What makes data last?

DANS Open Day
Open data, open science

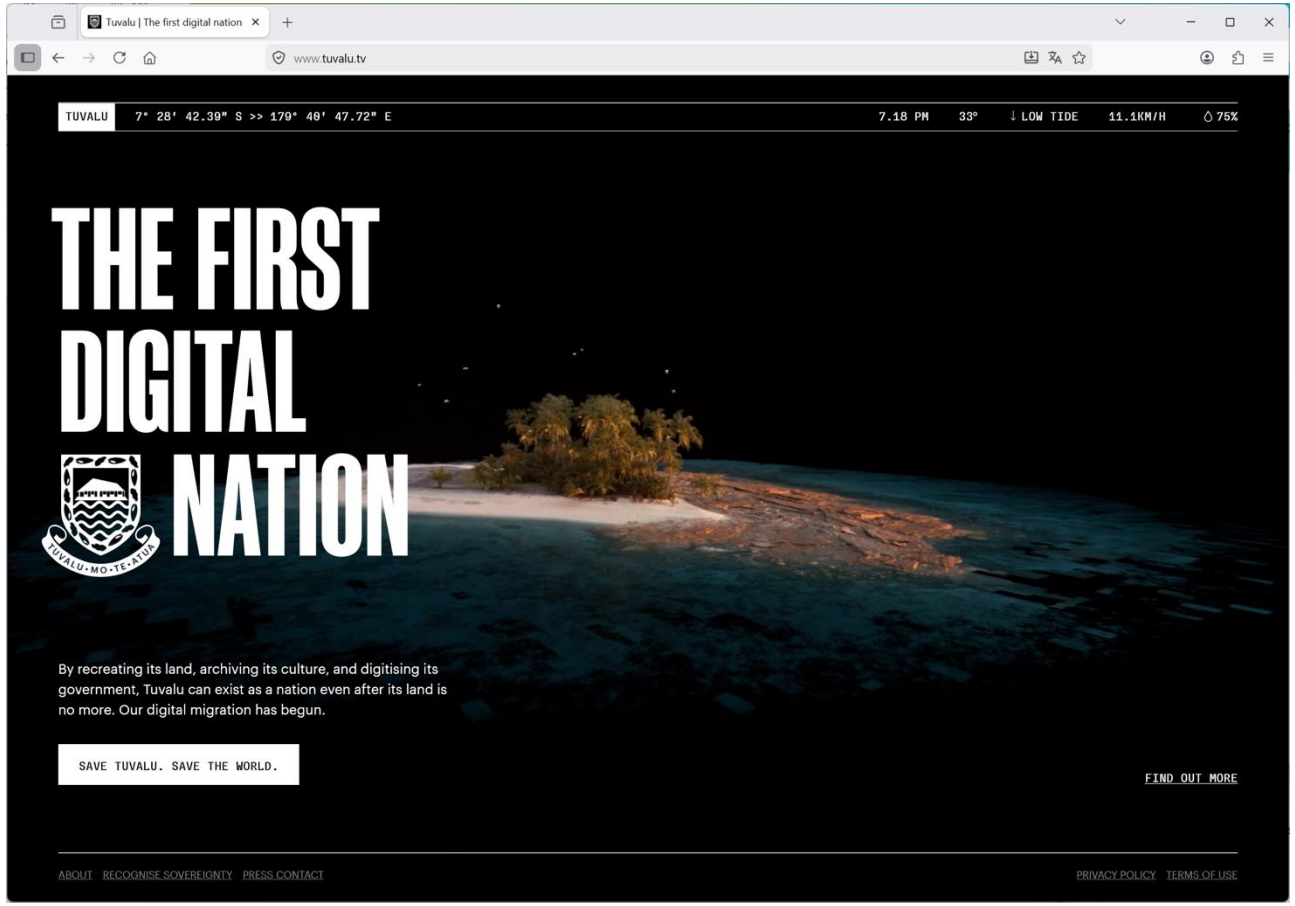
Curation, formats and guidelines

Valentijn Gilissen | Data Processing Team Leader /
Preservation Officer / Data Steward / Senior Datamanager

11 June 2026

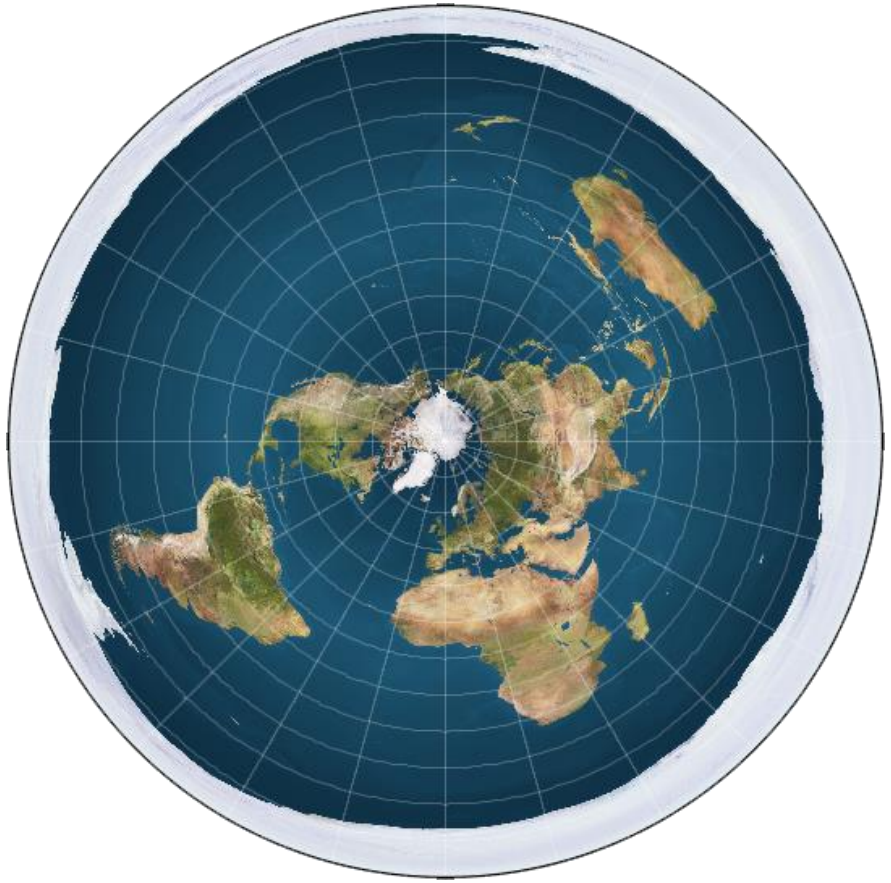
DANS

What makes our world last?



Left: Simon Kofe presenting the keynote 'Digital Nation' at iPRES 2025, Wellington NZ, 4/11/2025
Right: https://www.tuvalu.tv/

World of... science?!?

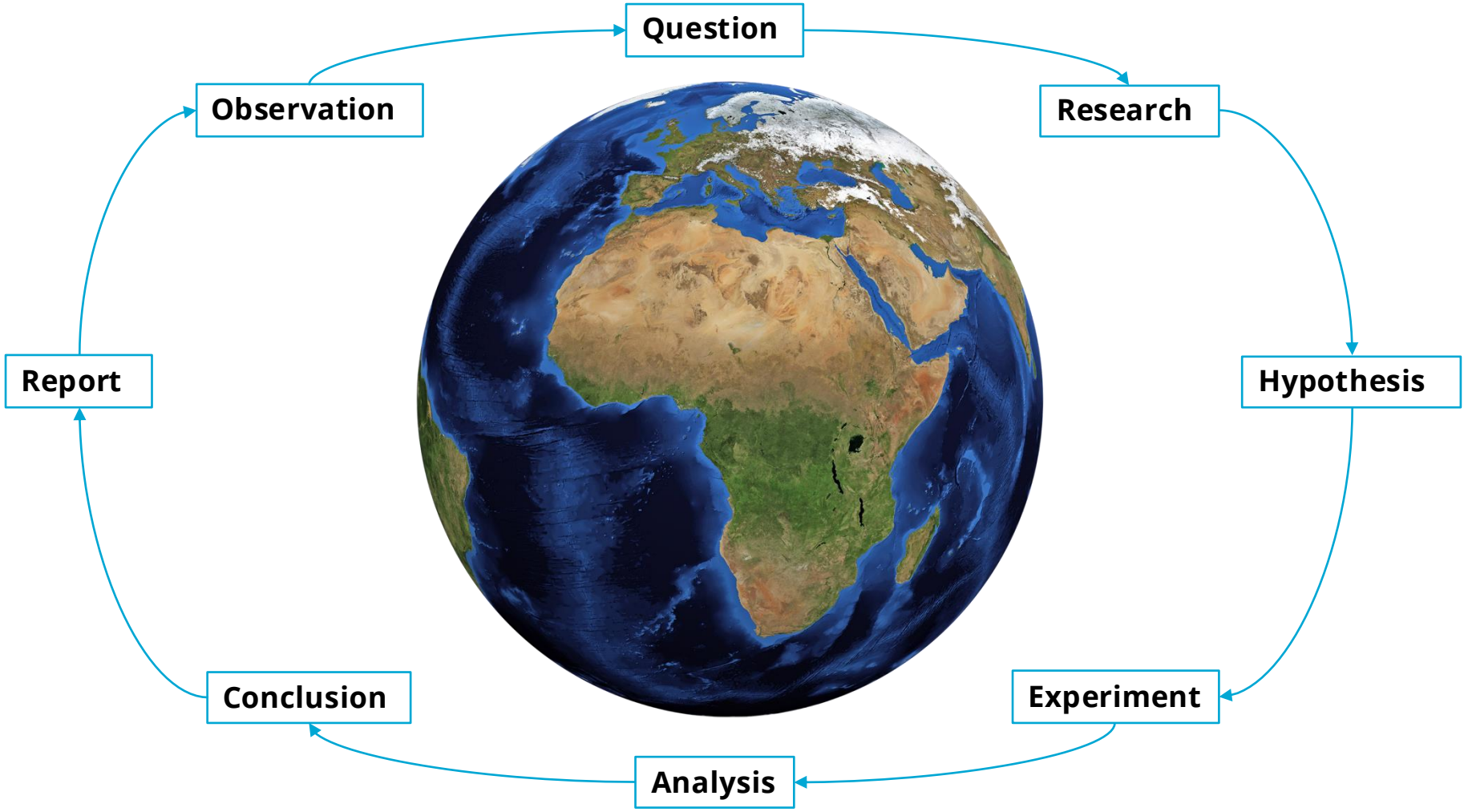


Trekky0623 at English Wikipedia ("I made this map myself"), Public domain, via Wikimedia Commons

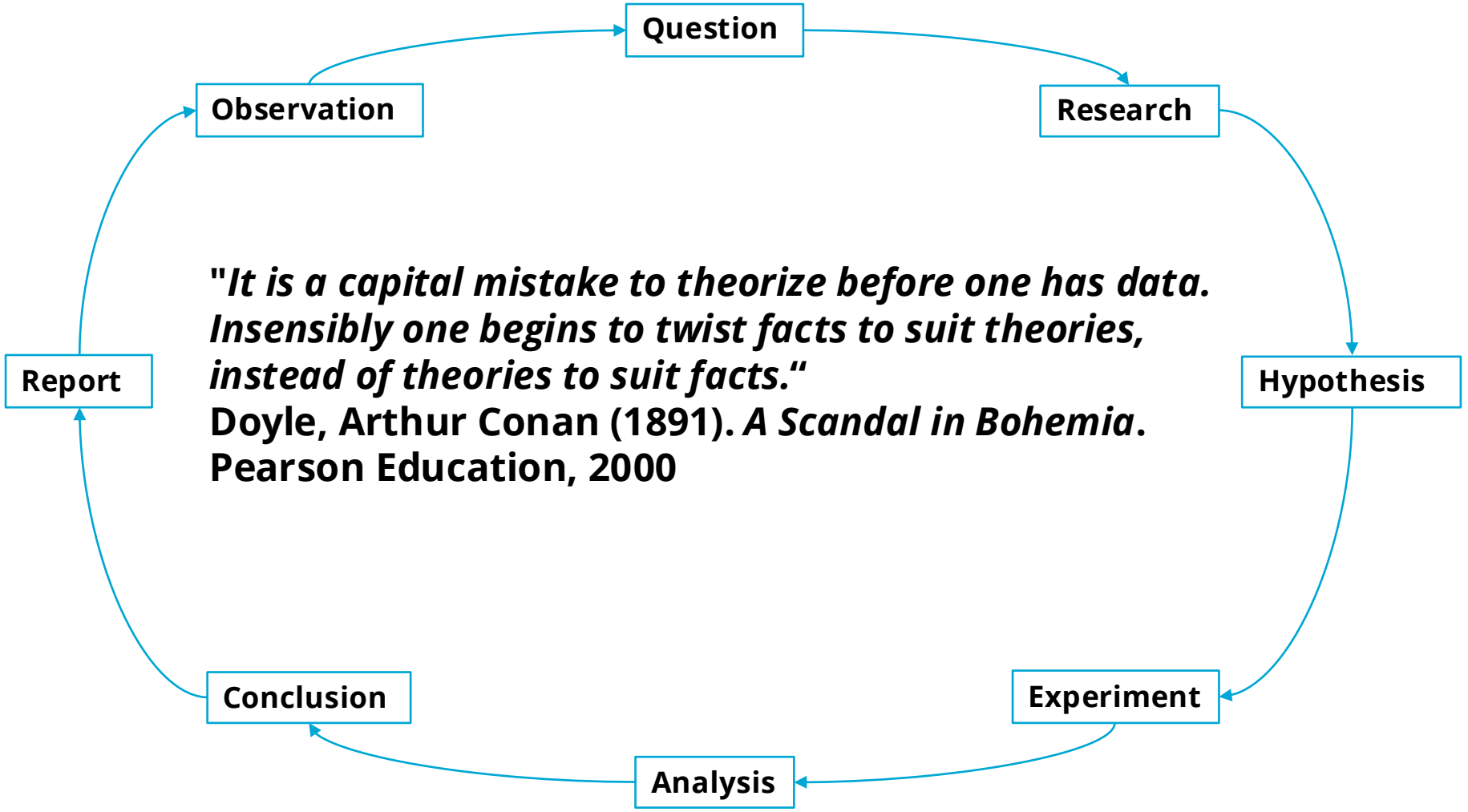
World of science!

Image by qimono via Pixabay

The Scientific Method

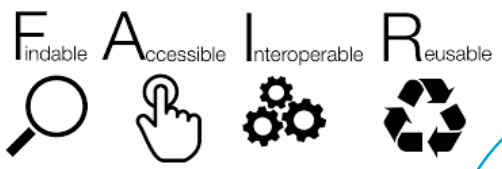


The Scientific Method (and Sherlock Holmes)



"***It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.***"
Doyle, Arthur Conan (1891). *A Scandal in Bohemia*. Pearson Education, 2000

The Scientific Method (and Sherlock Holmes)



Digital Preservation

Data reuse!

Question

Data reuse!

Observation

Research

Data reuse!
Data citation!

New data!

Report

Hypothesis

"*It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*"
Doyle, Arthur Conan (1891). *A Scandal in Bohemia*. Pearson Education, 2000

Data reuse!

New data!
Data citation!

Experiment

Conclusion

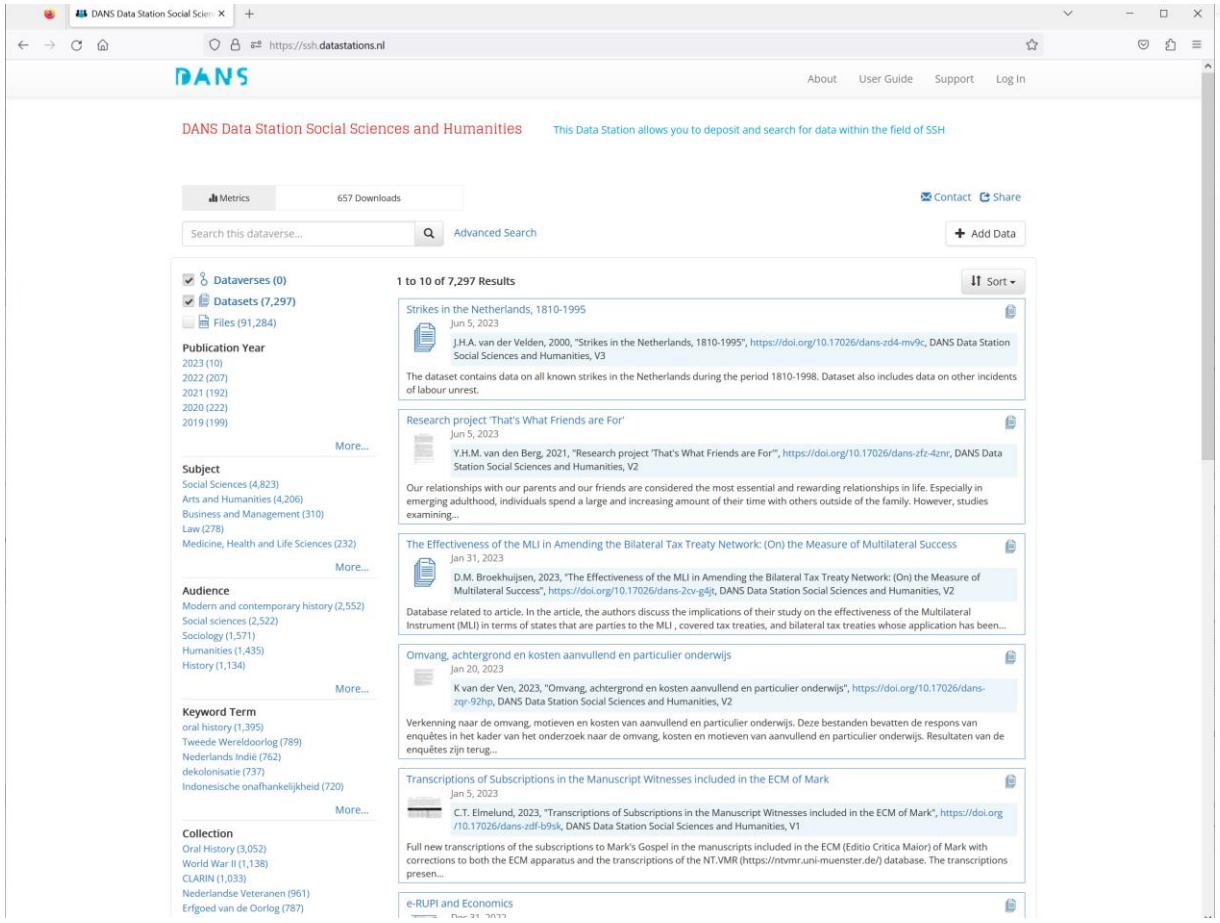
New raw data!

Analysis

New data!

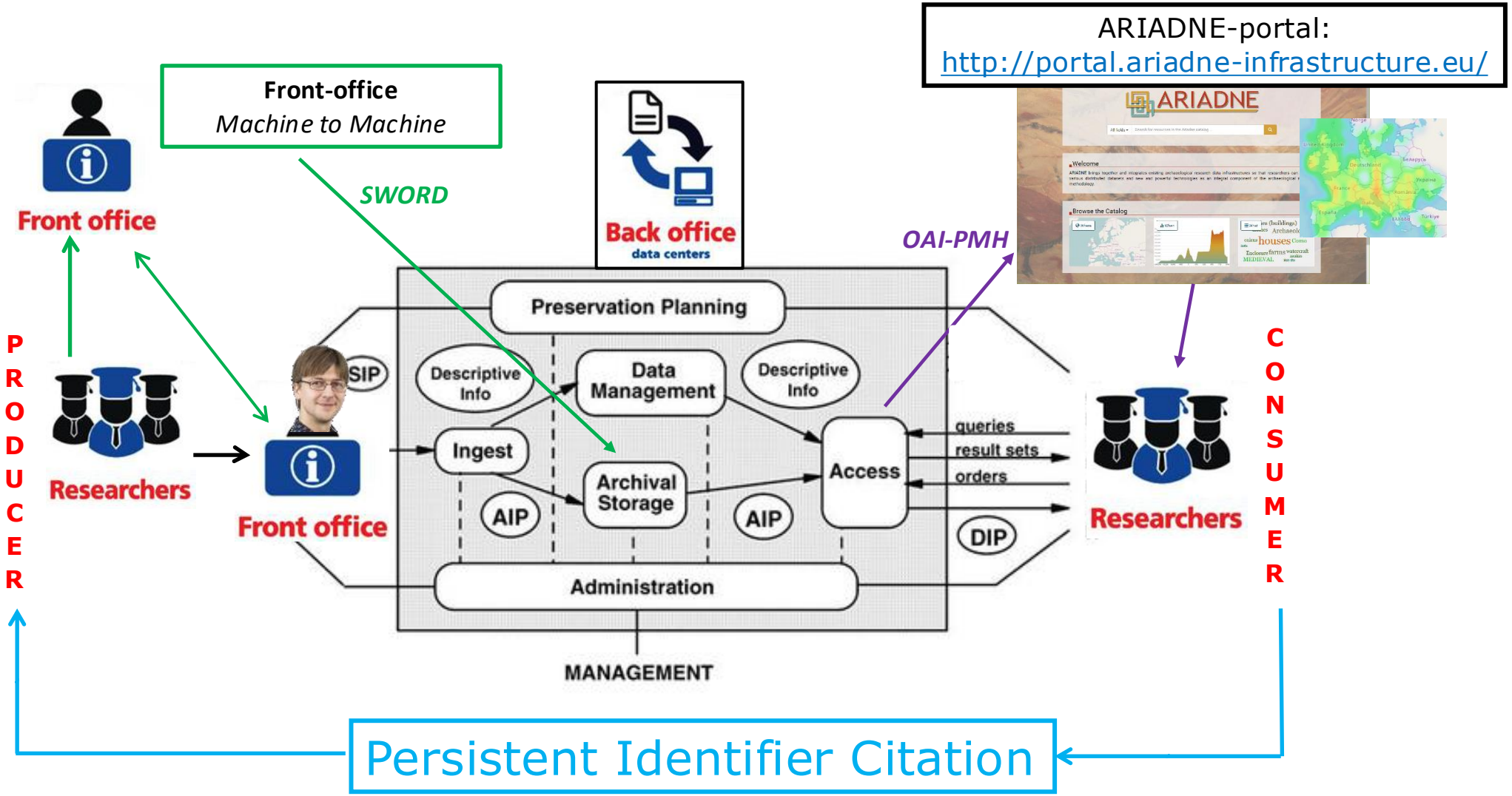
Research Data Management

DANS Data Stations



https://ssh.datastations.nl/; https://archaeology.datastations.nl/; https://lifesciences.datastations.nl/; https://phys-techsciences.datastations.nl/

The magic(?) box of the repository



Dataset curation

2023-2025

Levels of Curation

Select all relevant types from:

A. Content distributed as deposited

B. Basic curation *- e.g. brief checking, addition of basic metadata or documentation*

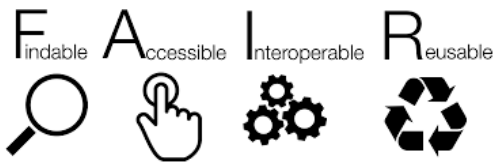
C. Enhanced curation *- e.g. conversion to new formats during ingest, enhancement of documentation and metadata*

D. Data-level curation *- as in C above, but with additional editing of deposited data*

- Inventorise deposit

Basic curation

- Check if the dataset is valid and complete
- Check metadata, edit/modify where necessary
- Check for privacy sensitive data
- Publish dataset



Enhanced curation

- Migrate files to Preferred Formats
- Modify directory structure
- Make a new version

Conform *Data Stations Policy*; *DANS Selection Policy*; *Data Processing Manual*

Dataset curation

2026-2028

Levels of Curation and Preservation

Select all relevant types from:

Z. Level Zero. Content distributed as deposited.

D. Deposit Compliance *with defined criteria, e.g. data formats, metadata elements, and compliance with legal and ethical norms*

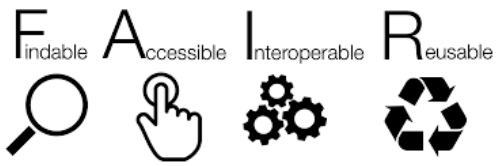
C. Initial Curation *to meet defined criteria. This may include, e.g., the correction or enhancement of metadata and/or data content, or the creation of dissemination formats*

A. Active preservation. *The repository takes long-term responsibility for ensuring that the data and metadata can be understood and rendered as required by the designated community for reuse.*

- Inventorise deposit

Initial Curation

- Check if the dataset is valid and complete
- Check metadata, edit/modify where necessary
- Check for privacy sensitive data
- Publish dataset

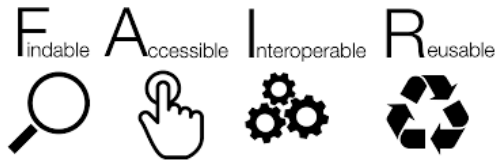


Active Preservation

- Migrate files to Preferred Formats
- Modify directory structure
- Make a new version

Conform *Data Stations Policy*; *DANS Selection Policy*; *Data Processing Manual*

Dataset curation



- **Validation -** DANS evaluates deposits on suitability for the Designated Community for the Data Station. DANS reserves the right to reject deposits submitted for evaluation if such datasets do not meet the criteria for acceptance.
- ***Submission Package* -** DANS always retains a copy of the data exactly as submitted by the depositor.
- **Initial Version -** The submitted data and curated metadata become the first version of a submitted dataset and form the basis for any preservation copies made of the dataset.
- **Enhancements -** DANS may make changes to the metadata and add conversion files to align them with community expectations, increase *FAIR* quality, and improve the long-term sustainability of the dataset. Changes to the data, and any changes to the metadata after the initial published version, will result in a new version of the dataset.
- **Publication -** If the dataset meets all criteria and after completion of the curatorial review, DANS will publish the dataset on behalf of the depositor and provide it with a *persistent identifier* (PID).

Conform *Data Stations Policy*; *DANS Selection Policy*; *Data Processing Manual*

Open Access metadata

Trustworthy repository

A

Standardised exchange protocols

Codebooks

Embargo when needed

Open Access when possible

Clear folder structure

F

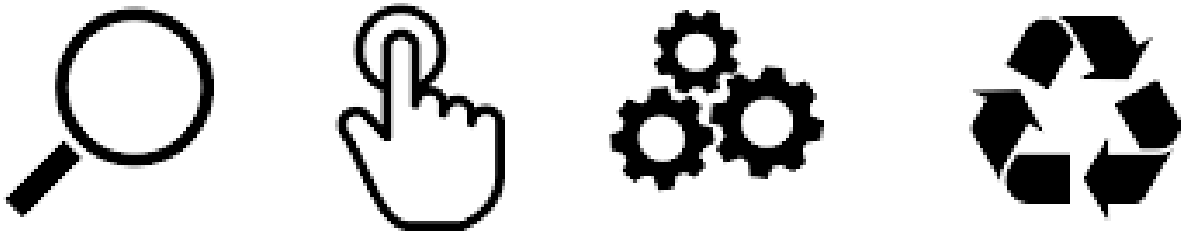
Data citation

Access and licence information

Licence

Clear names for files and folders

Findable Accessible Interoperable Reusable



Persistent Author Identifiers

Authenticity

Provenance

R

Metadata in scheme

Version control

Preferred formats

Common file formats

Workflows

Support metadata mapping

Data standards

Vocabularies

I

APIs

PARTHENOS, Hollander, H., Morselli, F., Uiterwaal, F., Admiraal, F., Trippel, T., & Di Giorgio, S. (2019). PARTHENOS Guidelines to FAIRify data management and make data reusable. Zenodo. https://doi.org/10.5281/zenodo.2668478

Metadata

Title
Subtitle
Alternative Title
Alternative URL
Other Identifier
Author (+ROR; ORCHID; ...)
Point of Contact
Description
Subject
Keyword
Notes
Language
Producer
Production Date
Production Location

Contributor
Funding Information
Distributor
Distribution date
Depositor
Time Period
Date of Collection
Data Type
Series
Software
Data Sources

SSH
Keyword Getty AAT
Keyword ELSST
Topic Classification CESSDA
Universe
Frequency
Sampling Procedure
Characteristics of Data Collection Situation
Actions to Minimize Losses
Weighting
Response Rate

Rights Holder
Personal Data in Dataset?
Language Of Metadata

Archaeology
Archis ID
Report (ABR)
Methods of Recovery (ABR)
Subject (ABR)
Artefact (ABR)
Temporal (ABR)
Keyword Getty AAT

Audience
Collection
Relation or Related Material

Temporal Coverage
Spatial Point
Spatial Box
Spatial Coverage

Access conditions & licenses

- CC0-1.0 is a licence with which the copyright of the dataset is given up, putting the dataset in the public domain. A user may copy, change or distribute the dataset without notification. Formal attribution is not required, but this is of course still common practice in science.
- With CC-BY-4.0 the copyright of the dataset remains with the rightsholder, but the user can share and change the dataset. This licence requires attribution and the user has to state if they have changed the dataset.
- With CC-BY-SA-4.0 the same rules apply as for CC-BY-4.0, but in addition the user has to distribute the dataset they have reused under the same licence as the original (CC BY-SA 4.0).
- With CC-BY-NC-4.0 the same rules apply as for CC-BY-4.0, but the dataset may not be used for commercial purposes.
- With CC-BY-ND-4.0 the same rules apply as for CC-BY-4.0, but the user is not allowed to distribute the dataset if they have changed or extended the dataset.
- The licences CC-BY-NC-ND-4.0 and CC-BY-NC-SA-4.0 are a combination of the aforementioned licences.

Access conditions & licenses

pseudonymisation/anonymisation

The User will act in accordance with the Netherlands Code of Conduct for Research Integrity, the GDPR and other applicable laws and regulations.

The User will always cite the dataset in the research results they publish, in whatever form, when it has been used in the research.

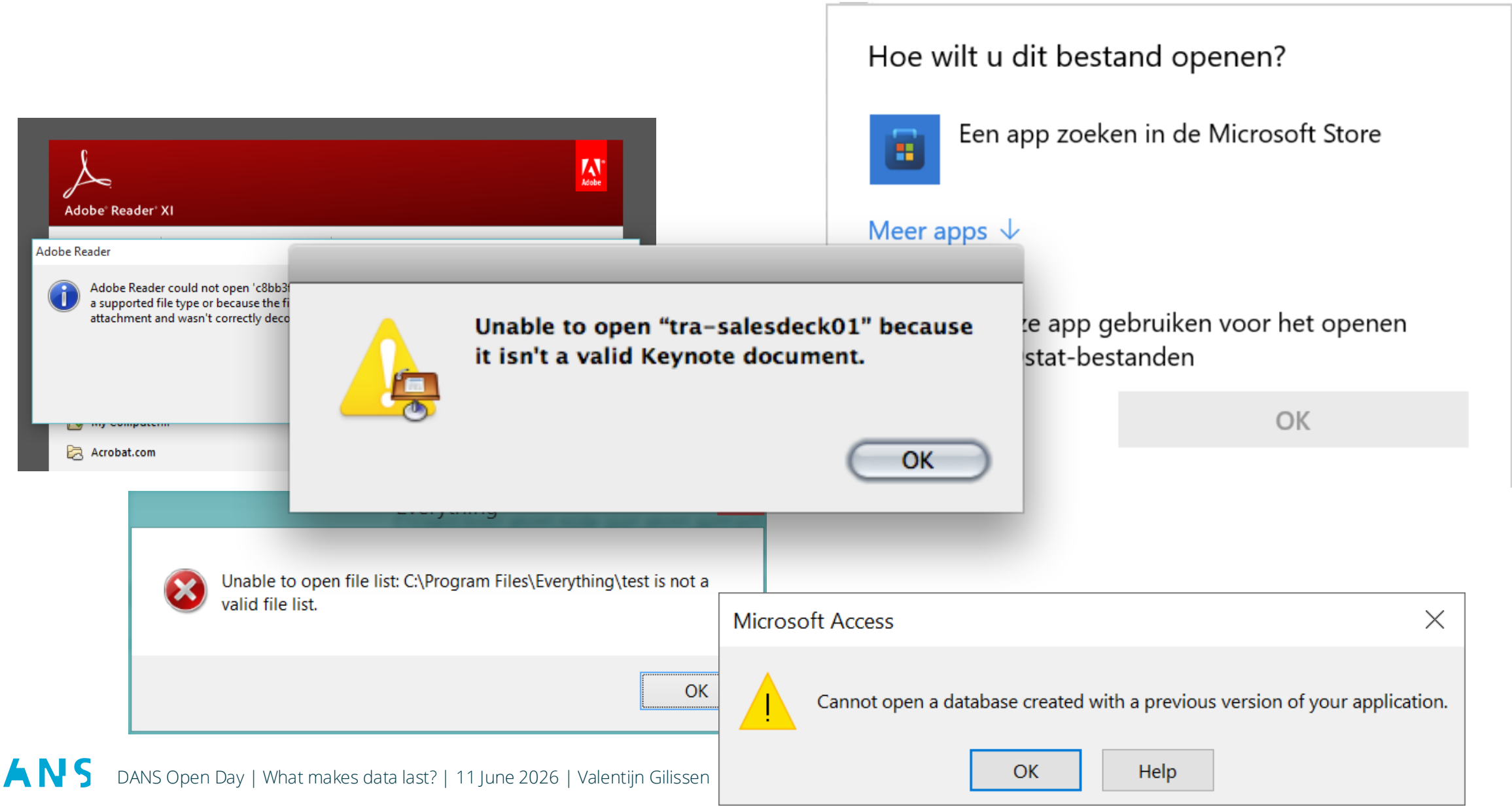
Embargo

For distribution or disclosure of the entire dataset or of substantial parts thereof, the User must first request permission from the holder of the rights to the dataset.

The User will always be responsible for the processing of personal data made available within the meaning of the GDPR and any other relevant privacy legislation, as well as for complying with any conditions set by the depositor.

Restricted access

Preferred Formats



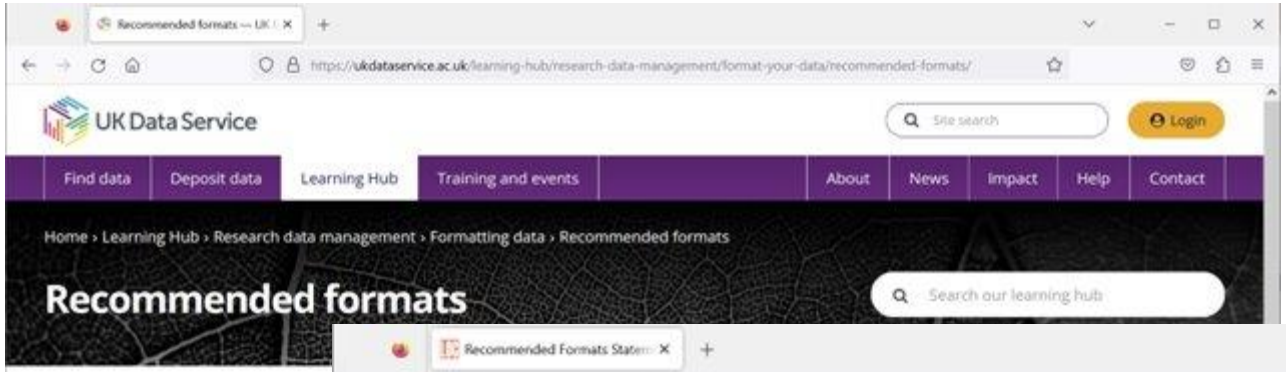
Preferred Formats

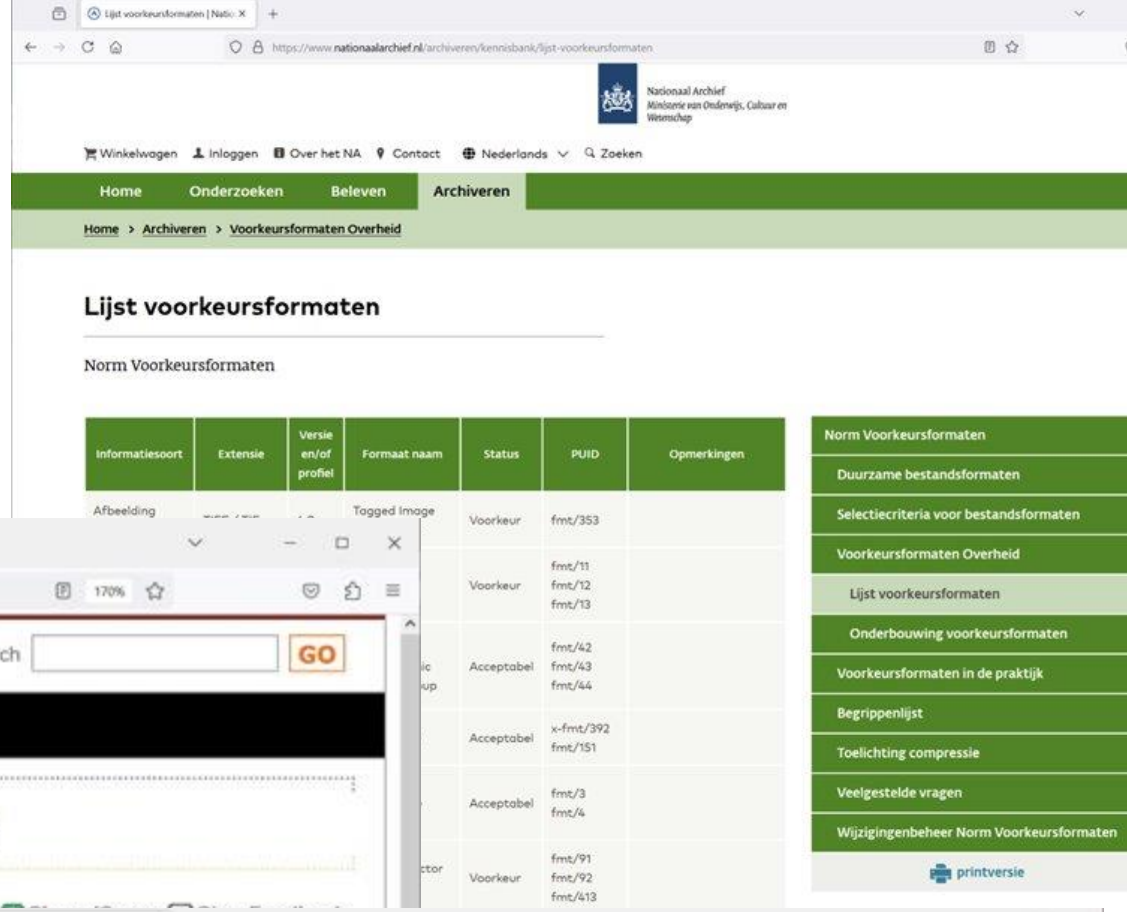
=> TGA; RAW; CDR

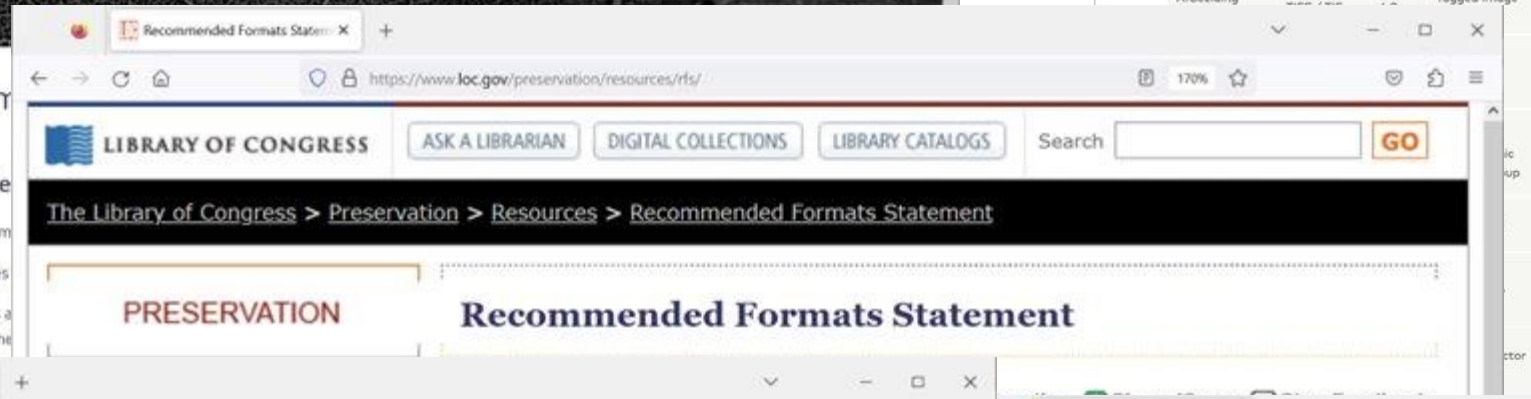
=> PCX; BMP; PSD

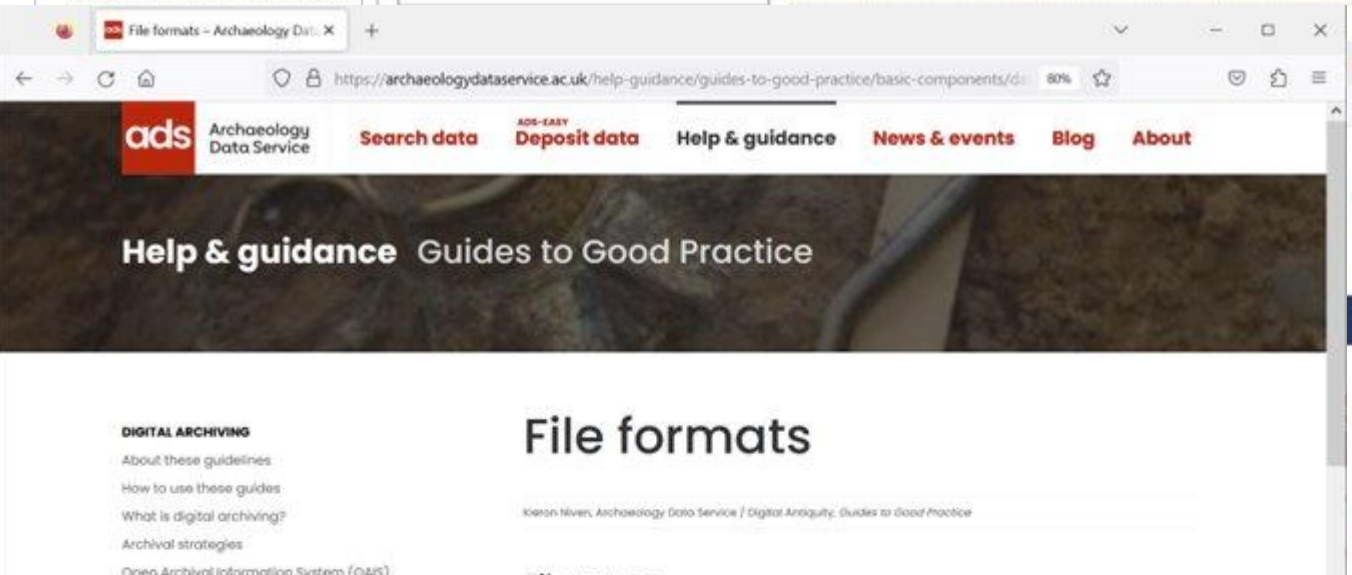
=> JPG; TIF; PNG

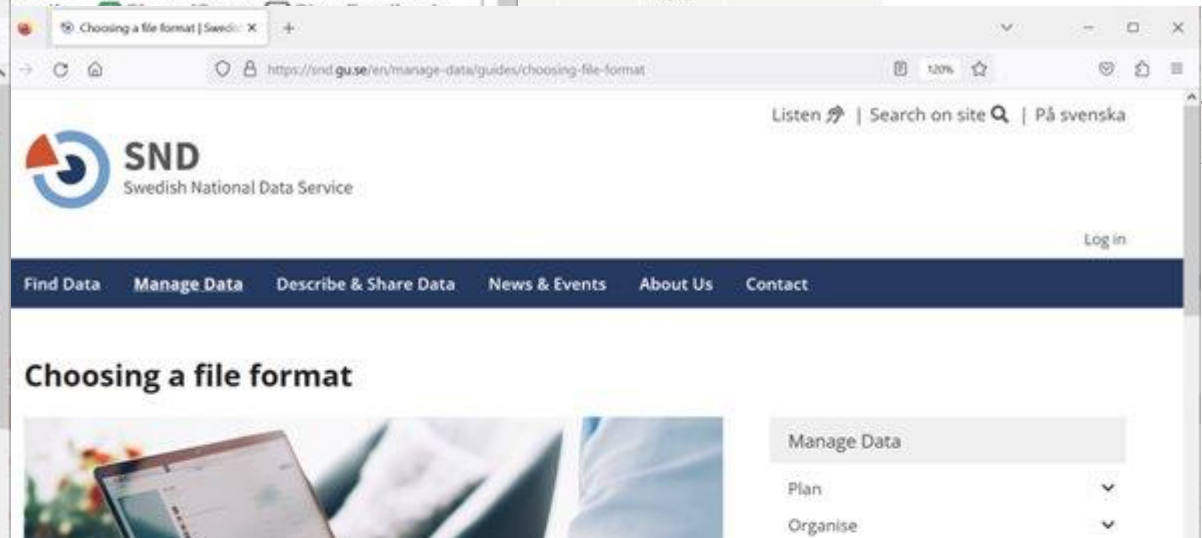
Preferred Formats











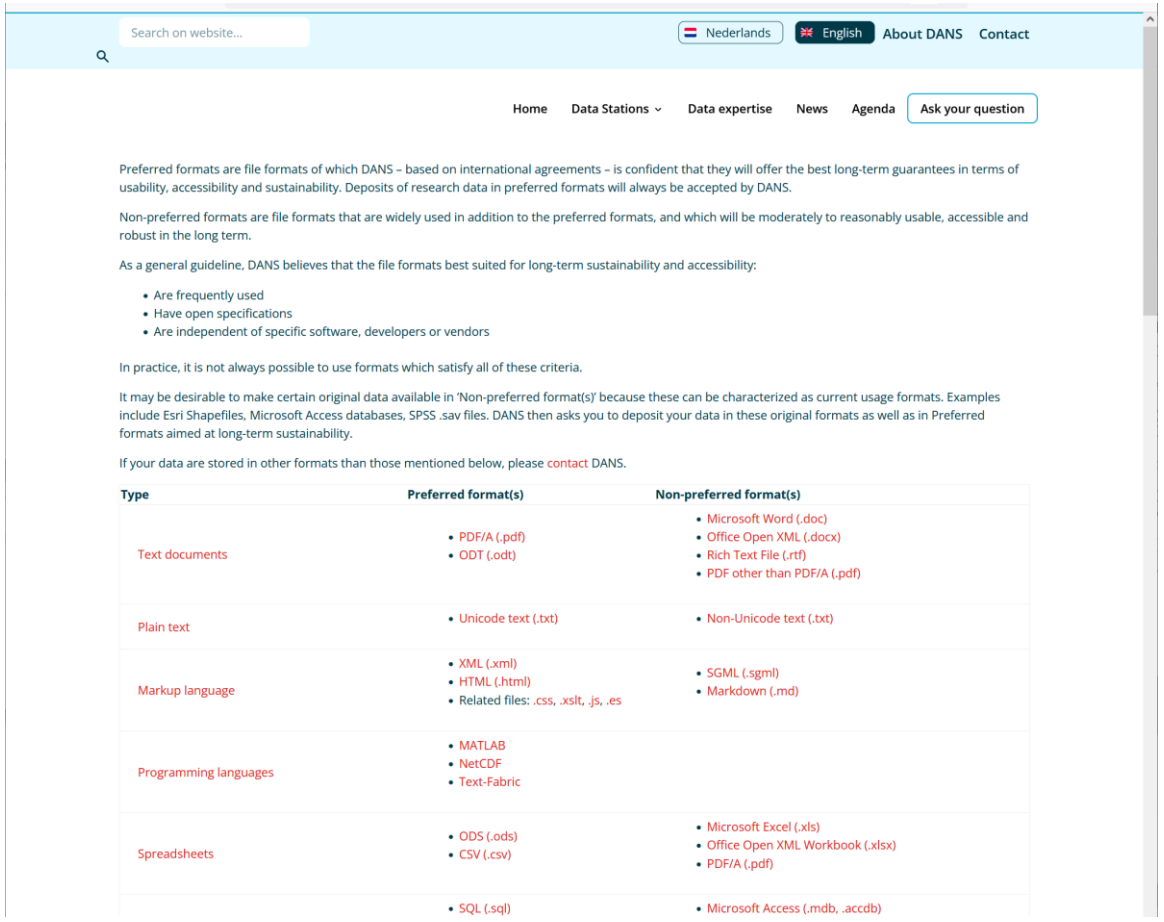
Preferred Formats

As a general guideline, DANS believes that the formats best suited for long-term sustainability and accessibility:

- are frequently used
- have open specification
- are independent of specific software, developers or vendors

In practice, it is not always possible to use formats which meet all the ideal characteristics of a preferred format.

https://dans.knaw.nl/en/file-formats



Search on website...

Nederlands | English | About DANS | Contact

Home | Data Stations | Data expertise | News | Agenda | Ask your question

Preferred formats are file formats of which DANS – based on international agreements – is confident that they will offer the best long-term guarantees in terms of usability, accessibility and sustainability. Deposits of research data in preferred formats will always be accepted by DANS.

Non-preferred formats are file formats that are widely used in addition to the preferred formats, and which will be moderately to reasonably usable, accessible and robust in the long term.

As a general guideline, DANS believes that the file formats best suited for long-term sustainability and accessibility:

- Are frequently used
- Have open specifications
- Are independent of specific software, developers or vendors

In practice, it is not always possible to use formats which satisfy all of these criteria.

It may be desirable to make certain original data available in 'Non-preferred format(s)' because these can be characterized as current usage formats. Examples include Esri Shapefiles, Microsoft Access databases, SPSS .sav files. DANS then asks you to deposit your data in these original formats as well as in Preferred formats aimed at long-term sustainability.

If your data are stored in other formats than those mentioned below, please contact DANS.

Type	Preferred format(s)	Non-preferred format(s)
Text documents	• PDF/A (.pdf) • ODT (.odt)	• Microsoft Word (.doc) • Office Open XML (.docx) • Rich Text File (.rtf) • PDF other than PDF/A (.pdf)
Plain text	• Unicode text (.txt)	• Non-Unicode text (.txt)
Markup language	• XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es	• SGML (.sgml) • Markdown (.md)
Programming languages	• MATLAB • NetCDF • Text-Fabric	
Spreadsheets	• ODS (.ods) • CSV (.csv)	• Microsoft Excel (.xls) • Office Open XML Workbook (.xlsx) • PDF/A (.pdf)
	• SQL (.sql)	• Microsoft Access (.mdb, .accdb)

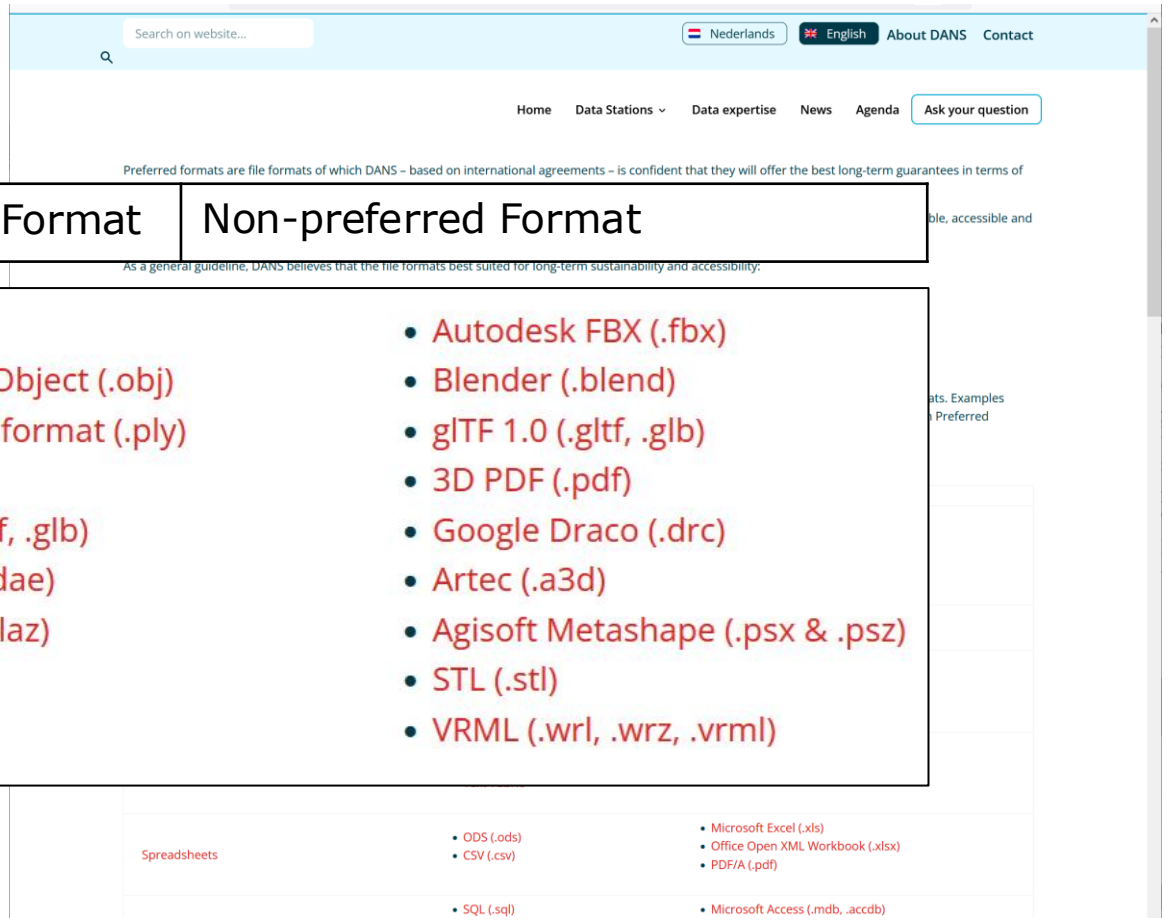
Preferred Formats

As a general guideline, DANS believes that the formats best suited for long-term sustainability and accessibility:

- are frequently used
- have op
- are inde
  develop

In practice, formats wh
characteristics of a preferred format.

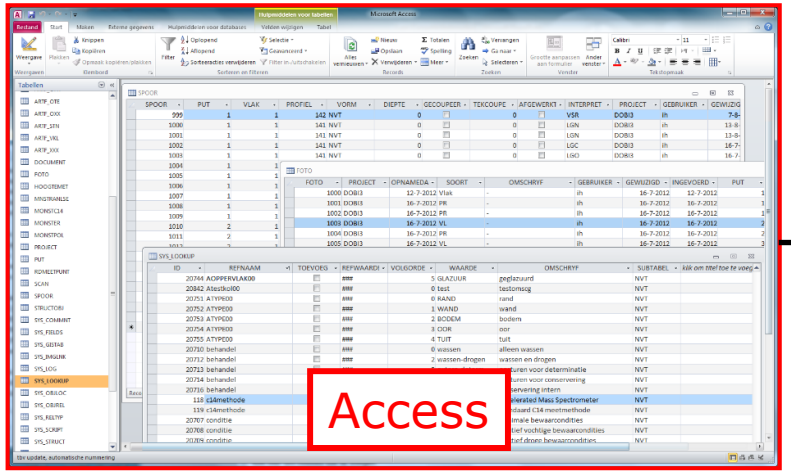
https://dans.knaw.nl/en/file-formats

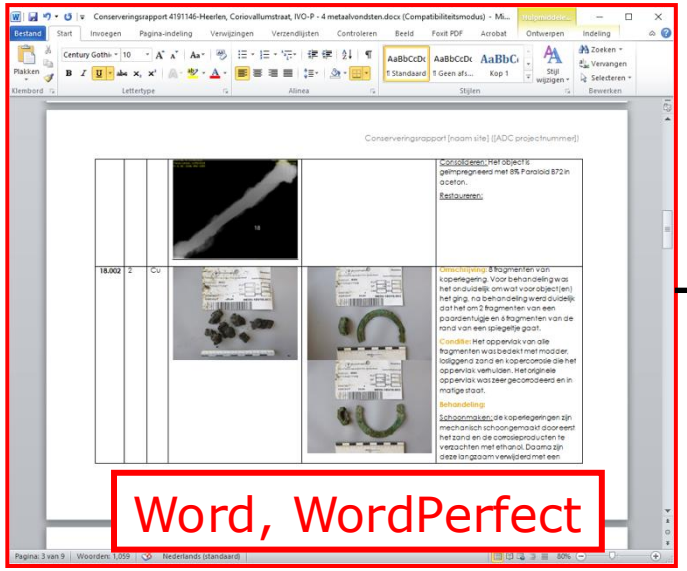


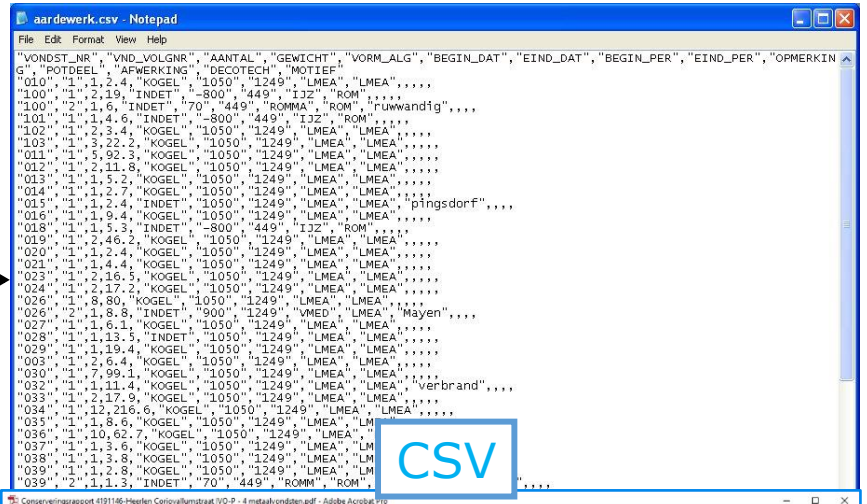
Preferred Formats

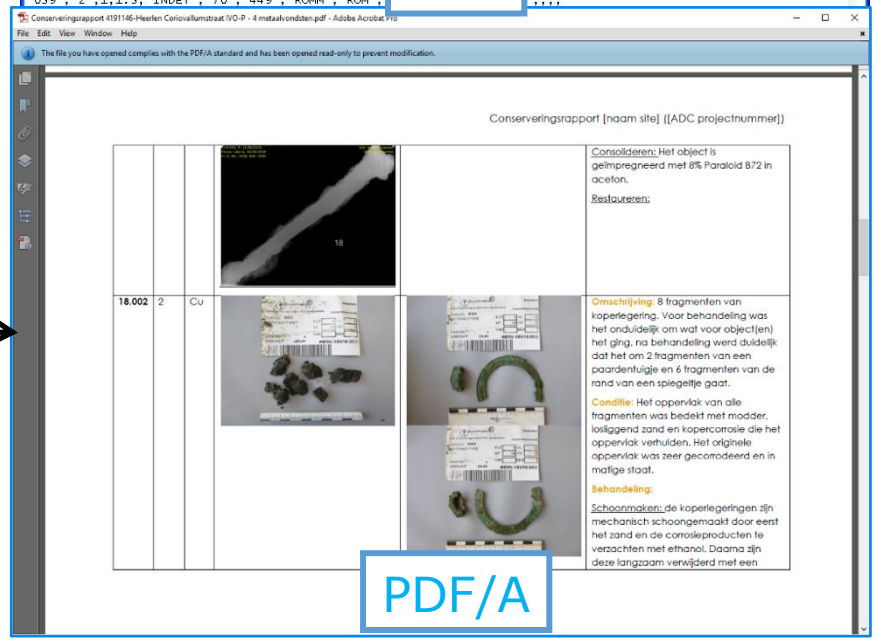
- Formats can still be in danger of falling into disuse and becoming unsupported. Long-term preservation should aim to make use of robust file formats
- If it is not possible to obtain ‘preferred formats’, non-preferred formats should still be accepted, although less long-term guarantees can be given
- When formats are converted/migrated, the original data should be preserved as well
- Certain non-preferred formats can be made available as current usage formats
- A policy with ‘non-acceptable formats’ risks receiving incomplete or lower-quality data / A policy with ‘acceptable formats’ risks a lack of effort towards sustainability
- Emulation may offer helpful solutions, but relying on emulation may simply shift preservation concerns from formats to software
- File format strategies need to be informed by expert input from data users and communities

Enhanced Curation

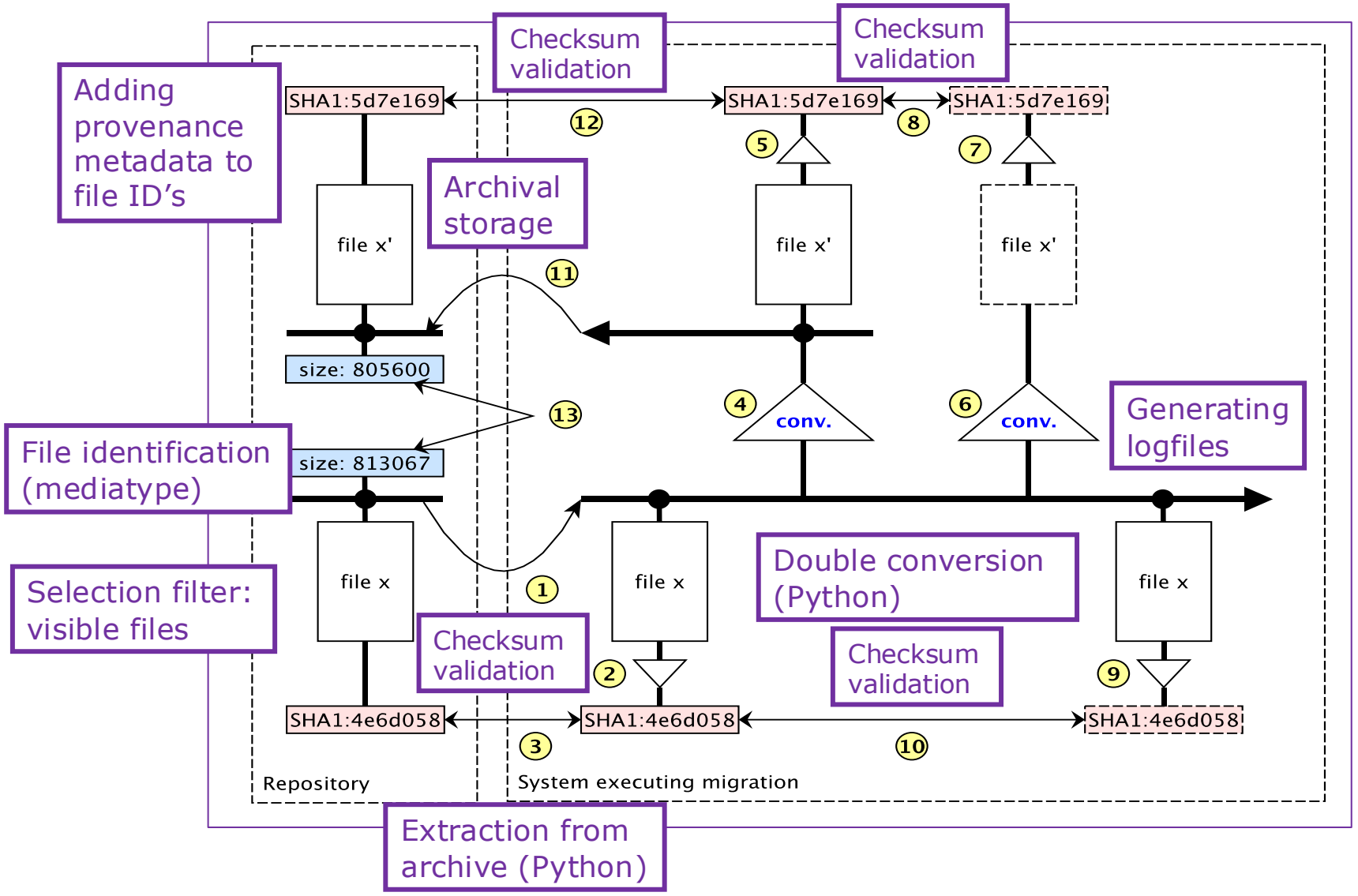




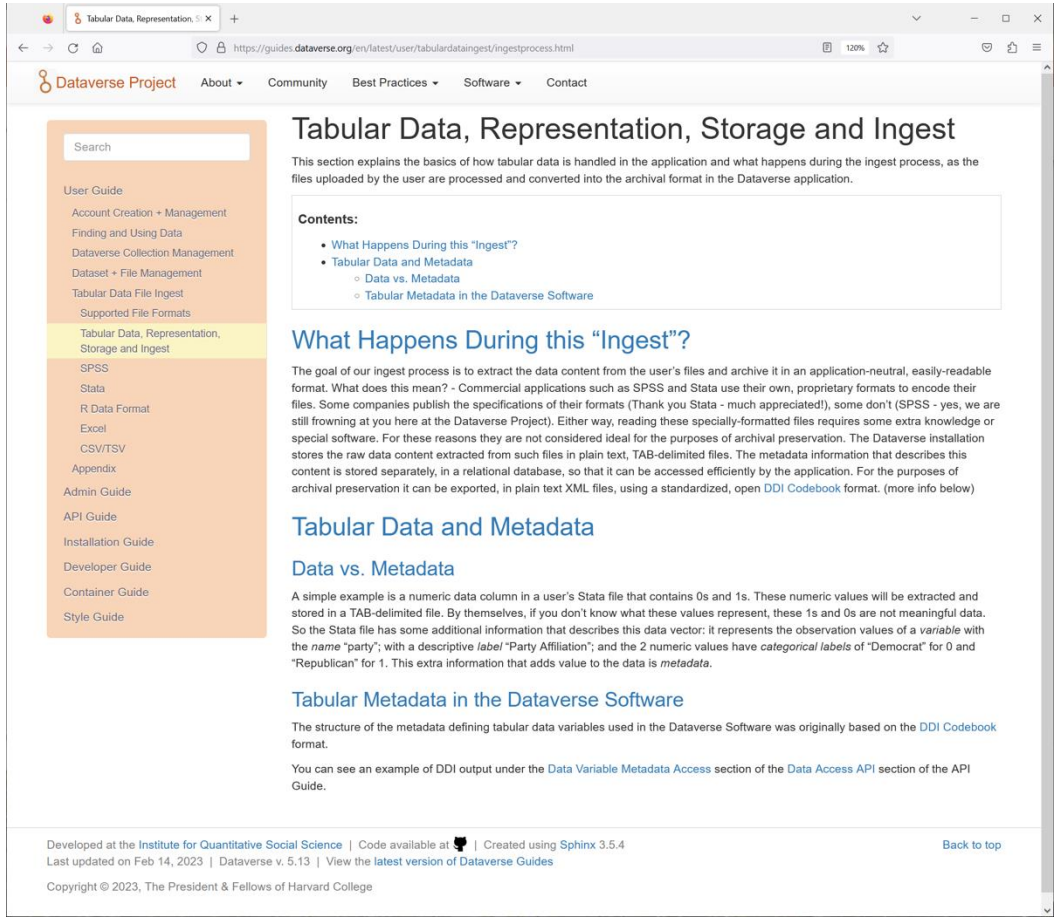




Enhanced Curation



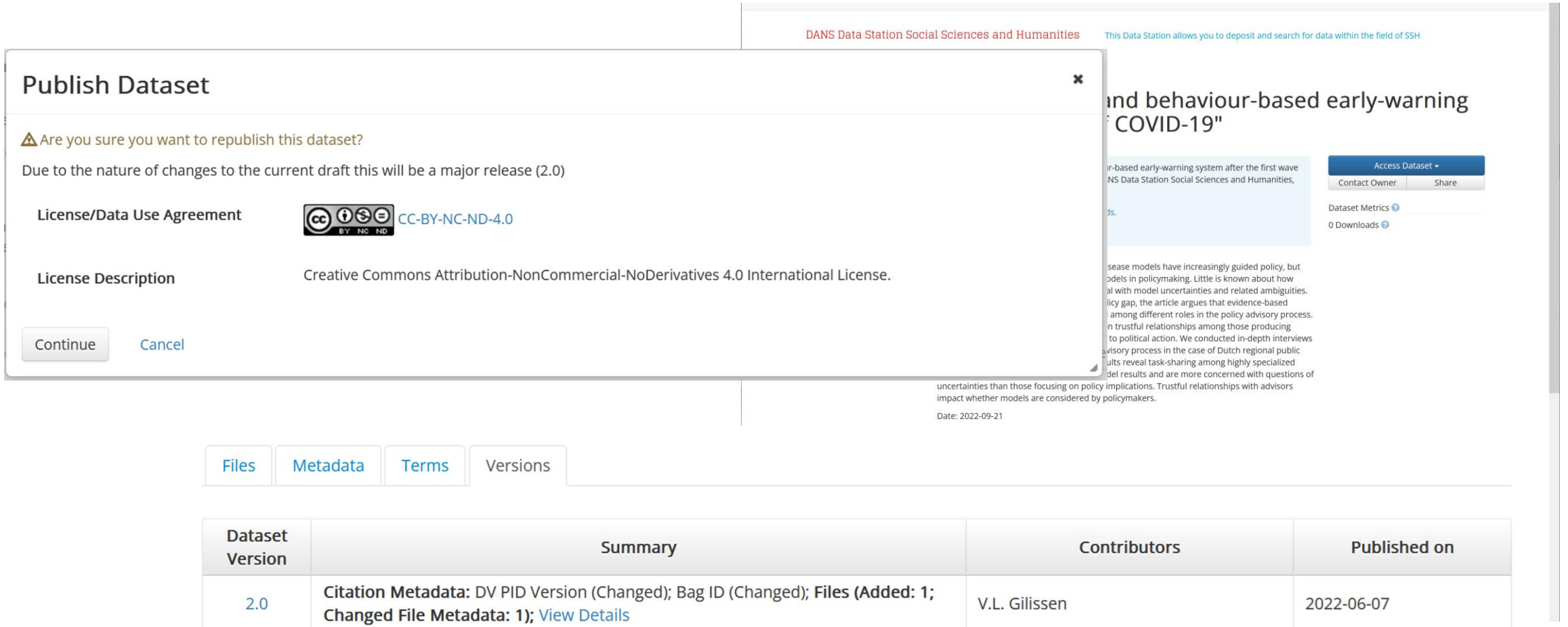
Tabular Ingest



- *The goal of our ingest process is to extract the data content from the user's files and archive it in an application-neutral, easily-readable format.*
- *The Dataverse installation stores the raw data content extracted from such files in plain text, TAB-delimited files. The metadata information that describes this content is stored separately, in a relational database, so that it can be accessed efficiently by the application. For the purposes of archival preservation it can be exported, in plain text XML files, using a standardized, open DDI Codebook format.*

Dataset versioning

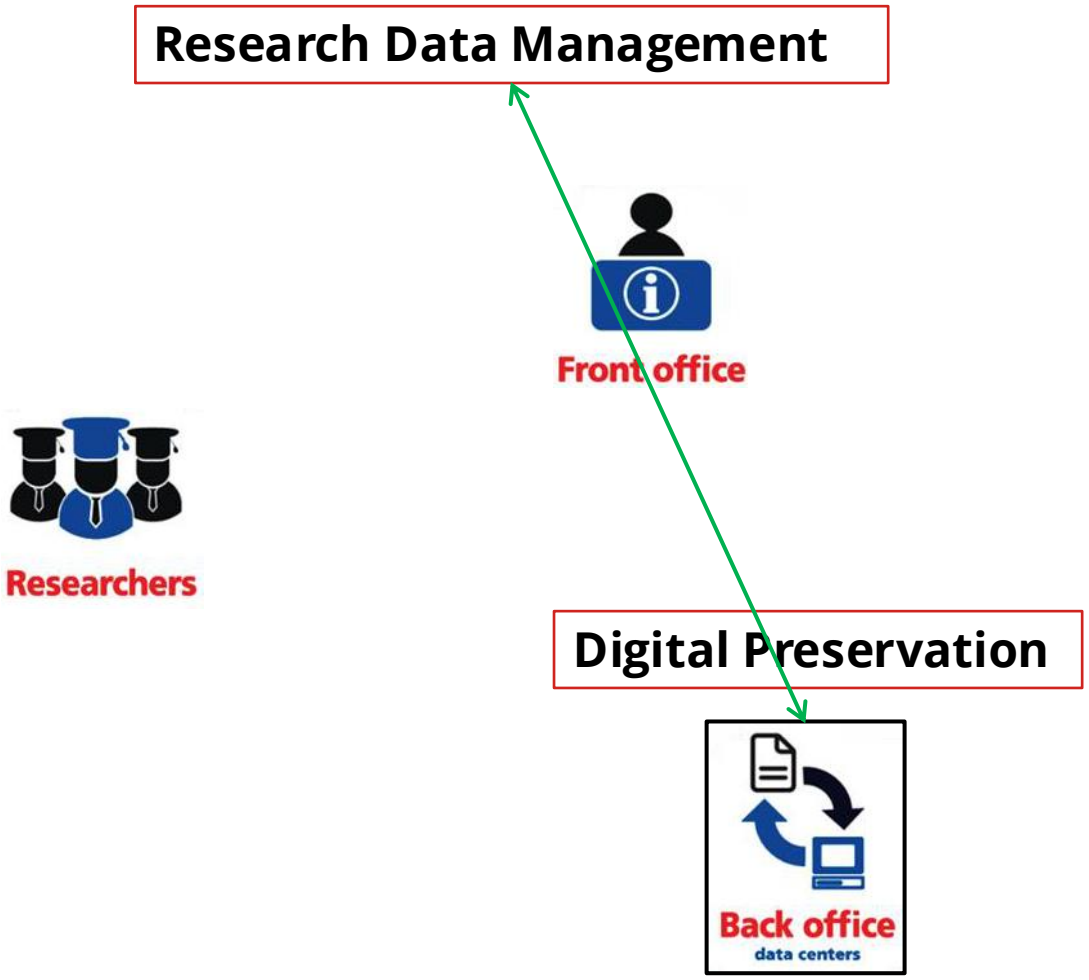
The Dataverse® Project



Files | Metadata | Terms | Versions

Dataset Version	Summary	Contributors	Published on
2.0	**Citation Metadata:** DV PID Version (Changed); Bag ID (Changed); **Files (Added: 1; Changed File Metadata: 1);** View Details	V.L. Gilissen	2022-06-07
1.0	This is the first published version.	Chris Baars, Chris Baars, EASY Migration	2022-05-27

Research Data Management vs Digital Preservation



Hickman, Jonathan (w), Steve Epting (p, i), Rick Magyar (i), Frank D'Armata (c), Joe Caramagna (l), Tom Brevoort (e), Lauren Sankovitch (e); Incursion from *Memento Mori*, New Avengers Vol 3 #1 (March 2013), Marvel Comics

Research Data Management

Personal data – pseudonymisation/anonymisation

Selecting a repository

File names – anonymisation

Discipline-specific requirements

Selection of relevant data files

Selection of relevant datasets

Data provenance

Choosing a licence for reuse

Personal data – awareness of rules and regulations

Organising data in folders

Digital Preservation

Providing dataset metadata

Choosing file formats

File formats policy

Restricting file access

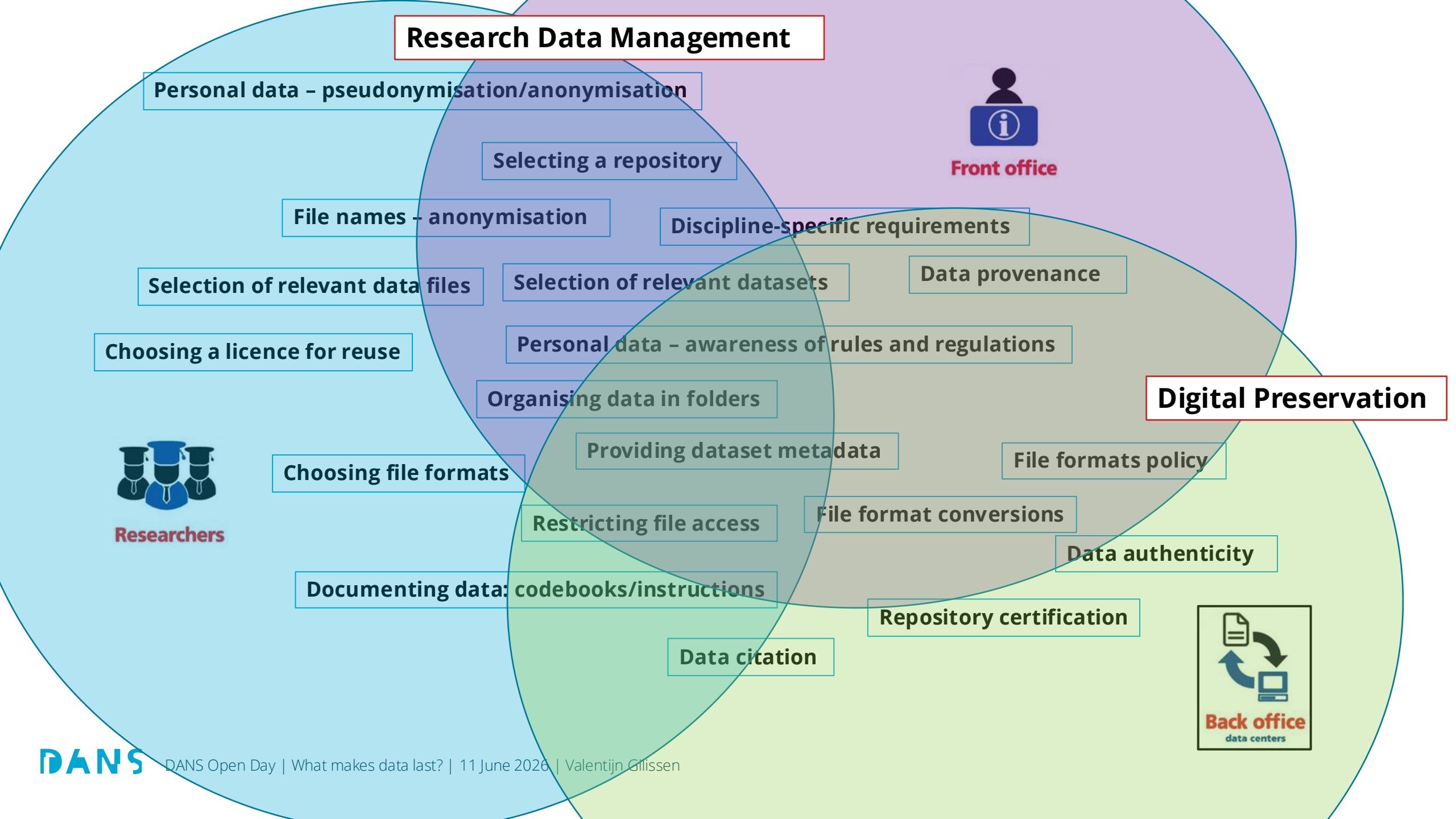
File format conversions

Data authenticity

Documenting data: codebooks/instructions

Repository certification

Data citation


Research Data Management
Personal data – pseudonymisation/anonymisation
Selecting a repository
File names – anonymisation
Discipline-specific requirements
Selection of relevant data files
Selection of relevant datasets
Data provenance
Choosing a licence for reuse
Personal data – awareness of rules and regulations
Organising data in folders
Digital Preservation
Providing dataset metadata
Choosing file formats
File formats policy
Restricting file access
File format conversions
Data authenticity
Documenting data: codebooks/instructions
Repository certification
Data citation
Front office
Researchers
Back office
data centers

Initiatives

European **O**pen **S**cience **C**loud
Enhancing **D**igital preservation strategies at **E**uropean and **N**ational level

eosc | EDEN
Enhancing Digital Preservation Strategies at European and National Level

Funded by the European Union

The EOSC EDEN project is funded by the European Union under the Horizon Europe framework programme and aims to support and promote digital preservation practices and standards at the European and national levels.

This support will be complemented by the development of user-centric tools, services, and standards designed to facilitate the creation of a distributed European infrastructure for digital data preservation, retention, and access.

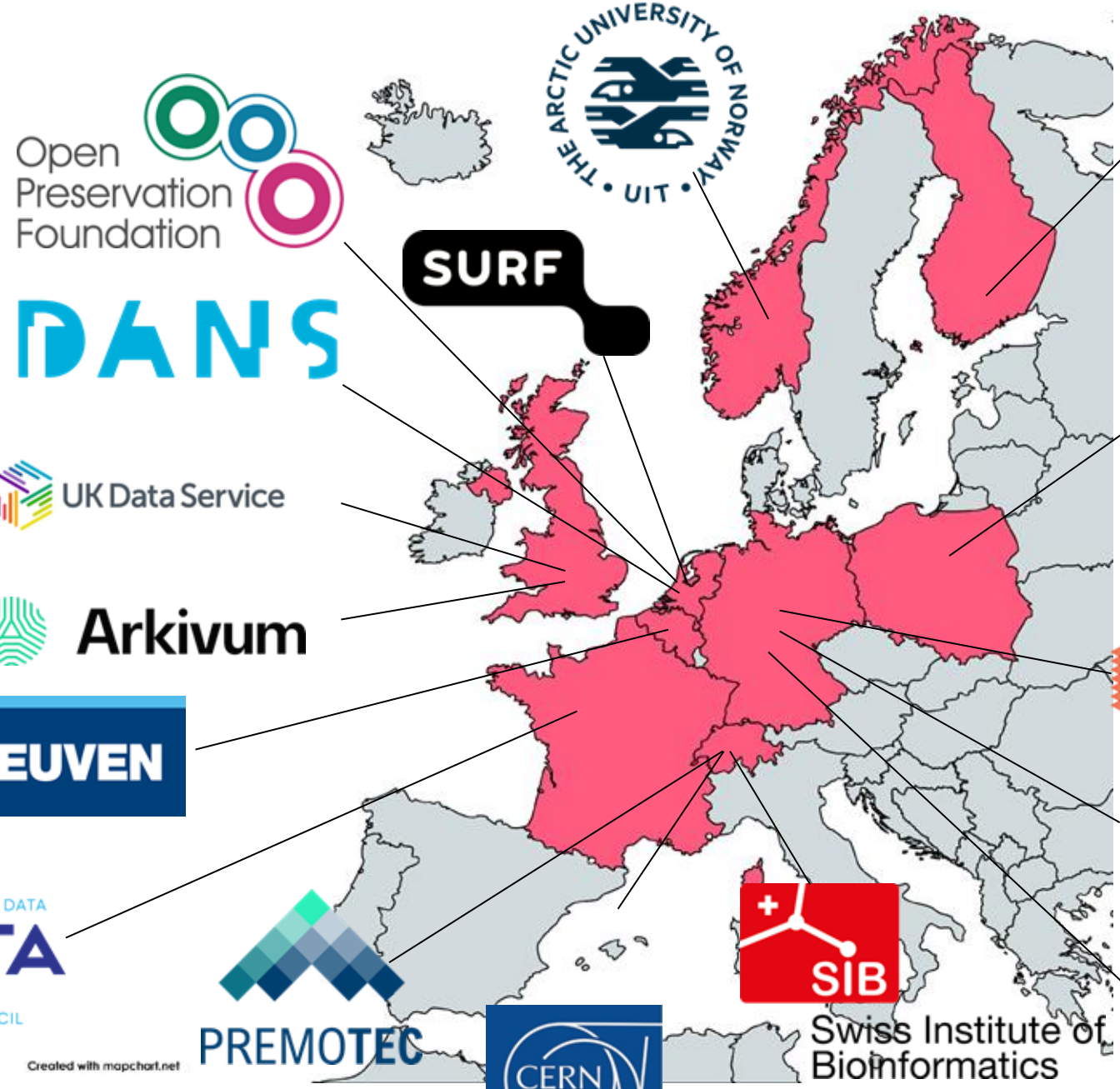
These tools will be developed following extensive stakeholder mapping encompassing individuals, groups, and organisations with an interest in or influence on, the project's development, implementation, and outcomes.

Initiatives

eosc | EDEN
Enhancing Digital Preservation Strategies at European and National Level

16 partners from nine countries with expertise in:

- digital preservation
- data quality
- curation
- FAIR data management and services
- hosting of repositories and archives



Initiatives

eosc | EDEN
Enhancing Digital Preservation Strategies at European and National Level

Core Preservation Process (CPP) = a specific action that every Trustworthy Digital Archive (TDA) should undertake (either directly or through an associated service)

CPP-001 Checksum Generation & Recording

CPP-002 Checksum Validation

CPP-003 Integrity Checking

CPP-004 Data Corruption Management

CPP-005 Identifier Management

CPP-006 AIP Batch Export

CPP-007 Virus Scanning

CPP-008 File Format Identification

CPP-009 Metadata Extraction

CPP-010 File Format Validation

CPP-011 Replication

CPP-012 Risk Mitigation

CPP-013 Object Management Reporting

CPP-014 File Migration

CPP-015 Emulation & Rendering Tools

CPP-016 Metadata Ingest & Management

CPP-017 Disposal

CPP-018 Community Watch

CPP-019 Data Quality Assessment

CPP-020 Rights Management

CPP-021 AIP Versioning

CPP-022 Significant Properties Definition

CPP-023 Risk Definition and Extraction

CPP-024 Enabling Discovery

CPP-025 Enabling Access

CPP-026 File Normalisation

CPP-027 File Repair

CPP-028 Creation of Derivatives

CPP-029 Ingest

CPP-30 Refreshment

Micky Lindlar & Bertrand Caron, 1/12/2025

Initiatives

eosc | EDEN

Enhancing Digital Preservation Strategies at European and National Level

Funded by the European Union

	eden-fidelis.eu
	linkedin.com/company/eosc-eden
	@eosc-eden.bsky.social
	@EOSC-EDEN

#EOSCEDEN

github.com/EOSC-EDEN

EOSC EDEN Zenodo Community

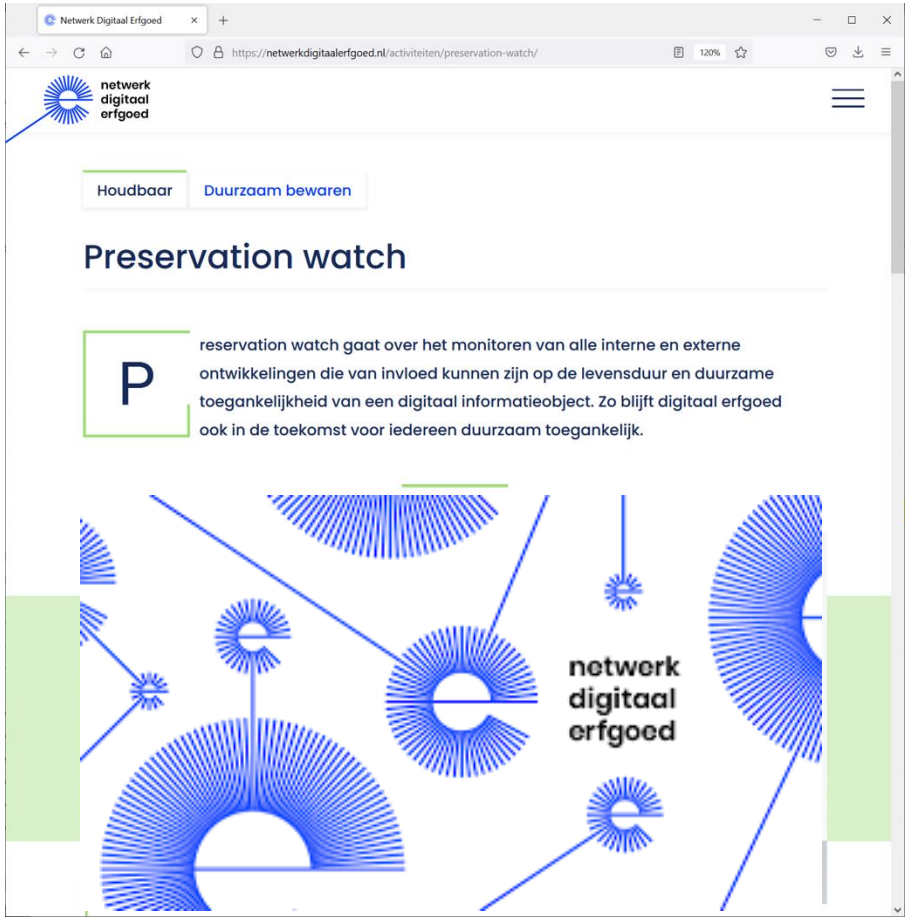
M1.1 Report on Identification of Core Preservation Processes

EOSC EDEN T1.2; Lindlar, Micky (Work package leader)[1]; Caron, Bertrand (Project leader)[1]; Benauer, Maria (Project member)[1]; Kylander, Johan (Project member)[2]; Dekeyser, Kris (Project member)[3]; Addis, Matthew (Project member)[4]; Levlin, Mattias (Project member)[2]; Laukkanen, Mikko (Project member)[2]; Lehtonen, Juha (Project member)[2]; Burger, Felix (Project member)[1]; Koho, Tiina (Project member)[2]; Schwab, Franziska (Other)[1]; Molloy, Laura (Other)[5]; Zhang, Fen (Other)[3]

https://doi.org/10.5281/zenodo.16992451

edu link:
https://edu.nl/4ecty

Initiatives



Dutch Digital Heritage Network

KNAW Humanities Cluster (includes DANS)
National Library of the Netherlands
National Archive
Netherlands Institute for Sound and Vision
Het Nieuwe Instituut
Netherlands Cultural Heritage Agency

Preservation Watch

Netherlands Institute for Sound and Vision
DANS
EYE Filmmuseum
National Library of the Netherlands
LIMA
National Archive
RHC Eindhoven
Amsterdam City Archives
The Utrecht Archives

Initiatives

www.wegwijzervoorkeursformaten.nl

Stappenplan

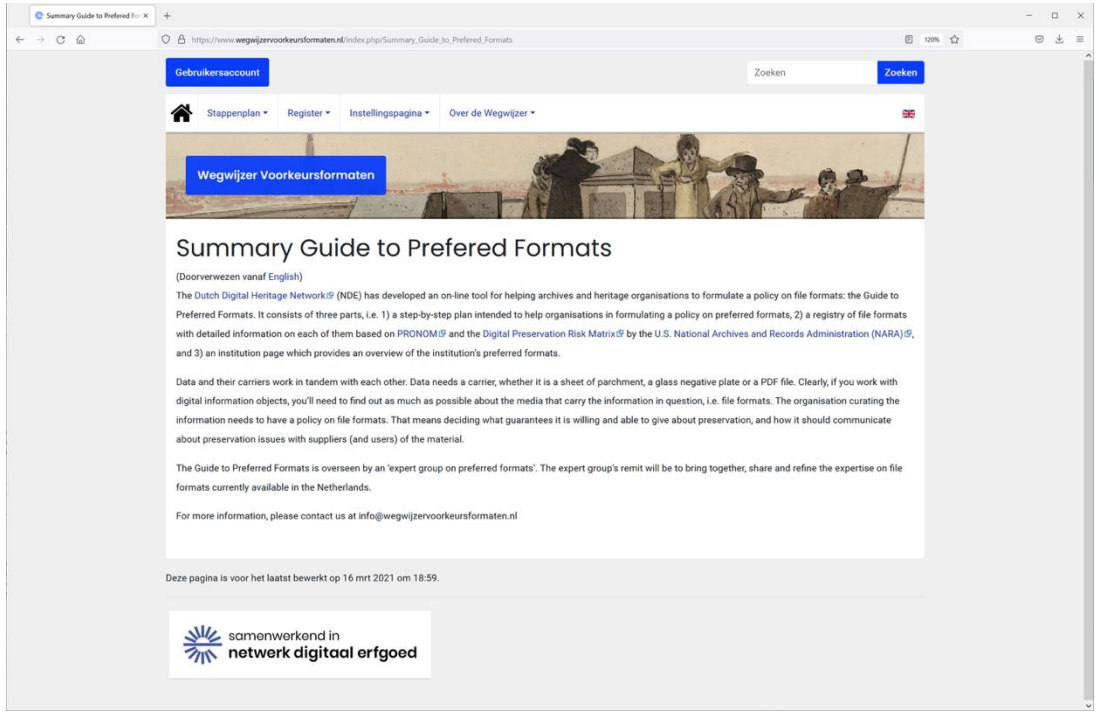
Interactive ***step-by-step module*** to help institutions create their own file formats policies through comparisons with other policies, via available information and wighing importance

Register

Register of file formats with information derived from international sources: PRONOM; Wikidata; Library of Congress; NARA; COPTR

Instellingspagina

Institution pages with overviews derived from the step-by-step plans and the register



Initiatives

https://coptr.digipres.org

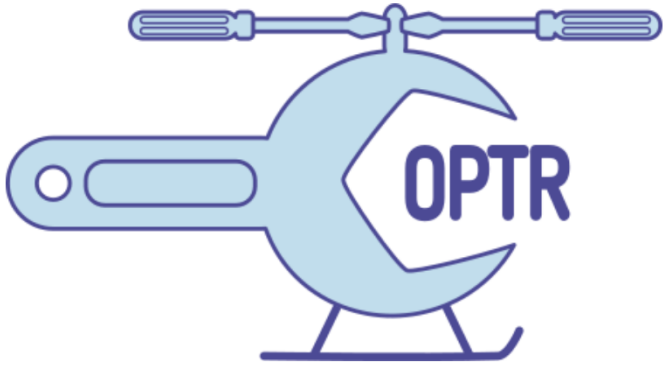
COPTR Home
Community Owned Workflows
Recent changes

Find tools
Tools grid
By lifecycle stage
By function
By content type
By file format
All tools

Help
About COPTR
Video guides to using COPTR
Mediawiki help
Wikitext cheat sheet
COPTR data structures

Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information
Cite this page
Browse properties

Community Owned digital Preservation Tool Registry (COPTR)



COPTR helps practitioners find tools needed for long term digital preservation tasks. It describes 608 tools and 36 workflows.

- Find a digital preservation tool using the **Tools Grid** (or by stage, function, content type or file format). Or watch this video to learn more.
- View digital preservation workflows
- Add a tool, or find out more about how you can get involved
- Add a workflow
- About COPTR, including how to Contact Us, how COPTR is structured and how to access COPTR data via the API.

NEW FOR JULY 2021: COPTR has been re-launched with new functionality. Find out all about it in this video presentation.

COPTR Partners

COPTR was created and launched with the support of the Aligning National Approaches to Digital Preservation initiative and has been populated and maintained by members of the COPTR partner organisations:

The Digital Curation Centre (DCC)

The Digital Curation Exchange (DCE)

The Digital Preservation Coalition (DPC)

Dutch Digital Heritage Network (DDHN)

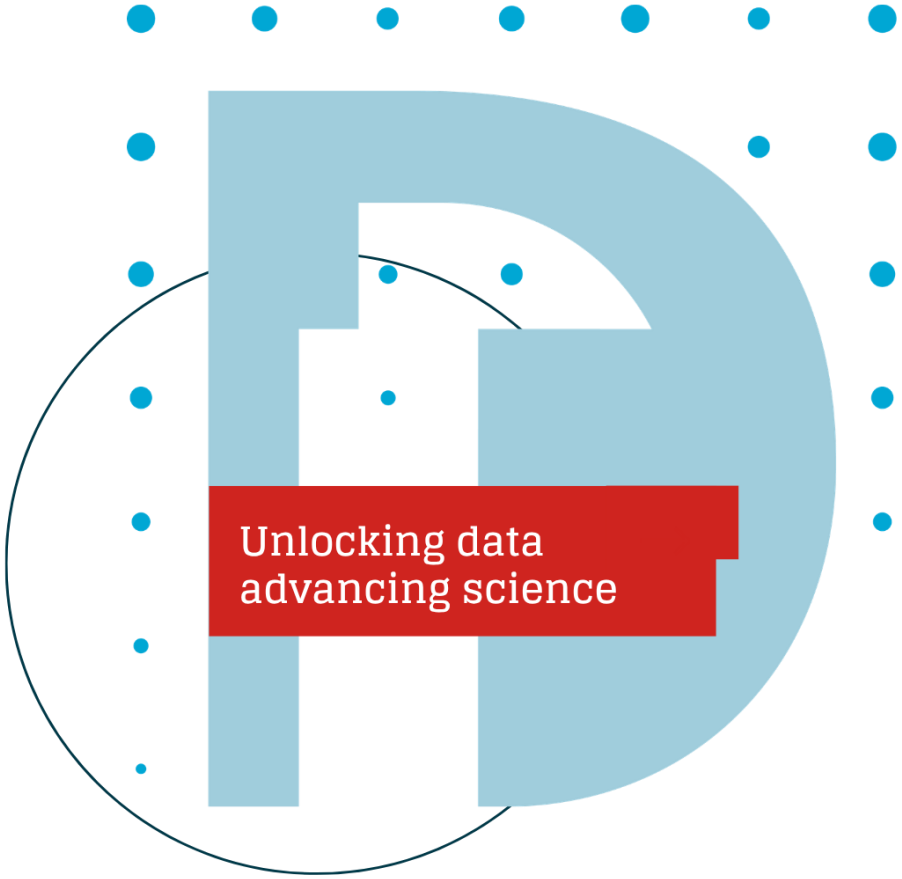
National Digital Stewardship Alliance (NDSA)

Network of expertise in long-term storage of digital resources in Germany (NESTOR)

The Open Preservation Foundation (OPF)

Thank you for your attention

DANS Open Day
Open data, open science



DANS

Anna van Saksenlaan 51 | 2593 HW The Hague | The Netherlands | +31 (0)88 003 46 66
National centre of expertise and repository for research data | An institute of the KNAW and NWO

DataLink LinkedIn BlueSky Mastodon www.dans.knaw.nl