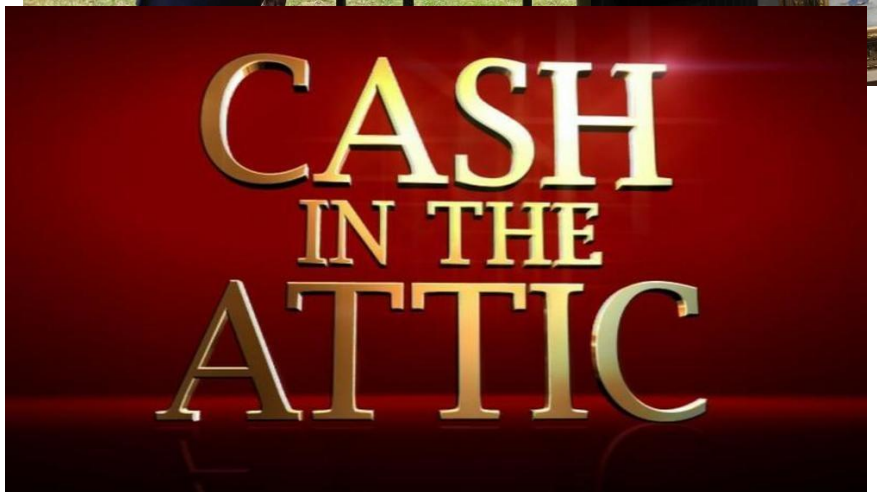
ILLUMINATING DARK DATA

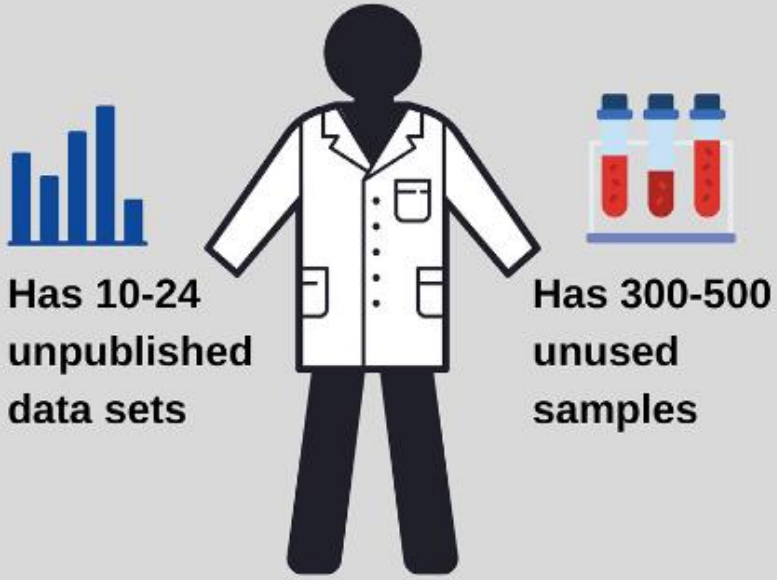
WHY YOU SHOULD STOP STORING DATA

ANTIQUES ROADSHOW



Survey responses indicate The typical US researcher:

Publishes about 60% of their publishable data



Has 10-24 unpublished data sets

Has 300-500 unused samples

Has $29,000 in unused resources

Unused resources at US Academic insitutuions:

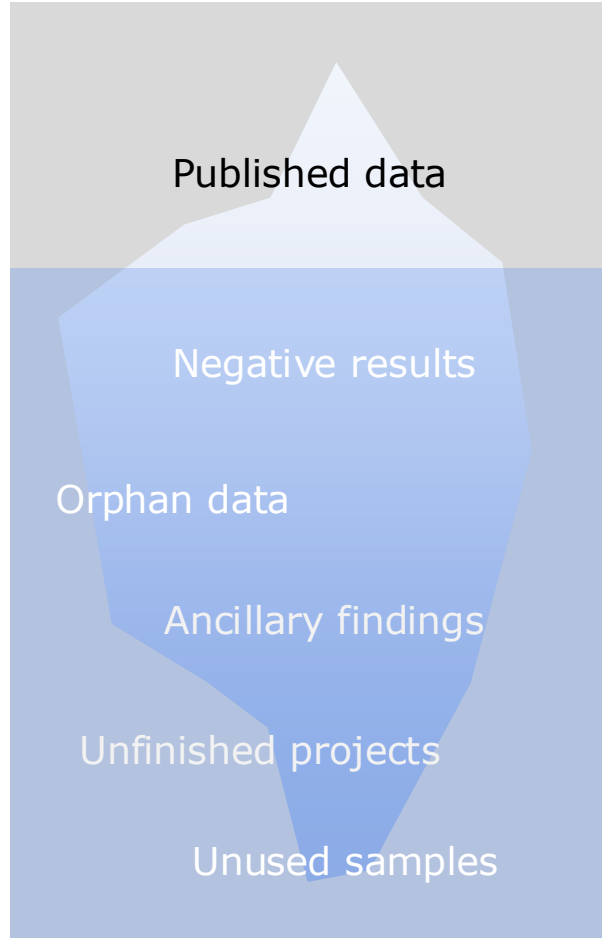
Highlights

Unpublished data and unused samples represent billions of dollars in stranded assets

Unfinished projects and orphan data were the primary sources of unpublished data

Unused samples are not shared because others don't know they exist

Addressing logistical problems can increase the efficiency of research



https://doi.org/10.1016/j.isci.2023.107166

WHAT IS DARK DATA?

- Researchers need to ensure their data and metadata comply with existing standards to be useful.
  - Incompatible data is essentially useless for further research.
- Researchers must allocate time to standardize their data and learn the current standards of repositories.
- Academia and institutions often do not support the time needed for these tasks.
- There are few incentives for researchers to share, format, and standardize data.
- This lack of support leads to reduced professional collaboration and wasted research efforts.
- Valuable scientific data can become forgotten and underutilized, despite significant costs associated with its creation and storage.
- Such neglected data is referred to as "dark data" due to its lack of visibility in research.

Are you storing dark data?

THE CONVERSATION
Academic rigour, journalistic flair

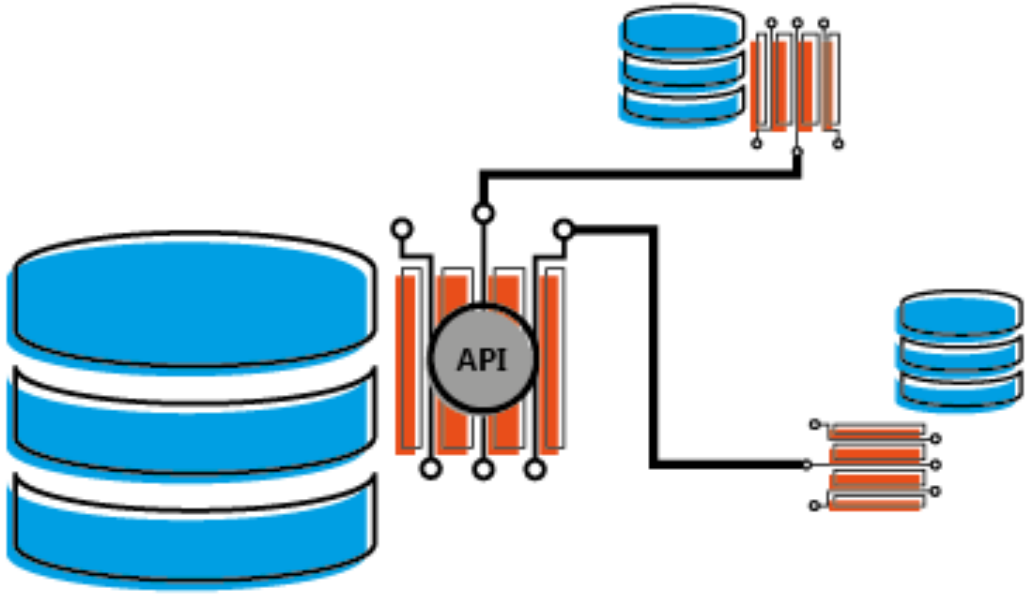
‘Dark data’ is killing the planet – we need digital decarbonisation

Published: September 29, 2022 6.12pm CEST

MAIN IDEA – DISTRIBUTE AND STANDARDISE

- Observation: Data will always reside in various places, and is usually slow to move
  - Also moving usually means copying, creating different versions
- Solution: leave data where it is, but make it accessible via *interfaces* (API's) so that users and applications can access them when needed
  - Improves control, findability and security

Stop storing data, and start *managing it!*



FAIR D

https://youtu.be/KWVCSwUNtBA?si=vqL-Q88JDGj_4pRS

FAIR PRINCIPLES

FINDABLE

Persistent unique identifiers enriched with metadata allow (meta) data to be searched for and found.

ACCESSIBLE

The (meta) data is accessible through standard communication (internet) protocols and available for research. Metadata is always accessible.

INTER-OPERABLE

Commonly used and open data formats, programming language, and vocabularies are used to allow easy exchange, integration, and reuse with other data and software.

REUSABLE

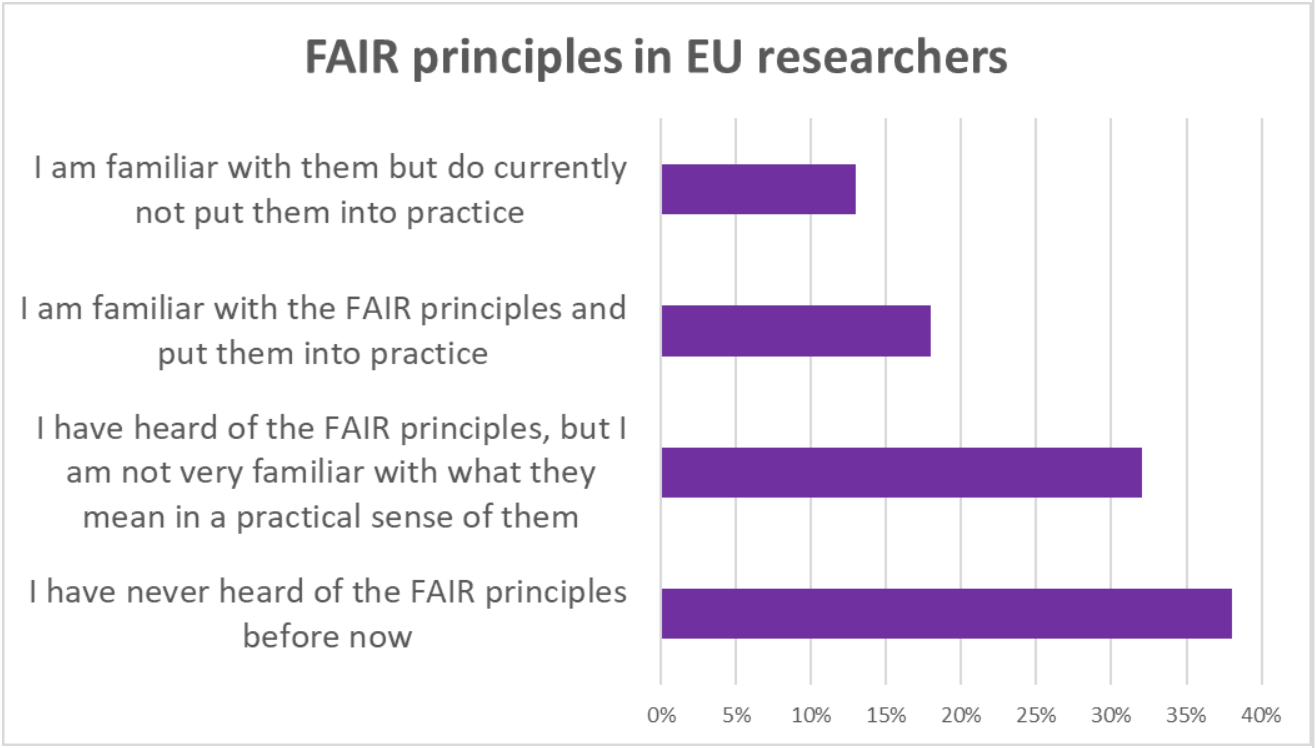
All data is clearly and elaborately documented so that the data can be correctly interpreted and reanalyzed (by others) under specified license.

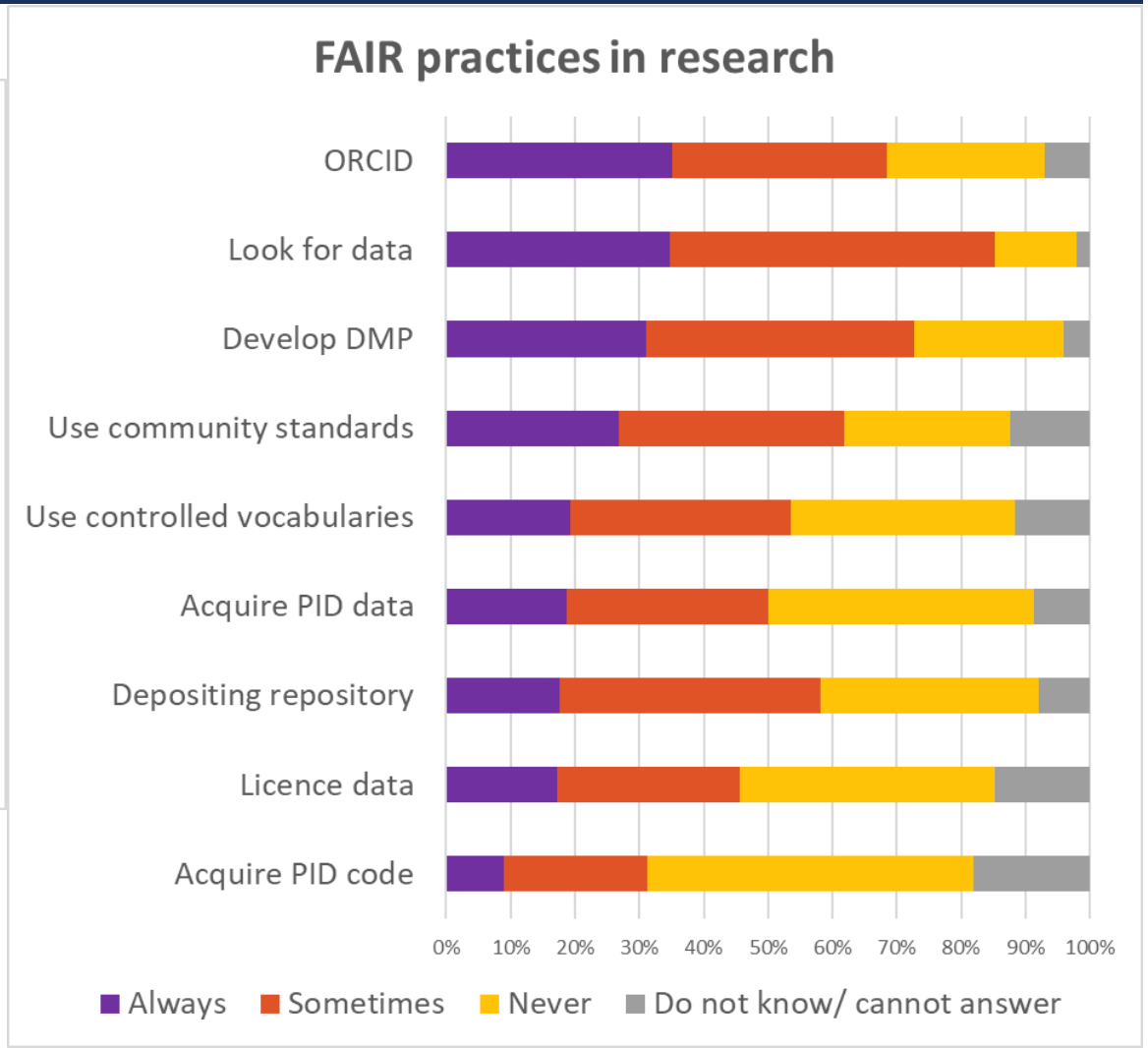
Modified from: 1) SangyaPundir, CC BY-SA 4.0; link https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg. 2) GO FAIR link https://www.go-fair.org/fair-principles/. 3)Health-RI link https://www.health-ri.nl/fair-principles. 4) OpenAIRE link https://www.openaire.eu/how-to-make-your-data-fair.

Expectation

Reality

LACK OF ADOPTION





https://zenodo.org/doi/10.5281/zenodo.6778743

WHY A LACK OF ADOPTION?

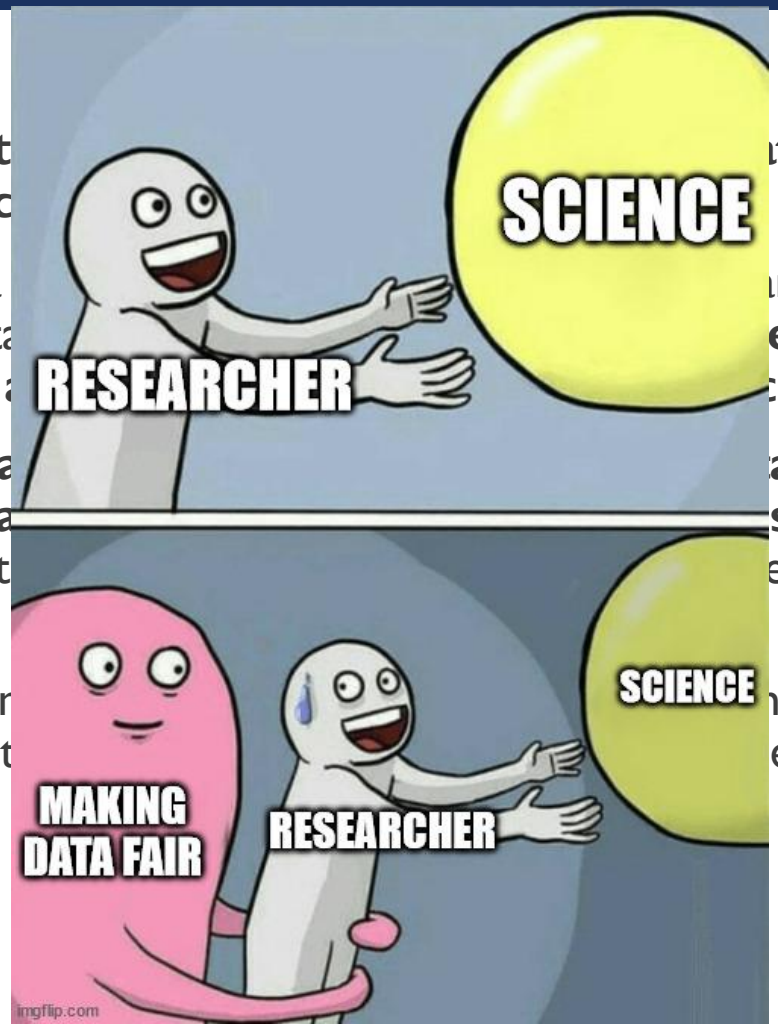
Schembera, B., Durán, J.M. (2019), "**Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer"**:

- Researchers who wish to share their data need to make sure that both the data and the metadata they are submitting are compatible with existing standards. **Incompatible data might be as good as noise, for it cannot be processed for further use and is therefore useless for research**.
- **Researchers must find time in their already busy agenda not only for standardizing their data and metadata but also for acquiring updated knowledge on the standards used by a given repository**. Unfortunately, academia and other institutions are not always receptive to their researchers spending qualitative time on such endeavors.
- As a result, these unacknowledged—but implicitly required—efforts give researchers very few incentives to share, format, and standardize their data and metadata for further use, with the subsequent loss of professional collaboration and research efforts.

WHY A LACK OF ADOPTION?

Schembera, B., Durán, J.M. (2019), "**Dark Dat … ata Science and the Introduction of the Scientific Data Offic**

- Researchers who wish to share their data … and the metadata they are submitting are compatible with existing sta… **e as good as noise, for it cannot be processed for further use … ch**.
- **Researchers must find time in their a… andardizing their data and metadata but also for acquiring upda… sed by a given repository**. Unfortunately, academia and other institut… esearchers spending qualitative time on such endeavors.
- As a result, these unacknowledged—but ir… ers very few incentives to share, format, and standardize their data and met… ent loss of professional collaboration and research efforts.



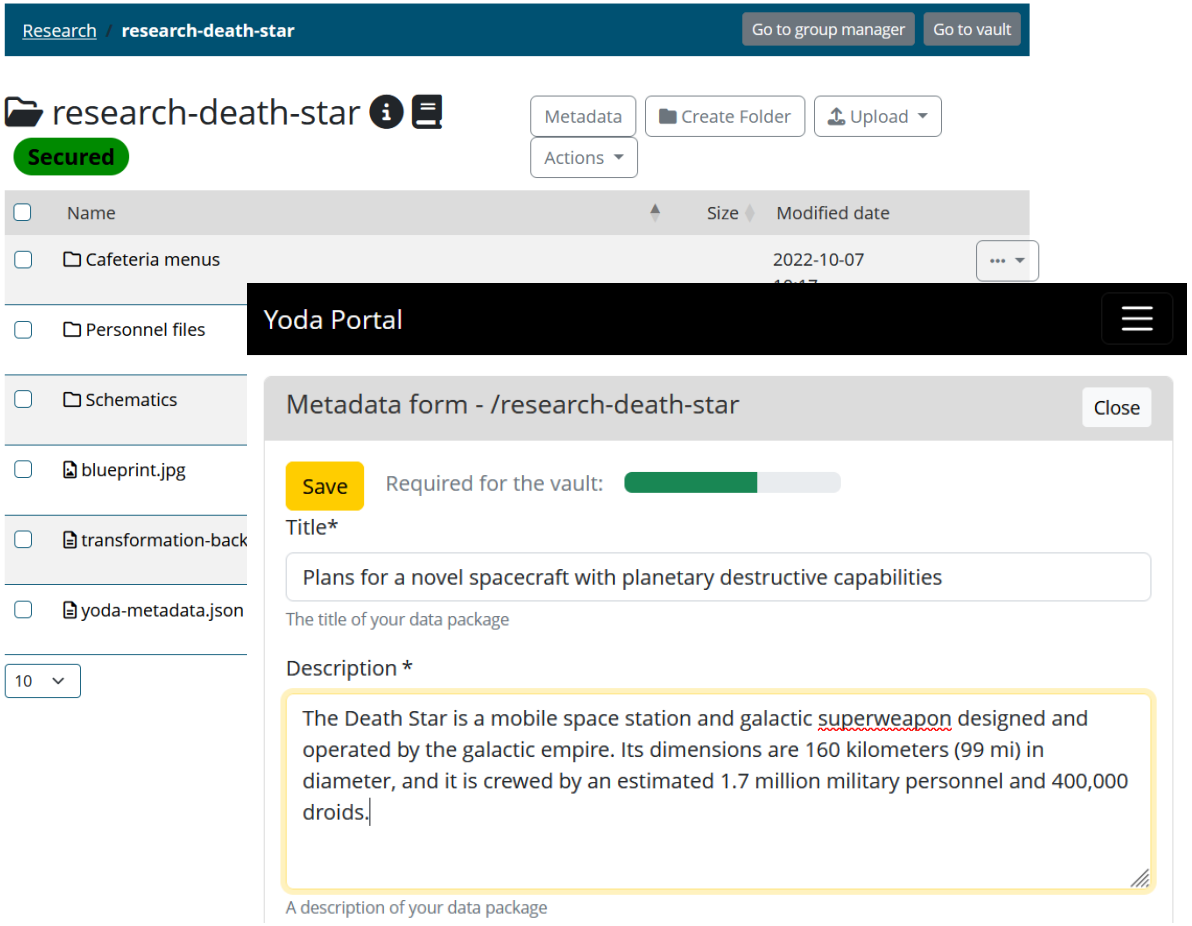
WHAT IS YODA*?

- System for FAIR research data management
- Based on iRODS
  - High performance scientific storage system
- YODA is a RDM focused GUI *on top* of iRODS

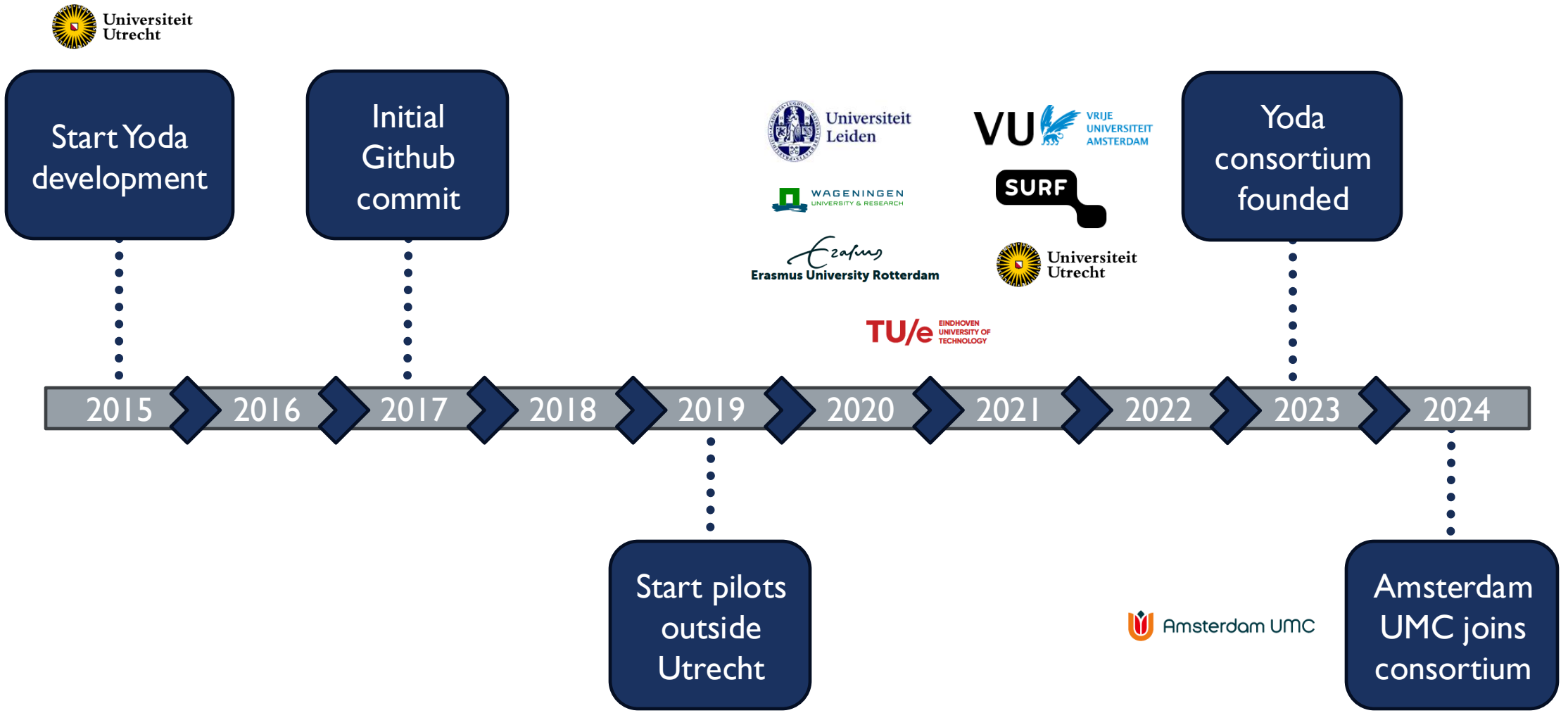
*Yoda is a work of data engineering. Any similarity to specific space opera characters, living or dead, or galactic events, is purely coincidental.

FEATURES

- Based on iRODS architecture and language.
- Yoda is the user-friendly interface on top of iRODS.
- Store and work with data in the Yoda Research environment.
- Archive/secure data in the Yoda Vault environment.
- Add metadata to folders.
- Easily share data within and outside the insititute.



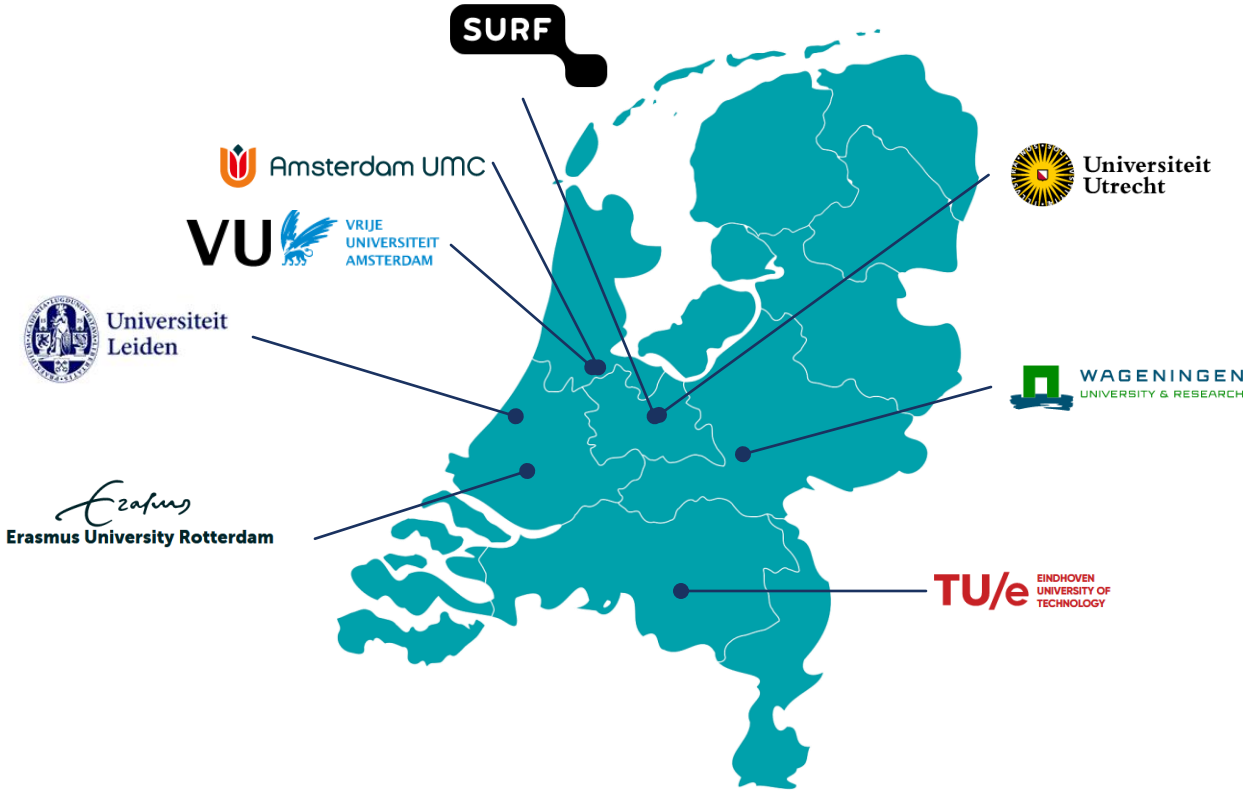
TIMELINE YODA & CONSORTIUM



YODA CONSORTIUM

- 6 out of 13 major universities in the Netherlands & 1 UMC
  - Additionally, 3 uni's use iRODS based solutions
- Partially funds and governs feature development of the Yoda software
- Representation for hosting of Yoda by SURF

- 15,000+ users
- 3,5+ PB of data



https://mapfight.xyz/map/nl/

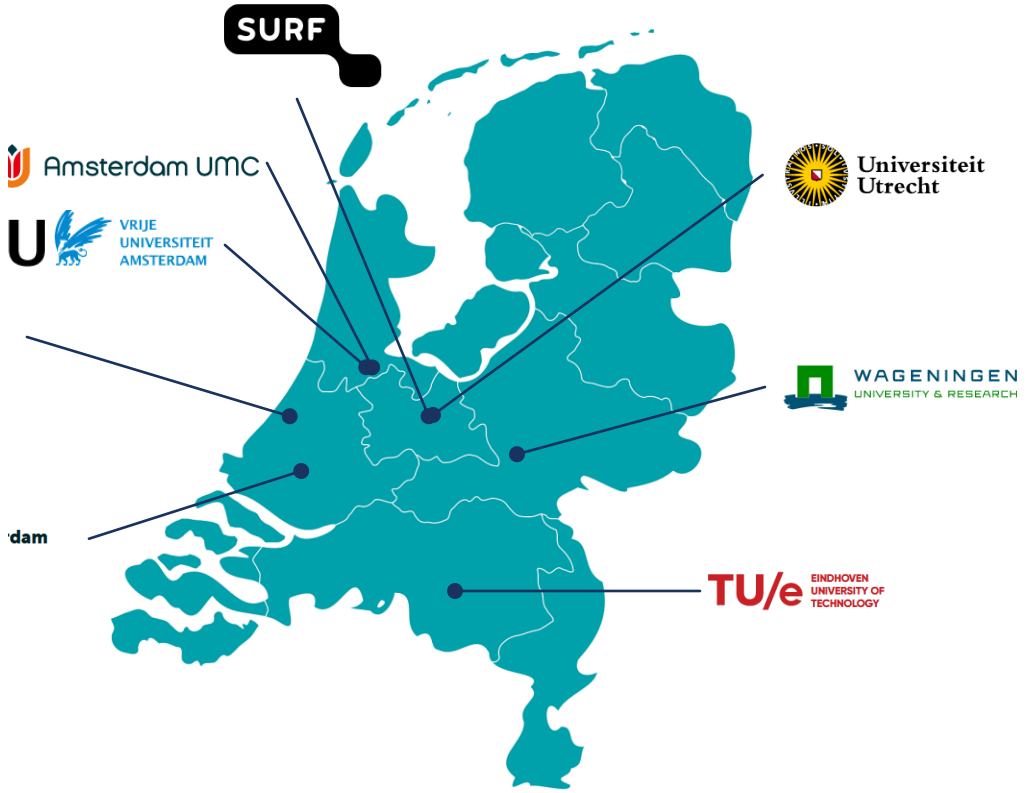
YODA CONSORTIUM

- 6 out of 13 major universities in the N
  & 1 UMC
  - Additionally, 3 uni's use iRODS based sc
- Partially funds and governs feature dev
  of the Yoda software
- Representation for hosting of Yoda by S

- 15,000+ users
- 3,5+ PB of data

North Carolina (US) is **3.36 times** as big as Netherlands

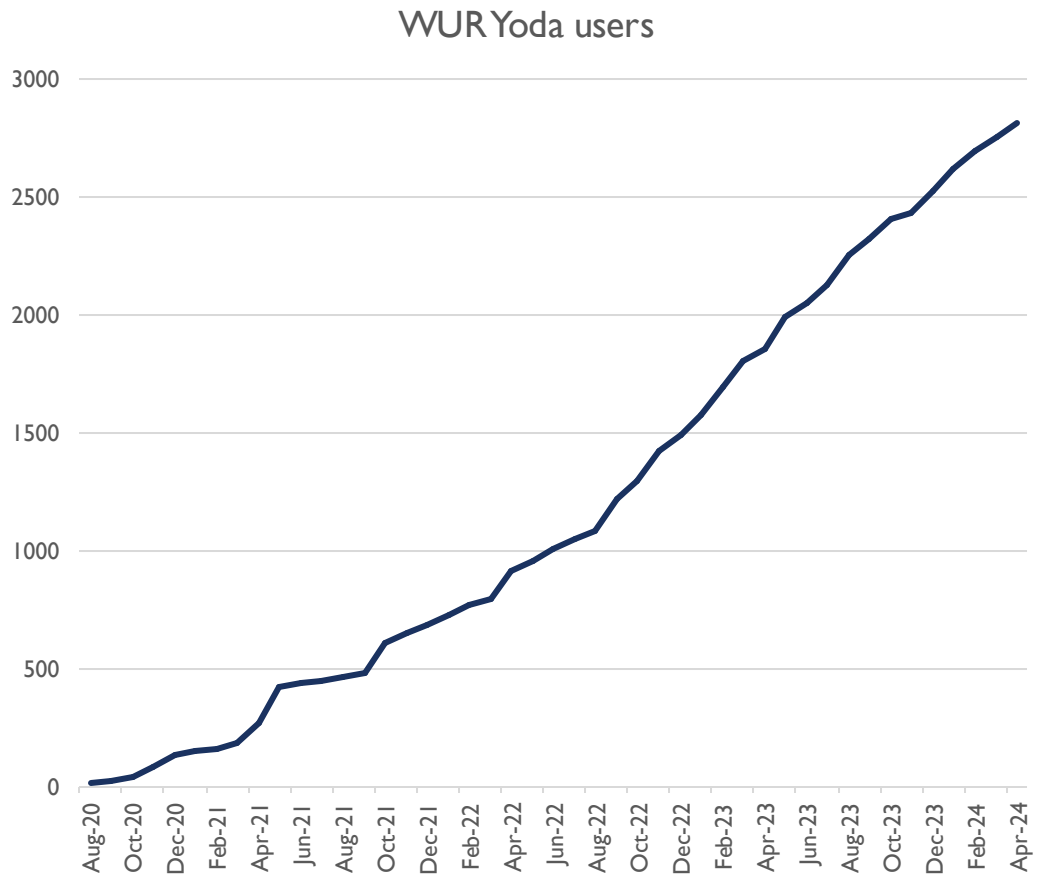
https://mapfight.xyz/map/nl/



BETTER TOOLS

- Growth of around 1000 users per year (within WUR)
- Similar growth in other consortium members
- User growth, without significant promotion or policy, indicates clear need
- Is part of their daily research process & environment

Yoda saves time and effort – it is practical RDM tooling

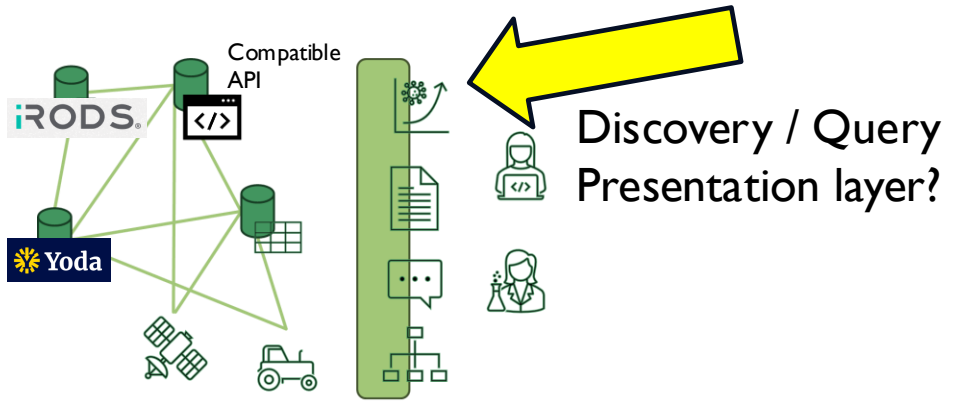


BETTER COLLABORATION

iRODS + Yoda + SURF SRAM

=

Dutch FAIR data train



RECAP & FUTURE OUTLOOK

- Dark data is a significant problem in –most- (Dutch) scientific institutions
- Yoda is a tool that supports researchers in their day-to-day data management activities, as well as enabling FAIRification of data
- Creates a large and active iRODS user base
- The Yoda consortium model creates a federated data network, without a single authority
- Expectation is that this model will work for most Dutch institutions
  - But also in a EU context (GaiaX, European Open Science Cloud and EU national activities)

THANKS!