

DCC Spring Training Days – Session 9

From researcher to repository and back again

Valentijn Gilissen – Data Processing Team Leader / Preservation Officer / Data Steward

Deborah Thorpe – Senior Research Data Management Specialist

08/07/2025

<https://dans.knaw.nl>



Mission: promote and provide permanent access to digital research resources



nestor
Seal
2016

Institute of Dutch Academy and Research Funding Organisation (KNAW & NWO) since 2005

First predecessor dates back to 1964 (Steinmetz Foundation), Historical Data Archive 1989

DANS Data Stations

The image displays two screenshots of the DANS Data Station Social Sciences and Humanities website. The left screenshot shows the search results page for the query "Interview data on 'Mobility- and behaviour-based early-warning system after the first wave of COVID-19'". The search results are filtered by "Dataverses (0)" and "Datasets (7,297)". The results list includes "Strikes in the Netherlands, 1810-19" and "Research project 'That's What Frier'". The right screenshot shows the dataset page for the "Interview data on 'Mobility- and behaviour-based early-warning system after the first wave of COVID-19'". The page includes a description of the dataset, a subject of "Social Sciences", a keyword of "COVID-19, evidence-based policymaking, policy advisory process, crisis management, mathematical modelling, uncertainty, Dutch Safety Regions", and a license of "CC-BY-NC-4.0". The dataset is available in "Files", "Metadata", "Terms", and "Versions".

DANS Data Station Social Sciences and Humanities

657 Downloads

Search this dataverse... [Advanced Search](#)

Dataverses (0)

Datasets (7,297)

Files (91,284)

Publication Year

2023 (10)

2022 (207)

2021 (192)

2020 (222)

2019 (199)

Subject

Social Sciences (4,823)

Arts and Humanities (4,206)

Business and Management (310)

Law (278)

Medicine, Health and Life Sciences (232)

Audience

Modern and contemporary history (2,552)

Social Sciences (2,522)

Sociology (1,571)

Humanities (1,435)

History (1,134)

Keyword Term

oral history (1,395)

Tweede Wereldoorlog (789)

Nederlands Indië (762)

dekolonisatie (737)

Indonesische onafhankelijkheid (720)

Collection

Oral History (3,052)

World War II (1,138)

CLARIN (1,033)

Nederlandse Veteranen (961)

Erfgoed van de Oorlog (787)

1 to 10 of 7,297 Results

Strikes in the Netherlands, 1810-19

Jun 5, 2023

J.H.A. van der Velden, 2000, Social Sciences and Humanities

The dataset contains data on all known of labour unrest.

Research project 'That's What Frier'

Jun 5, 2023

Y.H.M. van den Berg, 2021, Station Social Sciences and Humanities

Our relationships with our parents and emerging adulthood, individuals spend examining...

The Effectiveness of the MLU in America

Jan 31, 2023

D.M. Broekhuijsen, 2023, "The Multilateral Success", https://doi.org/10.17026/dans-z3c-b9sk, D

Database related to article. In the article: Instrument (MLU) in terms of states that...

Omvang, achtergrond en kosten aanpak

Jan 20, 2023

K van der Ven, 2023, "Omvang, achtergrond en kosten aanpak", DANS Data Station

Verkenning naar de omvang, motieven en kosten van het onderzoek naar de omvang, motieven en kosten van het onderzoek...

Transcriptions of Subscriptions in the Netherlands

Jan 5, 2023

C.T. Elmelund, 2023, "Transcriptions of Subscriptions in the Netherlands", https://doi.org/10.17026/dans-zdf-b9sk, D

Full new transcriptions of the subscription corrections to both the ECM apparatus and the ECM apparatus...

e-RUPI and Economics

Dec 31, 2022

Interview data on "Mobility- and behaviour-based early-warning system after the first wave of COVID-19"

Embargoed **Version 2.0**

F.A. Metz, 2023, "Interview data on "Mobility- and behaviour-based early-warning system after the first wave of COVID-19"", https://doi.org/10.17026/dans-z3c-54eq, DANS Data Station Social Sciences and Humanities, V2

Access Dataset

Contact Owner Share

Dataset Metrics

0 Downloads

Description

During the COVID-19 pandemic infectious disease models have increasingly guided policy, but also fueled a discussion about the role of models in policymaking. Little is known about how policymakers and public administrations deal with model uncertainties and related ambiguities. In contrast to the widely stated evidence-policy gap, the article argues that evidence-based policymaking is a collaborative effort shared among different roles in the policy advisory process. Whether projections guide policy depends on trustful relationships among those producing scientific knowledge and those translating it to political action. We conducted in-depth interviews with employees working along the policy advisory process in the case of Dutch regional public entities of disaster- and crisis response. Results reveal task-sharing among highly specialized roles. Knowledge producers tend to use model results and are more concerned with questions of uncertainties than those focusing on policy implications. Trustful relationships with advisors impact whether models are considered by policymakers.

Subject

Social Sciences

Keyword

COVID-19, evidence-based policymaking, policy advisory process, crisis management, mathematical modelling, uncertainty, Dutch Safety Regions

License/Data Use Agreement

CC-BY-NC-4.0

Files **Metadata** **Terms** **Versions**

Search this dataset...

Filter by

File Type: All Access: All

1 to 4 of 4 Files

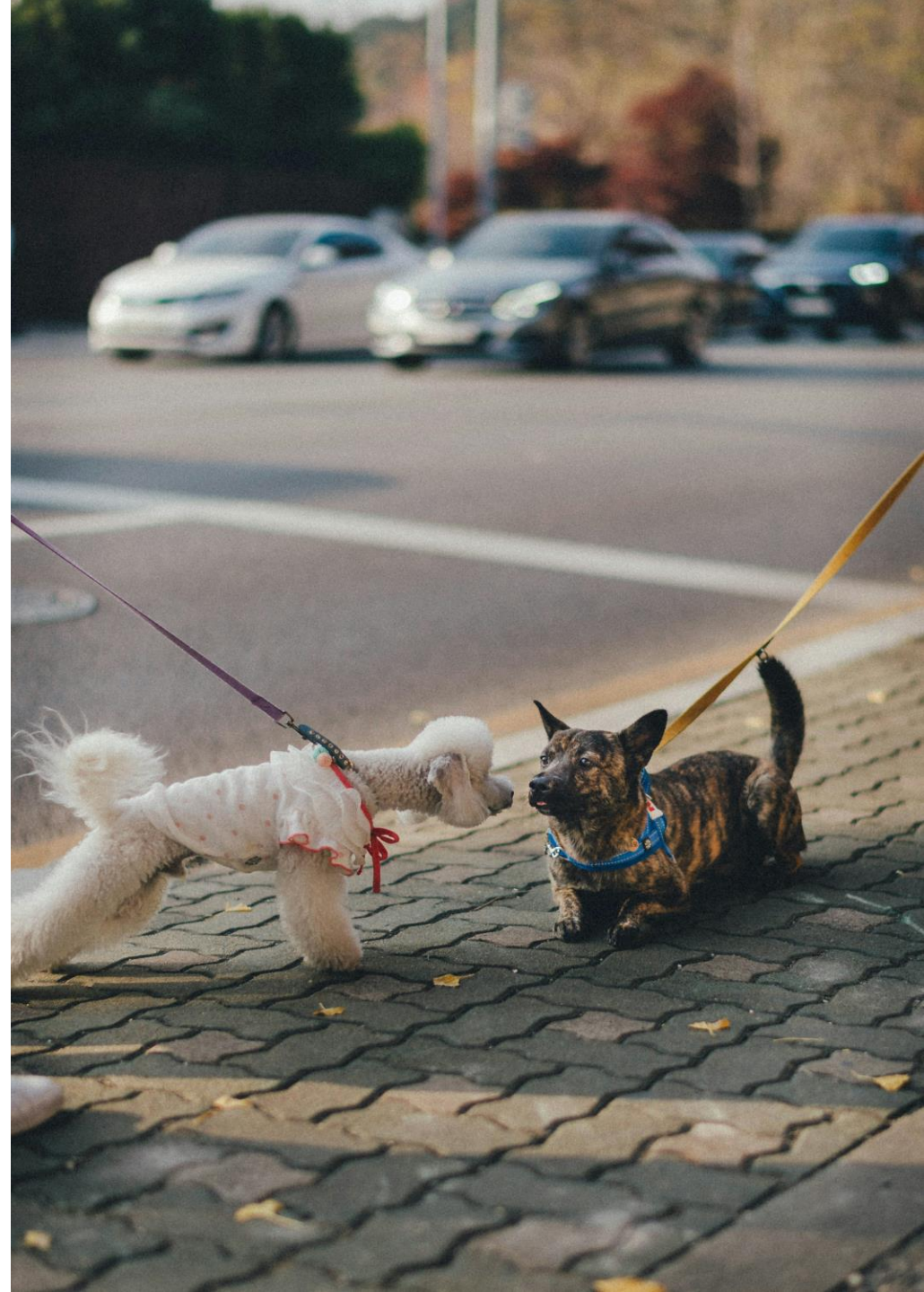
Documentation for the interview data.pdf

Adobe PDF - 318.0 KB

Metadata Released: Sen 22, 2022 Embargoed until 2023-12-31

Introduce yourselves

- Your **name**
- Your **role**
- What **types of data** do you collect / curate / advise / specialise in?



Personal data – awareness of rules and regulations

Personal data – pseudonymisation/anonymisation

File names – anonymisation

Restricting file access

Choosing a licence for reuse

Selection of relevant data files

Selection of relevant datasets

Organising data in folders

Documenting data: codebooks/instructions

Selecting a repository

Repository certification

Providing dataset metadata

Discipline-specific requirements

Choosing file formats

File formats policy

File format conversions

Data authenticity

Data citation

Data provenance



Front office



Researchers



Back office
data centers



Think, Pair, Share!

We will give you a discussion topic. Think for 5 minutes about your answer to the question.

Compare and combine your answer with the person next to you for 5 minutes

We will **discuss as a larger group** for 5 minutes



Think, Pair, Share!

How do you handle personal data?

What can a **repository** do to support the archiving of this? What does the **researcher** need to do / be aware of?



Licences and options

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.

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.

pseudonymisation/anonymisation

Restricted access

Embargo

Open Access licences

- [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.



Your ideas:

What metadata fields should be mandatory (most important for data reuasability)?



Metadata

Title
Subtitle
Alternative Title
Alternative URL
Other Identifier
Author
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

Rights Holder
Personal Data in Dataset?
Language Of Metadata

Audience
Collection
Relation or Related Material

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

Archaeology

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

Temporal Coverage
Spatial Point
Spatial Box
Spatial Coverage

-
-
-
-
-
-
-

Metadata

Title

Subtitle
Alternative Title
Alternative URL
Other Identifier

Author

Point of Contact

Description

Subject

Keyword
Notes
Language
Producer
Production Date
Production Location

Audience

Collection
Relation or Related Material

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

Rights Holder
Personal Data in Dataset?
Language Of Metadata

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

Temporal Coverage
Spatial Point
Spatial Box
Spatial Coverage

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

-
-
-
-
-
-
-



Your ideas:

Do you think anything is missing?

Would there be any additional mandatory fields that you would include?





Think, Pair, Share!

We will give you a question. Think for 5 minutes about your answer to the question.


Compare and combine your answer with the person next to you for 5 minutes

We will **discuss as a larger group** for 5 minutes





Think, Pair, Share!



What role should a
data supporter have in
**creating discovery
metadata** for a
researcher's dataset?





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).

Dataset curation

- Inventorise deposit

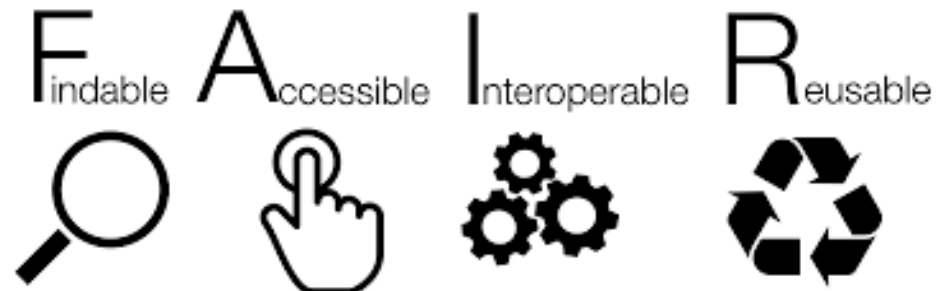
Basic curation

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

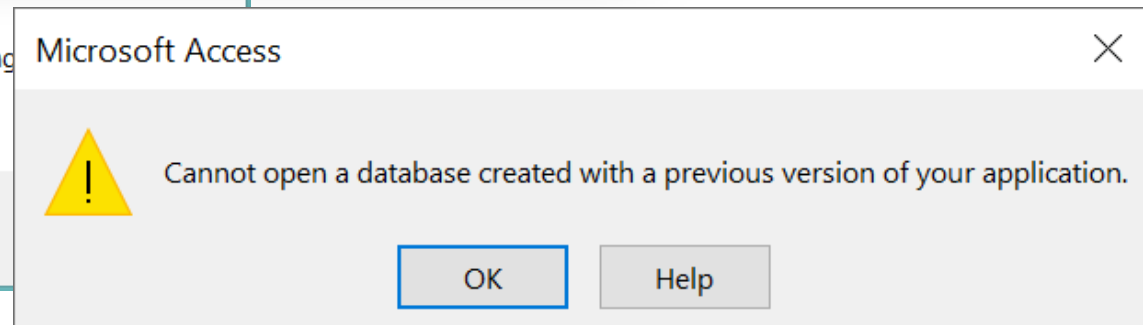
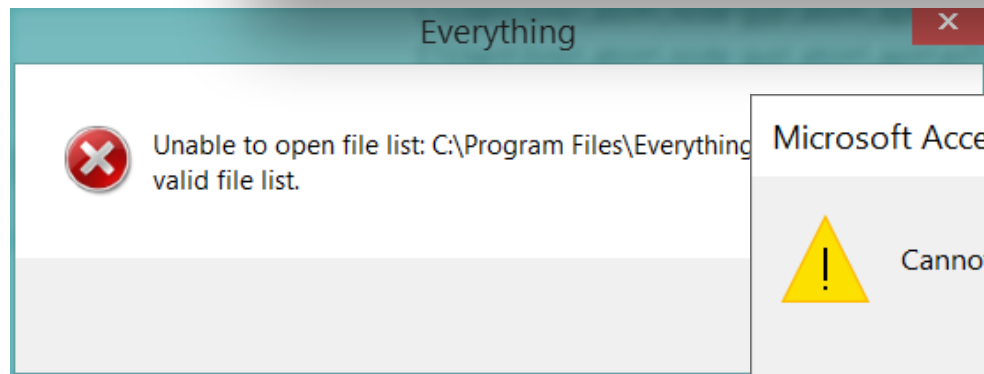
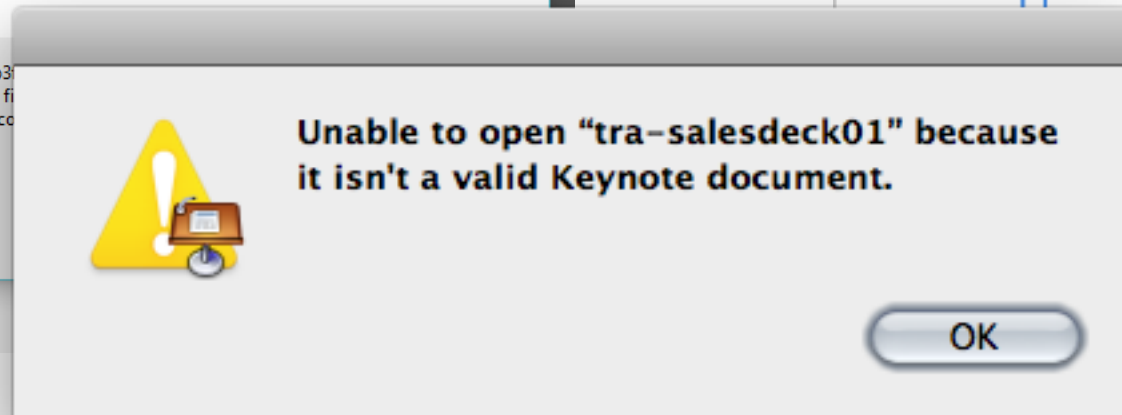
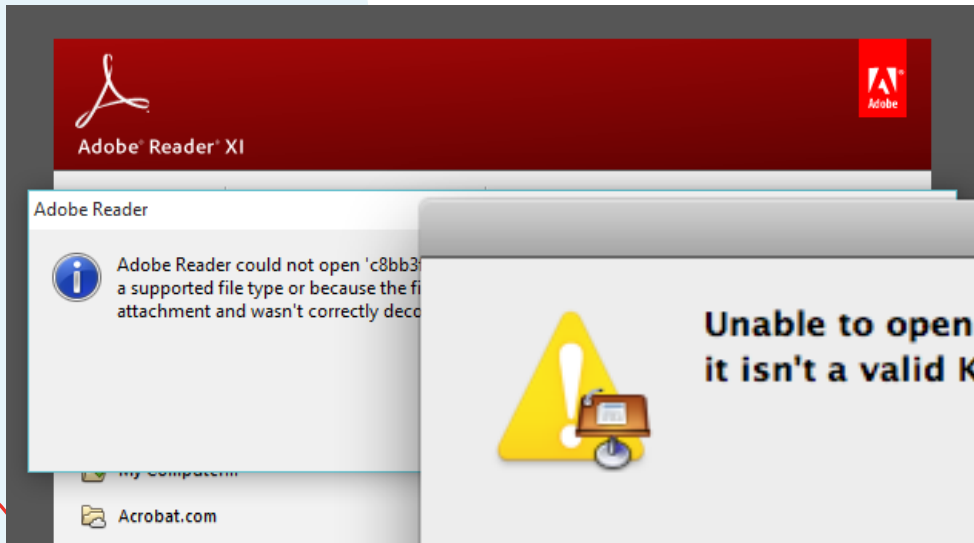
Enhanced curation

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

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



File formats

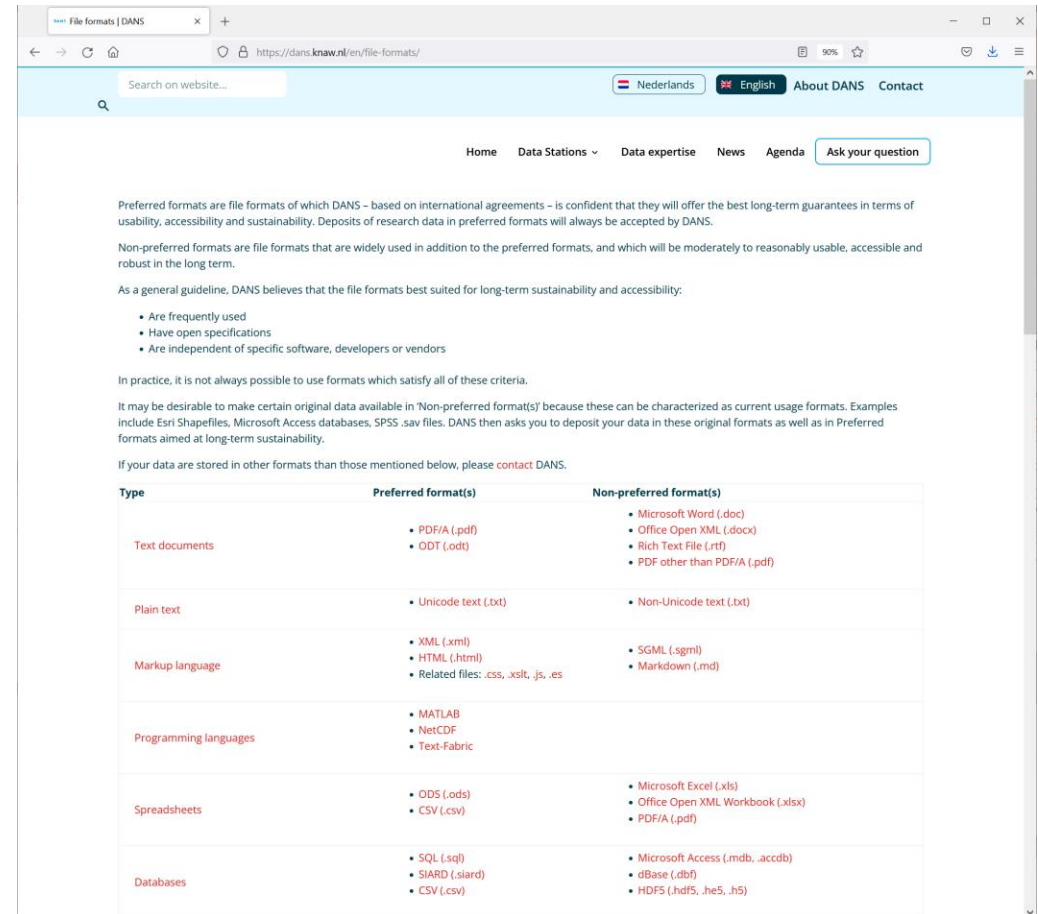


Preferred Formats

- 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>



The screenshot shows the 'File formats' page on the DANS website. The page has a light blue header with a search bar, language selectors for 'Nederlands' and 'English', and links for 'About DANS' and 'Contact'. A navigation menu includes 'Home', 'Data Stations', 'Data expertise', 'News', 'Agenda', and an 'Ask your question' button. The main content area explains the criteria for preferred formats: they must be frequently used, have open specifications, and be independent of specific software. It also lists non-preferred formats. A table summarizes these categories.

Type	Preferred format(s)	Non-preferred format(s)
Text documents	<ul style="list-style-type: none">PDF/A (.pdf)ODT (.odt)	<ul style="list-style-type: none">Microsoft Word (.doc)Office Open XML (.docx)Rich Text File (.rtf)PDF other than PDF/A (.pdf)
Plain text	<ul style="list-style-type: none">Unicode text (.txt)	<ul style="list-style-type: none">Non-Unicode text (.txt)
Markup language	<ul style="list-style-type: none">XML (.xml)HTML (.html)Related files: .css, .xslt, .js, .es	<ul style="list-style-type: none">SGML (.sgml)Markdown (.md)
Programming languages	<ul style="list-style-type: none">MATLABNetCDFText-Fabric	
Spreadsheets	<ul style="list-style-type: none">ODS (.ods)CSV (.csv)	<ul style="list-style-type: none">Microsoft Excel (.xls)Office Open XML Workbook (.xlsx)PDF/A (.pdf)
Databases	<ul style="list-style-type: none">SQL (.sql)SIARD (.siard)CSV (.csv)	<ul style="list-style-type: none">Microsoft Access (.mdb, .accdb)dBase (.dbf)HDF5 (.hdf5, .h5)

Preferred Formats

File formats | DANS

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 more 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 characterised by certain features. Examples include Esri Shapefiles, Microsoft Access databases, SPSS .sav files. DANS then asks you to deposit your data in these or other 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	<ul style="list-style-type: none">• PDF/A (.pdf)• Office Open XML (.docx)• ODT (.odt)	<ul style="list-style-type: none">• Microsoft Word (.doc)• Rich Text File (.rtf)• PDF other than PDF/A
Plain text	<ul style="list-style-type: none">• Unicode text (.txt)	<ul style="list-style-type: none">• Non-Unicode text
Markup language	<ul style="list-style-type: none">• XML (.xml)• HTML (.html)	<ul style="list-style-type: none">• SGML (.sgml)• Markup language other than XML/HTML
Programming languages		
Spreadsheets		
Databases		

Archaeology Data Service

Search data Deposit data Help & guidance News & events Blog About

Help & guidance Guides to Good Practice

DIGITAL ARCHIVING

About these guidelines

How to use these guides

What is digital archiving?

Archival strategies

Open Archival Information System (OAIS)

THE PROJECT LIFECYCLE

Planning for the creation of digital data

Project documentation

Project metadata

Data selection: preservation intervention points

File formats

Kieron Niven, Archaeology Data Service | Digital Antiquity, Guides to Good Practice

File types

The table below outlines some of the common formats used to create spreadsheets and databases and describes their associated applications and potential uses for long term preservation. As with many word processing applications, spreadsheets in particular have seen a move in recent years towards

Lijst voorkeursformaten | Nationaal Archief

https://www.nationaalarchief.nl/archiveren/kennisbank/lijst-voorkeursformaten

Nationaal Archief
Ministerie van Onderwijs, Cultuur en Wetenschap

Winkelwagen Inloggen Over het NA Contact Nederlands Zoeken

Home Onderzoeken Beleven Archiveren

Home > Archiveren > Voorkeursformaten Overheid

Lijst voorkeursformaten

Norm Voorkeursformaten

Informatiesoort	Extensie	Versie en/of profiel	Formaat naam	Status	PUID	Opmerkingen
Afbeelding (roster)	TIFF / TIF	6.0	Tagged Image File Format	Voorkeur	fmt/353	
				Voorkeur	fmt/11 fmt/12 fmt/13	
				Acceptabel	fmt/42 fmt/43 fmt/44	
				Acceptabel	x-fmt/392 fmt/151	
				Acceptabel	fmt/3 fmt/4	
				Voorkeur	fmt/91 fmt/92 fmt/413	

Norm Voorkeursformaten

Duurzame bestandsformaten

Choosing a file format | SND

https://snd.gu.se/en/manage-data/guides/choosing-file-format

SND
Swedish National Data Service

Find Data Manage Data Describe & Share Data News & Events About Us Contact

Choosing a file format

Listen Search on site På svenska

Manage Data

- Plan
- Organise
- Document
- Work with Data
- Prepare and Share
- Guides
- Checklist for Data Management Plan
- Choosing a File Format
- Training Resources

Recommended formats -- UK

https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

UK Data Service

Find data Deposit data Learning Hub Training and events About News Impact Help Contact

Home > Learning Hub > Research data management > Formatting data > Recommended formats

Recommended formats

Search our learning hub

Recommended formats

File formats recommended by the UK Data Service

The table contains guidance on file formats recommended and accepted by the UK Data Service for data sharing, reuse and preservation.

You may need to convert your data files to a preservation file format.

We welcome queries from researchers about appropriate file formats for working and preservation, particularly early in the research process. If you are unsure of the suitability of your file formats for the data you want to deposit with the UK Data Service, please [get in touch](#).

Type of data	Recommended formats	Other acceptable formats
Quantitative tabular data with extensive metadata.	Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta), .sas7bdat.	
A dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data.	Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information. Some structured text or mark-up file containing metadata information, e.g. DDI XML file.	SPSS portable format (.por). MS Access (.mdb/.accdb).

Valentijn's manifesto/DANS' position

- 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

*See: <https://openpreservation.org/blogs/monitoring-disappearing-file-formats-4-dans/>
and <https://doi.org/10.5281/zenodo.15123699>



Do you agree with Valentijn's manifesto?

What remaining questions do you have on file formats?



Tools and services

<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) [↗](#)



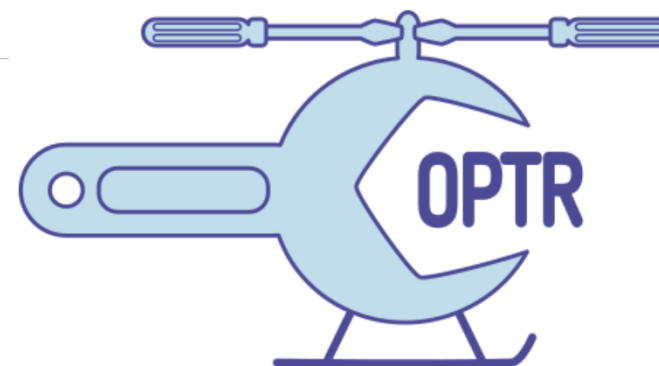
National Digital
Stewardship Alliance
(NDSA) [↗](#)



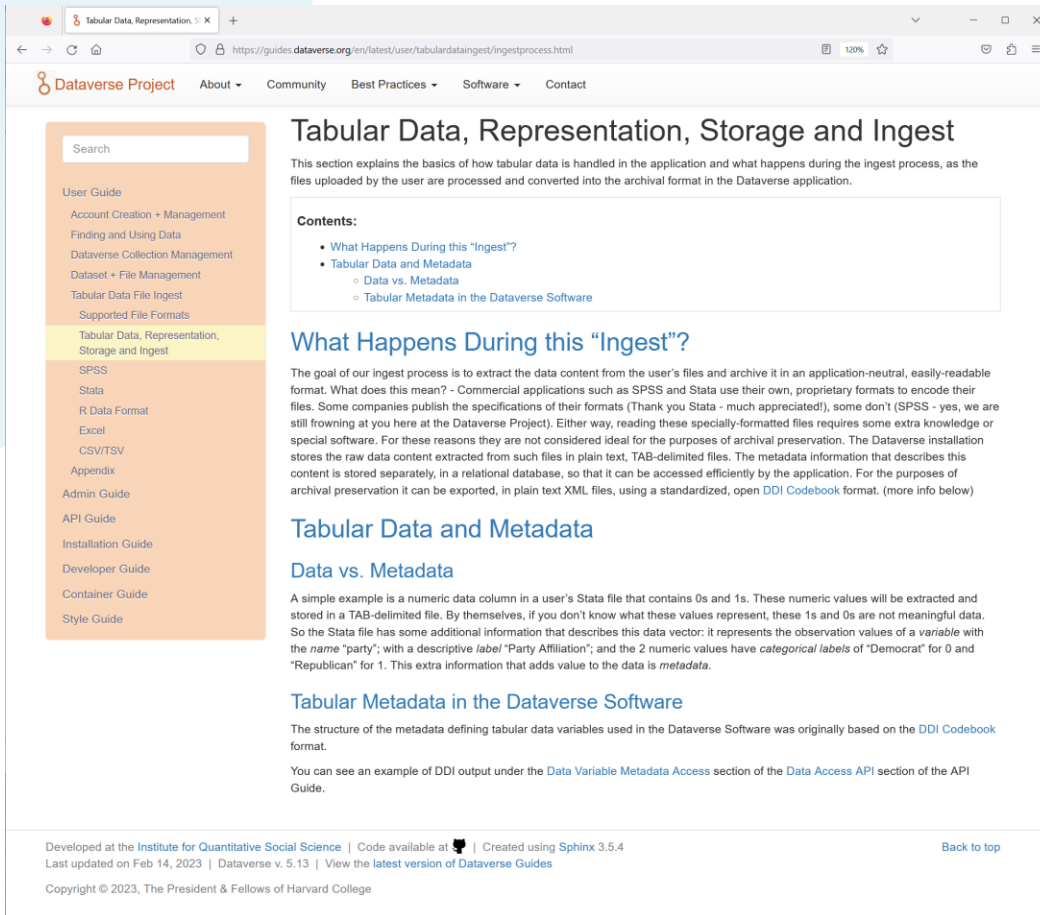
Network of expertise in
long-term storage of
digital resources in
Germany (NESTOR) [↗](#)



The Open
Preservation
Foundation (OPF) [↗](#)



Tools and services



The screenshot shows the Dataverse Project website. The main heading is 'Tabular Data, Representation, Storage and Ingest'. Below it, a search bar and a sidebar with navigation links are visible. The sidebar includes links for 'User Guide', 'Account Creation + Management', 'Finding and Using Data', 'Dataverse Collection Management', 'Dataset + File Management', 'Tabular Data File Ingest', 'Supported File Formats', 'Tabular Data, Representation, Storage and Ingest' (highlighted), 'SPSS', 'Stata', 'R Data Format', 'Excel', 'CSV/TSV', 'Appendix', 'Admin Guide', 'API Guide', 'Installation Guide', 'Developer Guide', 'Container Guide', and 'Style Guide'. The main content area has a 'Contents' section with links to 'What Happens During this "Ingest"?' and 'Tabular Data and Metadata'. Below this is a section titled 'What Happens During this "Ingest"?' which explains the ingest process. Further down is a section titled 'Tabular Data and Metadata' with a subsection 'Data vs. Metadata' and another section 'Tabular Metadata in the Dataverse Software'.

Tabular Data, Representation, Storage and Ingest

This section explains the basics of how tabular data is handled in the application and what happens during the ingest process, as the files uploaded by the user are processed and converted into the archival format in the Dataverse application.

Contents:

- What Happens During this "Ingest"?
- Tabular Data and Metadata
 - Data vs. Metadata
 - Tabular Metadata in the Dataverse Software

What Happens During this "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. What does this mean? - Commercial applications such as SPSS and Stata use their own, proprietary formats to encode their files. Some companies publish the specifications of their formats (Thank you Stata - much appreciated!), some don't (SPSS - yes, we are still frowning at you here at the Dataverse Project). Either way, reading these specially-formatted files requires some extra knowledge or special software. For these reasons they are not considered ideal for the purposes of archival preservation. 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. (more info below)

Tabular Data and Metadata

Data vs. Metadata

A simple example is a numeric data column in a user's Stata file that contains 0s and 1s. These numeric values will be extracted and stored in a TAB-delimited file. By themselves, if you don't know what these values represent, these 1s and 0s are not meaningful data. So the Stata file has some additional information that describes this data vector: it represents the observation values of a variable with the name "party"; with a descriptive label "Party Affiliation"; and the 2 numeric values have categorical labels of "Democrat" for 0 and "Republican" for 1. This extra information that adds value to the data is *metadata*.

Tabular Metadata in the Dataverse Software

The structure of the metadata defining tabular data variables used in the Dataverse Software was originally based on the [DDI Codebook](#) format.

You can see an example of DDI output under the [Data Variable Metadata Access](#) section of the [Data Access API](#) section of the [API Guide](#).

Developed at the [Institute for Quantitative Social Science](#) | Code available at [GitHub](#) | Created using [Sphinx](#) 3.5.4
Last updated on Feb 14, 2023 | Dataverse v. 5.13 | [View the latest version of Dataverse Guides](#)
Copyright © 2023, The President & Fellows of Harvard College

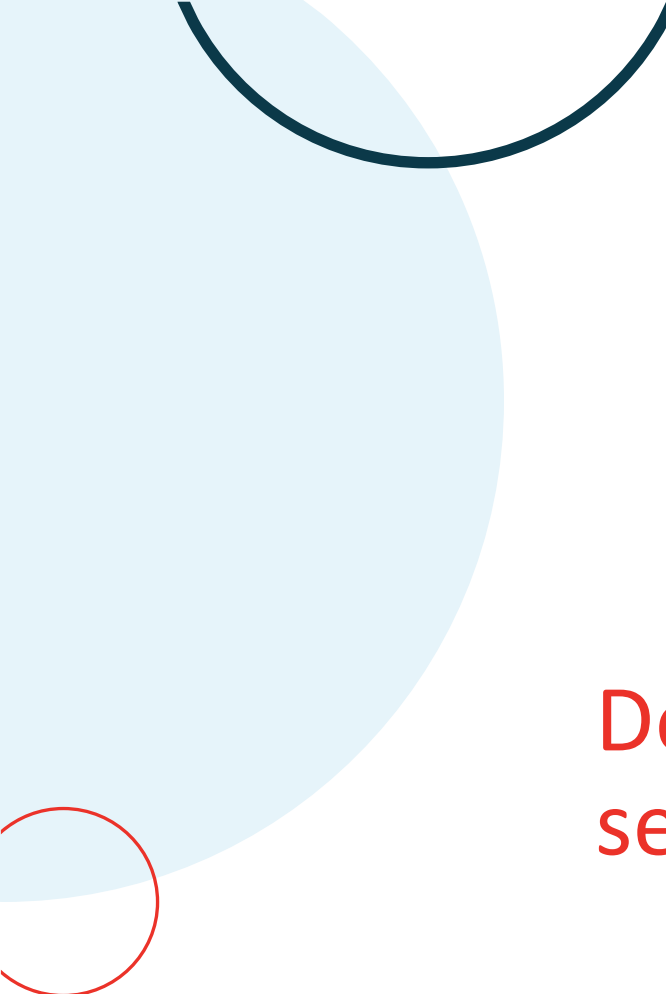
- *The goal of our ingest process is to extract the data content from the user's files and archive it in an application-neutral, easily-readable format.*
- *The Dataverse installation stores the raw data content extracted from such files in plain text, TAB-delimited files. The metadata information that 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.*

OpenRefine?



- Data is often very messy. OpenRefine provides **a set of tools to allow you to identify and amend the messy data.**
- It is **similar to spreadsheet applications**
- However ... **all actions are easily reversed in OpenRefine.**
- If you save your work it will be to a new file. **OpenRefine always uses a copy of your data and *does not* modify your original dataset.**



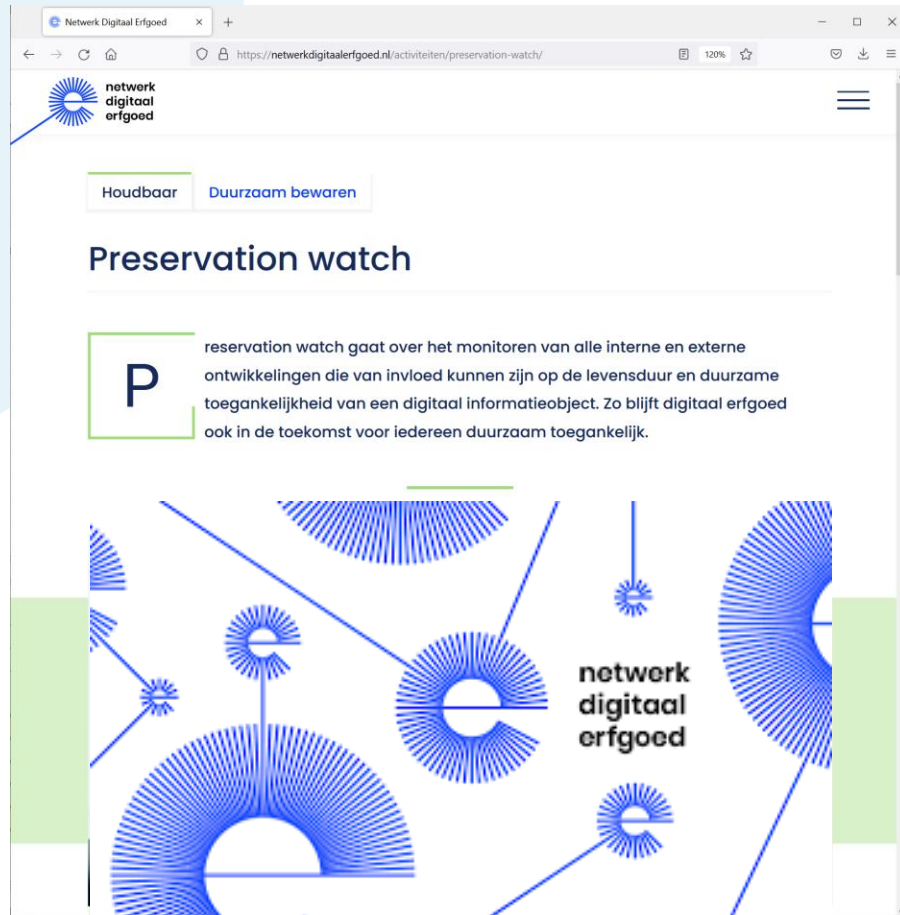
See: Data Carpentries Open Refine Training:
<https://datacarpentry.github.io/openrefine-socialsci/index.html>



Do you know of other tools and
services to share with the group?



Digital Heritage Network



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

Wegwijzer Voorkeursformaten

www.wegwijzervoorkeursformaten.nl

Stappenplan

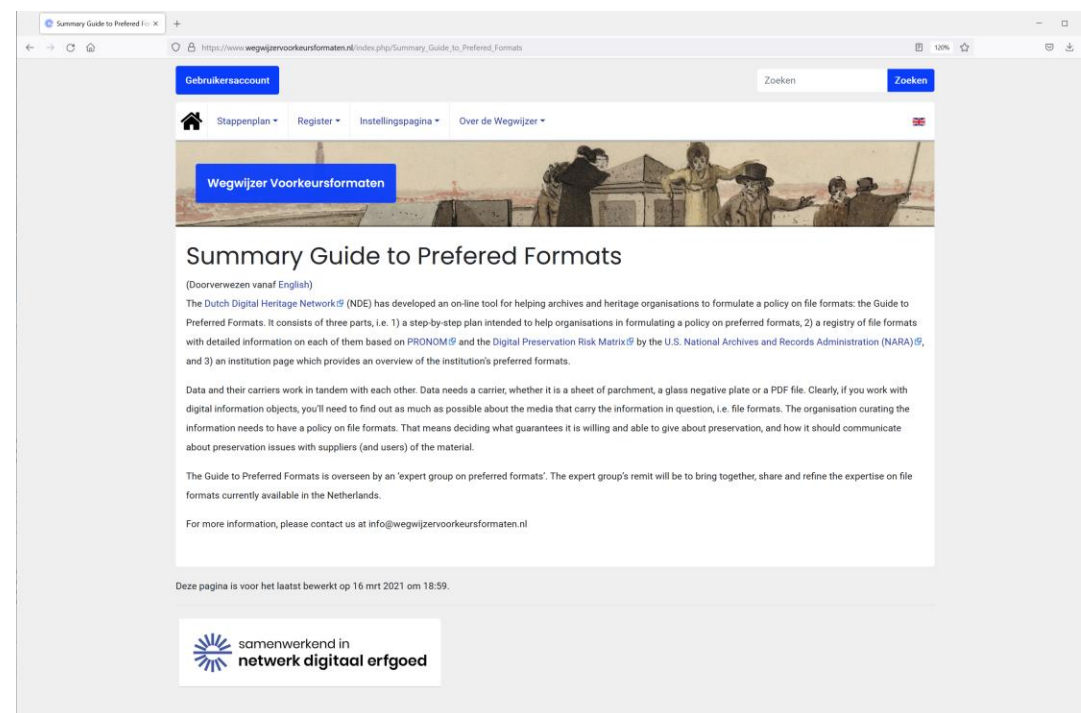
Interactive **step-by-step module** to help institutions create their own file formats policies through comparisons with other policies, via available information and weighing importance

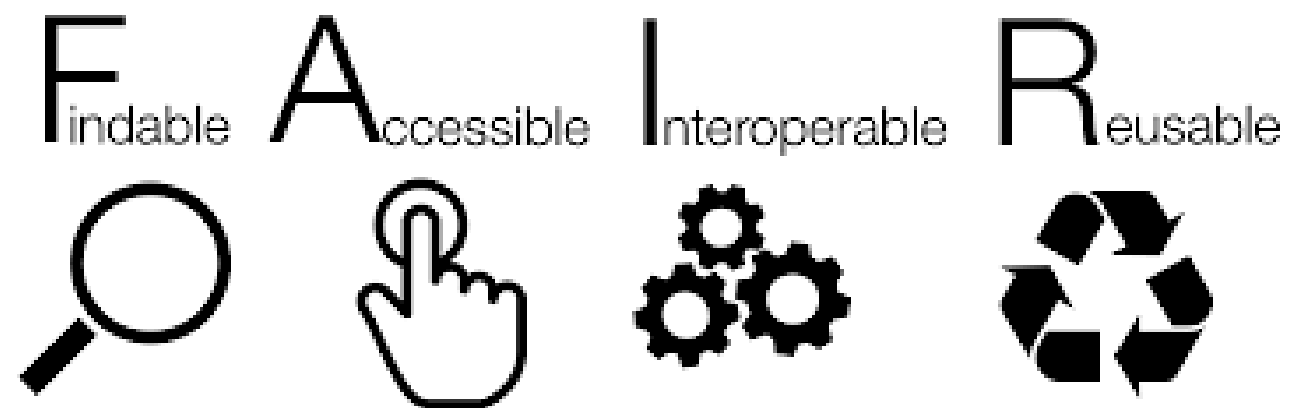
Register

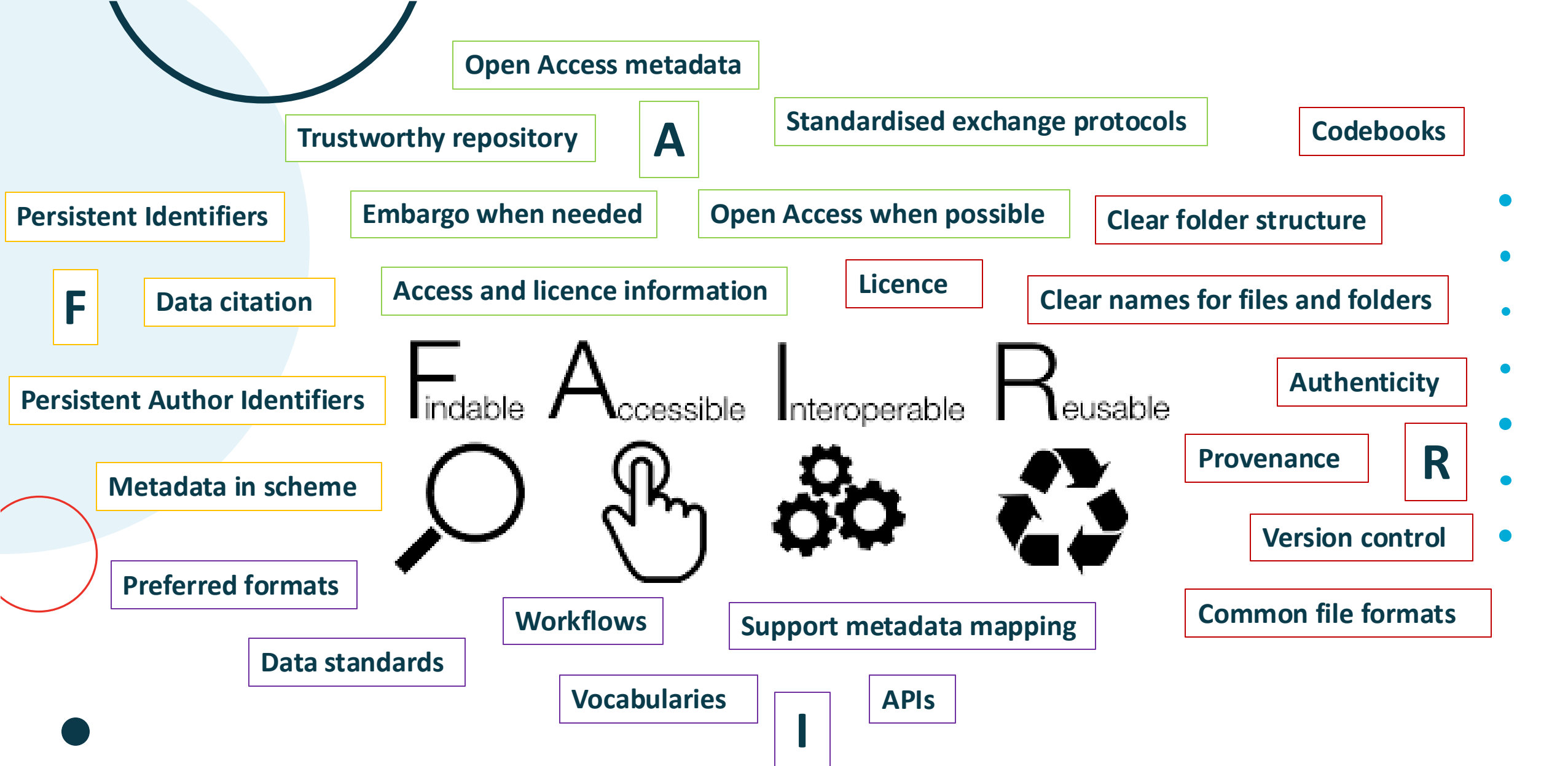
Register of file formats with information derived from international sources: PRONOM; Wikidata; Library of Congress; NARA; COPTR

Instellingspagina

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











Quick-fire PostIt notes: Take-home messages

Take a PostIt note

Answer the following questions on a PostIt and put it up on the sheet

What have you learned in this workshop that you can take into your daily work?





Quick-fire PostIt notes: Take-home messages

Take a PostIt note

Answer the following questions on a PostIt and
put it up on Sheet 3:

What do you still need to know /
learn?



Thank you!

More information

Visit our website **www.dans.knaw.nl**

And follow us online



Mastodon @DANS_knaw_nwo



LinkedIn @DANS



X @DANS_knaw_nwo

