DLCM: Comment être innovant en gestion des données de recherche

Pierre-Yves BURGI

Directeur adjoint DiSTIC et DLCM Project Director

Hugues Cazeaux

Responsable du pôle eResearch, DiSTIC et DLCM CTO

YARETA

OLOS SWISS MADE

DLCM | Mandated by swissuniversities

Agenda

1. What’s the problem ?
2. Design methodology: « under the hood »
3. What next ?
4. Questions

1. What’s the problem?

Four (good) reasons for practicing data governance (in research):

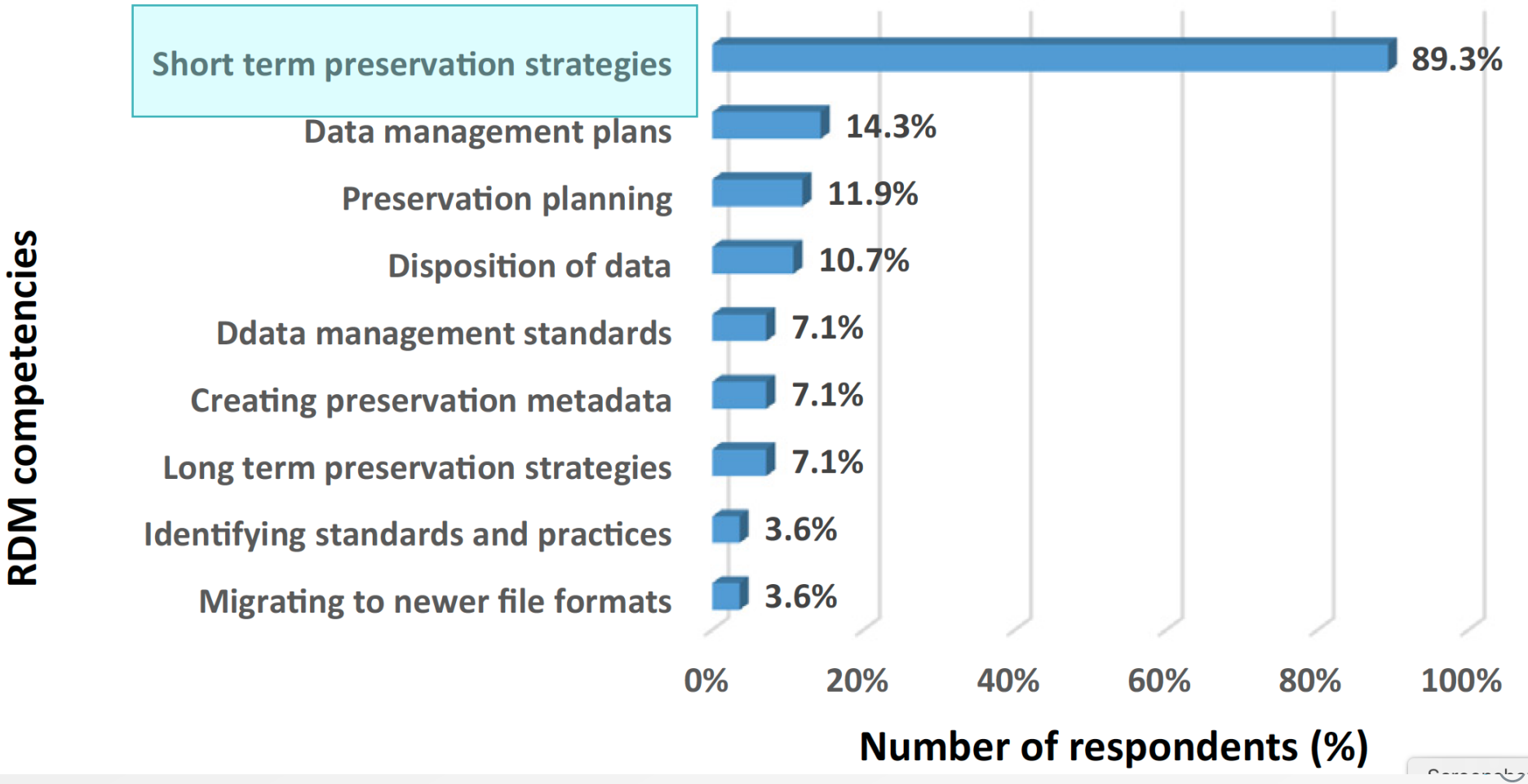
1) Avoid repeating experiments and thus reduces the costs
2) Contributes to global development by the study of phenomena other than those initially thought of
3) Guards against research fraud and improve reproducibility/replicability of results
4) Can be used by students in their learning activities

“**but,** researchers face challenges in organizing and sharing their data”

Factors affecting researchers' RDM (n = 84).

Challenge
Lack of policy frameworks
Lack of incentives
Lack of skills to create metadata
Lack of curation skills and training
Lack of storage network infrastructure
Lack of guidance and support
Lack of curation tools and software
Finding data produced by others
Lack of support from the university
Lack of incentives to share data
Most data is not trustworthy
Failure of re-users to cite my data
Prohibitive institutional policies
Obsolescence of technologies

Chawinga & Zinn, 2020



Chawinga & Zinn, 2020

Follow good practices !

Make Data FAIR DATA

- ✓ persistent identifier
- ✓ enriched metadata
- ✓ Searchable and findable online

- ✓ retrievable using standard communication protocols
- ✓ possibilily to define access rights

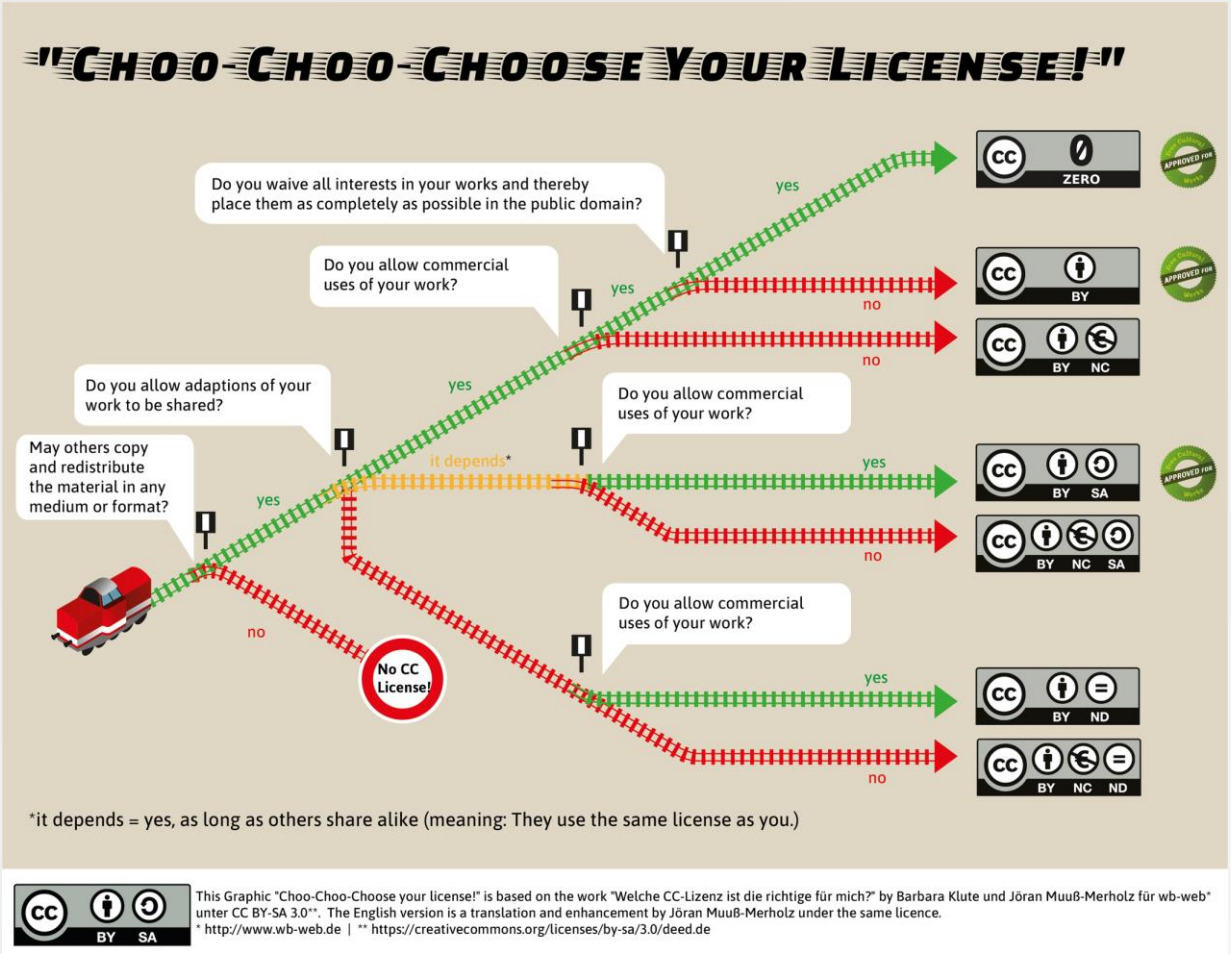
→ in a repository that ensures long-term preservation

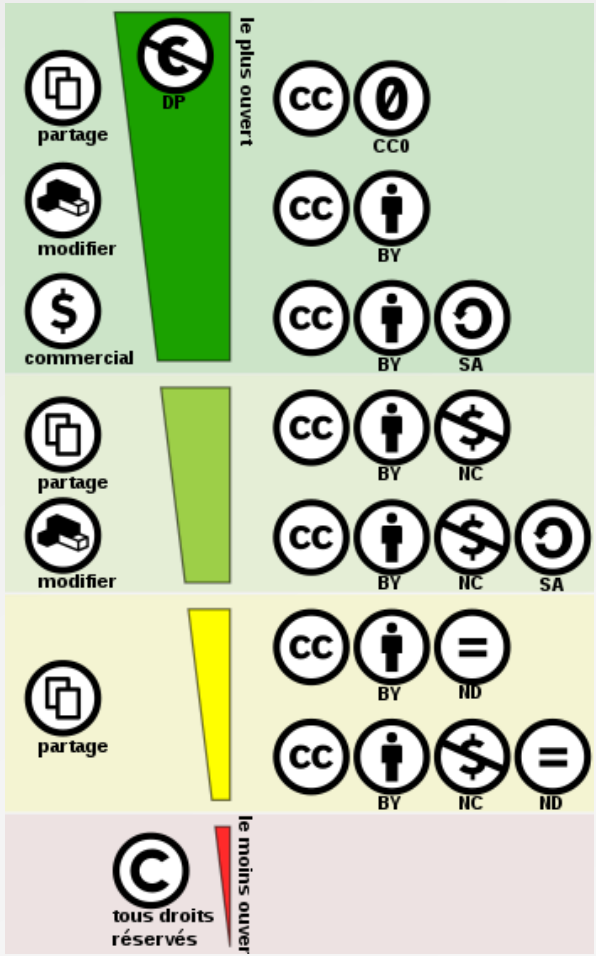
- ✓ standard formats
- ✓ controlled vocabulary to describe data

→ data will be compatible and combinable with others

- ✓ well-described & documented (e.g. in a README file)
- ✓ clear conditions to cite and reuse data (e.g. CC licenses)

→ allow data to be correctly interpreted and reused

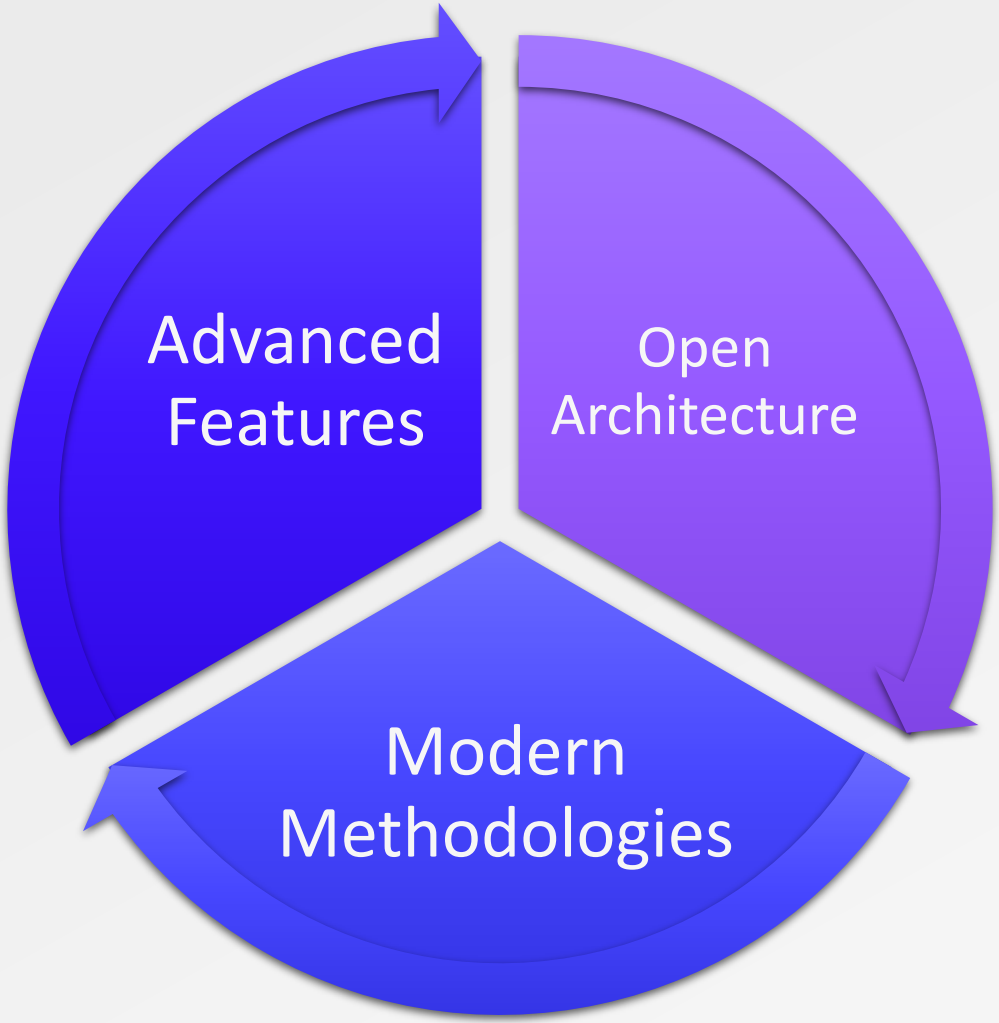




https://creativecommons.org/choose/

2. Design Approach: « *under the hood* »

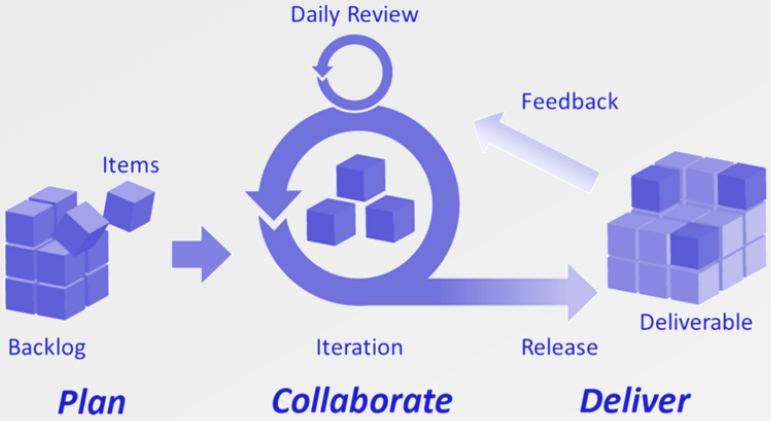
DLCM Project

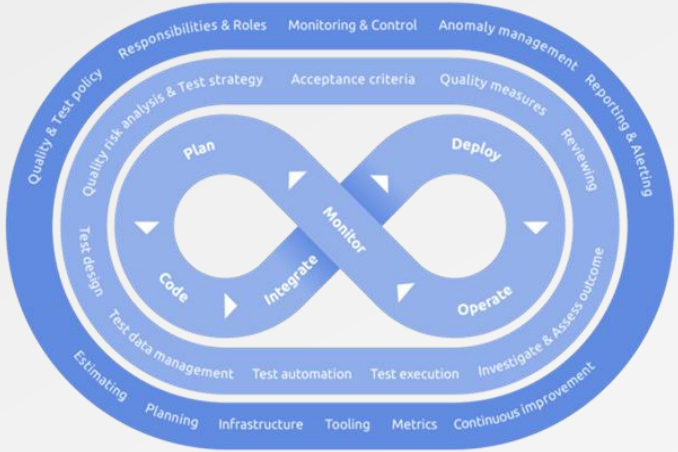

Advanced Features
Open Architecture
Modern Methodologies

2.1 Modern Methodologies

DLCM Project

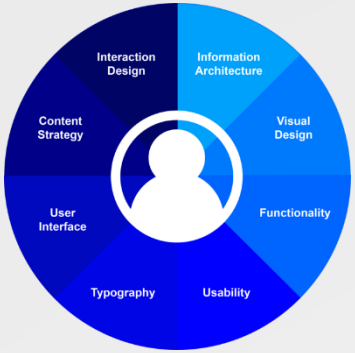
Software Agile Development





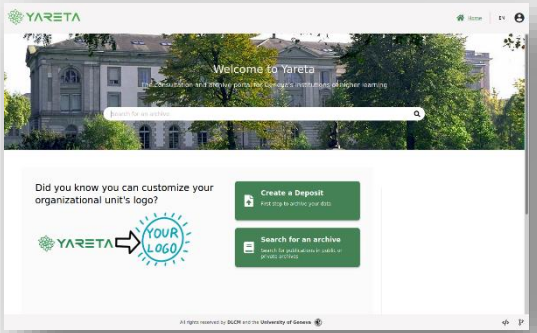
DevOps

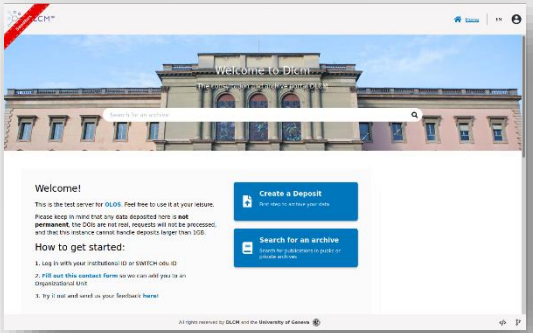
UX & UI

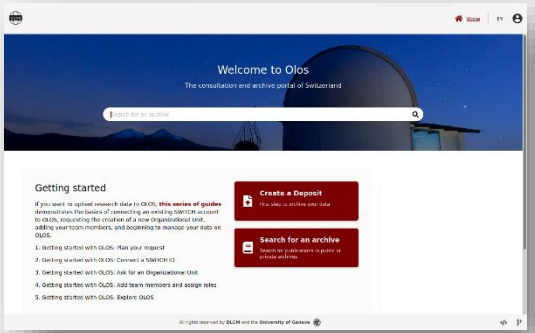

Interaction Design
Information Architecture
Content Strategy
Visual Design
User Interface
Functionality
Typography
Usability


USER
TEST

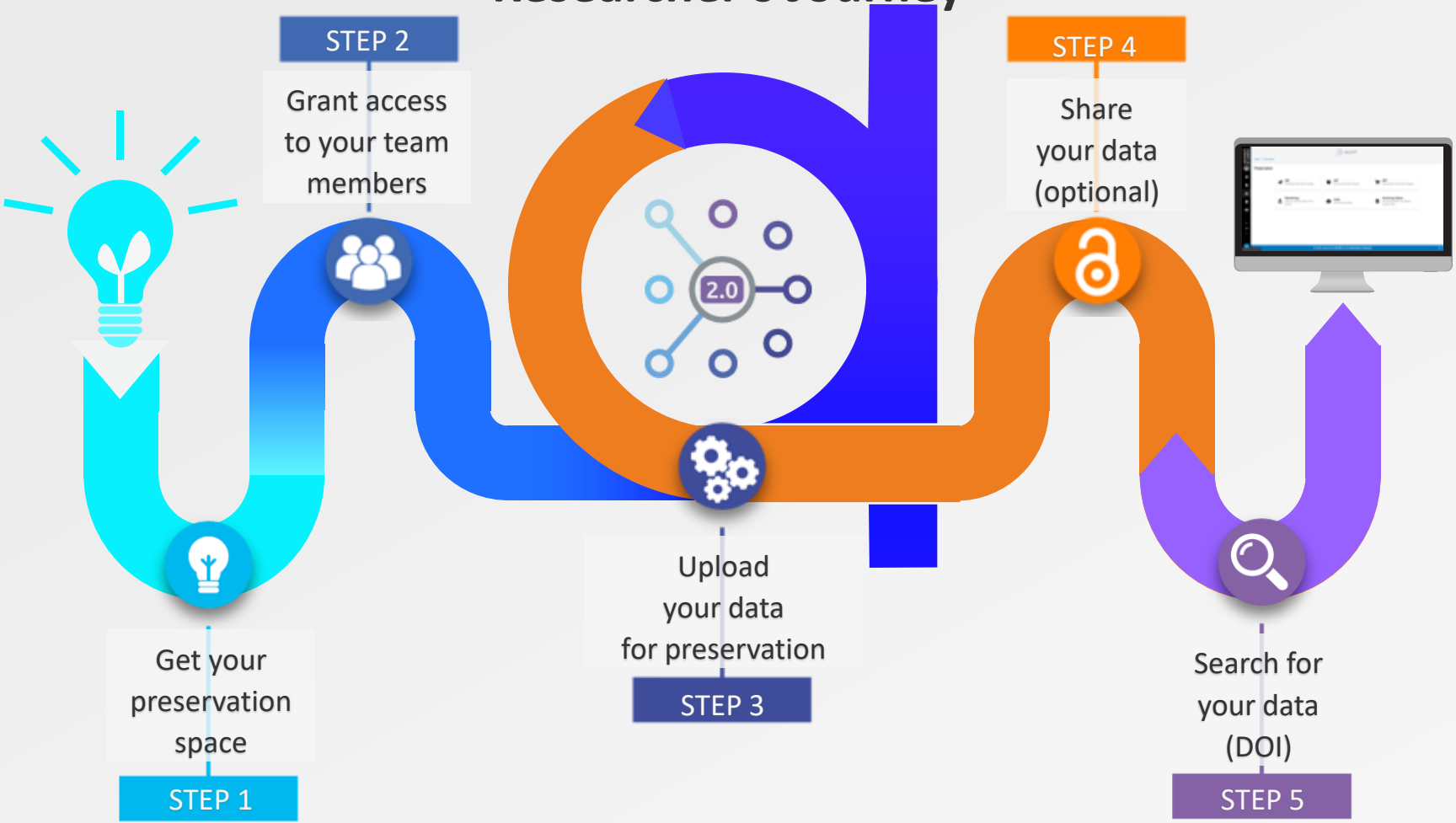

USABILITY
EXPERIENCE


YARETA
Welcome to Yareta
Did you know you can customize your organizational unit's logo?
YOUR LOGO
Create a Deposit
Search for an archive


Welcome to Dicrit
Welcome!
How to get started:
Create a Deposit
Search for an archive


Welcome to Olos
The consultation and archive portal of Switzerland
Getting started
Create a Deposit
Search for an archive

Researcher's Journey

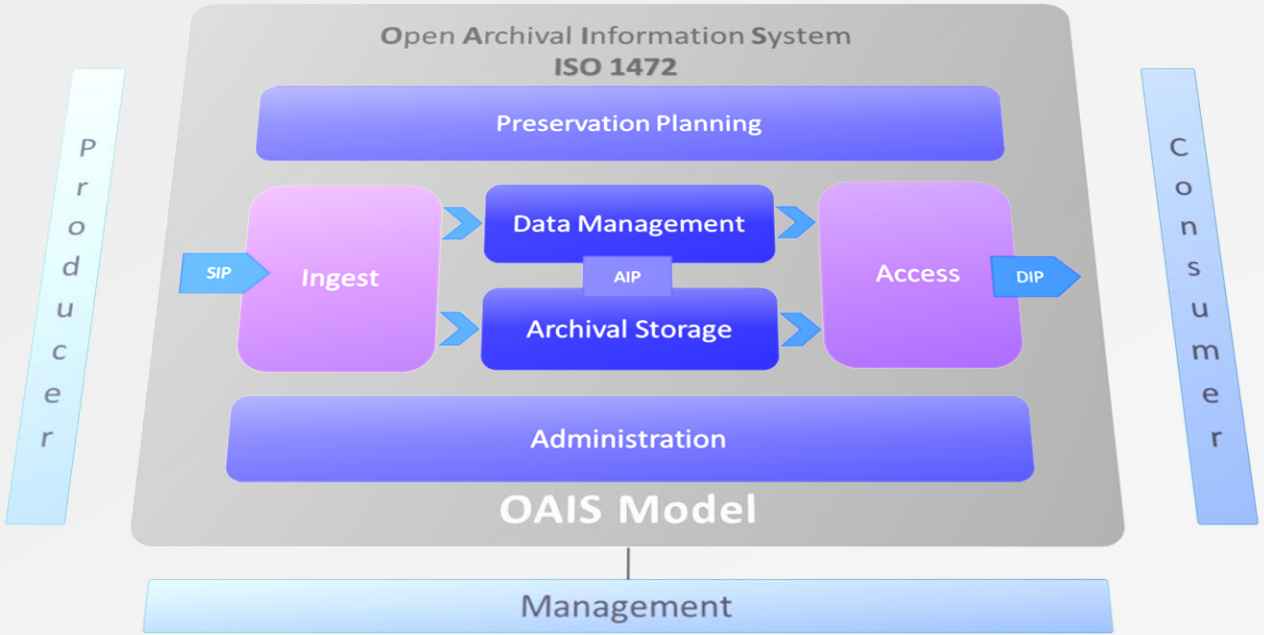


2.2 Open Architecture

DLCM Project

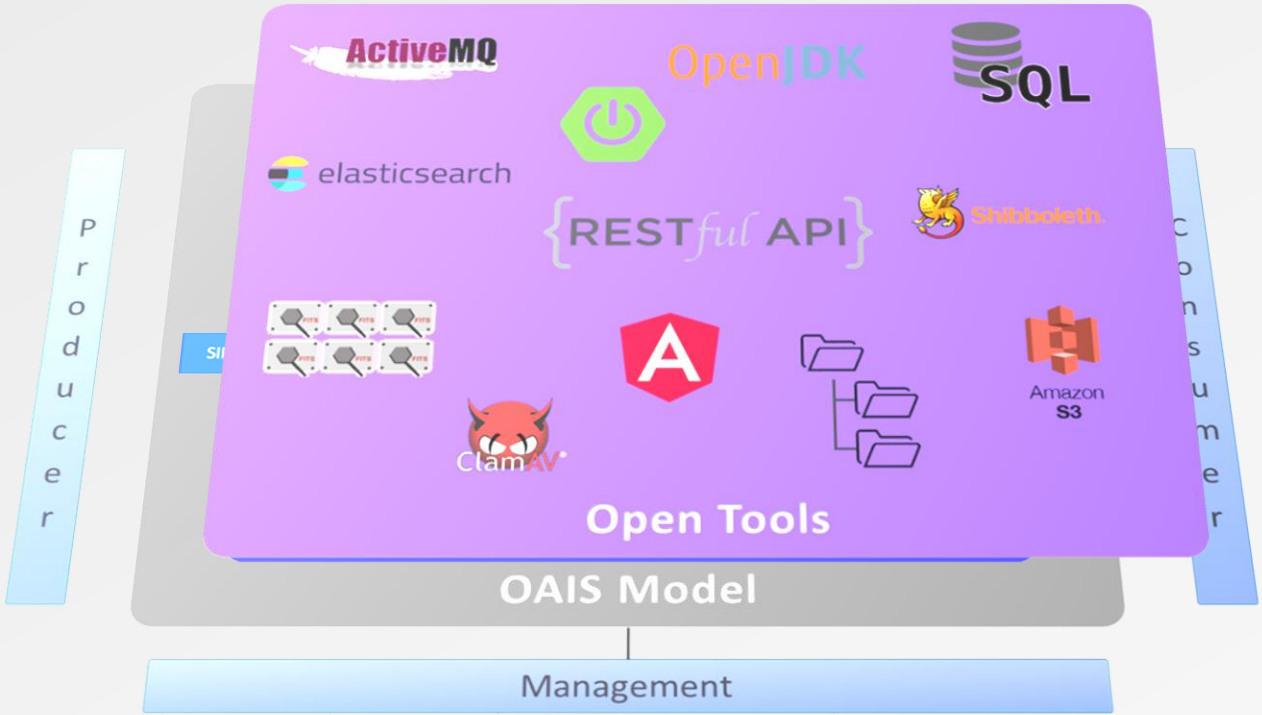
Innovative Design

Image by Painting Valley CC BY-NC 4.0



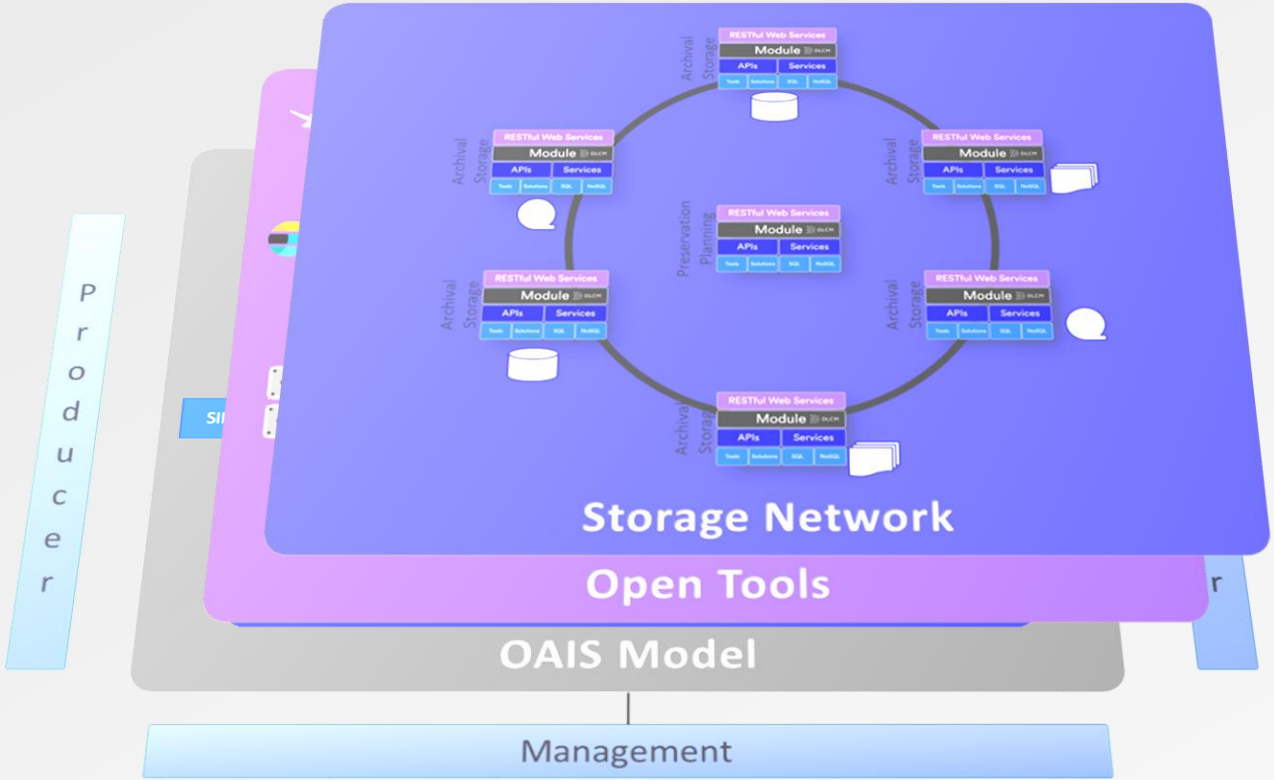
Innovative Design

Image by Painting Valley CC BY-NC 4.0



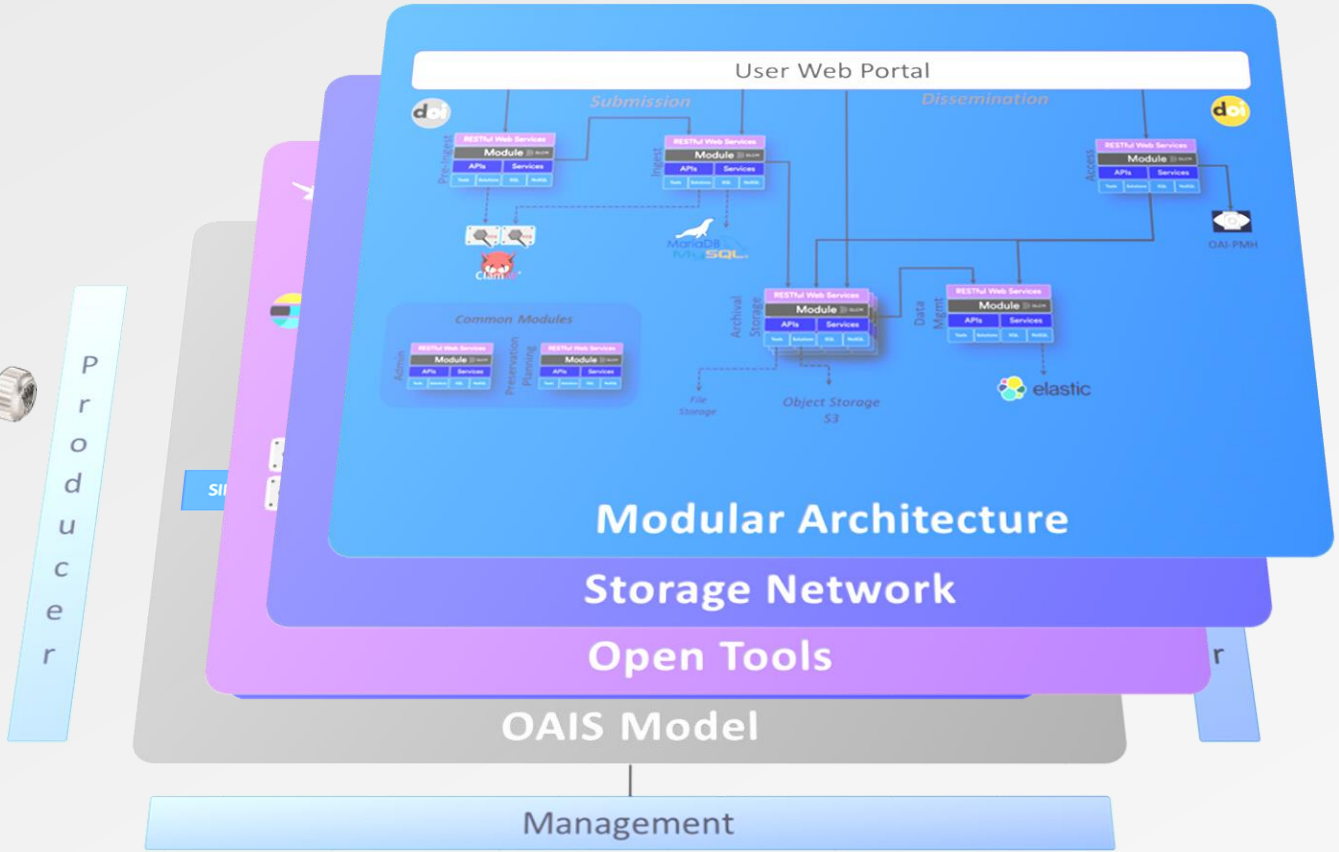
Innovative Design

Image by Painting Valley CC BY-NC 4.0



Innovative Design

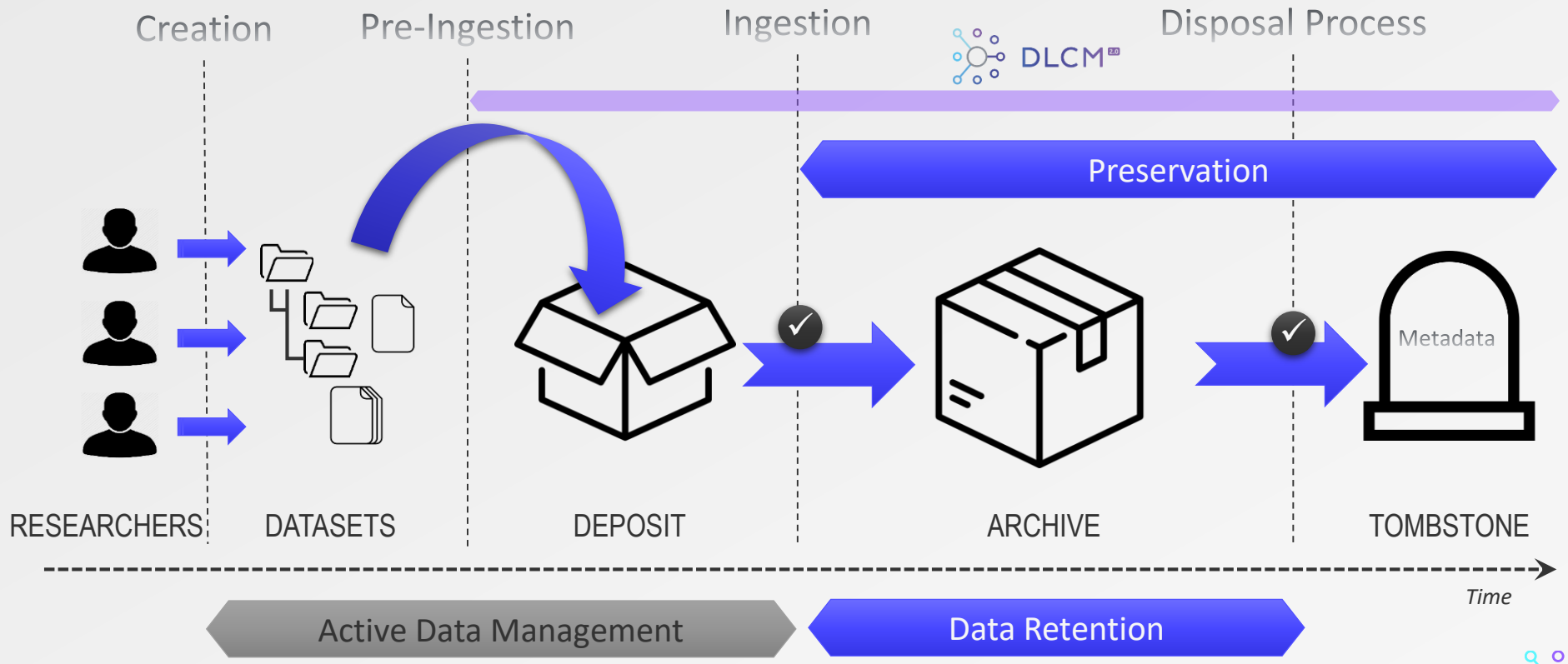
Image by Painting Valley CC BY-NC 4.0



2.3 Advanced Features

DLCM Project

What is Preserving Data?



Based on the OAIS Model (Open Archival Information System, ISO 14721)

Backup vs. Archiving

ARCHIVE

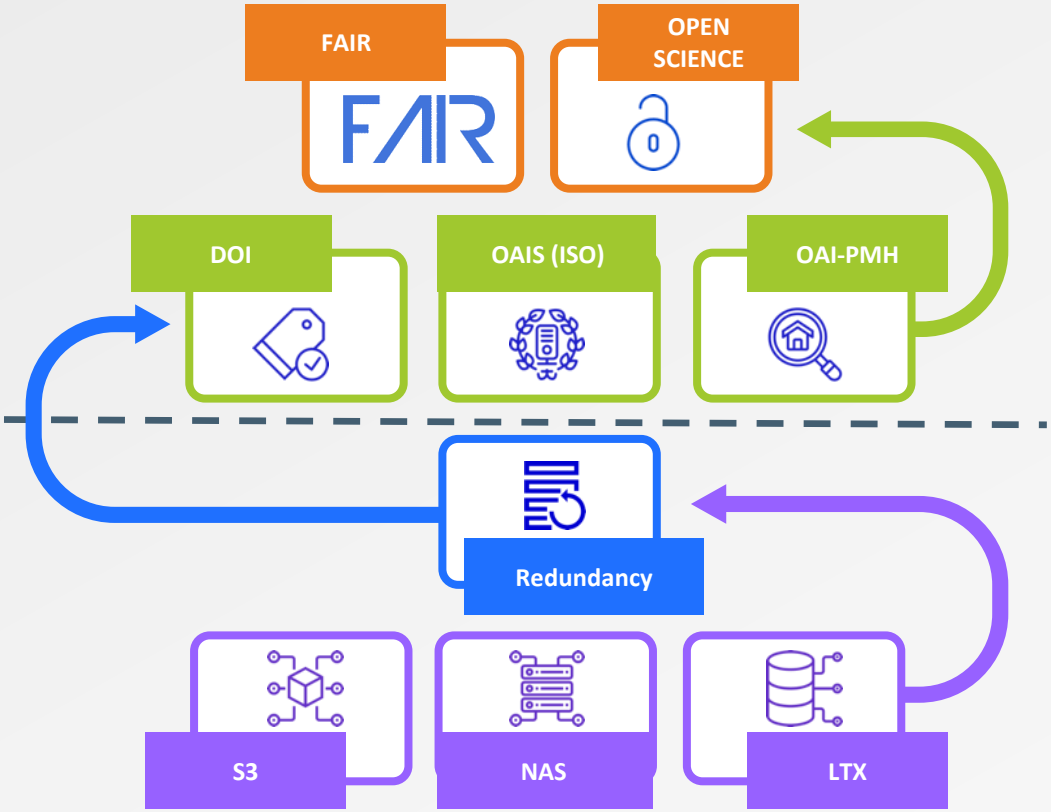
Preserve information as required by regulations and institutional policies

- Auditable
- Follows a life cycle
- Self-described
- Data integrity

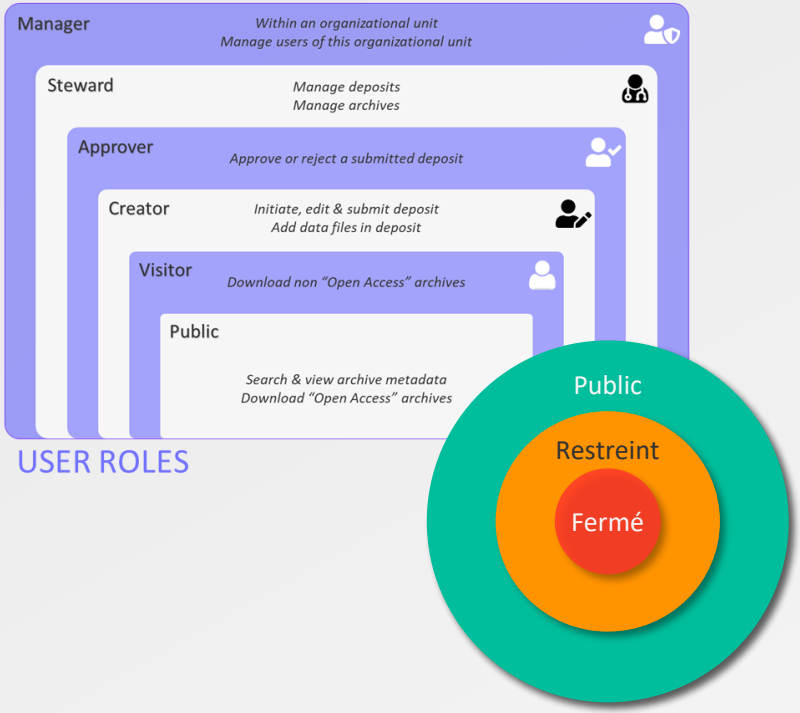
BACKUP

Insurance policy against unforeseen system failures

- Incremental
- Multiple snapshots
- Retained on short periods of time
- Not searchable



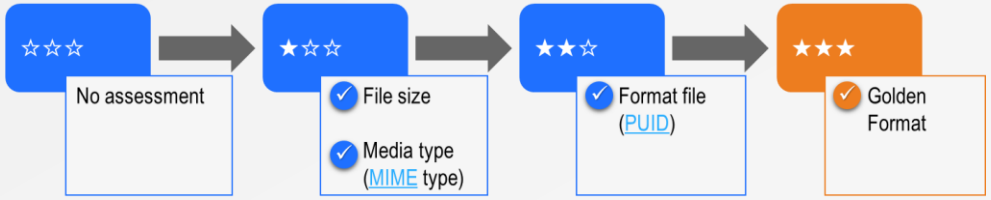
Features & Best Practices



Access Level

Data Sensitivity (Data tag) — Compatibility	*Public*	*Restricted*	*Closed*
Not defined	✓	✓	✓
Public	✓	✓	✓
Controlled public	×	✓	✓
Accountable	×	×	✓
More accountable	×	×	✓
Fully accountable	×	×	✓
Maximum restricted	×	×	✓

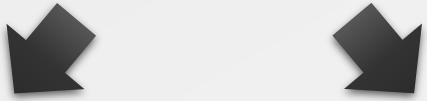
Compliance Level



PUID = PRONOM Unique Identifiers
The National Archive, UK
https://www.nationalarchives.gov.uk/PRONOM

Golden Format = recommended format for long term preservation
Library of Congress, USA
https://www.loc.gov/preservation/resources/rfs/

DLCM 2.0



YARETA

https://yareta.unige.ch/

https://olos.swiss/portal/

3. What Next ?

DLCM Project

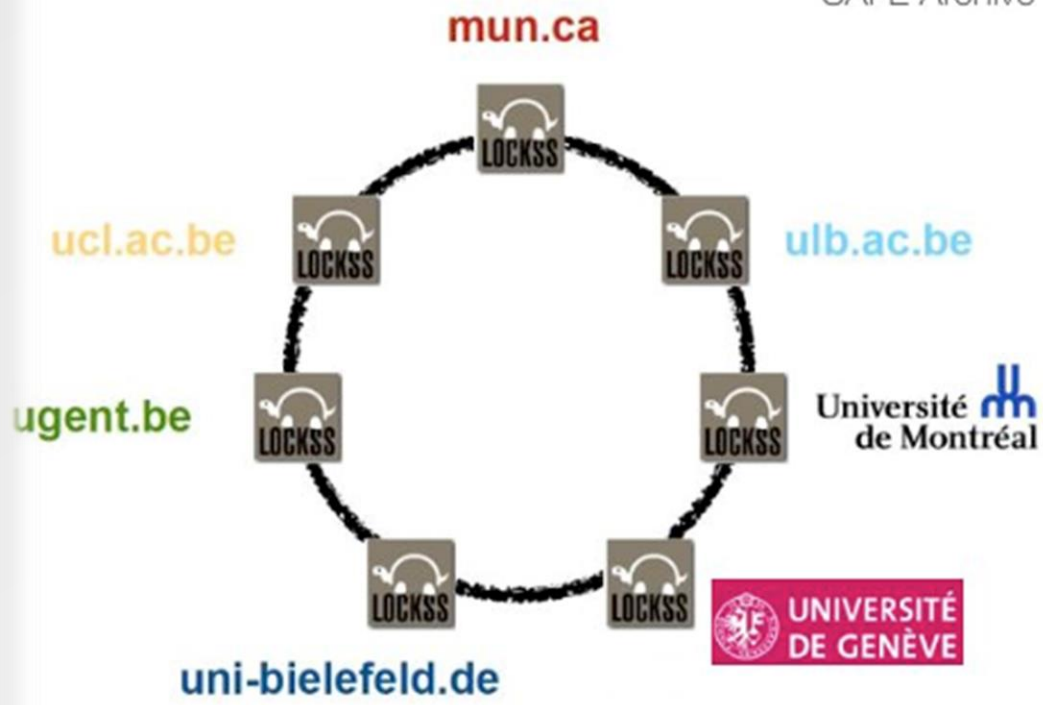
Adding Redundancy: Byzantine Agreement Problem

- First studied by Lamport et al. in 1982
- Distributed system with *n* replicas, which could tolerate *f* faults, $n = 3f + 1$
- LOCKSS* $n >> 3f + 1$ → from elections to opinion polls
- LOCKSS' design principles:
  - Cheap storage (unreliable), diverse technologies
  - Hash comparisons (and no stored checksums)
  - No long-term secret for identities (just for polling → sha1 enough)
  - Use inertia - running very slowly limits attacks and damage
  - Integrate intrusion detection intrinsically and assume a strong adversary

* Lot of Copies Keep Stuff Safe

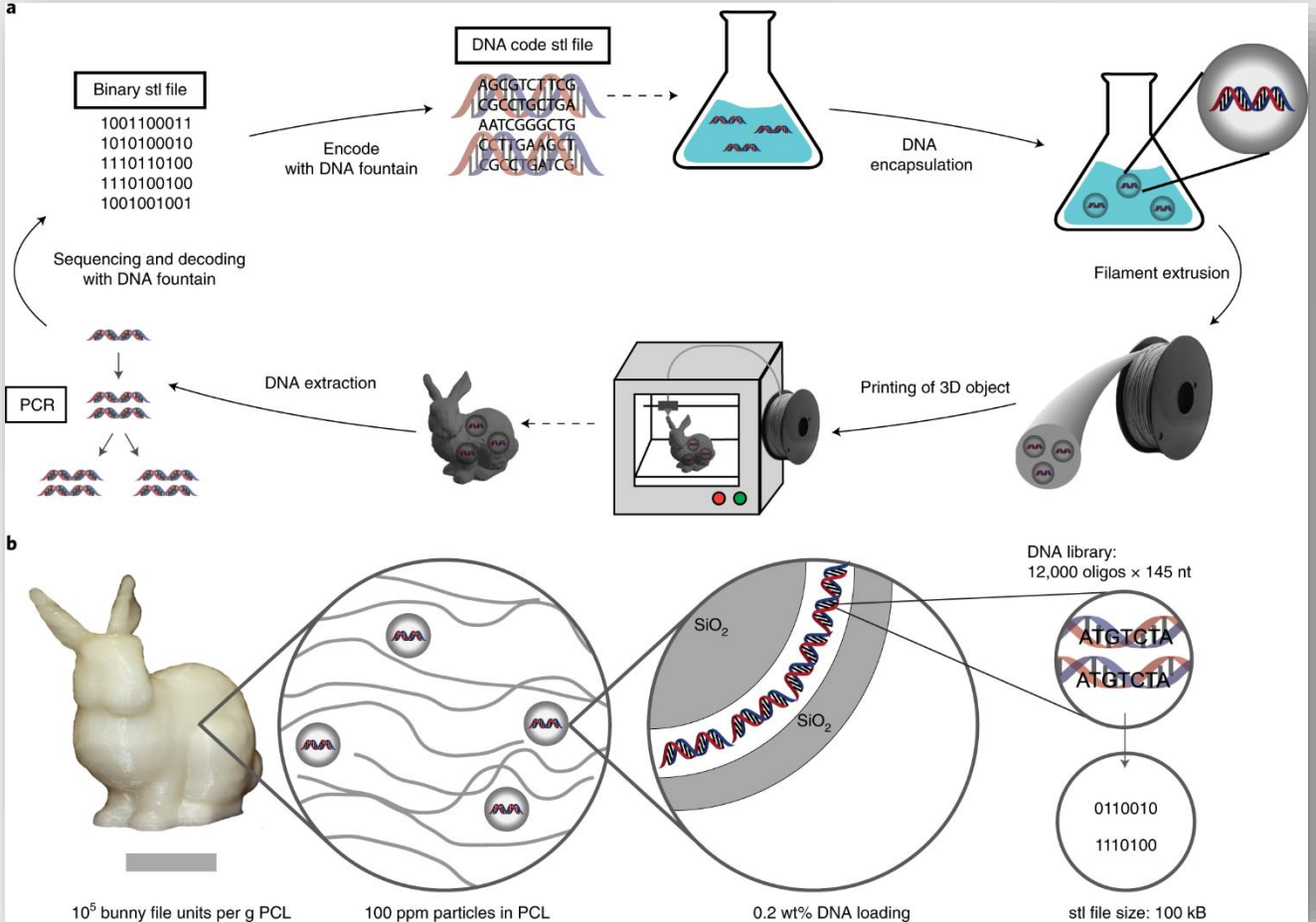
SAFE PLN

SAFE Archive FEderation



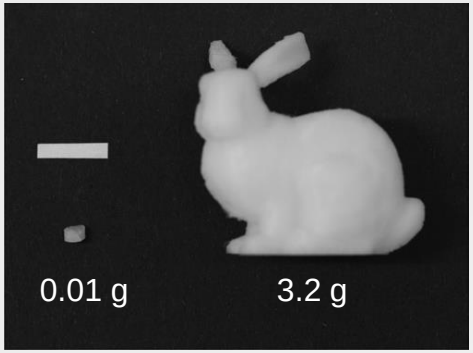
Used by UNIGE to archive electronic thesis

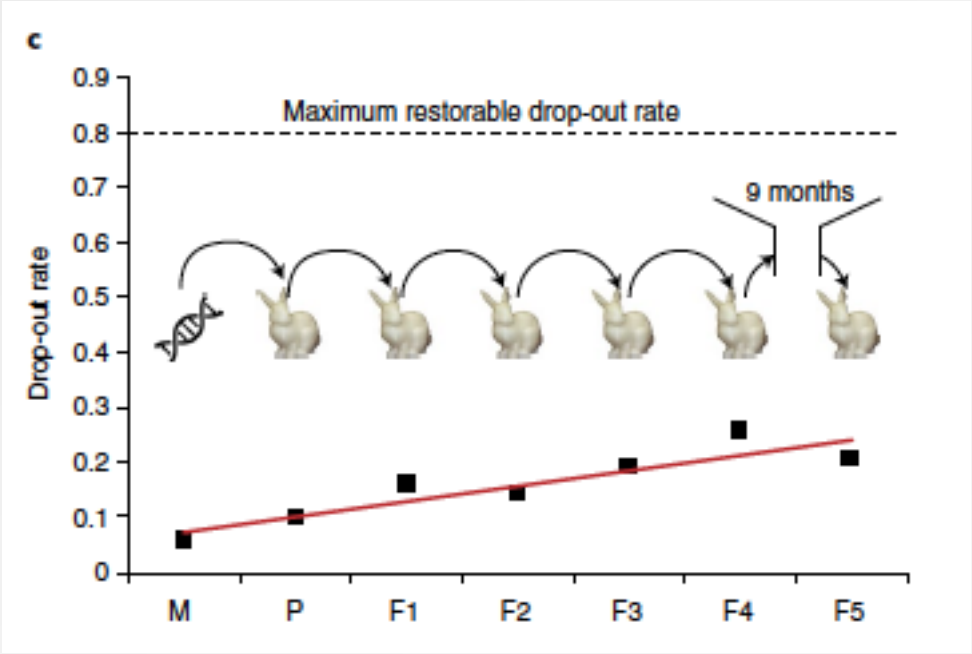
A DNA-of-things storage architecture to create materials with embedded memory



J Koch et al. (2019) Nature Biotechnology

… some numbers





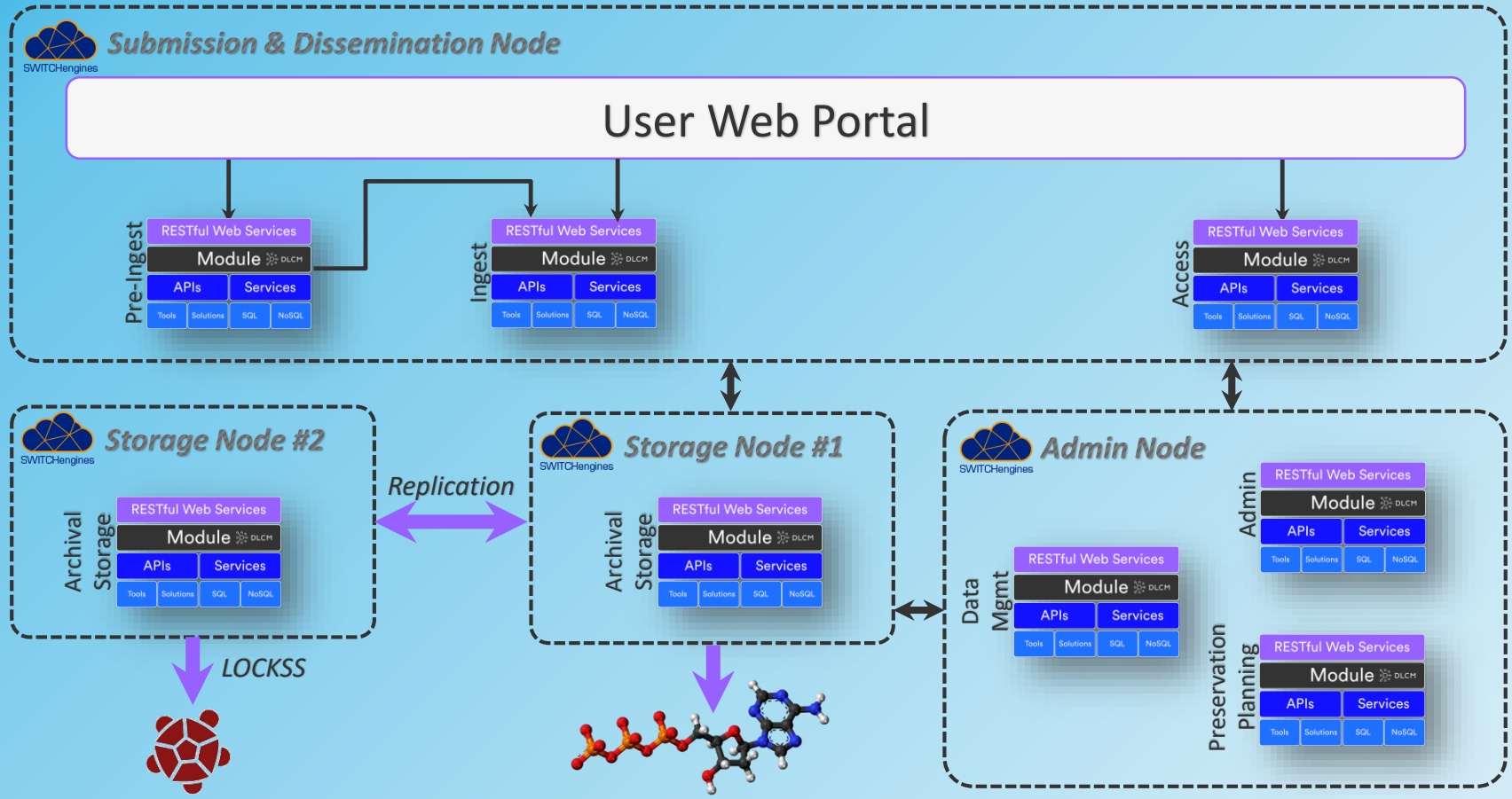
J Koch et al. (2019) Nature Biotechnology

0.01 g → 14’000 copies of the encoded file + redundancy

DNA storage capacity: 215 PB per gram

Half-life of several thousands years

Future Cloud Deployment



Questions?

info@dlcm.ch

Go to https://sandbox.dlcm.ch/