Leiden University Faculty of Archaeology

Research Data Management Handbook

Wouter Kool

Version 0.1

Contents

This Handbook is meant as a reference for researchers at the Faculty of Archaeology as to guidelines and best practices. In the first chapter, **Tasks**, the Research Data Management Process is described at high level. For detail references are made to the chapter **Themes, Tools, Regulations** and **Documents.** These can also be used as a reference in case of questions about specific subjects.

Tasks

Application

- Make sure you are familiar with **regulatory context** you are working in. -> **Regulatory.**
- If your research is carried out as part of a larger project, familiarize with those demands as well.
- Fill out Data Management questions on the **submission form**.
- Before doing this, familiarize yourself with the information in **Themes** and **Tooling**
- **Always fill out the DPIA** form -> **Themes/Ethics, Themes/Personal Data**
- The information you submit may be **rough and preliminary**. -> **Themes / Planning**
- **Consult the Faculty Data Steward** about your draft (well in advance of the submission date).
- Submit the final version of the DMP

Starting-up

- Write a final Data Management Plan (DMP) at the latest **two months after starting** the research project.
- Plan (coarsely) which part of your dataset to archive. Which data is eligible for reuse and which is not? -> **Themes / Selection**
- Re-evaluate the information in **Themes** and **Tooling**
- Provide a more concrete description of your activities: **who is going to do what, when/how often etc. ->** **Themes / Planning**
- Use the template provided by your funding agency, or else use the Leiden University template. -> **Documents**

Executing the plan

- Most importantly: **keep your promises!** Your DMP is like a contract.
- Make sure the activities in your DMP (backups…) are properly carried-out.
- **Your DMP is a living document.** Periodically review your DMP itself, to see if it still a applies, or needs to be updated.
- Be aware of where personal data is stored, who has access and whether it is anonymized/or pseudononymized in the promised way. Report data leaks. -> **Themes / Sensitive and Personal Data**
- If you have a finalized version of data for a sub-project, set the files aside, and already make a start with documenting them.
- **Keep your supervisor in the loop.** He/she is responsible for your actions and possible consequences.

Preparing for deposit

During the final phase of the project, you will be busy analyzing the, already finalized, data and writing. You can make a start selecting and documenting the data to be archived.

- Well before your contract ends, make agreements with your supervisor about which part of your dataset you are archiving.
- **Do a final review of your DMP**: which versions of the files did you plan to archive/make available? Is this still applicable. If not, change the DMP or adjust your activities.
- Make decisions about **access rights**. Which data is restricted for privacy reasons, copyrights, database rights, etc.? Do you set an embargo, for how long and why?
- Select which files contain the data to be archived. Criteria to take into account are described in -> **Themes / Selection**
- Review the metadata you added to the files. Is this really enough for a researcher outside your project. If there are any questions, get in touch with the Data Steward. -> **Themes / Metadata**

Depositing

- Publish your dataset in the **EASY archive of KNAW/DANS**, according to the metadata standards for eDNA.
- Make sure to **link the dataset to the relevant publication(s)** using a persistent ID.
- **Upload** the files that form your dataset, containing also the file lists and codebooks.
- **Send the persistent identifier to the data steward**, so that it can be included in reporting.

Finishing your research

- Delete identifying information after the retention period. -> **Regulatory / Retention periods**
- Submit the non-publishable part of your data to the faculty “data vault”. -> Regulatory / Data vault

Regulatory

VSNU Code of conduct

The VSNU code of conduct expects you to:

- “As necessary, describe how the collected research data are organized and classified so that they can be verified and reused. “
- “As far as possible, make research findings and research data public subsequent to completion of the research. If this is not possible, establish valid reasons[12] for their non-disclosure”

https://www.vsnu.nl/files/documents/Netherlands%20Code%20of%20Conduct%20for%20Research%20Integrity%202018.pdf

Leiden Data Management Regulation

https://www.library.universiteitleiden.nl/binaries/content/assets/ul2ub/research--publish/research-data-management-regulations-leiden-university_def.pdf

Leiden Policy on Personal Data

https://www.medewerkers.universiteitleiden.nl/ict/privacy-en-gegevensbescherming/algemene-verordening-gegevensbescherming-avg/procedures-privacy/archeologie/fda-bestuur-bureau?protected=true&cf=archeologie&cd=fda-bestuur-bureau

Faculty of Archaeology Data Management Guidelines

Research Ethics

https://www.organisatiegids.universiteitleiden.nl/en/faculties-and-institutes/archaeology/committees/ethics-committee?_ga=2.87050000.940610080.1581581781-1251370249.1571748397

EDNA

Since 2007, archaeologists in the Netherlands have been formally obliged to deposit their data via DANS in accordance with the Dutch Archeology Quality Standard (KNA). The data is stored in EASY, the online archiving system of DANS.

https://dans.knaw.nl/nl/over/diensten/easy/edna

DANS guidelines for archaeological data

Data deposited to eDNA has to be accompanied by the following metadata:

https://www.itc.nl/library/research/data-metadata-preparation-v2.pdf

Personal Data

AVG/GDPR

https://www.autoriteitpersoonsgegevens.nl/nl/over-privacy/wetten/algemene-verordening-gegevensbescherming-avg

Data protection laws of the world

DLA Piper's Data Protection Laws of the World Handbook: https://www.dlapiperdataprotection.com

DNA-material

Het verwerken van modern genetisch materiaal valt onder het Nagoya-protocol.
https://www.nvwa.nl/onderwerpen/nagoya-protocol

Anthropology

Dutch Anthropological Association

Indigeous peoples

United Nations Declaration on the Rights of Indigenous Peoples

https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html

Themes

Planning

When doing Data Management planning, think about the issues below. While your project continues, think about the issues more concretely. After all, it is research. For instance, at first, you do not know exactly how much data you will produce, or you might change to other tooling later. However, it is important that you get a realistic sense of the **necessary activities, effort and cost** involved with your plans! For instance, in the application phase your estimates may be broad, but when your research becomes more concrete, you can be more specific. **Planning is an iterative process**, therefore, your DMP is a living document.

Activity Planning

- Plan what needs to be done, as specific as possible

Resource Planning

- Storage:
    - Estimate the size of your data set. Take into account the information under **Themes / File formats.**
    - Estimate tooling to use. Take into account the information under **Themes / Storage** en **Tools / Storage.**
    - Determine how you retrieve the relevant data. What searching capabilities would you need.
    - What storage facility will you use after the project?
    - This not necessarily only include digital data. Also take into account where and how you store ***physical data.***
- Software:

    - Determine whether you need specific software, and its licensing conditions and fees.
    - Take into account the information under **Tools**
    - Also think whether you would need bespoke software development, for instance to create a website to reach out to the general public.
    - Determine whether you store personal data and whether a tool is fit for this and whether a processing agreement is needed -> **Themes / Personal data**
- Staff
    - Time needed and staff allocated for Data Management
    - IT-services and support needed, including storage.
    - Staff responsible for backups, version management, backups etc.

Budget planning

- Based on the above determine proper costs. Determine if there are costs resulting from specific recourses and activities.

Open Science and Archiving

The FdA's policy on Open Science is "**Open when applicable**". This means you are required to publish your data in Open Access when possible. See **Access Rights** for exceptions

eDNA

The publishable part of your data set should be archived in the EASY archive of KNAW/DANS, according to the metadata standards for eDNA[1] (the eDepot Nederlandse Archeologie). From our Faculty perspective, this also applies for Archaeological projects executed outside the Netherlands. Your funder's requirements or contracts with authorities may oblige you to deposit with another repository. The faculty then prefers you to store the data set as in DANS/EASY as well (you can set the access rights to "not accessible". If this is not allowed, metadata in the repository should be functionally equivalent to eDNA. Please consult the Data Steward.

Faculty Data Vault

Apart from the deposit at eDNA, you need to publish the remainder of your data in the faculty "data vault". This is meant for the evaluation of your results in case of doubt or another eventuality where more data is needed that the subset you selected for archiving. Only the Faculty data manager has access. Handover your internal dataset completely (including the documents you created for administration and communication purposes) to your supervisor.

Access Rights

However complete open access may not be possible because of various reasons: research ethics, privacy legislation, intellectual property, continuing research. The Easy archive at KNAW/DANS provides various access levels to deal with these issues. Please also read **Themes /Personal and sensitive data**, because the notion of "context" may also impact the possibility to provide data as open access.

When publishing in Open Access, we advise you to use a proper cc-by license.[1] Please do not use the CC0 waiver, because it will allow others to use your data without citation, which is in violation of proper research practice.

[1] Get in touch with a data steward which of these icenses is best applicable.

If you hesitate to provide the current dataset in open access because it is not published yet, or you need to continue your research on ityou can publish it under embargo. Nobody except you has access. The embargo period you can set with DANS/Easy is 2 years, but you can ask to extend this. A further option is only to grant access if a request for access is submitted to the researcher. The researcher then knows who is consulting the data and can reach agreements about its use and reuse.

Data Management Plan

Include your Data management plan in a separate folder. The DMP has to archived longer that the data itself.

Retention periods

At the Faculty of Archaeology we maintain the following retention periods:

- Indefinite for the published dataset
- 5 years for non-personal and anonymized data.
- 6 months after the research for identifying information
- 10 years for the data management plan

Personal and Sensitive data

Personal data is ***any data that contains personal information***. This data has to be treated with special care. Data containing for instance political/religious convictions or health information, or containing direct identifiers for people (such as passport numbers), has to be treated with even more care.

Sensitive / special personal data

This is a special category of personal data from which can be deduced either:

- Race or ethnic origin
- Political views
- Religious or philosophical beliefs
- Membership of a trade union:
- Genetic data
- Health (both physical and mental)
- Sexual behavior or sexual orientation

As well as:

- Identifiers like passport numbers an SSID's
- Biometric data
- Data about children / minors:[2]

[2] In the Netherlands the AVG limit is of 16n the context of the AVG. In other countries this may differ, so always check this.

Identifying information

This is the subset of your data identifying people; for instance names, addresses, portrait etc.. This will usually concern a combination of data elements. A name can lead to a group of people, but may by having also the address, the combination identifies an individual.

Indirect identification

Indirect identification is also identification. For instance, a pub owner in a specific village may in combination with some other characteristic lead easily to an identification. It may also be that the identified group is so small that with very little effort Please, note that the AVG also takes ***context*** into account. If you allow your data to end-up in a context where people can combine it with other data or skills so that they can identify the individuals, your data is still personal data. For instance, if you publish your data in open access, allowing anyone with big data skills to combine it with other open data, this may form a risk.

Anonymous data and anonymization

Anonymous data is personal data that has been stripped from all *(directly and indirectly*) identifying information, so that it is ***impossible*** to trace the information back to the individuals it concerns. This may done in two ways:

- *Aggregating;* meaning reworking the data about individuals so that it only provides information about classes of individuals. In this case the aggregates have to be so large that it is impossible to identify the individuals. For instance the category of people in the university earning more than $100.00 may easily lead to an identification.
- *Pseudonymization* (see below).

The Data Privacy Impact Assessment (DPIA)

If you handle sensitive data you are obliged to perform a so-called Data Privacy Impact Assessment (DPIA) fill out the DPIA form. A DPIA is an assessment to identify the risks of processing personal data. It consists of a number of questions on the basis of which you determine whether the processing of personal data in your research project is legitimate and which measures should be taken to make sure this processing is compliant with the GDPR.

The first version of the DPIA should be completed before you start collecting or using personal data. Without a DPIA, your research will not be compliant, and will not receive a statement of compliance from the University's Data Protection Officer.

To assist you in performing a DPIA, a template has been composed for your convenience. After you filled-out the form, send it to the Privacy Officer. Incorporate any suggestions you receive back. During the research, you update the DPIA every time your research changes. In addition, send a copy to the Privacy Officer before you start working with personal data. -> **Documents / DPIA Form**

Consent and withdrawal

When you process personal information, you need a legal ground. This might simply be the fact that you are doing scientific research, which is classified as a legitimate interest. However, to be safe, especially with handling sensitive data, we advise you to always do your research based on "consent". Before you want to do a processing based on consent, you must request this consent.

The consent request must be unambiguous, and thus be written as simply as possible and tailored to the audience. It must clearly state which data is collected and for which purpose. It must also inform the subject he/she can withdraw the consent, resulting in the deletion of the data. Consent of the person may not be used as a ground for processing if the person (possibly) feels as if he or she has no choice (such as in the case of an employer / employee relationship).

Without further notice, you may use the data only for the research you obtained it for, not for other future research, without a renewal of the consent. If you intend to use the data for future reearch, you should explicitly ask permission for this.

Use the template consent forms. -> **Documents / Leiden University Consent forms** These forms are examples, which need to be rewritten to the specific processing for which you want to request consent. Everything that is going to happen with the personal data is clear to the person who gives the consent. It should be clear exactly what consent is given. It is should also be clear how the person can withdraw the consent.

Consent may be withdrawn at *any* moment. This means that the identifying information] has to be deleted. This does not mean that all data must be deleted, but this does mean that for that specific person everything that can lead back to this person (such as name, address, etc.) must be deleted. The information may need to be anonymised even further to prevent indirect identification. The anonymized data may be continued to be used as well as published.

Retention period

In the consent form you specify the retention period. This is the period during which you keep the identifying information. After this date the data has to be anonymized. Take note that after this data you have to make sure that all copies of the identifying information have to be deleted. -> **Themes / Archiving**

Pseudonymization

If you use personal data (and especially sensitive data, you should go through a process called pseudonymization.[3]

Anonymize sensitive data. Separate / remove identifying information

- Determine which information in your data identifies the individuals involved.
- Create an anonymized copy of the sensitive data, with all the identifying information removed.[4]
- Create another dataset with just the identifying information and refer between the two with a code.
- When you have running text (for instance transcribed interviews) replace identifying information with […] or a similar marker.
- Treat these two versions of the data *very differently*:

[3] In certain cases, it is possible to obtain participant's consent to use and share the data unaltered, with additional access controls if necessary.

[4] This is tricky. Also information indirectly identifying individuals might still be considered identifying information.

    - Keep the identifying information inside the University network at all times, on a place you only have access to (for instance the P:\-drive. Destroy personal data after the legal term (6 months after last use for your research)
    - Treat the pseudononymized data in the way you treat other research data. This data may be published.
- Some information is difficult to pseudonymize. [5] Digital manipulation of audio and image files (voice alteration, image blurring) can be labour-intensive and expensive and might damage research potential of the data. Treat this data *very cautiously*, especially if there exists a version where identifying information and all other research is combined. Just to state the obvious:
    - Keeping sensitive data on your personal share in the university network or on a laptop with hard disk encryption. Do not carry around (for instance) unencrypted USB-sticks.
    - On the computer you store this data on, do not download software from the internet you do not know the provenance of.
    - When using e-mail, encrypt your files, or use the *SurfFileSender* (instead of services like WeTransfer).
- Pseudonymized data for which the identifying information and the linking keys is no longer available is anonymous data. Please take into account the notions of "indirect identification" and "context". If the data might end-up in a context where individuals can be indirectly identified (for instance openly online where data any consumer of the data might combine the data with other datasets and use AI-skills to further process it!), the data is not anonymous. So stay on the safe side.

Processing Agreement

If you allow personal data to be handled by an external party you should have a proper processing agreement. The Processor Agreement states the legal position of the controller (you) and processor (the service), which and how personal data is processed and the measures that have been taken to protect this data and how to act if things go wrong. The AVG requires the controller to prepare the Processor Agreement.

This applies for any cloud software you use and has consequences you might not think of at first:

- Cloud services, though provided by the University, might not have a proper processing agreement for sensitive data and are thus not fit for storing that type of data.
- You might also need a processing agreement for data you do not provide yourself. For instance, if you use a could service where students need to create an account and provide personal details, this is also in need of a processing agreement. This is even the case when no account is created, but the platform stores IP-addresses.
- It is completely disallowed to use commercial cloud services (like Dropbox) for any personal data. For starters because there is no processing agreement.

When in doubt, get in touch with the privacy officer, to make sure this is the case.

[5] Besides a spreadsheet containing both types of information this might be interviews on video or where people mention their names.

Fieldwork

Please note that although your research is not about persons as such, you might collect personal data during your fieldwork, for instance:

When you perform fieldwork abroad, also read **Going abroad and non-EU citizens.**

Local contacts

During fieldwork you might keep notes about local contacts, people you meet during surveys. Apart from names and addresses, you might be recording summaries of conversations with them. These are also personal data according to the AVG, and might contain sensitive information. You should perhaps ask for consent. When in doubt, perform a DPIA.

Daily Reports

Daily reports usually list names of people and their presence or activities at a given point in time. In the case of researchers or students this is unproblematic, since this data is part of the lawful task of conducting research. For external visitors, this is different, especially when it concerns minors (school classes) or you record religious beliefs (Christian school classes). Think of a consent procedure.

Photographs during fieldwork

During Fieldwork a lot of photographs are taken containing people. The purpose of these photo's is not always the research proper but more to convey the atmosphere at the excavation. These photo's inevitably show some special personal details (ethnicity, possibly biometric data, expressions of a religion etc.) Some photos might be children under the age of 16 years, which qualify as sensitive data. A number of conditions must be met before these photos can be published.

- Publishing of photographs revealing sensitive information and children under 16 years should be avoided as much as possible. It is important to scan each photo individually before publishing.
- Notify the participants before the excavation that photo's will be taken.
- The photographer should ask for permission before a photo is taken. This is not necessary when a large group of people is photographed.
- If someone objects to the publication of a photo, this should be removed.

Information about participants

It is probably sensible to take data on dietary demands and medical conditions about students or other participants. Ask for permission and specify the retention period. This is sensitive information. Store a single copy of this information in an encrypted file, folder or device. Make sure to delete the data after the retention period (presumably directly after the fieldwork). -> **Tools / Encryption.**

Going abroad and non-EU citizens

If you transport personal data to a jurisdiction other than the one the personal data was collected and you take the data (either on your laptop/harddrive or by downloading it) you are exporting the data. This might for instance easily take place during fieldwork. You should explicitly ask permission for this, and not take the data otherwise.

Also personal data collected abroad and about non-EU citizens is subjected to the GDPR.

Planning

Plan the data management of personal, and especially sensitive data very carefully. As with a Data Management Plan in general, with personal data it is even more important to keep your promises. Be aware that the European law for "data leaks" is very strict: any personal data falling in the hands of unauthorized parties must be reported ***and this may potentially lead to serious penalties for the University***.

Archiving and publishing

If you are sure the data set has been anonymized, it can be archived in DANS/Easy. DANS will check the anonymization as a general procedure. But even then, it is best *not to make this dataset public access*.

Alternatively you can archive a dataset only containing aggregates, not the individual observations.[6] In that case the dataset containing the individual (anonymized) observations has to be archived in the Faculty data vault.

Research Ethics

The Ethics Committee gives binding advice on ethical issues in research. This may be the case when you handle sensitive data. But also in other cases, like handling human remains, the Ethics Committee may require you to take steps to protect your data, which will implicate research Data Management.

Make sure you took proper measures to ensure ethical approval of your research. When in doubt, get in touch with the Ethics Committee:

https://www.organisatiegids.universiteitleiden.nl/en/faculties-and-institutes/archaeology/committees/ethics-committee

Information Security

Backups

You can't archive something that has been lost.

If you use a Faculty network share (like MyDocuments folder), you are OK. If you store data elsewhere (for instance on your unmanaged laptop or private computer), you are responsible for making backups yourself.

Structuring data

Follow best practices for data structure

For structuring data proper, you can use some best-practices to ensure that the data are easy to manipulate, analyze and visualize. Some clear guidelines are offered by the Tidy Data framework, that can be summarized as follows:[7]

[6] Also aggregate data might be considered sensitive because it identifies people indirectly.
[7] If you want more guidelines Hadley Wickham, Tidy Data (Journal for Statistical Software, 2014): http://vita.had.co.nz/papers/tidy-data.pdf.

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Directory Structure and file naming

Think of a clear and consistent directory structure

It is very important to organize your data in a clear directory structure. The structure aids users of your data set to navigate and get n overview of your data quickly. Directory names should be readily understandable, but not too long. If you need more text, put a "readme.txt" file in a directory explaining (in at most a couple of sentences) the contents of the files and the purpose they serve in your research.

We prefer you to have your data structured in one of the following structures:

- Site
    - General
        - Data files (spreadsheets / databases)
        - Photographs
        - Drawings / maps
        - Etc.
    - Material type
        - Data files (spreadsheets / databases)
        - Photographs
        - Drawings / maps
        - Data files (spreadsheets / databases)
        - Etc.
- Part of research or Publication
        - Data files (spreadsheets / databases)
        - Photographs
        - Drawings /maps
        - Etc.
    - (Sub-research)
        - Data files (spreadsheets / databases)
        - Photographs
        - Drawings /maps
        - Etc.

If one of these structures does not work for you, please drop me a note before you submit!

Think of a clear file naming convention

A proper file naming convention makes clear the category, purpose and version of a file. Proper file names:

- are unique (not the same names in different folders)
- are consistent (upper / lower case)
- are permanent
- reserve the 3 letter extension (preceded by a dot) for the file type
- do not contain other dots than the one above
- indicate the version of the file in format v1, v2-3 or similar (if applicable)
- are meaningful but brief
- classify the file
- do not have spaces and special characters
- use hyphens '-' or underscores '_' to separate logical elements
- use the format YYYYMMDD (easy to find and sort chronologically)
- only use codes that are documented

Examples of good filenames:
Rim_sherd_ElFlacco_v1-6.mdb
Survey_DominicanRepublic_2015-05.xlsx

Examples of bad filenames:
Mysurvey.xls
Article_jcr.def.def
Wouter Kool's research.doc
00015.MTS

Version Management

Determine how you keep track of the final version and intermediate versions

Having a good overview of the versions of the various files in your dataset will greatly improve the speed in the archiving phase:

- If possible, use tools with proper version management included.
- Incorporate versions in your file naming convention and methodically main those.
- Think of a strategy to correct proliferation of files.
- Decide how to handle files that are (temporarily) located elsewhere (home, fieldwork).
- What is the procedure to make sure changes are incorporated in the current version? How to prevent multiple versions from deviating?
- If you work on data collectively, make sure to make agreements who maintains the primary version of a file, and where it is located.
- Determine who is responsible for the final archiving of the shared piece of data, so that it is not archived for various projects (possibly in multiple versions).

Storage and sharing

Size

Estimate the amount of data you have created and what will be added? Choose an appropriate storage facility -> **Tools / Storage.** Split directories of over 400 files into separate (logical) directories

Sharing

Think about how and with whom you will share data. Determine a location of shared data that is accessible for all project members. Make agreements with team members about the usage of the data. And determine access rights accordingly.

Never (never!) use generic services like Dropbox for sensitive data

Commercial Cloud Services

Commercial could services like Dropbox or Google are seriously advised against. Firstly, these services allow themselves to reuse your data (to enhance their products or selling it) or transport the data to other providers and jurisdictions. Especially for sensitive data using these services is not allowed.

File formats

Notify the data manager when you have non-standard file formats

For people to reuse your data, the file formats you publish it in must be "open" as well. This means:

- Long-term data formats that are supported by data archives should be used if possible.
- The format should preferably "human readable", not binary (this means that if you open the file in e.g. notepad, you should see readable text (possibly interspersed with codes), not weird glibberish. These formats are usually "proprietory", which means that only a specific company understands the codes.
- If the format is binary, but it is widely used in your field and can be opened by multiple softwares (for instance, Shapefiles), submit it to the archive but provide a converted version in a more open format.
- It should be able to use the formats independently of the underlying hardware, such as microscopes, scanners or recording equipment.

Notify the data manager if you have files that are not according to these criteria.

https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats

Project Management

Create an inventory of all the people/roles (potentially) involved in the research project, including their roles and responsibilities (with regard to the research data), how they relate to the university and what they would need to perform well (instructions, supervision, equipment etc.). Get advice about the planned involvement of people involved who are not employed by Leiden University. From data management perspective the most important topics would be how to deal with information security, privacy, choice of infrastructure and intellectual property.

Once you have a clearer picture of the data management needs in your project, consider that data management and certain tools might be new to the people you work with. Discuss with the Faculty Data Manager what would be the best way to ensure that all people involved have the information they need and are working according to the project procedures. Developing good habits and learning to use new tools usually requires some time.

Field work and going abroad

Make sure you are familiar with the legal and regulatory situation in the country you travel to, especially for personal and sensitive data. Read the information in **Personal and sensitive data / Field work and going abroad**

Think about these issues as well:

- How to access your data during fieldwork? Preferably use some kind of remote access to the University environment.
- If this is not possible, think clearly how do you transport them. Is the data safe during transport? What is the procedure when they come back?
- If you change files during fieldwork, how do you maintain version control, preventing people at home to change the same file?
- How to get data from proprietary devices like TS- or GPS-equipment. Usually these devices store in proprietory formats only accessible from the device or specific software.
- In your backup strategy take into account you are working outside or in unclean environments, which increases the risk of equipment breaking down.
- Create a proper strategy how to incorporate newly gathered material in the field into datasets you need for analysis.

Metadata

Keep in mind a fellow researcher: someone who knows your area of expertise, but not necessarily as well as you do. Are your files really understandable as they are, or do they need cleaning enhancement? Keep the right balance: make sure the metadata is sufficient for reuse, but do not get carried away and spend valuable time on minor issues. **Enough is enough**.

eDNA metadata

You should add metadata according to the criteria of eDNA:

https://www.itc.nl/library/research/data-metadata-preparation-v2.pdf

Codebooks

> ***Start creating a codebook for your spreadsheets, databases etc.***

For spreadsheets, databases and other complex data files (for instance SPSS), you need to create a codebook. At codebook adds information to make your data understandable for other people than yourself. Creation of codebooks can start well in advance of the archiving, but preferably when you have finalized your date to save yourself spending too much time updating the codebook whenever you change structure of the data file.

Codebooks should be of the following format:

- For MS Access or MS Excel files, for each table/worksheet in the file, add a table "codebook_[table name]"
- For SPSS, use the codebook function.
- For other programs , create a separate MS Excel file called "[database file name]_codebook.xlsx."

Codebooks should contain the following information:

Variable	Description	Values	Scale
Name of a variable	It's description	List of possible values (if applicable (only when it is a class variable)	Scale used (if applicable (only if it is a numeric variable and the is a particular scale)

- Which files need a codebook to be understood? Create the codebooks and make sure they can be found through the file list. Compile the codebooks in such a way that a future archaeologist will be able to interpret your data independently and correctly with little effort.

Intellectual property[8]

Determine whether data you reuse is subject to intellectual property rights

As a general rule, research data is free from intellectual property restrictions. However, there are many exceptions, and the actual situation is not always clear. The rightsholder[9] has to be asked for permission for reuse of the dataset.

If you obtained data from another party, first check any licenses, stating what is allowed. If there is no explicit license and you intend to publish data, or share it outside the research project, you have to check copyright law and database law.

Legislation is not always straightforward and copyright law and database law can be conflicting. In the following cases you will probably need to ask for permission:[10]

- The data possesses its own original character and bears the personal stamp of the author[11] **AND** the reuse does not take the form of a "citation".[12]
- The data is a collection of independent materials that have been arranged systematically **AND** there has been a "substantial investment" in obtaining, verifying or presenting the materials **AND** you reuse a "substantial portion" of the data.

[8] These are very general terms. If you have any doubts about your copyright situation, please contact the data manager.

[9] To counter a persistent misunderstanding: ***ownership of data does not exist legally***. What does exist is a rights holder for the data.

[10] Actually the law is applicable for the country where the data was created, which is a difficulty with the Caribbean involved The indications below are based on EU law, which is quite strict.

[11] Creative wording or drawing, subjective choices in processing, specific formatting or presentation etc.

[12] Be especially cautious when you reuse photographic material. Photographs are usually subject of copyright because of artistic choices involved. You should be especially sensitive when photographs comes from heritage institution, even without specific licenses or if they are taken by yourself.

Finally, to stress the obvious: whatever the legal situation, it is a ***scholarly best practice to inform and cite the source of your data.***

Selection

Start thinking in advance about the selection for long-term preservation

You select the subset of your dataset for long term preservation. The actual choices for selecting often depend on the specifics of the research involved, but some guidelines can be given:

Data archiving concerns the final version of the original data, as far as they are relevant for verification of the conclusions or for scholarly reuse.

As a general rule, the processed data that was used for a publication is archived, as well as the original source data. With the "processed data" is meant the data responsible for creating maps, graphs, tables or other analysis in publications. With "source data" we mean (for example) the original excavation database, any photographs, interviews, satellite images. Sometimes one or more intermediate stages are also worthwhile preserving, for instance when they can be reused more easily than the actual source data or when effort is needed to create them (like interview transcriptions).

General criteria to take into account are:

- Is it really Data? Files like spreadsheets, source code (like R) is probably something else and not needed for long-term preservation.
- Are the data already archived as part of another project, or with a publication?
- Are the data useful for future research?
- Are the data directly relevant for your publications?
- Are the data unique (impossible to recreate)?
- Are the data valuable (for instance effort to recreate, cultural heritage value)?
- Are there any obligations for long-term storage?
- Are the data subject to intellectual property rights (see below)?
- Are the data subject to privacy limitations etc. (see below)?

Keep in mind as a general rule that the data is to be reused by a researcher about 10 years from now.

Tooling

Getting tooling

Standard software is provided through the Software Centre. You can find Software Center by clicking on the Windows start button on the bottom left corner and begin typing "Software Center". immediately. Other software (either commercial or free software) should be

Encryption

The tool Veracrypt can be requested through the University helpdesk portal.

Archiving

- Archie helps you to create datasets compliant with eDNA guidelines. It is still under construction.

https://github.com/NCMulder/Archie

Storage and sharing

Small data sets can be stored on the faculty network, this is also the safest. Various alternatives:

- Local P:\ drive for personal data, but not for sharing
- J:\ Workgroups: for sharing data with a restricted group, also for personal data.
- J:\ Research Data[13] the cost for this is Euro 410 per Tb/Year, charged on the research project, so you should be able to fund this. Not for personal data.

Sharing inside the University

- Sharepoint/Microsoft teams (for personal data?)

Sharing outside the University

- The easiest option is probably synching your local drive with a Surfdrive.[14] You can share folders with arbitrary users and synchronize your harddrive. But it is not a full collaboration environment. Not for personal data.
- Surf Research Drive (currently in pilot).
- Never use generic Cloud Services like Dropbox or Google Drive!
- OneDrive provided by the University. (personal data?)

Transfer

- Use SurfFileSender, for personal data you can encrypt your data.
- Never use commercial services like WeTransfer.

Interview data

- Atlas TI is the software product of choice.
- Do not use unauthorized tools.

Documents

Leiden University DPIA form

https://www.staff.universiteitleiden.nl/ict/privacy-and-data-protection/general-data-protection-regulation-gdpr/procedures-privacy/archaeology/fda-board-office?protected=true%3Fprotected&cf=archaeology&cd=fda-board-office

Leiden University Consent form

https://www.staff.universiteitleiden.nl/ict/privacy-and-data-protection/personal-data/informed-consent/archaeology/fda-board-office?cf=archaeology&cd=fda-board-office

[13] http://issc.leidenuniv.nl/formulier-bulkdata

[14] http://www.issc.leidenuniv.nl/ict/samenwerken/surfdrive.html