DANS Data Stations –

Data Processing Team Manual

Handling personal data

The following agreements apply to the Data Processing Team for processing personal data:

Limit access to data during data processing

For the purposes that require personal data processing, access should only be granted to the people directly responsible for the processing of the data.

Measure: DANS uses an authorisation structure that establishes roles and permissions. Access will always be granted following this structure. See Appendix 1.

Keeping data up to date (correctness – 'Accuracy')

As per the Terms of Use, DANS may correct metadata and convert files for the purposes of sustainable archiving and/or following the wishes of the end user. It is essential that personal data is not accidentally edited during this process.

Moreover, the Terms of Use state that DANS may not change the content of files. The anonymisation of data that may, in DANS's opinion, have been accidentally included in the files is therefore not allowed as per the above stipulation.

Measure:

- When DANS carries out data conversions, it will only do so in regard to the file format to ensure the long-term sustainability and/or accessibility of the dataset. The data manager will check whether the contents have remained unaltered by opening the converted files afterwards (or using a representative sampling for large file quantities).
- Editing the content of data files is not a part of the procedure; the focus is on form and file format. In exceptional cases, DANS can alter the contents of a data file if a depositor/claimant gives permission to do so. The file may then only be changed in the agreed manner.

Data preservation and removal

Data may not be stored longer than necessary. This also reduces the chances of a data breach occurring. The Data Processing Team stores data on different media.

Local:

- Laptops/PCs: (locally managed)
- DANS server: (locally managed)
- Microsoft Office (locally managed)
- Open Office (locally managed)
- Libre Office (locally managed)
- StatTransfer (locally managed on a licensing basis)
- Python scripts (locally managed)
- FFmpeg (locally managed)
- Adobe Suite (locally managed)

External:

- Vancis (as per Terms of Use, indefinitely where possible)
- Microsoft 365 (in the future): as per KNAW's retention policy (to be drafted)
- SURFfilesender: SURF adopts a standard retention period of 4 months

Measure

- For datasets containing personal data, local data storage facilities must be cleaned up as soon as possible and no later than one month after dataset publication.
- SURFfilesender: data is automatically removed by SURF after four months. This retention period is appropriate for DANS, giving them sufficient time to download the data.

Adequate protection regarding the use of SURFfilesender

For sending personal data and other restricted data of a sensitive nature, the encryption option within SURFfilesender must be used.

Measure: When requesting data via SURFfilesender, the Data Manager will always point out to the depositor the extra protection measure through encryption when sending personal data.

Deposition

Steps that the client undertakes for depositing a dataset: the curator must know this process well. See also the Data Deposition Manual.

- The client logs in
- The client clicks on 'Add Data' => 'New Dataset'
- A form appears for filling in the metadata. This form is based on a 'template' - the Host Dataverse and the Dataset Template used are displayed at the top of the form. DANS has set a standard template as a default for each Data Station.
- The client describes the dataset using the metadata fields and has the option of adding files.
- The client indicates whether the dataset contains personal data when filling out the form ('Personal Data in Dataset?') using a selection menu with the following options: Yes/No/Unknown. This selection is checked by the curator (see Curation). The depositor remains ultimately responsible for the dataset.
- After selecting the 'Save Dataset' option, the client can choose a different licence using the 'Terms' => 'Edit Terms Requirements' tab.
- After selecting the 'Save Dataset' option, the client can add descriptions to each uploaded file or adjust accessibility via the File Options menu (the three dots to the right of the file). This menu includes an option to 'Restrict' the file, with or without 'Enable access request'. If the client deselects the 'Enable access request' option, the client must supply the *Terms of Access for Restricted Files.* DANS promotes the reuse of research data; selecting the option for working with 'access requests' is standard procedure. Deviating from this option is only allowed in consultation with and after approval from the Data Station Manager.
- The File Options menu also has an option to set an embargo on files. This can be set separately per file. DANS recommends setting a maximum embargo of two years, but this is not technically enforced.
- After selecting 'Save Dataset', the client can edit the metadata (Add + Edit Metadata) and/or upload additional files (Upload Files).
- The client submits the dataset using the 'Submit for review' button. After this, the client is no longer able to make adjustments to the dataset.

Points to note:

- A dataset can be added to a collection by selecting the right collection via autofill under 'Relation Metadata' - 'Collection' in the metadata. This functionality was not available in EASY: it replaces the *thematic collection* functionality (these were usually empty datasets with an HTML front page with lists of collections and links to associated datasets). It is something that we will need to point out to clients, that will take some getting used to, and that we need to continue to

be aware of during curation: are the datasets already assigned to the right collection, or do we still need to do that?

- When adding files, each file (or all files by selecting them) can be set as 'Restricted' or left unrestricted via the 'Edit' option. Edit gives the following options: Delete/Restrict/Unrestrict. If you choose the Restrict option, you must indicate separately whether the possibility of an access request should be in place for the file: 'Request Access' - 'Enable access request' (tick box).
  Restricted data without this option is effectively Closed Access: the files are visible but cannot be downloaded. It is *not* possible to make files invisible/hide files in Dataverse (visibility:NONE, as in EASY).
- When multiple files are set to Restricted, it is *not* possible to make any part of these files accessible via an access request and keep others closed. **When the option 'Enable access request' is selected for even just one Restricted file, this option is selected for *all* Restricted files in *all* dataset versions.**
  Therefore, data managers should always thoroughly check whether the correct options have been selected. Are the files that should be set to Restricted Access truly set to Restricted, and can they be requested? Restricted Access data should be available upon request as a standard. For exceptions where making specific Restricted Access files not available upon request is desirable (informed-consent files, for instance), see *Curation - Restricted Access.*
- It won't be immediately apparent to everyone how a dataset licence must be adjusted and that this needs to be set *after* 'Save metadata' under 'Terms'. Data managers should be aware of this so they can offer the proper guidance and check the licence selected.

Curation

See also the Data Stations Policy. The data managers appraise and provide basic curation for all manually deposited datasets submitted for review. After logging in to the Data Station, the data manager can view these datasets by applying the **Publication Status - In Review** filter. All datasets that are 'in review' are also assigned the 'Draft' status until DANS publishes the dataset. Entirely new depositions are also assigned the 'Unpublished' status; submitted versions of existing datasets are not set to the *Unpublished* status. These must be reviewed, curated, and published by the data manager as well.
Datasets with the 'Draft' status but *not* the 'In Review' status are still being processed by the depositor. They must submit the dataset before the data manager can start reviewing the data.

Inventory

Each Data Station has internal overviews available for registering datasets.

Appraisal

A dataset must be appraised on its general content to determine whether the dataset can be accepted for archiving. See also the DANS selection policy.

Scientific research

A dataset should be the result of scientific research or research for policymaking. This appraisal is, in principle, carried out by the data managers. In case of doubt, consultation with the Data Processing Team Leader is required.

New content

A situation may arise in which the depositor does not realise that a dataset has already been archived at DANS, but the data manager becomes aware that this is the case. A second, separate dataset deposition with identical data will be refused. However, creating a new version of an existing dataset is possible.

Licence

Check the licence under Licence/Data Use Agreement. If it concerns a CC0 licence, check the corresponding inventory spreadsheet to see whether this is a standard option within the organisation. Consult with the depositor if you're not sure. Depositors are not always aware of how to convert a licence, and some depositors may have forgotten to do so.

If the dataset contains Restricted Files, the Terms tab will indicate the Terms of Access and whether the option for using Access Requests is selected. If the files are restricted and the terms state that 'Users may **not** request access to files,' check to see whether a good reason has been given under 'Terms of Access for Restricted Files', and consult, if needed, with the Data Station Manager about whether or not this is acceptable.

Become at least somewhat familiar with the dataset subject and contents by viewing the descriptions and checking the documents and data files that may potentially contain personal data. If there is a suspicion that the dataset may contain personal data but the depositor has chosen an open licence, contact the depositor to check and make sure.

Basic Curation (Metadata)

The metadata may be edited before publication by navigating to the Metadata tab and clicking the 'Add + Edit Metadata' button.

- Check to make sure the entered metadata is complete and correct. Minor adjustments can be made if this benefits the dataset's FAIRness.

Citation Metadata

- Check to make sure all the **Authors** and co-authors have been filled in and add any missing details.
- The names must be recorded as follows: **initials, surname prefix, surname**
- Check ORCIDs and other Author Identifiers: they should not be entered as links (https://orcid.org/ID) but solely as ID.
- Check the **Description:** Can another researcher understand what the dataset is about, given this description? Only make adjustments (or check with the depositor) if the description is too short or too lengthy. It is customary to place the report summary or conclusions in this field; the aim should be a succinct description.
- Does the text in the **Description** contain hard returns? If it does, you may clean this up: copy the text to an application such as Notepad++, remove the hard returns, and copy the text back into the description.
- Note the type of research and conclusions mentioned in the **Description** field: keywords or standardised terms from vocabularies (*Archaeology-Specific Metadata*; *Social Sciences and Humanities Metadata*) may be extracted from this description.
- Has a **Date** been filled in under the Description? That should only be the case if there are multiple descriptions. If not, this can be removed (or moved to Production Date).
- Has the **Language** been filled in? If not, fill it in. This concerns the language of the data.
- Has the **Production Date** been filled in? If not, fill it in. This date can be the same day as the final publication/dataset delivery date.
- Has a **Distribution Date** been entered that is set for the future? If so, then this is probably meant as an embargo date. However, the file settings are not automatically adjusted accordingly. In such instances, check whether an embargo is set on the files that matches the Distribution Date.

Rights Metadata

- Has the depositor indicated that the dataset contains personal data? If so, is that correct? Does it concern personal data as outlined in the GDPR? See also the subheading *Personal data* regarding this item. Names of authors, collaborators and clients are considered to be 'bibliographic data' and can be included in the metadata; that is not a reason to mark this field with 'Yes'. If the dataset does contain personal data, consult with the legal counsel.

- If the depositor has not filled in a language under 'Language of Metadata', add the language. This concerns the metadata language; if multiple languages have been used in the description, you can add multiple languages.

Relation Metadata

- Check whether the dataset has been assigned to the proper **Collection.** If not, adjust this accordingly.
- If you notice or discover any related datasets, you can indicate these in the **Relation** field.

Temporal and Spatial Coverage

- Make sure that, where applicable, spatial coordinates are marked in the metadata under **Spatial Point** (central coordinate) or **Spatial Box** (N-S-E-W boundaries). For most archaeological projects, coordinates are recorded using the Dutch Rijksdriehoekstelsel coordinate system (scheme = RD), written out in full to the metre, without decimals. X has a value between 0 and 300000, and Y has a value between 300000 and 600000.
- If Latitude/Longitude coordinates are submitted, note that the form first asks for an X coordinate followed by a Y coordinate. However, that's not how Lat/Lon is usually annotated: Latitude = Y and Longitude = X.
- The **Spatial Coverage (free text)** field is the preferred field for submitting location data. Each location aspect should have a separate field (+ for repeating fields): toponym (street; placename)/town/municipality/province. We enter the word Municipality before the municipality name to separate municipalities from cities.

Archaeology-Specific Metadata

- Important for archaeology: Ensure that the **Archis Zaakindentificatie** (ZaakID, or case ID number) has been filled in. Most depositions will have a ZaakID assigned. All archaeological research is registered in the Archis system of the RCE (Dutch Cultural Heritage Agency) and is assigned such a number. The current Archis 3 system (since approx. 2015) has a 10-digit number. Older research projects may have the old Archis research report number, which usually has 5 digits (sometimes 4).
- The **Archis ID** field is where identifiers other than the Zaakidentificatie from the Archis system may be entered, with values accompanied by the corresponding type. This could concern an observation, a discovery report, a monument, or a research project. ('Onderzoeksnummer' – Research number – is different to a 'Onderzoeksmeldingsnummer' – Research report number: in Archis 2, when you reported research, you were given a research report number; when the research was completed and signed off, you would get a separate research number).
- Check the use of the ABR vocabulary fields, and complete where necessary: **Report (ABR Rapporten)** with the associated **Report number; Methods of Recovery (ABR Verwervingswijzen)** (see the report subtitle or the beginning of the summary/Description);

Subject (ABR Complextypen) / Artefact (ABR Artefacten) / Temporal (ABR Periodes) - only if anything has been ascertained with certainty; see the conclusions/Description for this.

Social Sciences metadata

- Check the subject classifications with one (or multiple) terms from the European Language Social Science Thesaurus (ELSST) and a term from the CESSDA topic classification, a list of topics drawn up by CESSDA.
  If these fields have been left empty, but the provided description or supplied keywords make it clear which terms from the ELSST or CESSDA Topic Classification apply, add them.

- If you have any questions, contact the depositor (Contact Owner).

OH-Smart

OH-Smart is a separate deposit module for Oral History datasets. Depositions for OH-Smart datasets are submitted to the Data Station Social Sciences & Humanities. Datasets deposited with the OH-Smart deposit module contain five files with OH-specific metadata that cannot be integrated into the DS metadata: 1 JSON-LD file with all metadata, 1 JSON-LD file with all restricted metadata, 1 JSON-LD file with all public metadata, 1 TXT file with all metadata, 1 TXT file with all public metadata. The datasets are easily identifiable through these metadata files. For these datasets, the following important points apply:

- Three metadata fields cannot be transferred through SWORD: subtitle, grant agency, grant ID - these are registered but need to be entered manually into the DS metadata.
- Ensure that the files that should have restricted access do have restricted access.

Files

- Check file availability: a green lock on the file icon means the file has been set as Restricted. For multiple files, use the filter to check which files have been set as Restricted. Does this correspond to the selected or intended licence? The files must be marked as Restricted under the DANS Licence licence option. ‘Access Request’ should also be enabled; otherwise, no one can request access to the files. Note: each file can be individually set to Restricted or not, but the ‘Access Request enabled’ toggle affects all the dataset files that are set as Restricted.

Data managers may not alter any files before publication unless the depositor expressly requests them to do so. If something needs to be adjusted at a file level, the data manager will return the dataset: ‘Return to Author’.

Documentation file completeness

The dataset must have the necessary documentation to make reuse possible by third parties. This concerns two types of documentation: the research documentation and the data documentation.

Research documentation

The research documentation pertains to the research, including any tools or instruments used and the documentation on the files included in the dataset. The latter is only required for datasets with large numbers of files or a complex structure. The documentation is essential for the correct interpretation of the dataset.

The following must be present in the research documentation:

- General information about the context of the research
- Information about the research setup and data collection method
- (Where applicable) instruments and tools used
- (Where required) Information about the files and/or structure: number of files, relations between files, etc. This is supplied in the form of a file list

A (published) article can be submitted as research documentation, but this can also be an informal document such as a readme file. The documentation can consist of one or more files.

The documentation must be complete in terms of content (for example, a report with all the required chapters and appendices) and final (no draft versions).

The documentation must be included in the dataset as files. This allows the documentation to be archived sustainably along with the dataset.

Data documentation

Data documentation is the documentation at the data level. Examples include an explanation of the terms and variables used in a data file. This explanation can be included in a data file (as labels in an SPSS file, for instance) or submitted separately. This documentation must be present to interpret all the values and codes correctly.

Incorrect or missing documentation files

A data manager contacts the depositor to add or replace the required documentation or have incorrect information removed. The data manager may also do this after consultation with the depositor.

Data file completeness

The number of data files must be complete. All the files referred to in the documentation must be present (for example, files A through D must be present in a dataset where the documentation refers to files A through D). If this is not possible, a clear explanation must be included. The depositor will be contacted if there are any questions or doubts.

Incorrect or missing data files

A data manager contacts the depositor to arrange the necessary additions or adjustments before the dataset is published.

Submissions outside the Data Station

The data manager may complete the dataset after consultation with the depositor. In that case, the depositor will submit the necessary files outside the Data Station. Small files can be attached as an email attachment; SURFfilesender is more suitable for larger files or large file quantities.

File curation and enhanced curation

- Check whether material has been submitted in file formats that are not set by DANS as a preferred format as per our guidelines. If this is the case, check whether we can convert them ourselves *(see separate protocols - currently found in the internal documents for Data Processing Team Student Assistants)* or whether we should contact the depositor to discuss the data format.
- If a large number of files are submitted, check to see if they have been clearly structured. If not, we can decide to adopt a (better) folder structure.
- If the research data contains personal data or looks like it may contain personal data, check whether the metadata, the licence and the accessibility options reflect this. If you have any questions, contact the depositor. More generally, contact the depositor if you suspect personal data may unintentionally be included in the dataset (metadata and files). This does not need to be actively monitored; it concerns situations where suspicions may spontaneously arise.
- Check whether all submitted data is relevant: no draft versions, temporary working files, system files, etc.
- Check to make sure that no ZIP files or other compressed files have been submitted. We prefer not to have ZIP files in the archive, only as a dissemination format, if this is the best form of dissemination. In that case, we also want to include the uncompressed data in the dataset.
- Check that the file names are correct and that they do not contain any personal data.

If it is deemed necessary to convert files to preferred formats, the new formats may *not* be added to the deposited version of the dataset. **In that case, the dataset must first be published, after which we create a *new version.***

Adding converted files in preferred formats or adjusting the dataset structure is called **Enhanced Curation.**

Indicate in the inventory whether Enhanced Curation is deemed necessary.

Dataset restructuring can be done by selecting the files under the Files tab and then editing the 'File Path' under Edit Files => Metadata. If the dataset contains so many files that it's not efficient to adjust this via the interface, the dataset may be downloaded and structured locally in folders. A new version of the dataset may be created by removing the current data and re-uploading the data with the new folder structure by compressing the structured dataset in a ZIP file and uploading the ZIP file. The ZIP file is automatically extracted.

When we convert files to preferred formats, we download the original data locally and convert the data as per protocol *(see separate protocols - currently found in the internal documents for Data Processing Team Student Assistants).* Create a new dataset version by removing the original file and uploading the converted file.

NOTE!! When adding new data, please double-check the file accessibility settings for the original Version 1 dataset files. All 'Restricted' data must be set as Restricted in the uploaded new data. For all data that falls under an embargo, the embargo must be re-set for the newly uploaded files.

Publication

Publish the dataset using the 'Publish Dataset' button.

Any changes made to a published dataset will result in a new version.
Dataverse works with minor version increments (like version 1.1) and major version increments (like version 2). A superuser (see the authorisation matrix in Appendix 1) can make changes and overwrite the dataset without creating a new version.
Our standard:
- Published dataset, changes to the metadata = Minor revision
- Published dataset, changes to the files = Major revision
- Changes to the licence or data accessibility = Major revision. Please note that with such changes, earlier versions of the dataset must not be left with undesirable settings. If that's the case, the earlier versions must be marked as *deaccessioned.*

Depositors may submit new versions where the data is more open than previous versions. The other way around works differently: in theory, a more restricted licence may not be selected after publication

(you cannot retroactively impose additional restrictions on open use) and making data less accessible conflicts with DANS' objective of making research data reusable. Such requests must be discussed first.

Whenever a depositor submits a new version, check whether they have indicated that it concerns a Major revision when it concerns changes in the data or the licence. If not, the new version must be rejected.

Datasets with Restricted Files

When Access Request is enabled

Because DANS publishes the dataset, any Access Requests will also be sent to DANS. By assigning the depositor the additional role with the permission 'ManageDatasetPermissions' enabled, the client also receives the Access Request and can approve or reject the request accordingly.
We may only do this when it has been established that the client is authorised to process Access Requests.

When Access Request is enabled, but specific files may not be published under any circumstances

As mentioned earlier, Dataverse cannot process a combination of Restricted files where some files can be requested and others cannot. When the option 'Enable access request' is selected for even just one Restricted file, this option is selected for *all* Restricted files in *all* dataset versions.
If specific files may not be requested under any circumstances (for example, when an original file contains personal data and only anonymised file versions may be requested), then the first version of the dataset must be set to Deaccessioned.
Edit Dataset => Deaccession Dataset => choose the first version. Standard reason: 'Legal issue or Data Usage Agreement'.

Accounts and rights

Sometimes, an account must be given rights to a dataset. For instance, when someone is not the depositor but is authorised to submit a new version of the dataset. Or when a dataset is created on an expired depositor account, and the depositor wishes to gain access to the dataset via a new account. Or when an additional person needs to start reviewing access requests for Restricted Data.
We can assign these permissions by navigating to Edit dataset=>Permissions=>Dataset, Assign Roles to Users/Groups, look up an account and assign them 'Contributor with ManageFilePermissions' rights.
If it concerns a situation in which the dataset is 'taken over' from the original depositor by a new account, also check that the correct account is linked in the *Dataset Contact* field and edit where needed.

Policy differences between Data Stations and EASY

Several differences exist between the policies and contracts governing the Data Stations and EASY. The contract differences are outlined here on the website.

Datasets submitted to EASY will continue to follow EASY's terms of use unless the depositors explicitly wish to transfer this to the Data Stations' terms of use. We're currently reviewing whether the latter is possible.

Restricted access is required for datasets in EASY that contain personal data; therefore, if transferred, the same conditions must be maintained and this setting may not be altered.

APPENDIX 1 Data Stations authorisation matrix

Role \ Permission (per object)	Anonymous visitor	Logged-in visitor	Data manager	Functional manager	Application manager	IT support manager
View metadata (published datasets)	X	X	X	X	X	X
Create and submit dataset		X	X	X	X	
Publish dataset			X	X	X	
Overwrite dataset version				X	X	
Deaccession dataset			X	X	X	
Download non-restricted files (published datasets)	X	X	X	X	X	
Download restricted files (all datasets)		X*	X	X	X	
Grant access to restricted files (self-deposited datasets)		X				
Grant access to restricted files (all datasets)			X	X	X	
Deaccession (dataset (version))			X	X	X	
View account details (all accounts)				X	X	
Secure Shell access with limited rights (data station server)			X		X	X
Secure Shell access with unlimited rights (data station server)						X

*After access is granted by an authorised user (usually following an access request, but can also be granted proactively)