SURF super day 2019

The data frontline in academia

Marc Galland, 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

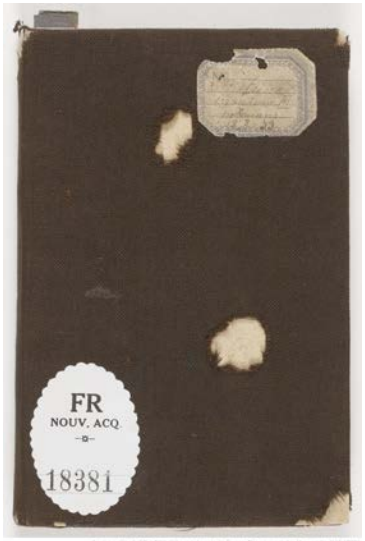
Outline

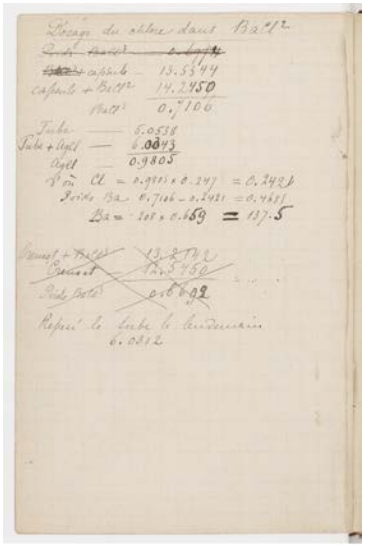
1. **Current Research Data Management in academia**
2. **The “Green Data Vault” project**
    a. History: Seed Valley
    b. Vault workflow and architecture.
    c. Milestone: decision to adopt iRODS & YODA (May 2019)
    d. Test of YODA at the Green Life Sciences cluster (UvA)
3. **Future plans (also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

Data are recorded in paper laboratory notebooks

Marie Curie Sklodowska
(in 1921)

Laboratory notebooks



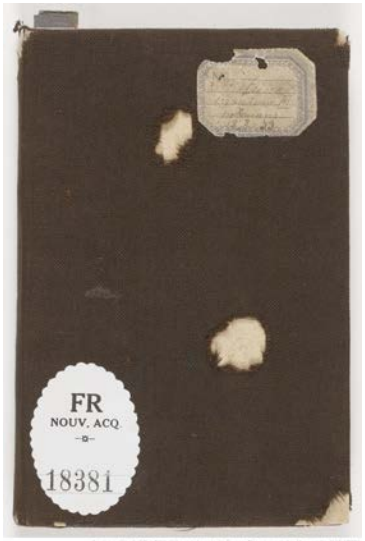


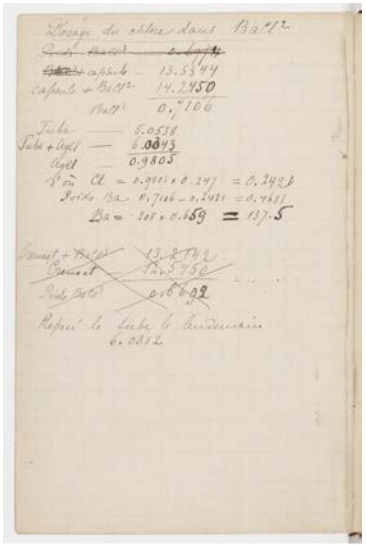
Bibliothèque Nationale de France: www.gallica.bnf.fr
Jean-Luc Pasquier: https://6bisruedemessine.wordpress.com/

Data are recorded in paper laboratory notebooks

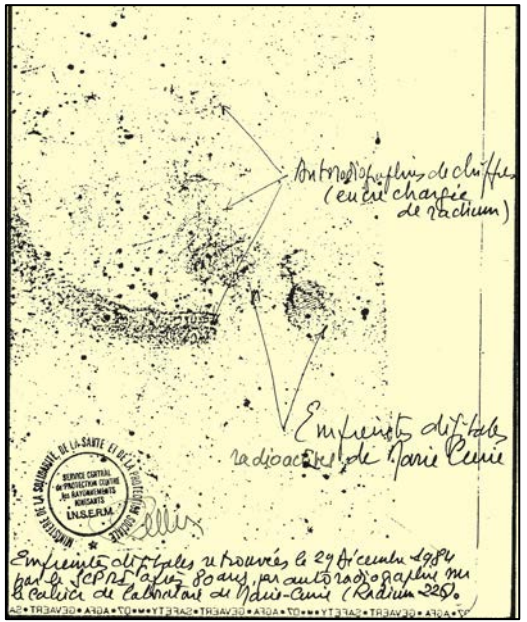
Marie Curie Sklodowska (1921)

Laboratory notebooks





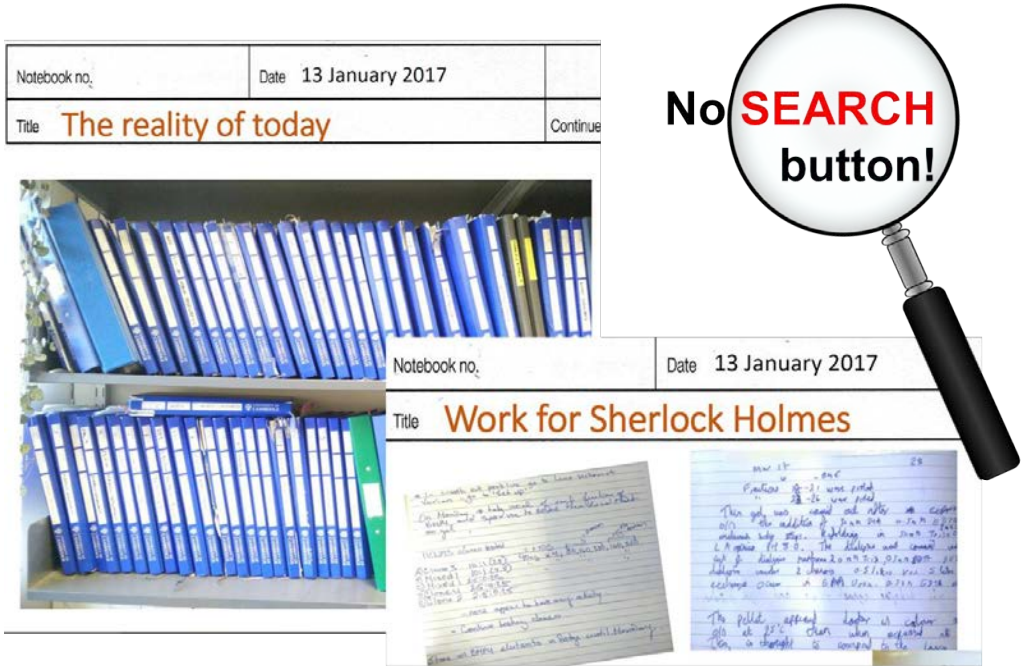
Radioactive lab notebook



Bibliothèque Nationale de France: www.gallica.bnf.fr
Jean-Luc Pasquier: https://6bisruedemessine.wordpress.com/

Laboratory notebooks have no search button

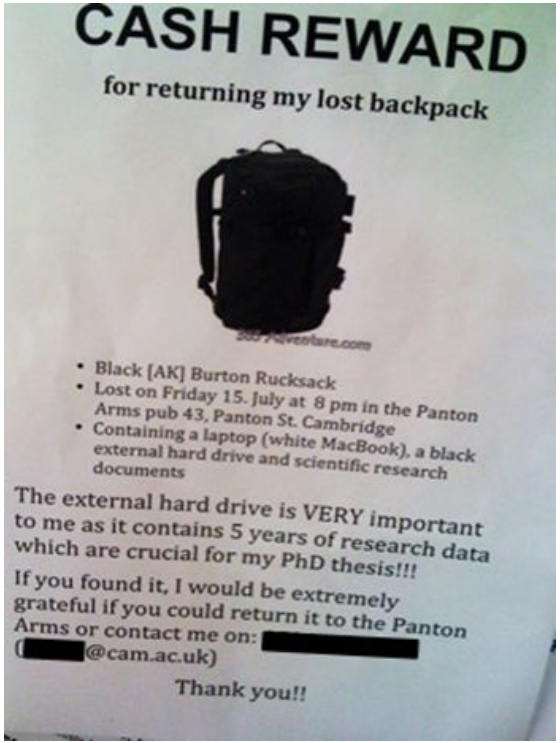
Experimental notes today…



Presentation by Dr Marko Hyvonen

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

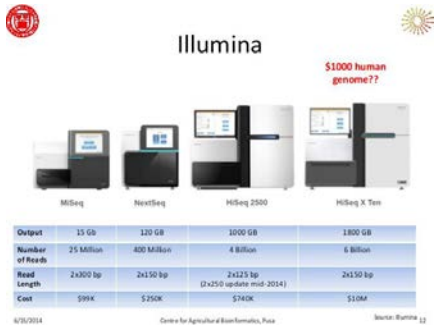
Data are (often not) backed up.



Reward for the whole PhD!
Unspecified amount

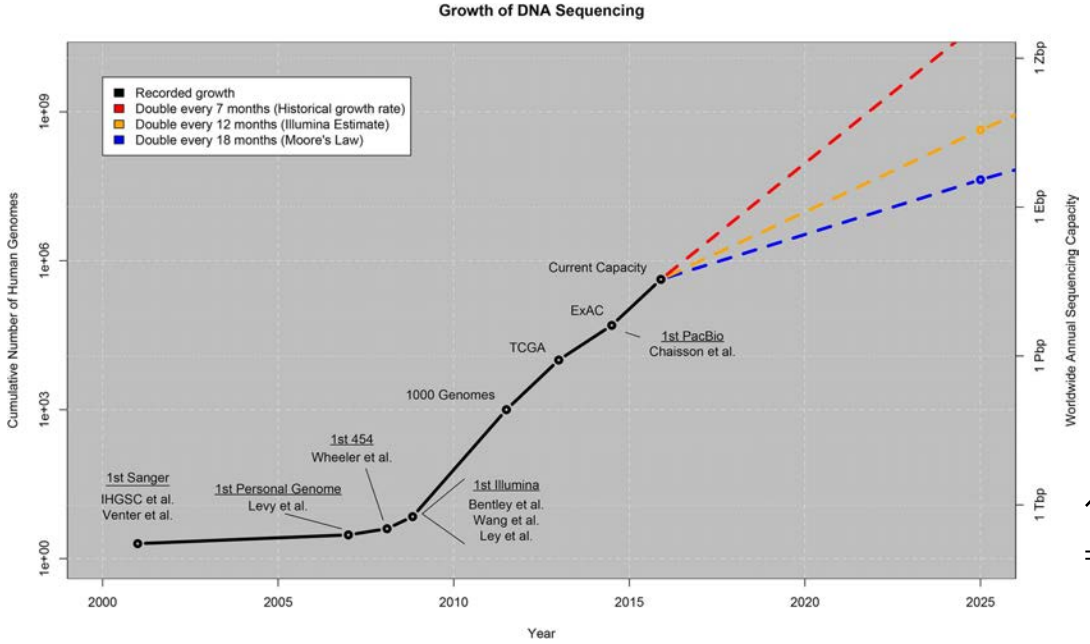
Photo credit: Peter Murray-Rust

A genomic data deluge in the Life Sciences



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

A genomic data deluge in the Life Sciences



1 Zbp
= 10^{21} base pair

1 Tbp
= 10^{12} base pair

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.

Photo credits: British Library, Museums Victoria on Unsplash

Academics sent joyfully to the data battlefield

PhD students, postdocs, etc.

“Good luck with all these data”

Photo credits: British Library, Museums Victoria on Unsplash

Academics sent joyfully to the data battlefield

They get harmed:

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

Photo credits: British Library, Museums Victoria on Unsplash

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...

Photo credits: British Library, Museums Victoria on Unsplash

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

Photo credits: British Library, Museums Victoria on Unsplash

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

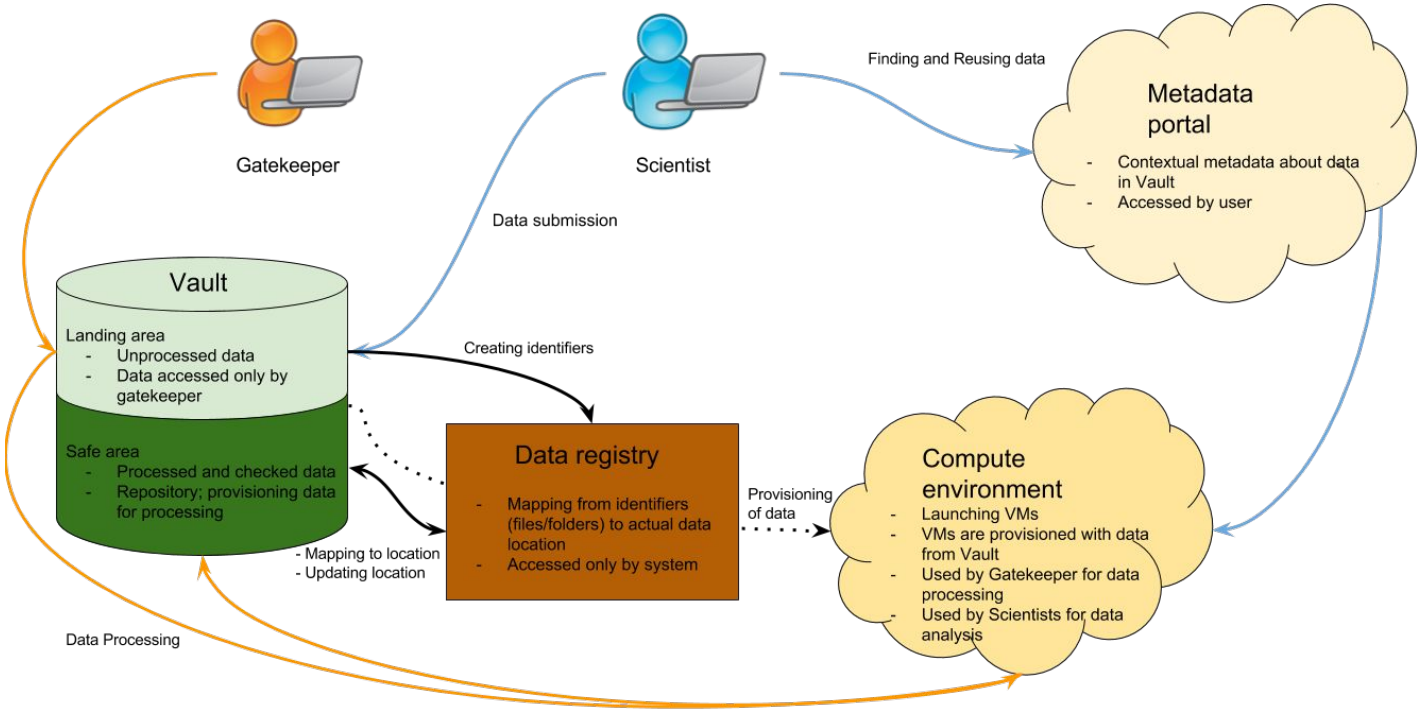
“Green Data Vault” project

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

“Green Data Vault” project

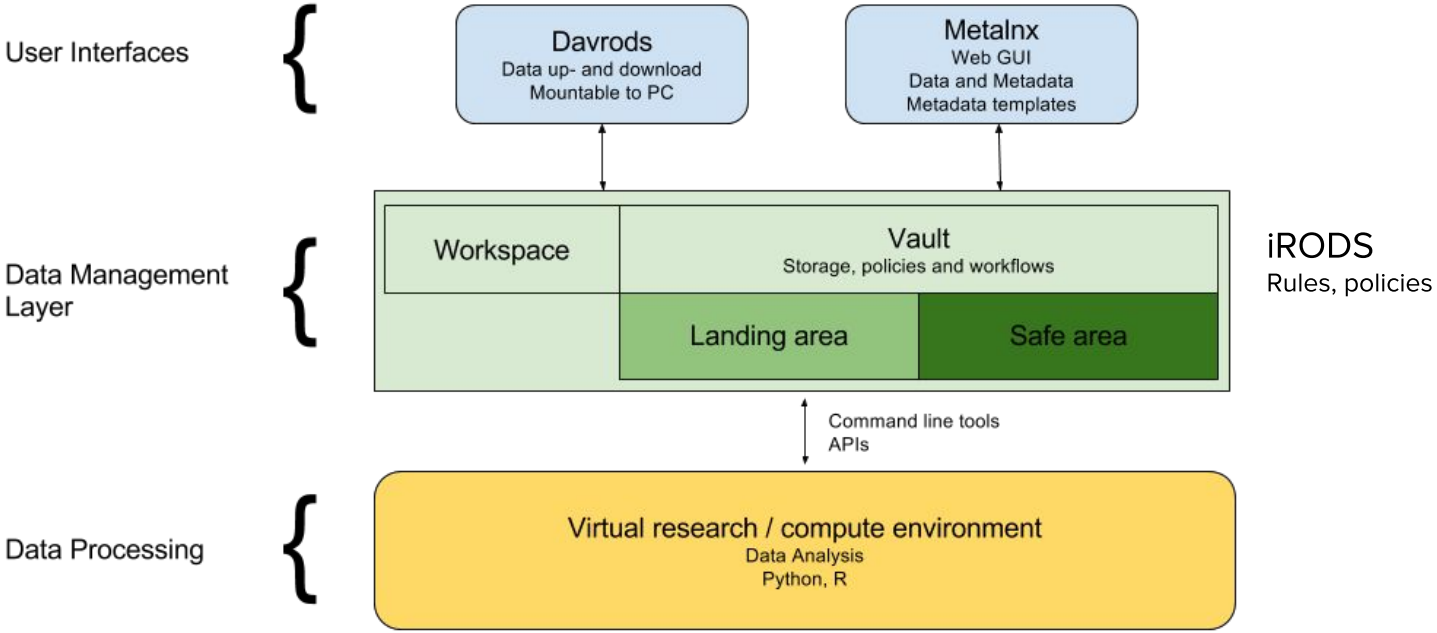
- **Seed Valley:** a consortium of plant breeders in North Holland.
- **Goal:** Share datasets to fasten and strengthen findings.
- Eventually, facilitate the obtention of new vegetable cultivars.

Definition of the “Green Data Vault” workflow



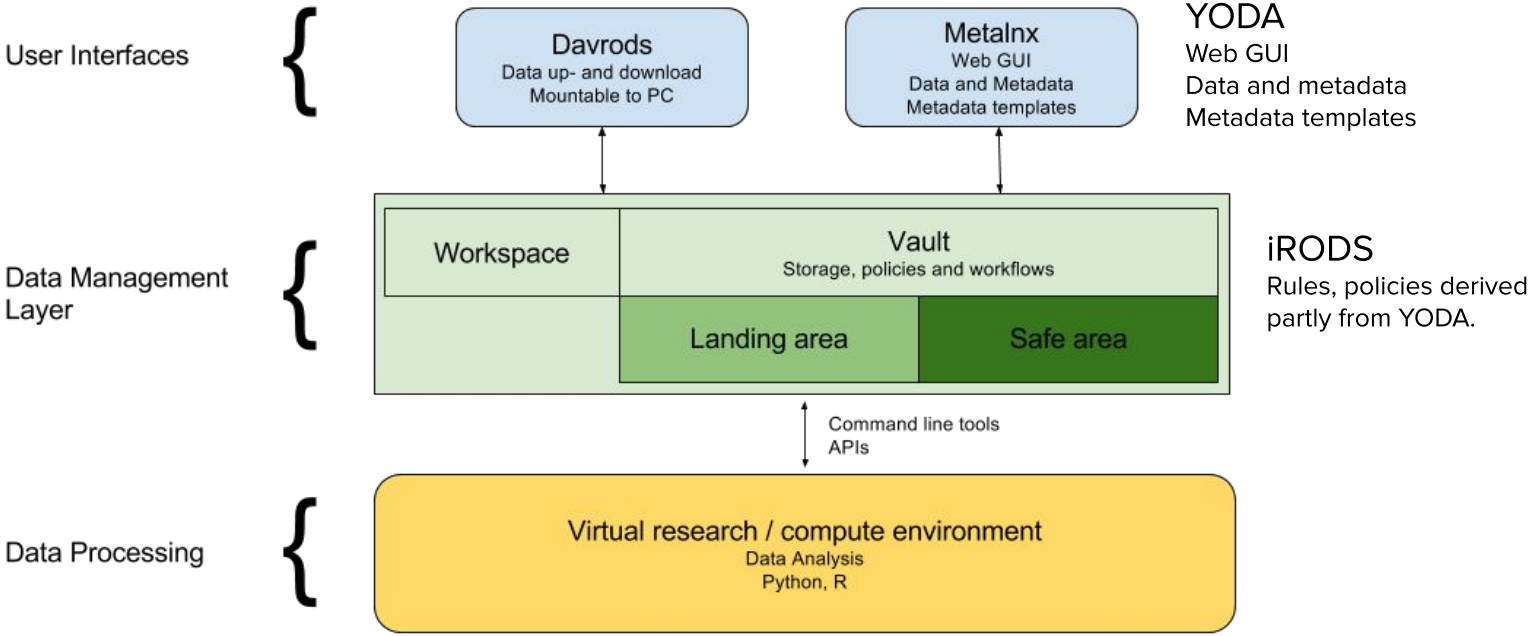
Scheme by Christine Staiger (2017)

Definition of the “Green Data Vault” system architecture



Scheme by Christine Staiger (2017)

Definition of the “Green Data Vault” system architecture



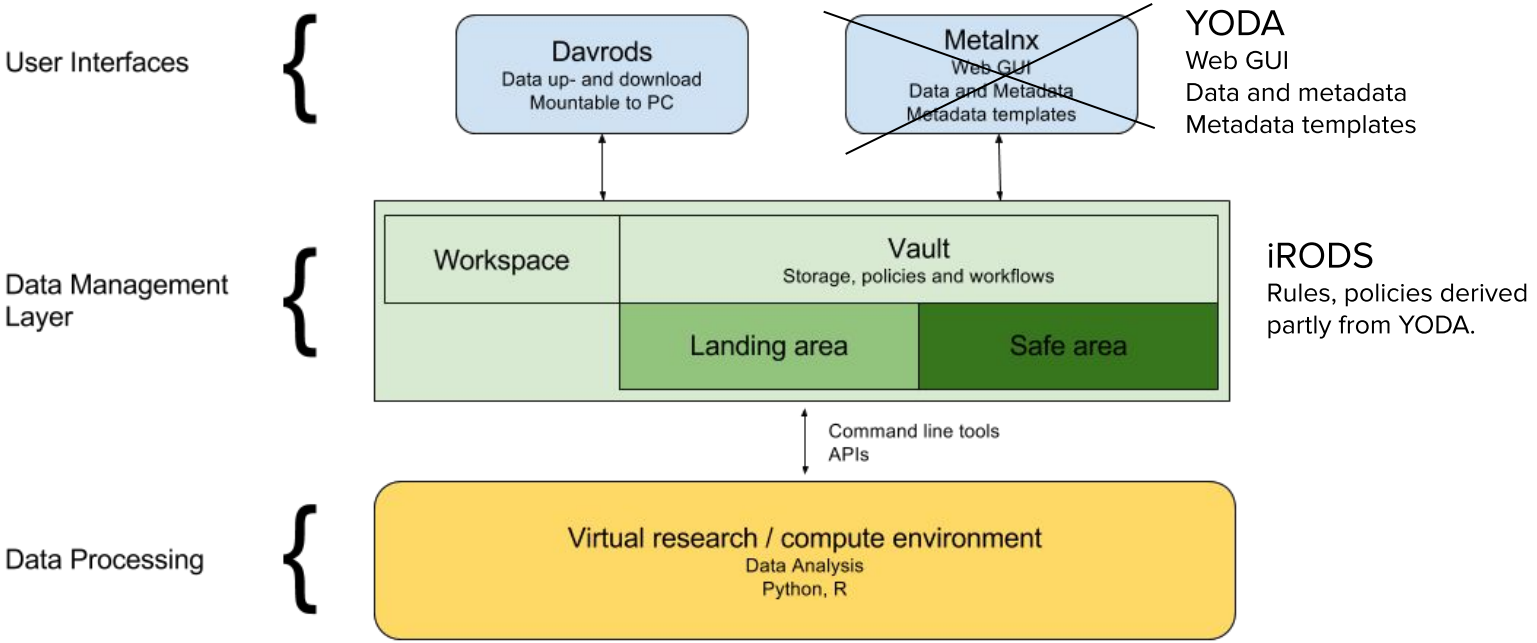
User tests (April 2019)

- 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: MetaInx (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”.

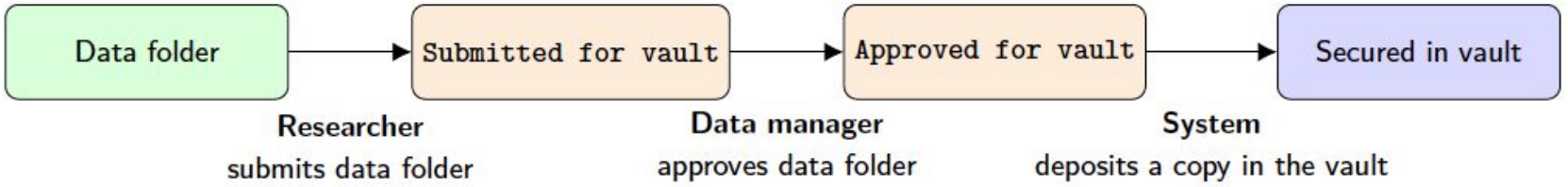


Figure 2. The workflow to deposit a data package into the vault

Smeele and Westerhof, 2018. Proceedings of the iRODS User Group Meeting 2018

Two main workflows supported by YODA

2. Data publication (outside world)

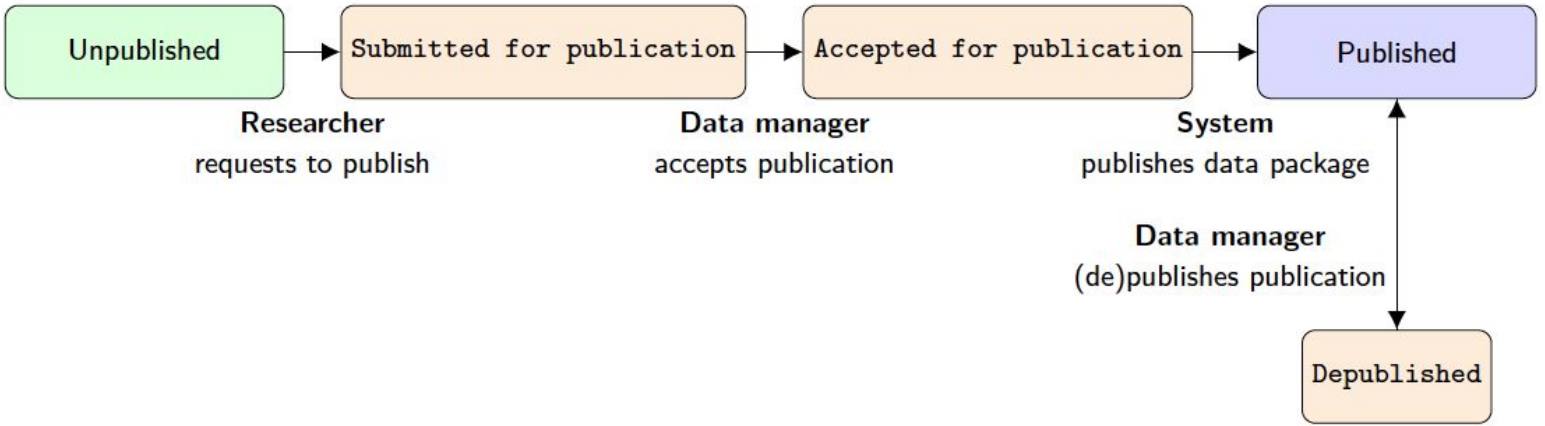


Figure 3. The workflow to publish a data package

Smeele and Westerhof, 2018. Proceedings of the iRODS User Group Meeting 2018

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.

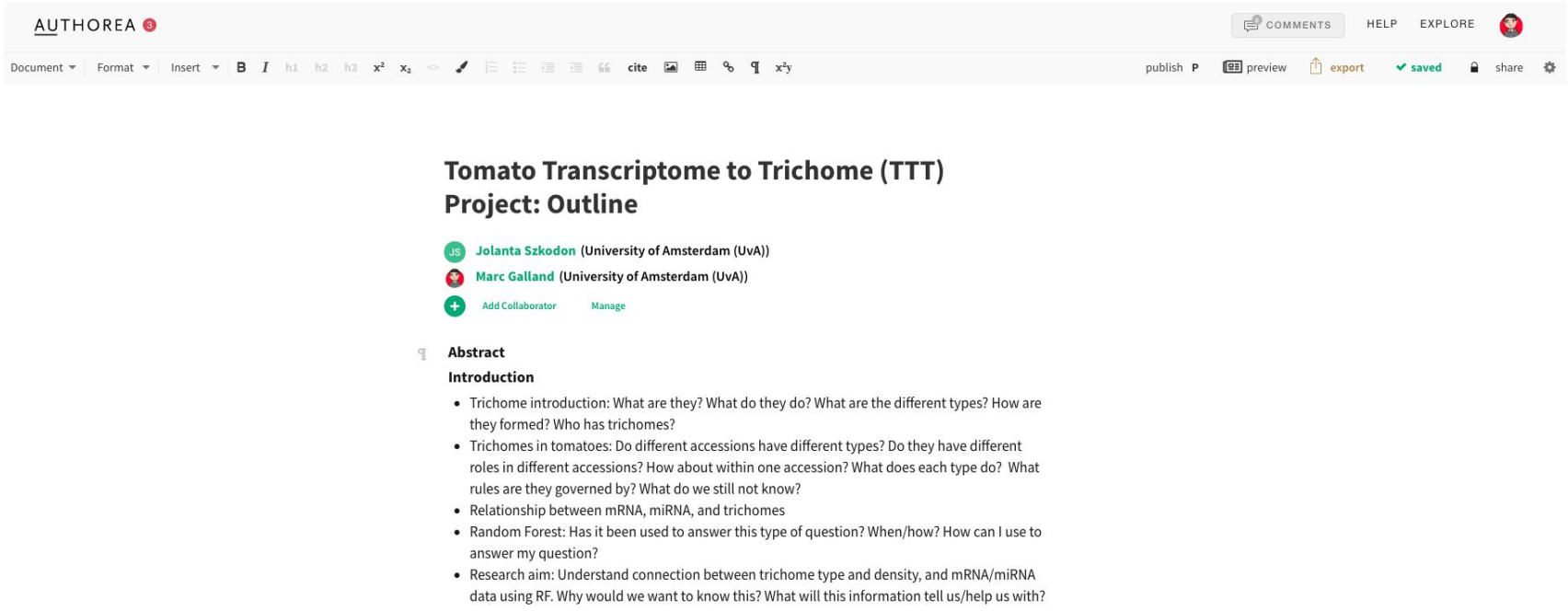
Query
ENA search API
iRODS API

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

Analysis
Versioned
genomic pipeline

Traceable and
reproducible results

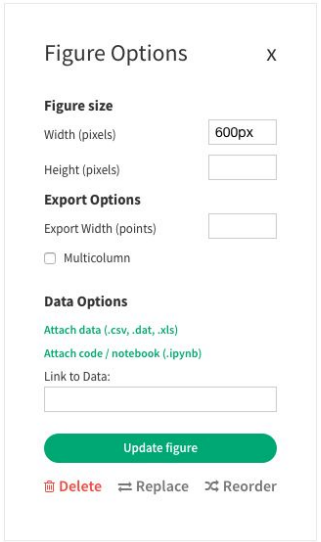
New ways to perform research and publish

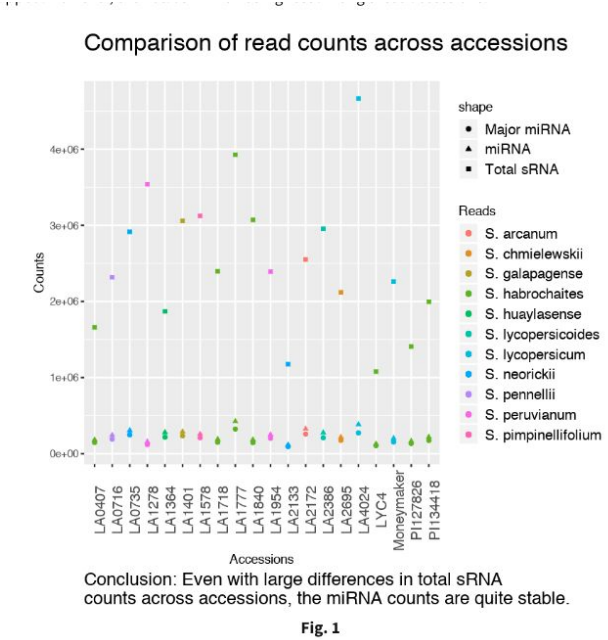


Alyssa Goodman, Josh Peek, Alberto Accomazzi, et al. The "Paper" of the Future. *Authorea*. February 21, 2017. **DOI:** https://doi.org/10.22541/au.148769949.92783646

Link to the data and code underlying a figure

Link datasets





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"

24 views | 5 downloads | 0 citations

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)

Newcastle University

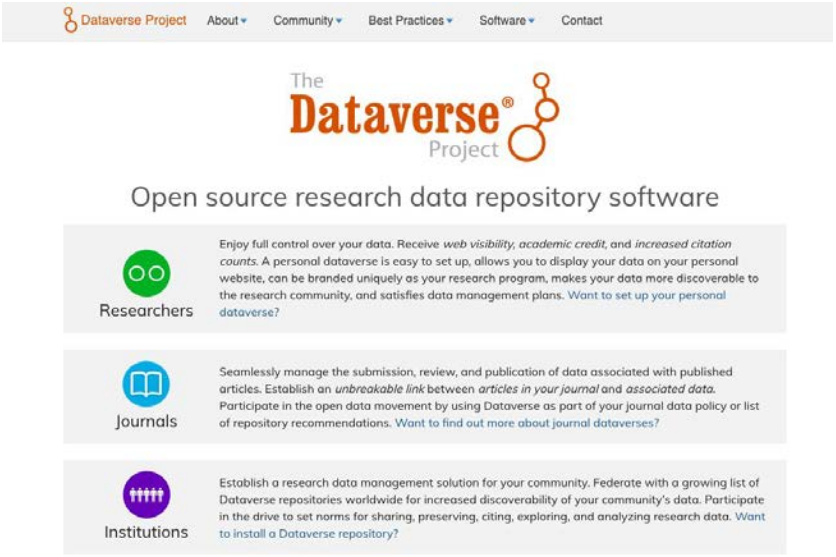
CATEGORIES
- Applied Computer Science
- Artificial Intelligence and Image Processing
- Health Informatics

KEYWORD(S)
experiment results | machine learning | multi-class classification | imbalanced data | clinicial decision-making | patient selection | knee osteoarthritis

LICENCE
CC0

Why current solutions are not sufficient (DataVerse)

Harvard University: IGSS institute



“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.

Comment | Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

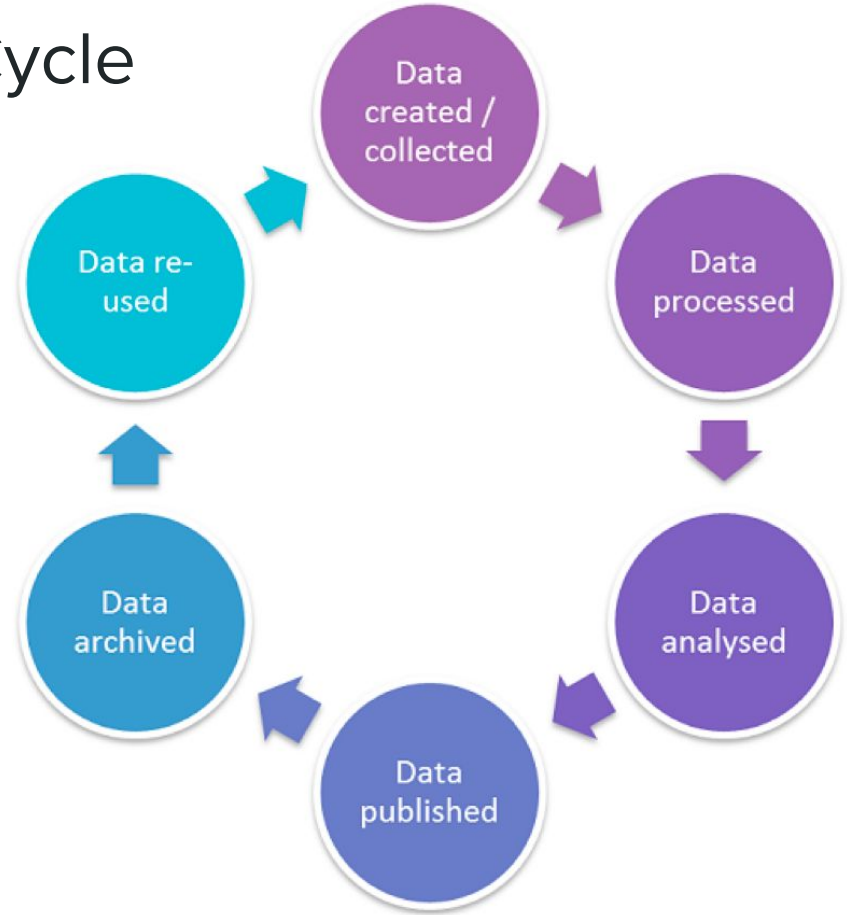
Scientific Data **3**, Article number: 160018 (2016) | Download Citation

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 measureable 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

The Research Data Life Cycle

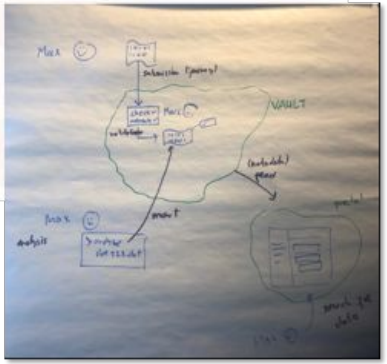


Definitions of the tool functionalities and scope

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		

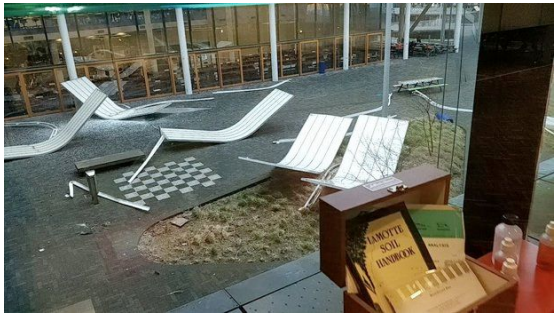
OUT OF SCOPE FOR POC



Summarized from detailed description at: https://docs.google.com/document/d/1J-XF5Rg_Q2xRoHMcofILVeDblBYAlZDxqef6UcARWnE/edit#

Physical damage to data

“Hot” data (TU Delft, 2008)

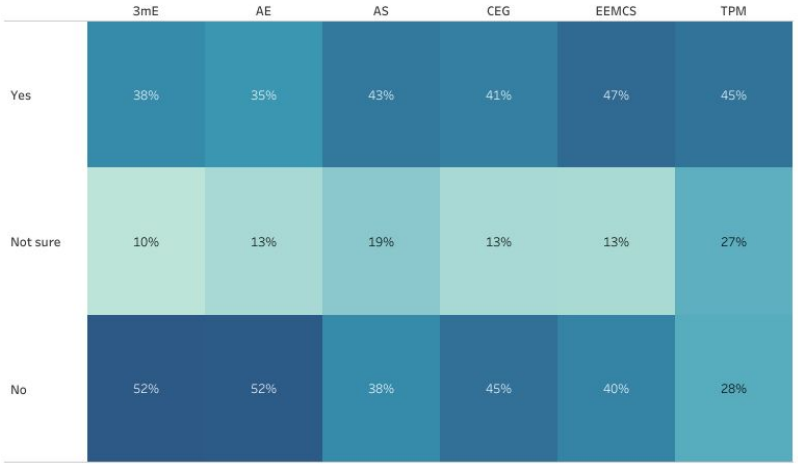


Data could have “vanished” in thin air (UvA 2018)

Source: Wikipedia. Folia

But it can be improved relatively easily

Is your research data automatically backed up?



Is your research data backed up?

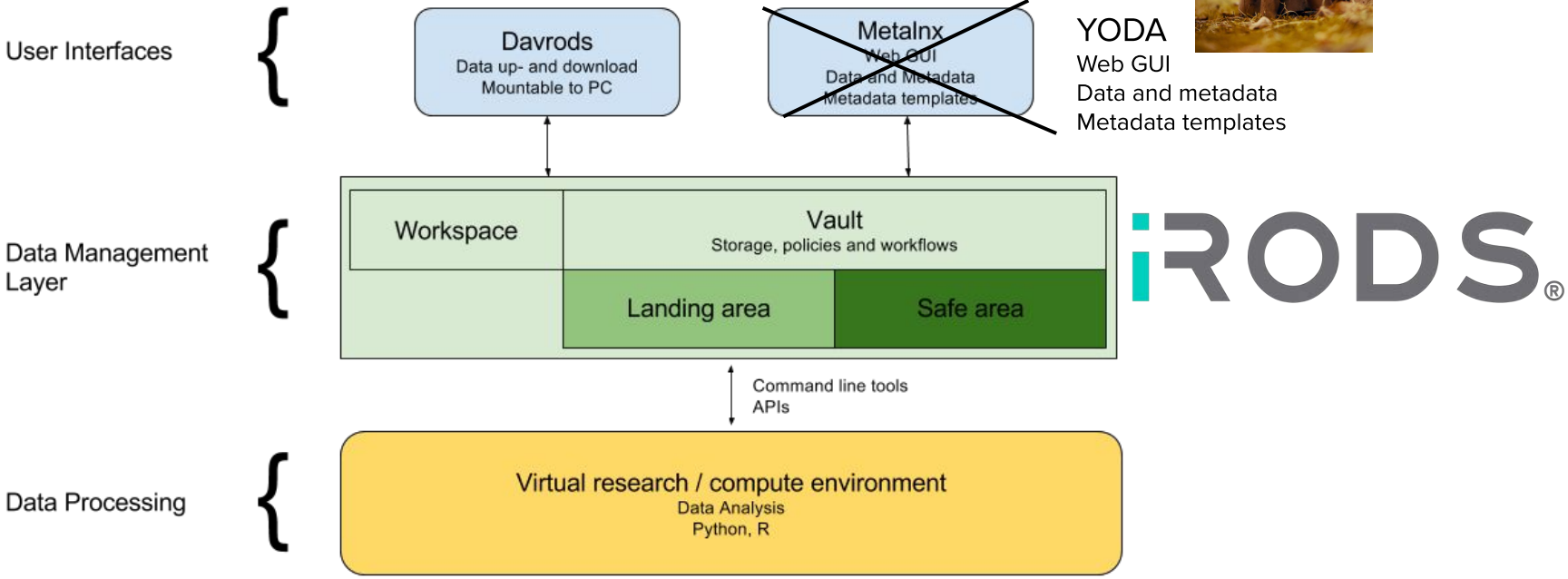
TU Delft 2017/8
No ≅42.5%

8 Data Stewards hired

TU Delft 2019
No ≅6%

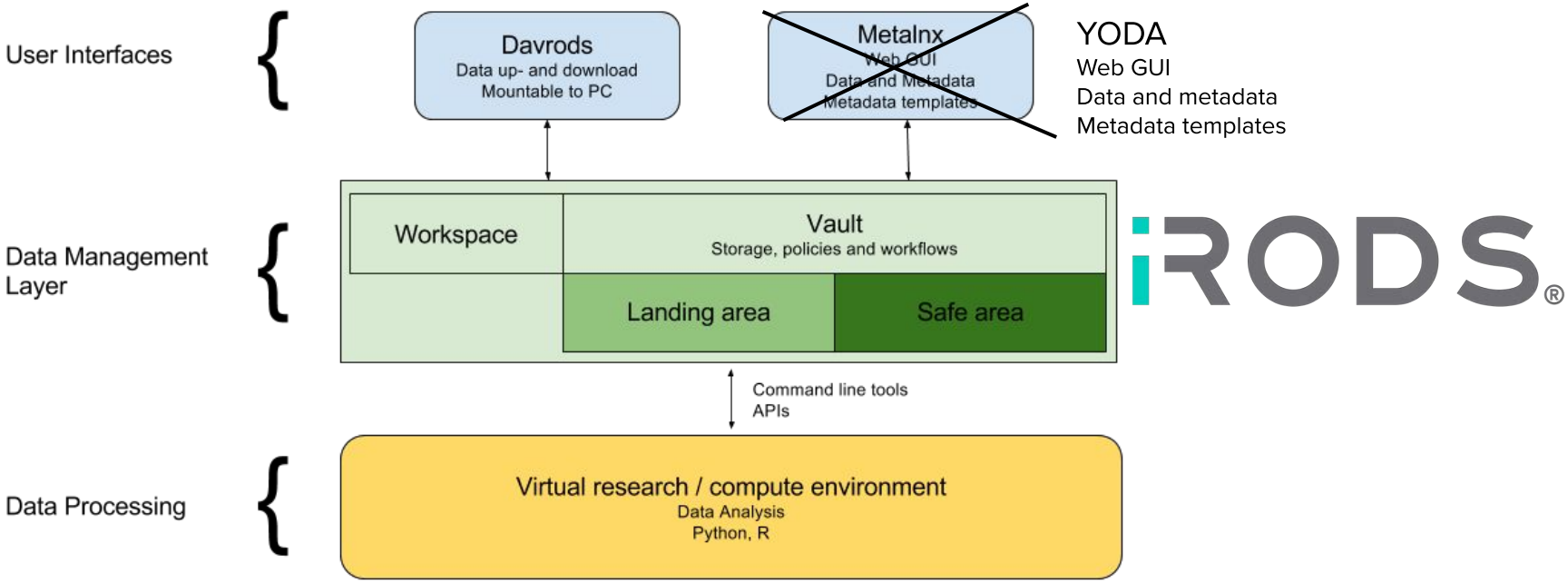
TU Delft survey (2017/18 and 2019)

YODA is the car body



YODA
Web GUI
Data and metadata
Metadata templates

iRODS is the engine under the hood



And then...

Seed Valley

Confidentiality issues.

Lack of trust that the system effectively protects datasets.

Other reasons?

University of Amsterdam

SURF SARA

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)

Seed Valley

Universiteit van Amsterdam

SURF SARA

Project members “Green Data Vault” (2019 -)

Michel Haring (UvA)

Maarten Noom (HvA-UvA)

Petra Bleeker (UvA)

Joyce Nijkamp (HvA-UvA)

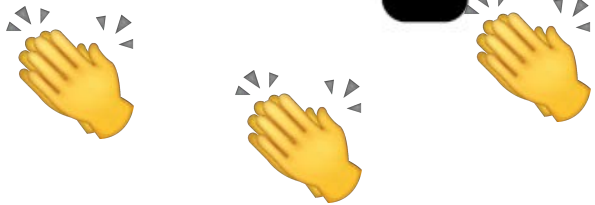
Hylke Koers (SURFsara)

Arthur Newton (SURFsara)

Marc Galland (UvA)

UNIVERSITEIT VAN AMSTERDAM

SURF SARA



Traceability and provenance (and much more)

Provenance

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