DaSCH - Swiss National Data and Service Center for the Humanities

How to archive your data with DaSCH at the end of a research project

Are you a researcher who has collected a dataset, and you would like to archive it with DaSCH? Then you're right here. This manual guides you through the process.

Table of Contents

DaSCH - Swiss National Data and Service Center for the Humanities

1 Some basics first

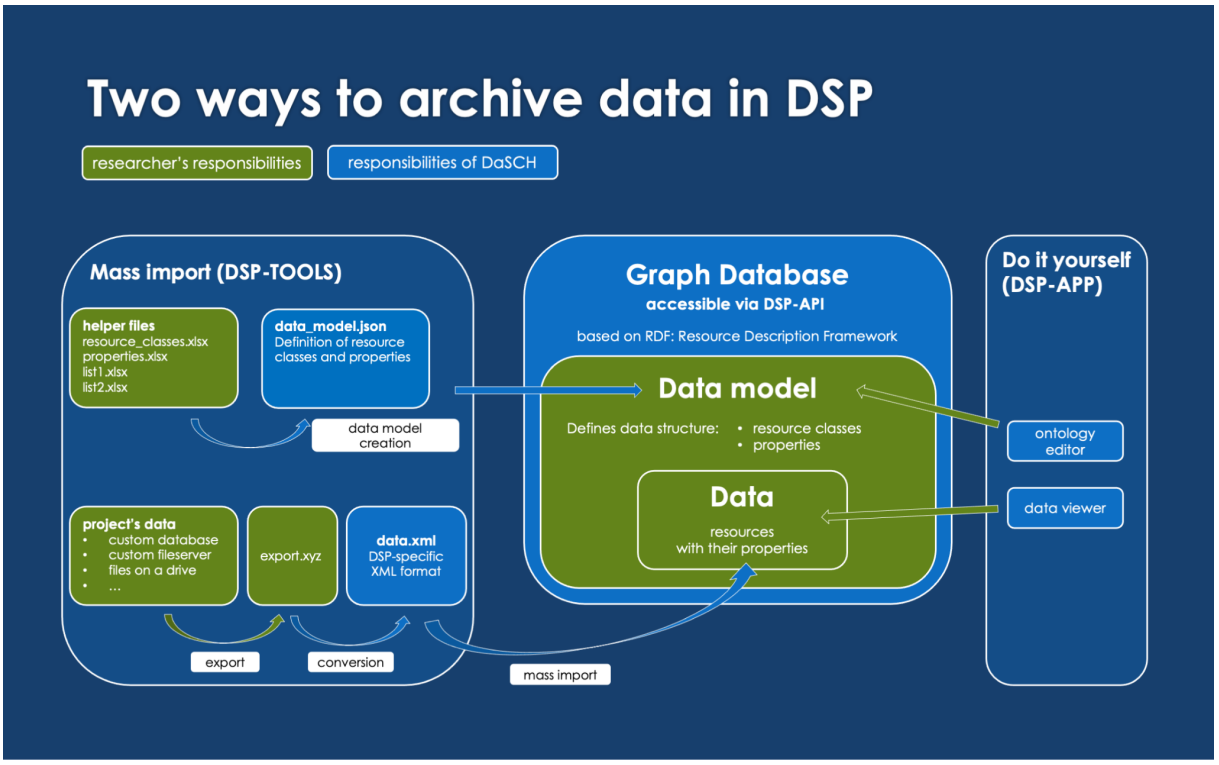
DaSCH is a repository for research data in the Humanities. We are specialised in archiving data itself, not its presentation. This means that we cannot host and preserve research tools or VREs (Virtual Research Environments) with their functionality and GUIs (Graphical User Interfaces) which you may have developed during the lifetime of your project.
We store data in our archiving software called DSP (DaSCH Service Platform). DSP is a graph database based on RDF (Resource Description Framework). It runs on our servers and is available at https://admin.dasch.swiss.
We at DaSCH want to archive your data in a well-structured way. For this reason, you first need a data model that defines the structure of your data. The structure is described by resource classes (e.g. "book", "author") and their properties (e.g. a "book" has the property "was written by" that points to an "author"). Once your data model is ready, you can add actual data, e.g 10 authors and 100 books that were written by one of the authors.

There are two ways to work with your data on DSP:

- **Automated technical access:** Our DSP-API offers access for programs and applications. This allows other online tools or applications to access data stored at DaSCH in an automated way.
- **Browser access:** You can use our DSP-APP as an online tool in your browser. This is useful to work on your data model with some mouse clicks, look at the data, edit and enrich it. Check it out on https://ark.dasch.swiss/ark:/72163/1/082E/YEwhVysoSxmcR01tjSFwiAb.20220413T085845535117Z.

For this reason, there are two ways to archive data in DSP, as visualised in the graphic below:

- **Mass import via DSP-TOOLS** (described in this manual): Send us your dataset, so that we can import it into DSP with a program called DSP-TOOLS.
- **Do it yourself** (not part of this manual): Use DSP-APP in your browser to create a data model and to type in your data. This is done manually and works for small datasets.



The method depicted on the left, "Mass import (DSP-TOOLS)", is the subject of this manual.

2 The archiving process

2.1 Create a data model

Technically, a data model consists of a JSON file that will be uploaded to the DSP server by DaSCH staff. Fortunately, you as a researcher don't have to write it by hand, but instead you are going to fill in Excel templates that will be converted to JSON by DaSCH staff.
A data model file is composed of three main sections:

- **Properties:** a list of the properties that your resource classes need (e.g. "was written by", "has home country")
- **Resource classes:** a list of the resource classes that your data consists of (e.g. "book", "author")
    - Cardinalities: For every resource class it must be defined what properties it has, if they are mandatory, and how many values a property can have at most. The resource class "book", for example, has the property "was written by" with a cardinality of 1-n, because every book must have one or more authors. The resource "author" has the property "has home country" with a cardinality of 0-1, because every author was born in exactly one country, but that country may be unknown, and the field left empty.
- **Lists:** Oftentimes, you want to restrict possible values to a closed set of options, e.g. in order to avoid spelling variants of one and the same thing. In the case of "has home country", you might want to define a list of countries, and one of them has to be chosen.

There are three Excel templates that correlate to these three sections. They are available as downloads at https://docs.dasch.swiss/latest/DSP-TOOLS/dsp-tools-excel/. There, you can also find a detailed description on how to fill them in.

In order to tailor the data model to your data, it is important to know about the different types of resource classes and properties. You can specify the type of a resource class/property

- by deriving it from a DSP base resource class/base property (defined in the "super" field)
- in case of properties: by defining what target it points to (defined in the "object" field).

The two tables below give you an idea of the most important types of resource classes/properties. Please consult the documentation for more types: https://docs.dasch.swiss/latest/DSP-TOOLS/dsp-tools-create-ontologies.

Most resource classes (like the mentioned "book" and "author") are just "normal" resource classes derived from "Resource". There is no special typing needed. But as soon as your resource class represents a multimedia file (e.g. "image of book" or "audiobook"), you want your resource class to be a "Representation", derived from one of the base classes below:

Resource types		
Kind of resource class	**DSP base resource class ("super")**	**Supported file formats**
Generic resource class (to be used in all cases when your resource class is none of the special cases below)	Resource	--
Text file	TextRepresentation	TXT, XML, XSL, XSD, CSV
Document file	DocumentRepresentation	PDF, DOC, DOCX, XLS, XLSX, PPT, PPTX
Image	StillImageRepresentation	JPG, JPEG, JP2, TIF, TIFF, PNG
Audio	AudioRepresentation	MP3, MP4, WAV
Video	MovingImageRepresentation	MP4
Compressed folder	ArchiveRepresentation	ZIP, TAR, GZ, Z, TAR.GZ, TGZ, GZIP, 7Z

Discover more details at
https://docs.dasch.swiss/latest/DSP-TOOLS/dsp-tools-create-ontologies/#super_1.

For properties, we also offer a variety of options:

Property types		
DSP base property ("super")	**"object"**	**Example value**
hasValue	BooleanValue	TRUE/FALSE
hasColor	ColorValue	#FF8000 (hexadecimal colour code)
hasValue	DateValue	01.01.2022
hasValue	DecimalValue	3,14159
hasValue	GeonameValue	Geographical place (geonames.org identifier)
hasValue	IntValue	5
hasValue	ListValue	List node defined in the "lists" section
hasValue	TextValue	Text field
hasValue	TimeValue	01.01.2022, 13:55:00 GMT
hasValue	UriValue	www.wikipedia.org

hasLinkTo	the resource class the link points to	
isPartOf	the resource class of the compound object	
isSequenceOf	the resource class of the audio/video	

Discover more details at
https://docs.dasch.swiss/latest/DSP-TOOLS/dsp-tools-create-ontologies/#super.

Examples of complete data models are available at

- https://docs.dasch.swiss/latest/DSP-TOOLS/dsp-tools-create/#fully-fleshed-out-example-ontology or
- https://github.com/dasch-swiss/082E-rosetta-scripts/blob/main/rosetta.json.

2.2 Send us your data

There are basically three pieces of information that we need to import your data:

- the data model
- the data itself
- documentation about how the data relates to the data model

After having learned how to create a data model, we will now examine the data itself. Research data typically consists of two parts: Files like MS Office documents or images on the one hand, and a logical data structure on the other hand. The logical data structure is often stored in a database application.
The files can easily be sent to us by cloud sharing (e.g. Dropbox link). But for the logical data structure, some more explanations are necessary: Nowadays, a variety of database technologies are in use for research data, and each one has its own export mechanism and export format. Unfortunately, it is not possible for DaSCH to process every export format that is around. Rather, we need to receive it in a relatively easy text-based format, that can be processed with standards-based means. Therefore, we kindly ask you to convert your data into a simple tabular format like CSV or Excel. We can also process XML. Our staff will then transform your data into our DSP-specific XML format, to make an import into our database.

☞ **Our preferred data formats: CSV, XLSX, XML**

Together with the data, we also need some documentation that allows us to understand how it relates to the data model. This can simply be a list of instructions like “The first column of this Excel sheet contains the ‘books’ of the data model, and the second column the ‘authors’”.

Unfortunately, we often encounter data which is in a poor condition: it is incomplete, incorrect, or not well formatted. A simple example might be an Excel sheet that should contain numbers in a certain column, but some of the cells contain letters instead. Such

data is not clean, and cannot be processed or archived with DaSCH. **Cleaning data is one of the key challenges in the archiving process.** Therefore, we advise you to strive for clean research data from the very beginning on.

☞ To learn more about clean data, watch the first section (3 minutes) of "Understanding Clean Data" (https://www.youtube.com/watch?v=kCP-H8VRDCw).

2.3 Import into DSP

As soon as we receive your data model, your dataset and the documentation, we start to write an import script, i.e. a program that transforms your data into our DSP-specific XML format. If your dataset is small, clean and well documented, this doesn't take long – but with growing project size and messy data, it easily costs us several months. If we discover inconsistencies that we cannot clean up ourselves, we need to get back to you, and wait until you resolve the inconsistencies.

☞ **This is an iterative process and requires your active participation.**

When the import script is ready, we upload your data on a test server. There, you can play around, make yourself acquainted with our user interface, and give us feedback if everything is correct. If you realise that your data model or the data is not as you wish, we will adapt the data model and/or the import script, and upload your data again on the test server, until you are happy.

The final stage is the import onto our main server.

☞ **As soon as your data is on the main server, the only way to modify it is by hand, via DSP-APP. At that stage, an automated mass modification is not possible any more.**
☞ **All modifications via DSP-APP are restricted by your data model. For example, if you have configured a field as mandatory, you cannot simply delete a value for that field in a record.**

2.4 Fill in the metadata form

Every dataset published in DSP must have a comprehensive project description, namely metadata about the dataset. The minimum requirements in this respect consist of information such as a title, abstract, keywords, discipline(s), the principal investigator, funding institution, the time period covered, how the data was collected, etc. These metadata are necessary for other researchers to find and understand your dataset.
We will send you a form to fill in. When your dataset is published on our main server, we are going to publish your metadata on https://meta.dasch.swiss, so that your dataset is findable by other researchers.

3 Checklist: steps to archive your project with DaSCH

Responsibilities of DaSCH	Responsibilities of research project
	create data model: fill out the 3 Excel files from https://docs.dasch.swiss/latest/DSP-TOOLS/dsp-tools-excel/
convert the Excel files to JSON **This service is subject to a charge.**	
	send us a copy of your dataset (you may continue working with your dataset)
	fill in the metadata form
write import script **This service is subject to a charge.**	
	Have a skilled person ready who can resolve inconsistencies between data model and data
	send us the final copy of your dataset From now on, you can't work anymore on the data until it's available on the website of DaSCH.
import the data on a test server **This service is for free.**	
	check if the data on the test server is OK
import to main server **This service is for free.**	
	pay the bill for the services that are subject to a charge

Congratulations! From now on, we guarantee that

- the data and their permanent identifiers (ARKs) stay available in the future
- the data is still readable in the future[1]
- the data is interoperable according to FAIR principles
- the data can be found based on its metadata
- the general public has access to your data

All these services are for free!

[1] Compressed archives like ZIP are excluded from this warranty, because they can contain data of any format that may not be readable in the future.