

# Case studies for the data retention workshop

Open Science Festival 2025

## Do the re-appraisal criteria work?

These case studies / examples are meant to test the re-appraisal criteria. Using the provided re-appraisal criteria sheet and workflow, with your group assess the two assigned case studies. Do you retain or discard the datasets? Do the criteria work? Specifically:

- Are the criteria for a re-appraisal of data retention / discarding as we are suggesting them suitable and sufficient?
- Are criteria missing?
- Are there criteria that should be rephrased?
- Are some criteria perhaps superfluous or not suitable?
- Are some criteria more key than others? If so, which ones? How should we weigh the criteria?
- And does this hold for all case studies you look at, or are there differences between data types or disciplines?

*The case studies are based on real-life examples but have been adapted for the purposes of this workshop.*

## Group 1 (Cultural heritage theme)

### 1. Archaeological excavation datasets

Imagine you are the data curator(s) at your institutional repository, where research datasets from the whole university are deposited, either openly available or with restricted access. Often there are no clear instructions by the researchers/depositors about the retention of the data in the long term. The university, following (inter)national guidelines stipulates that all underlying research data are to be kept for a minimum of 10 years. For each of the datasets below, this 10-year mark has now been reached and you have to assess whether to retain or discard the dataset.

#### 1a. Archaeological excavation dataset A

The Bronze Age to Iron Age archaeological site of Deir 'Alla in the Jordan Valley has been excavated by archaeologists from Leiden University and Yarmouk University for years, from the 1970s to the present. Documentation was, and is, initially done on paper forms, notes, as sketches, plans, photographs/slides and more, which were (and are) kept at the Faculty of Archaeology at Leiden University. At some point in the 2010s, the whole archive to that date was digitised by scanning the documents to PDF and the photos and slides to JPG. This complete archive was then, with some basic metadata and documentation, deposited at the repository, where it is openly available. The archaeological site is preserved, with the excavated layers of course gone, and the artefacts and samples mostly in storage in Jordan and partly in Leiden.

#### 1b. Archaeological excavation dataset B

The Neolithic and Late Bronze Age archaeological site Tell Sabi Abyad in northern Syria was excavated by archaeologists of the Leiden Museum of Antiquities and Leiden University from the 1980s until the war started. As with the above dataset, there is a digitised analogue archive with field notes, field forms, plans, and photographs, as well as born-digital material like digital photographs and elevation models. The digital archive has been deposited at the repository with restricted access (only the standard metadata are visible); one can contact the depositor (the researcher, now retired) for access. What remained of the site after excavation seems to have mostly survived the war, but unfortunately most artefacts and samples that were kept in Syria have been destroyed.

## 2. No longer part of the mission

An extensive book on cultural heritage, containing appendices with data, was published as hard copies by a small publisher. This being during a time when e-books and open access were just coming up, the publisher did not have options to make the book available online (open access or not) and it was agreed that the researchers could deposit the book in a discipline-specific repository, where the PDF was then available open access. At the time, the repository mostly published 'grey literature', or unpublished reports describing built heritage and archaeological sites and such, and a book fitted in well with these. In the meantime, with many other options existing for making publications (openly) available online with publishers or in institutional publication repositories, the repository has adjusted its mission and has become a research **data** repository only. As such, it has taken the book offline, with only the metadata still visible. The book is now not available online anywhere. The small publisher has been bought up by a large publisher, who will only make the book available online against a hefty fee.

## Group 2 (Fieldwork theme)

### 1. Anthropological fieldwork-based data

A former PhD researcher at a social sciences faculty in the Netherlands conducted intensive ethnographic field research in Kenya 20 years ago. Investigating power relations between men and women in a small rural village, she lived there for a year, making extensive notes on conversations she had or observed and on other observations. Before doing this, she got the consent of the head of the village and other village elders, and she also made it clear to all participants what she was doing and that she was documenting her observations. Back in the Netherlands, she wrote up her thesis and included a good number of (pseudonymised) observations and direct quotes from conversations. She digitised her original field notes, as well as the photos she took and sketches she made – these were not published but archived on the university drive as well as on a personal hard drive that the researchers still keeps, although she has moved on from research and is no longer connected to this or any university. The university is cleaning up their storage drives and the current data steward is asked what to do with the dataset.

### 2. Linguistic dataset

A linguistic study recorded a rapidly disappearing language of in Indigenous community in the Amazon region, using audio recordings, which were then transcribed. Informed consent was obtained, and the dataset (recordings plus transcriptions) was published in the Endangered Languages Archive with contextual information. When assessing after 10 years if the dataset should be kept, it seems an obvious case. But then a researcher not related to the original study but working in the same area contacts the repository; they are worried because they feel the Indigenous community was not involved enough in the decision-making, and the situation does not meet the current thinking about indigenous data sovereignty.

## Group 3 (Social and Behavioural Sciences theme)

### 1. Clinical psychology

Imagine you are the data curator(s) at your institutional repository, where research datasets from the whole university are deposited, either openly available or with restricted access. Often there are no clear instructions by the researchers/depositors about the retention of the data in the long term. The university, following (inter)national guidelines stipulates that all underlying research data are to be kept for a minimum of 10 years. For the dataset below, this 10-year mark has now been reached and you have to assess whether to retain or discard the dataset.

The dataset underlying a paper into long-term effectiveness of online self-help intervention for people with HIV and depressive symptoms is held in the repository with restricted access. The study was done using questionnaires into patient health and anxiety symptoms with open and closed questions. The dataset consists of the raw results of these (pseudonymous) questionnaires, the results processed using SPSS. The questions used and the scripts used to analyse the data are also included, as well as a short README file. Summary data were published in the paper, in which the dataset DOI is also referenced.

### 2. Earlier discarding?

You are the scientific director at a Social and Behavioural Sciences Faculty. The regulations stipulate that the rawest form of the data need to be archived for at least 10 years, or 15 / 25 / 35 years for medical, “WMO-plichtig” research depending on the type of WMO project. Especially in your field this is seen as important for research integrity, after some media scandals (luckily at other universities) where famous professors turned out to have fabricated data. However, now there is a case where a researcher is going to interview children involved in the “Toeslagenaffaire” (benefits scandal), in order to write a governmental report. The researcher would like to have permission to destroy the interviews (recordings and transcriptions) immediately after writing the report, because these children are in a vulnerable position, and because they have a very low trust in the government.

## Group 4 (Natural sciences theme)

### 1. Building and testing machines

As part of a Horizon Europe funded research project, a new kind of microscope was developed at Leiden University, the Optical Near-field Electron Microscopy (ONEM). It brings together the best of both worlds – the high resolution of electron microscopy, and the ability to examine samples without damaging them, like you can with a traditional light microscope instrumentation. The most important project result was the instrumentation itself, and a patent was filed. To judge if the instrumentation functioned as intended, several thousands of images were produced during the project, each 1-100 MB in size. In addition, the project data consisted of Python code and Matlab files, with the results stored in text files. While the instrumentation itself was patented, the dataset could be openly published.

You are the data curator at your institutional repository, where this dataset was deposited. No clear instructions were left by the researchers about the long-term retention of this very large dataset. The university, following (inter)national guidelines stipulates that all underlying research data are to be kept for a minimum of 10 years. This 10-year mark has now been reached and you have to assess whether to retain or discard the dataset.

### 2. Chemistry data produced in collaboration with a company

During a project taking place within the Chemistry department, a group of researchers collaborated with an external, commercial company. A clear collaboration agreement was signed between the involved parties, and it was agreed that part of the data was to be deposited in the institutional repository, partly open access and available for reuse while the other confidential part of the data would be kept by the commercial company for their internal use only. The deposited dataset contains some very large elements, such as images and modelled data. More than 10 years on, the repository data manager in collaboration with the scientific director and the former project PI have agreed that it is not necessary or useful to retain all the data. The company, however, no longer exists as such. What to do?

## Group 5

### 1. Palaeoenvironmental data

A stalagmite from Oman was analysed for its geochemical composition and stable isotope composition, which forms a 60,000-year long palaeoenvironmental sequence of moisture. It was also dated using Uranium-Thorium dating. The data was submitted as several CSV files and a few images (of the stalagmite and images of stalagmite in the cave it was sampled from and its surroundings). No README file or similar was submitted, but a related publication described the methods used. The stalagmite itself is still available at the university and could be re-analysed, if sufficient money and time were available (the analyses are not cheap and depending on the resolution it would take several weeks to several months of labwork).

You are the data curator at your institutional repository, where this dataset was deposited. No clear instructions were left by the researchers about the long-term retention of this very large dataset. The university, following (inter)national guidelines stipulates that all underlying research data are to be kept for a minimum of 10 years. This 10-year mark has now been reached and you have to assess whether to retain or discard the dataset.

### 2. The researcher objects

During re-appraisal, following the various steps in the workflow and based on the re-appraisal criteria, it is decided that a dataset is to be deleted from the repository, retaining only the metadata. However, when the researcher is contacted about this, they object: They find the dataset is still useful and moreover, it is one of their research outputs that they want to retain as a whole. Do they get a veto here?

## Group 6

### 1. 3D model dataset

As part of a PhD study aiming to enhance water, food, and energy security in Zambia, a 3D model of part of the Lower Zambezi river was made. The dataset contains close to 500 images in JPG, Python scripts, bathymetry TIFF files, a georeferenced model in LAZ, river and waterline track files, a volumized bin file, and a X, Y, X coordinate TXT file. The dataset was deposited with a CC0 license, and all files can be opened using specific, but open software, at least as available at the time of depositing (this software itself is not included). The total size of the (zipped) dataset is 3 GB.

You are the data curator at your institutional repository, where this dataset was deposited. No clear instructions were left by the researchers about the long-term retention of this very large dataset. The university, following (inter)national guidelines stipulates that all underlying research data are to be kept for a minimum of 10 years. This 10-year mark has now been reached and you have to assess whether to retain or discard the dataset.

### 2. Sensitive data on the institutional drive

A study on the Dutch governmental approach in dealing with terrorist events was carried by interviewing MPs and involved members of the government at the time of the events. Consent was obtained before the interviews. The interviews were recorded. After transcription, the recordings were destroyed. The interviews are impossible to pseudonymise and contain very sensitive information. They were stored on an encrypted hard drive placed under lock and key in a non-disclosed location. 10 years have passed and the question is asked whether this storage is an appropriate solution for the dataset.