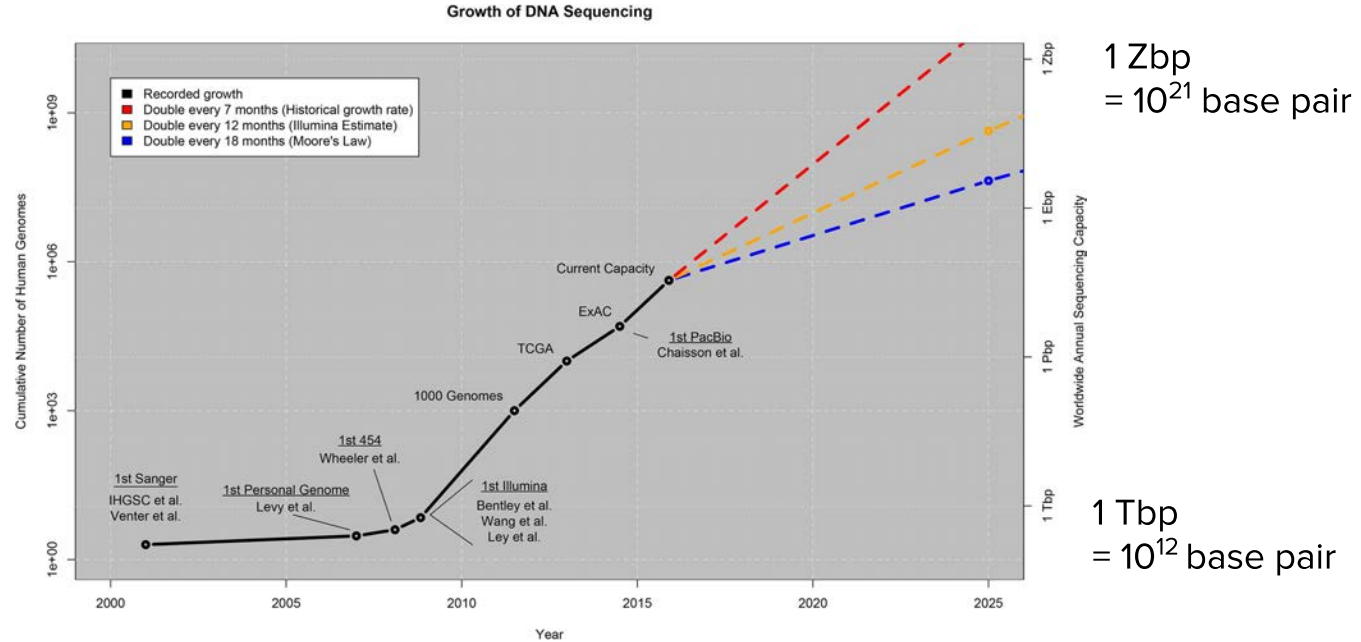
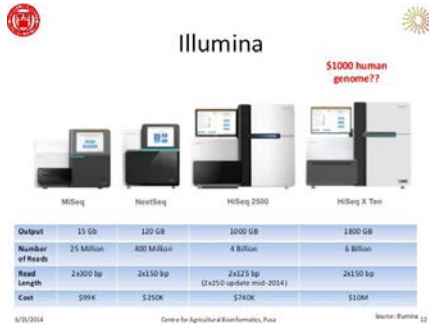
Reward for the whole PhD!  
Unspecified amount

# A genomic data deluge in the Life Sciences





# A genomic data deluge in the Life Sciences



Stephens et al. Big Data: Astronomical or Genomical?. *PLoS Biol.* 2015;13(7):e1002195.

# Academics sent joyfully to the data battlefield



PhD students, postdocs, etc.

# Academics sent joyfully to the data battlefield



PhD students, postdocs, etc.



“Good luck with all these data”

# Academics sent joyfully to the data battlefield



They get harmed:

- data loss.
- data corruption.
- data provenance is lost.

# Academics sent joyfully to the data battlefield



They get harmed:

- data loss.
- data corruption.
- data provenance is lost.

Good practices in RDM are not very rewarded by the academic system...

# Academics sent joyfully to the data battlefield



They get harmed:

- data loss.
- data corruption.
- data provenance is lost.

Good practices in RDM are not very rewarded by the academic system...

...but it's changing!

- Citation of archived datasets (Zenodo, etc.)
- Data papers (*Scientific data*)
- Data steward positions

# Outline

1. **Current Research Data Management in academia**
2. **The “Green Data Vault” project**
  - a. History: Seed Valley
  - b. Vault workflow and architecture.
  - c. Milestone: decision to adopt iRODS & YODA (May 2019)
  - d. Test of YODA at the Green Life Sciences cluster (UvA)
3. **Future plans (also with SURFsara):**
  - a. Building use cases: compiling genomic datasets (iRODS)
  - b. Academic writing 2.0

# “Green Data Vault” project

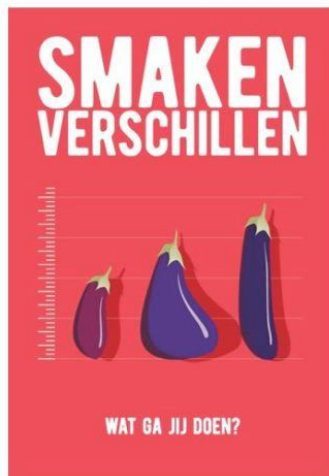
- **Seed Valley:** a consortium of plant breeders in North Holland.
- **Goal:** Share datasets to fasten and strengthen findings.



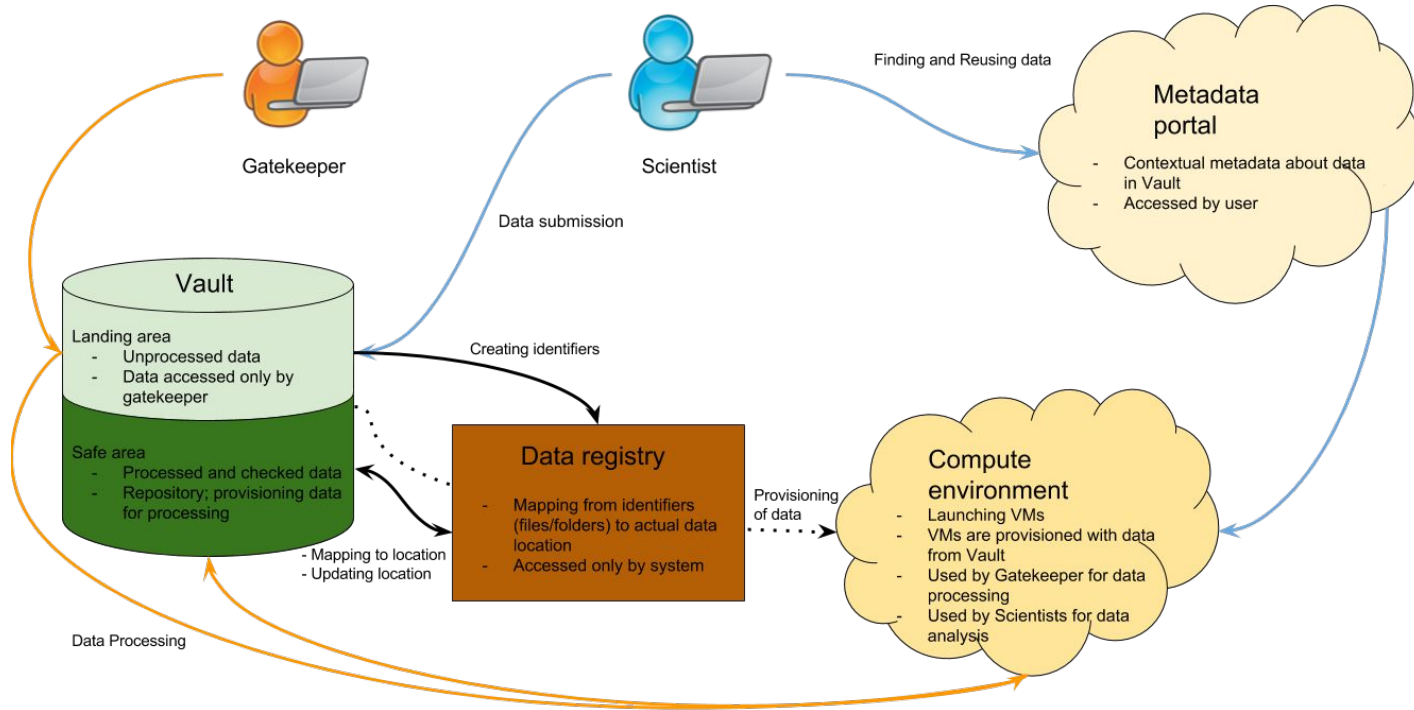


# “Green Data Vault” project

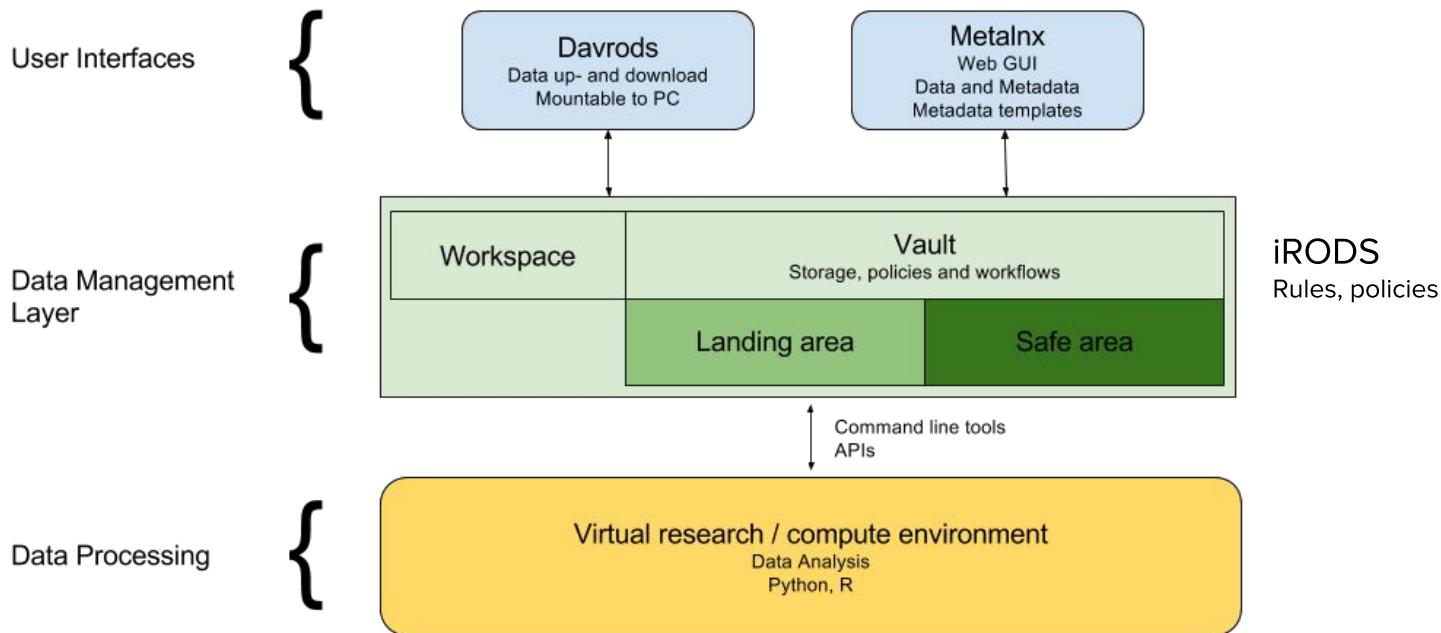
- **Seed Valley:** a consortium of plant breeders in North Holland.
- **Goal:** Share datasets to fasten and strengthen findings.
- Eventually, facilitate the obtention of new vegetable cultivars.



# Definition of the “Green Data Vault” workflow

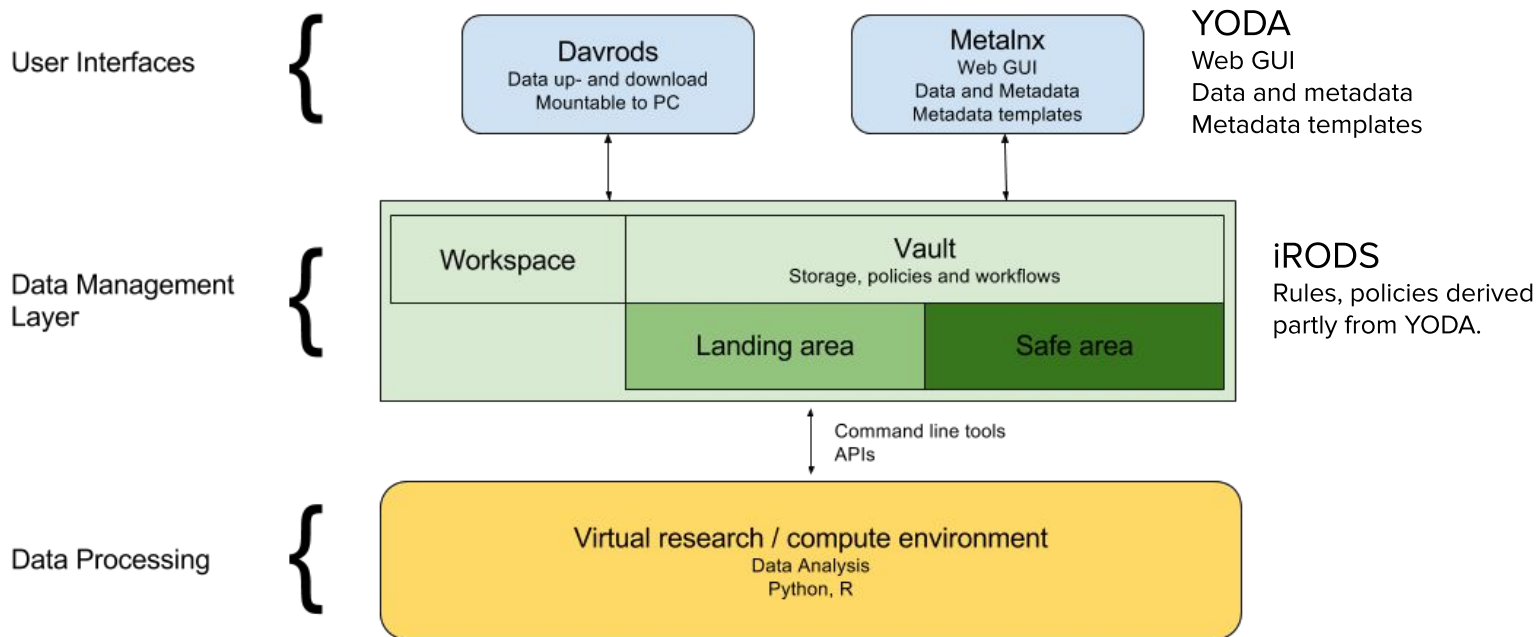


# Definition of the “Green Data Vault” system architecture



Scheme by Christine Staiger (2017)

# Definition of the “Green Data Vault” system architecture



# User tests (April 2019)

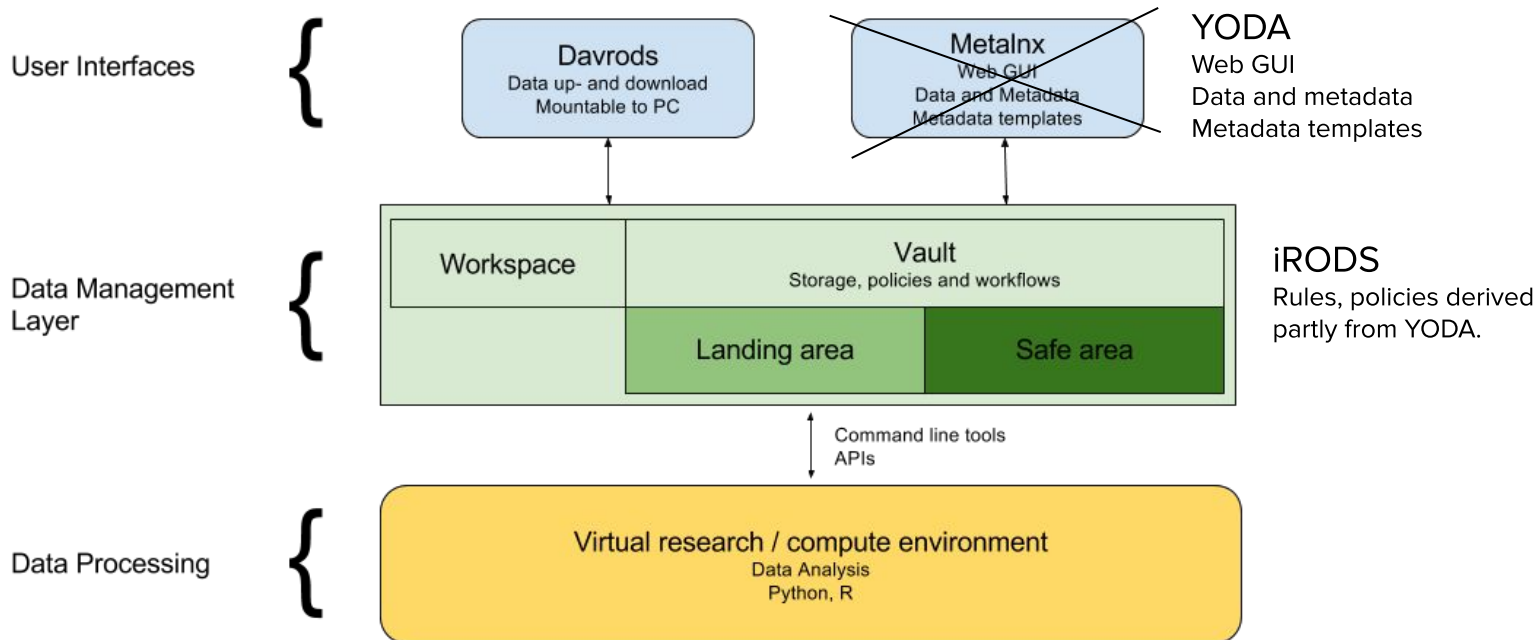
- Maria (Mol. Plant Phytopathology)
- Ruy (Plant Physiol.)
- Benjamin (Plant Hormone Biol.)
- Mehran (Plant Hormone Biol.)
- Bora (Plant Hormone Biol.)
- Saskia (Mol. Cytology)

Support from Utrecht University

Comparing two solutions: Metalnx (DELL) and YODA (Utrecht University).



# Definition of the “Green Data Vault” system architecture



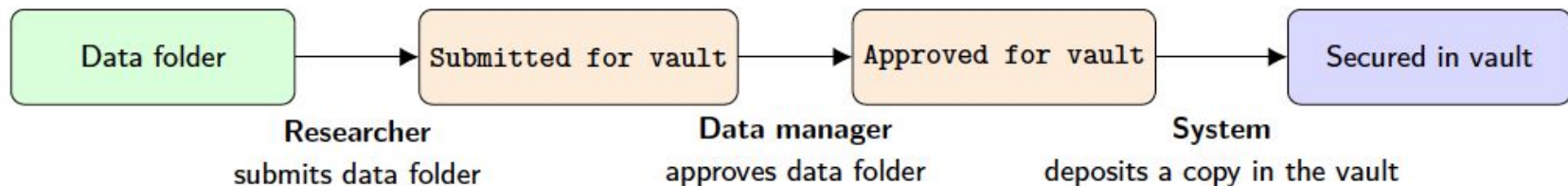
# YODA: Your Data

- Research Data Management system
- Aimed at researchers
- Facilitate their data management, sharing and publication.



# Two main workflows supported by YODA

1. Data deposition in the vault: from the research workspace to the “vault”.

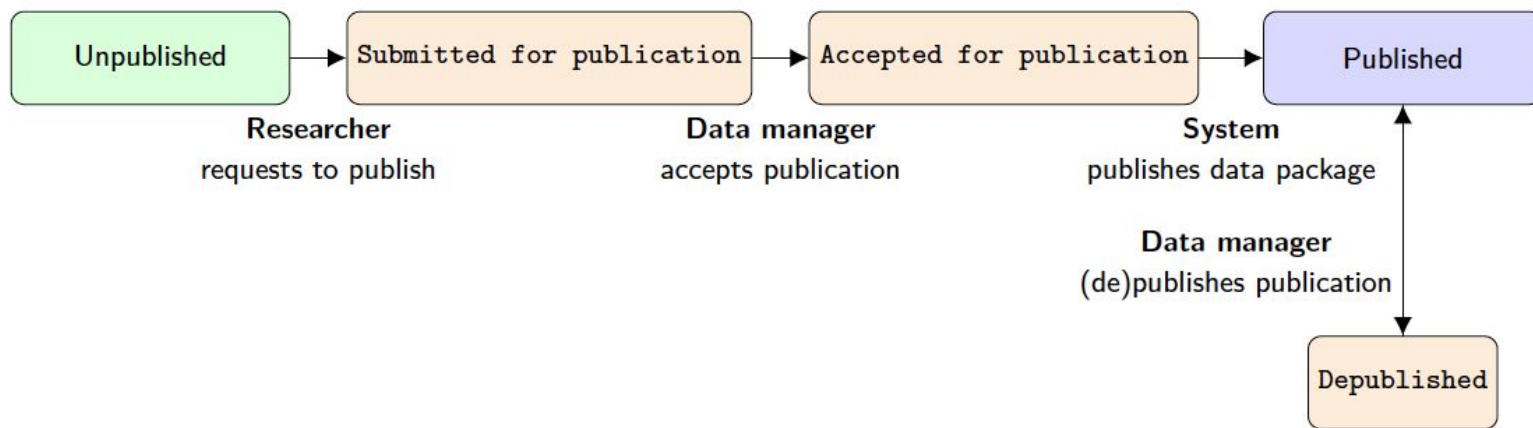


**Figure 2. The workflow to deposit a data package into the vault**



# Two main workflows supported by YODA

## 2. Data publication (outside world)



**Figure 3. The workflow to publish a data package**

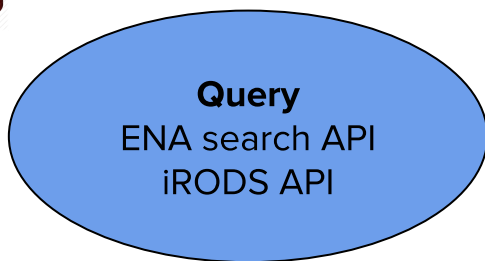
# Outline

1. **Current Research Data Management in academia**
2. **The “Green Data Vault” project**
  - a. History: Seed Valley
  - b. Vault workflow and architecture.
  - c. Milestone: decision to adopt iRODS & YODA (May 2019)
  - d. Test of YODA at the Green Life Sciences cluster (UvA)
3. **Future plans (also with SURFsara):**
  - a. Building use cases: compiling genomic datasets (iRODS)
  - b. Academic writing 2.0

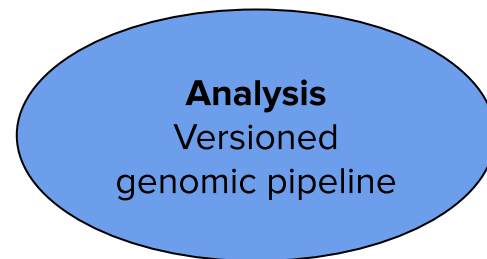
# iRODS powered genomic data analyses



Public repositories e.g. ENA and/or iRODS “in-house” vault.



sample	file	condition
s001	../s001.fastq	control
s002	../s001.fastq	treated



Traceable and  
reproducible results

# New ways to perform research and publish



## Tomato Transcriptome to Trichome (TTT) Project: Outline

 **Jolanta Szkodon** (University of Amsterdam (UvA))

 **Marc Galland** (University of Amsterdam (UvA))

 Add Collaborator  Manage

### Abstract

#### Introduction

- Trichome introduction: What are they? What do they do? What are the different types? How are they formed? Who has trichomes?
- Trichomes in tomatoes: Do different accessions have different types? Do they have different roles in different accessions? How about within one accession? What does each type do? What rules are they governed by? What do we still not know?
- Relationship between mRNA, miRNA, and trichomes
- Random Forest: Has it been used to answer this type of question? When/how? How can I use to answer my question?
- Research aim: Understand connection between trichome type and density, and mRNA/miRNA data using RF. Why would we want to know this? What will this information tell us/help us with?

Alyssa Goodman, Josh Peek, Alberto Accomazzi, et al. The "Paper" of the Future. *Authorea*. February 21, 2017. DOI: <https://doi.org/10.22541/au.148769949.92783646>

# Link to the data and code underlying a figure

Link datasets

### Figure Options

Figure size

Width (pixels)

Height (pixels)

Export Options

Export Width (points)

Multicolumn

Data Options

[Attach data \(.csv, .dat, .xls\)](#)

[Attach code / notebook \(.ipynb\)](#)

Link to Data:

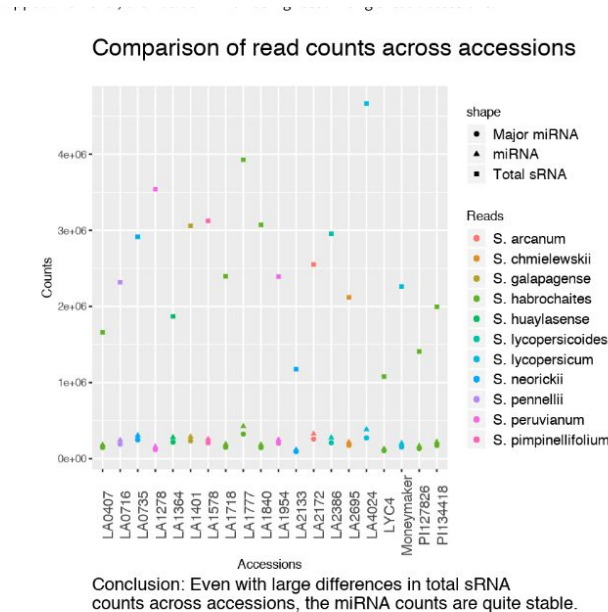


Fig. 1

Alyssa Goodman, Josh Peek, Alberto Accomazzi, et al. The "Paper" of the Future. *Authorea*. February 21, 2017. DOI: <https://doi.org/10.22541/au.148769949.92783646>

# Thank you for your attention.

## Acknowledgments

### SURFsara Data Services team:

- **Hylke Koers (2019 - )**
- **Arthur Newton (2019 - )**
- (2017-2018) Ander Astudillo
- (2017-2018) Christine Staiger

### University of Amsterdam

- Michel Haring
- Maarten Noom (HvA)
- Petra Bleeker
- Joyce Nijkamp (HvA).

### Data Steward Interest group (DTL):

- Jasmin K. Boehmer (Utrecht UBC)
- Mateusz Kuzak (eScience centre)
- Christine Staiger (DTL Liaison)

### TU Delft Data Stewards

- Martha Teperek

### Vrije Universiteit Amsterdam

- Brett Olivier (AIMMS)
- Maria Cruz (Library)

Extra slides

---

# Why current solutions are not sufficient (Figshare)

Figshare: open access repository provided by Digital Science (UK company)



## Results of machine learning experiments for "Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data"

Version 2  Dataset posted on 30.10.2019, 16:59 by Pawel Widera

The archive file includes results of machine learning experiments performed for the article "Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data". The hypothesis of the article is that prediction models trained on historical data will be more effective at identifying fast progressing knee OA patients than conventional inclusion criteria.

For all experiments the first level folder hierarchy indicates the *method* used. Where parameter tuning is performed, the second level folders indicate *algorithm parameters*. Each experiment output is stored in a xz compressed text file in **JSON** format.

In experiments measuring the **learning curves** (*training-\**), each results file describes:

- \* experiment setup (algorithm, number of subsets, down-sampled class size)
- \* list of training set sizes
- \* performance measure statistics for all subsets at each training size (flat list) including min, median and max score, and median deviation from median (mad), given for both test and training set instances

In **parameter tuning** experiments (*prediction-multi-\**), each results file contains:

- \* experiment setup (method / algorithm, number of CV repeats, number of model runs)
- \* imputer parameters (*not important* - kept constant in all experiments)

24 views | 5 downloads | 0 citations



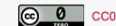
### CATEGORIES

- Applied Computer Science
- Artificial Intelligence and Image Processing
- Health Informatics

### KEYWORD(S)

- experiment results
- machine learning
- multi-class classification
- imbalanced data
- clinical decision-making
- patient selection
- knee osteoarthritis

### LICENCE





# Why current solutions are not sufficient (DataVerse)

Harvard University: IGSS institute



Open source research data repository software



Researchers

Enjoy full control over your data. Receive *web visibility*, *academic credit*, and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. [Want to set up your personal dataverse?](#)



Journals

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data*. Participate in the open data movement by using DataVerse as part of your journal data policy or list of repository recommendations. [Want to find out more about journal dataverses?](#)



Institutions

Establish a research data management solution for your community. Federate with a growing list of DataVerse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. [Want to install a DataVerse repository?](#)



“DataVerse is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows you to replicate others' work more easily. Researchers, journals, data authors, publishers, data distributors, and affiliated institutions all receive academic credit and web visibility.”

# Why current solutions are not sufficient

No peer-review of datasets and metadata.

No use of ontologies and controlled vocabulary (tomato, Tomato, *S. lycopersicum*, etc.)

No separation between a working space and a curated controlled area.

Different read-write permissions and user group management (?)

Not easy to “mount” datasets to a computing environment for analysis.

MENU ▾

SCIENTIFIC DATA

Comment | [Open Access](#) | Published: 15 March 2016

# The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons *Scientific Data* **3**, Article number: 160018 (2016) | [Download Citation](#)  An Addendum to this article was published on 19 March 2019

## Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and



Search



E-alert



Submit



Login

Download PDF



1048

Citations

1405

Altmetric

[Article metrics](#) >>

## Associated Content

Collection

[Scientific data](#)

Collection

[Metadata Quality](#)

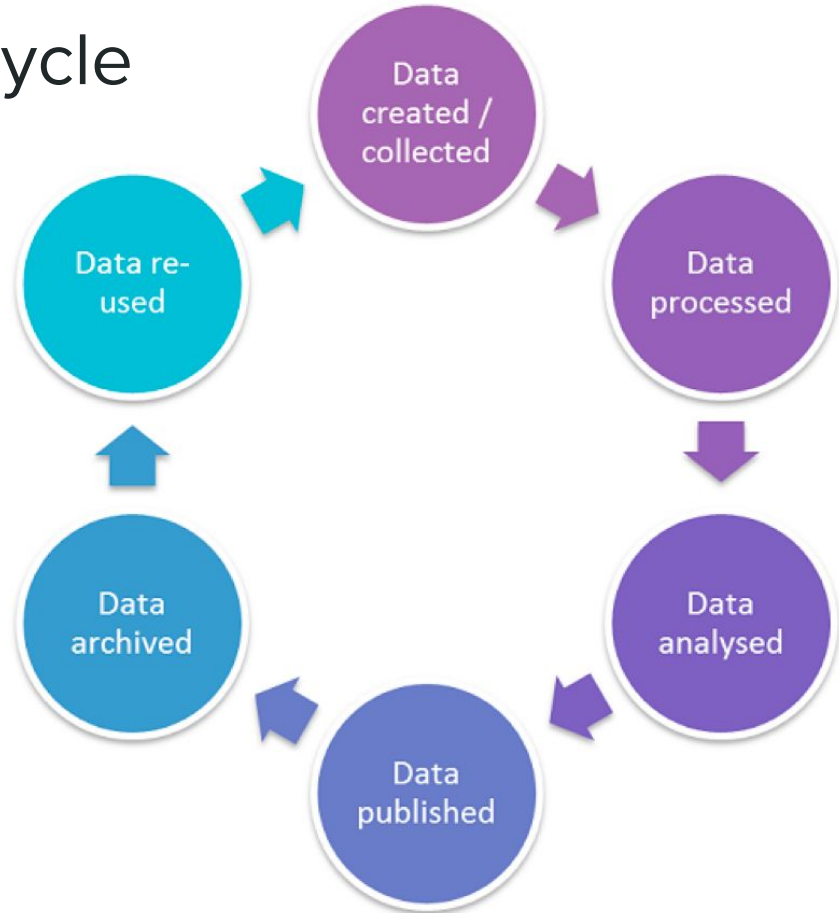
Sections

References

Abstract

[Comment](#)[Additional Information](#)[References](#)[Acknowledgements](#)[Author information](#)[Rights and permissions](#)[About this article](#)

# The Research Data Life Cycle

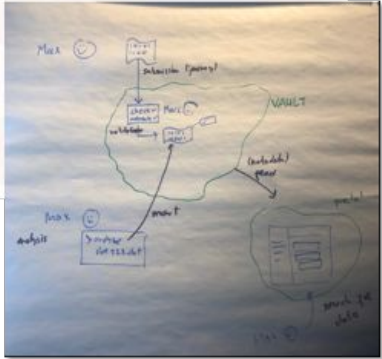


# Definitions of the tool functionalities and scope

## User journeys

	Max (researcher)	Marc (data steward)	Petra (group leader)	Michel (department professor)	Lisette (student)	Alain (external collaborator)
Data submission & curation	X	X				
Data search	X		X		X	
Data analysis	X	X				
Data publication	X					
Status overview			X			
Invite external collaborator			X			
Reproduction / traceability				X		

OUT OF SCOPE FOR POC



Summarized from detailed description at: [https://docs.google.com/document/d/1j-XF5Rg\\_Q2xRoHMcofILVeDbIBYAIzDxqef6UcARWnE/edit#](https://docs.google.com/document/d/1j-XF5Rg_Q2xRoHMcofILVeDbIBYAIzDxqef6UcARWnE/edit#)

# Physical damage to data



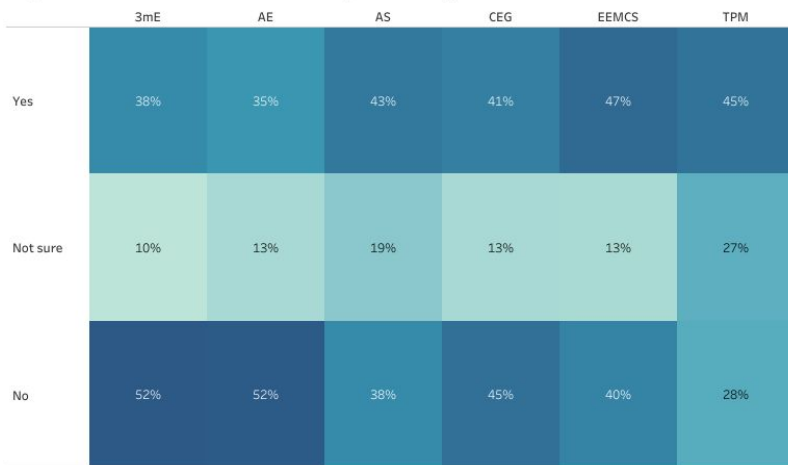
“Hot” data (TU Delft, 2008)



Data could have  
“vanished” in thin air  
(UvA 2018)

# But it can be improved relatively easily

Is your research data automatically backed up?



Is your research data backed up?



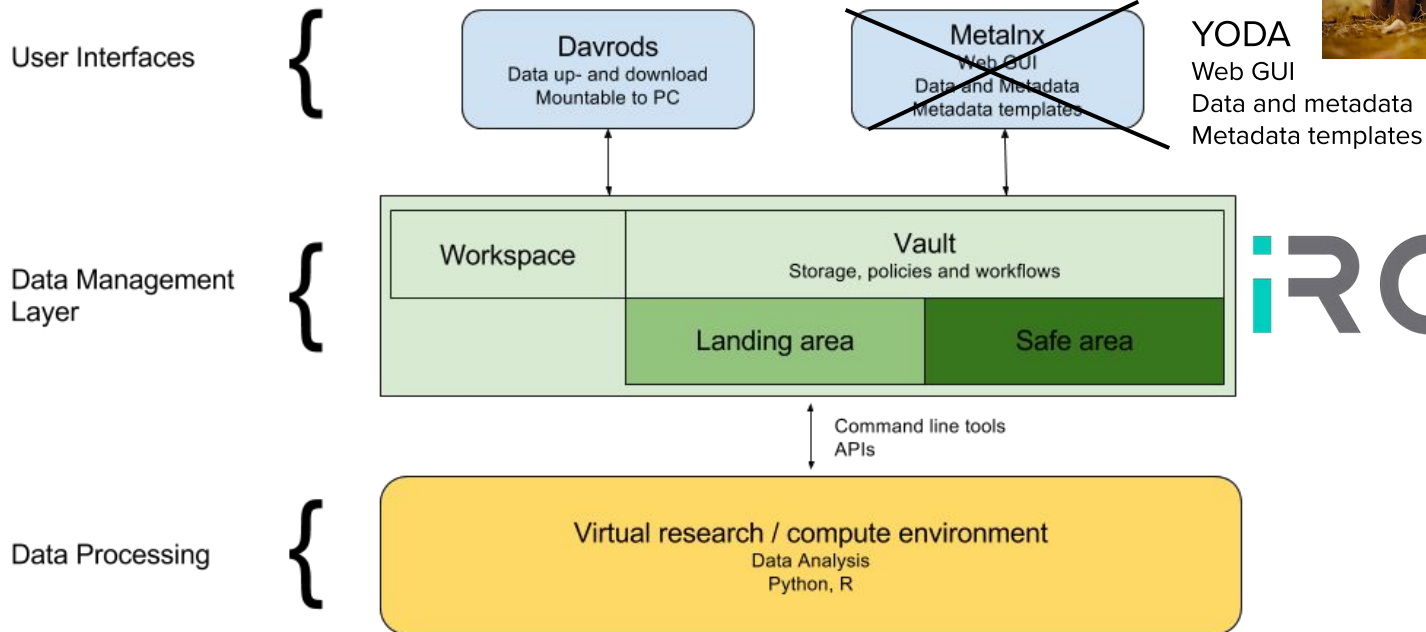
TU Delft 2017/8  
No  $\approx$  42.5%

8 Data Stewards hired

TU Delft 2019  
No  $\approx$  6%

TU Delft survey (2017/18 and 2019)

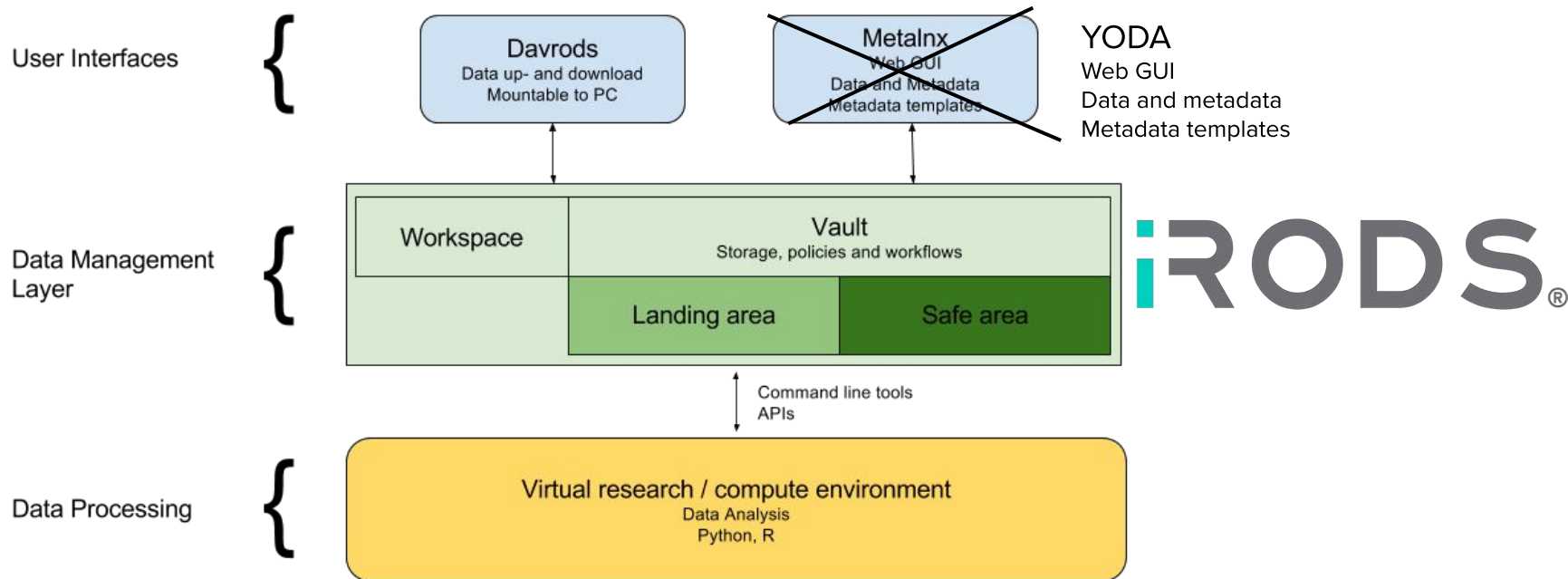
# YODA is the car body



iRODS®



# iRODS is the engine under the hood



# And then...

Confidentiality issues.

Lack of trust that the system effectively protects datasets.

Other reasons?



University of Amsterdam

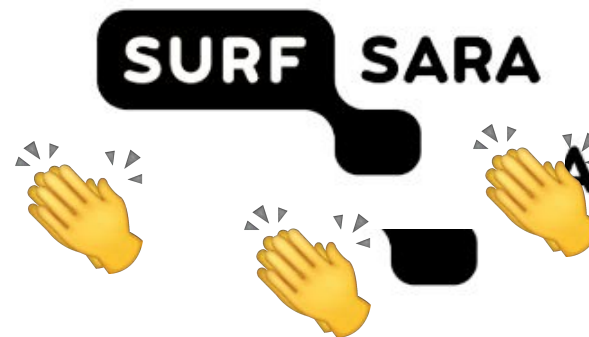


Photo by Analise Benevides on Unsplash

# iRODS: integrated rule-oriented data system

Open Source data management software

Main advantages:

- **Data virtualization:** virtualize the data storage resources: several files = one data object
- **Data discovery:** metadata catalog describing files, directories, etc.
- **Data workflows:** actions can be automatically implemented based on defined rules.
- **Secure collaboration between users**

# Project members (2017-2018)

## People involved in the “Data Vault” pilot project



Jeroen Rouppe  
van der Voort  
(Seed Valley)



Michel Haring (UvA)



Maarten  
Noom  
(HvA-UvA)



Petra Bleeker  
(UvA)



Joyce  
Nijkamp  
(HvA-UvA)



Christine  
Staiger  
(SURFsara)



Ander  
Astudillo  
(SURFsara)



Marc Galland (UvA)



UNIVERSITEIT VAN AMSTERDAM



# Project members “Green Data Vault” (2019 - )



Michel Haring (UvA)



Maarten Noom (HvA-UvA)



Petra Bleeker (UvA)



UNIVERSITEIT VAN AMSTERDAM



Joyce Nijkamp (HvA-UvA)



Hylke Koers (SURFsara)



Arthur Newton (SURFsara)



Marc Galland (UvA)



# Traceability and provenance (and much more)

## Provenance

Clicking the *book* next to the package will show the provenance.

Home / vault-myresearch-data / analysis[1508406722]

analysis[1508406722] ⓘ 📖

Metadata

Published



Provenance information:

2017/10/18 21:04:02 - **Submitted for vault** - testresearcher

2017/10/18 21:04:36 - **Unsubmitted for vault** - testresearcher

2017/10/19 11:50:26 - **Submitted for vault** - testresearcher

2017/10/19 11:51:29 - **Accepted for vault** - testdatamanager

2017/10/19 11:52:02 - **Secured in vault** - system

2017/11/07 11:56:01 - **Submitted for publication** - testdatamanager

2017/11/07 12:00:02 - **Approved for publication** - testdatamanager

2017/11/07 12:00:07 - **Published** - system