Introduction on how to create and upload data in the ATTO Data Portal

Authors: Marcus Guderle

Contact ATTO Data Management Team: attodbm@bgc-jena.mpg.de

Table of content

1 Overview

Under Collect you find tools to create datasets. The basic functions are (Figure 1):

- Create Dataset
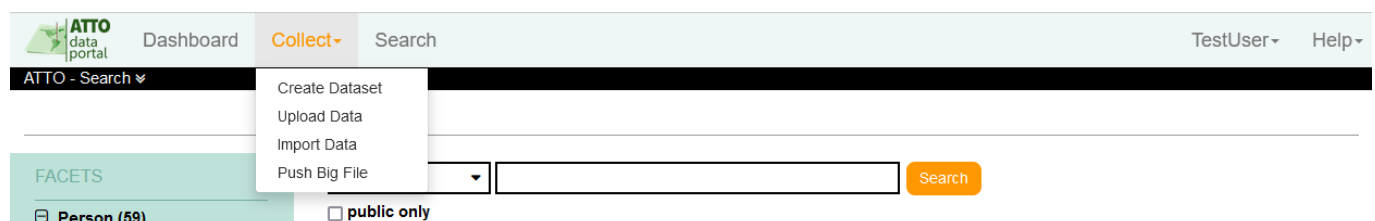- Upload Data
- Import Data
- Push Big Files

Figure 1: Functions of the upload menu

2 Defining a data structure

Data structures contain all variables, which are part of the dataset to be uploaded. Prior to uploading a dataset, the data structure has to be defined in order to have a template for the upload process. Defined data

structures can be used and should be used multiple times for long term data. This prevents that the same variable is defined several times with different names and units (e.g. time in seconds and in Universal time). This is essential to enable an extensive and consistent data management as well as exact search results. For this reason, the definition of data structures will be done by the data management team.

Figure 2: Example of a data structure

It is possible to create data structures for tabular (structured) data such as Excel tables, CSV-Files, etc. and for files such as images, videos, etc. Please send a request for a data structure including the header of your dataset with a clear description of each variable to the data management team (). The respective data structure will be created according to the example in Figure 1. The data management team will send you an Excel Template similar to Figure 2, which can be filled with the respective data and used for their upload. To open this template, you have to enable macros. Macros automate frequently-used tasks. Depending on what Microsoft version you use, enable or disable macros is a bit different. Macro security settings are generally located in the Trust Center.

Please insert the data in the respective column below the grey shaded area (Figure 3).

Figure 3: Example of an Excel Template for the data structure in Figure 2

3 Data collection

The Data Collection Module provides tools to create new datasets, enter metadata, upload data to the system, and import metadata structures. There are some workflows available under the "Collect" tab (see Figure 1):

- Create Dataset - Upload Data - Import Data - Push Big File

Creating a new dataset with tabular data includes the following steps:

1. Create a new dataset
2. Fill out metadata
3. Upload primary data
4. Add data or update dataset
5. Add attachments to datasets

3.1 Create a dataset

To create a new dataset, please click on the "Collect" tab and choose "Create Dataset" (Figure 1). You will be directed to an upload wizard, which leads you through the process step by step. When you hover the mouse pointer over the orange "Select" buttons in the wizard (see Figure 4) an information field about the respective option will appear.

If you want to upload a new dataset, please select "New Dataset" in the "Dataset" field. In case you want to use an already existing dataset (for example form your time series), you have two options to make a selection. First, you can click on the "Select from table" button and choose the respective dataset. Second, you can click on the text field above the "Select from table" button and choose the preferred dataset from the drop-down menu. In the next step you can choose whether to upload a new tabular data or a file. In both cases a new data structure will be created. In case you have been assigned a data structure by the data management team, please use the option "Existing tabular data structure" or "Existing file data structure" and select the appropriate data structure from the drop-down menu. The next step is to define the metadata structure. In our case there is only one predefined one called "ATTO".

Figure 4: Overview of Data Collection Wizard

In order to continue the data creation process please click the "Next" button. You will be directed to the metadata form, which will be explained in the following section.

3.2 Fill out metadata

The following section introduces the metadata schema for the ATTO consortium and gives an over-view of the individual information required. In case you want to upload a new dataset to an existing data structure and comprehansive metadata information have thus already been uploaded previously, the respective metadata sheme can be requested from the data management team. You will receive a .xml file, which can be imported via the "Import" button (see Figure 5). You then have to change the appropriate information for your new dataset that differ from previous ones, such as collection dates, layers etc.

Figure 5: Overview of metadata form

The metadata structure contains four main sections:

- Dataset Level Metadata: Basic information on the data like title, abstract, owner, methods etc.
- Environmental Level Metadata: Detailed georeferenced information on the location, where the data were collected.
- Published In: DOI if dataset is already published in a data repository or in context of a scientific publication.
- Project: Information of the project, in which context the dataset was collected, including funding for the project.

Some fields are a free-text field, which have to be filled in by the data creator. Others have, drop-down menus and an autocomplete function that helps to fill in the form swiftly.

> Note: Completely filled metadata are essential for finding datasets within the ATTO data portal and allows an easy publishing process through the Max Planck Digital Library (MPDL).

Table 1 in the Appendix gives an overview and explanation of each variable of the metadata form. You also get this information by hovering the mouse pointer over the respective variable name left side of the text boxes and the green header of the subsections. All fields marked with a red asterisks * are mandatory for a valid metadata form.

After you complete the form, you can validate (see Figure 6) the entries and the system will give you further information in case any entries are wrong or whether more entries are required.

Figure 6: Bottom of Metadata form with "Validate" and "Save" button

When you completed and validated the metadata form, please click "Save" in order to finish the creation of the dataset. If there are metadata entries missing, the window shown in Figure 7 will pop up. In this case it is recommended that you click on "Cancel" to fill in the missing. However, while clicking on the "OK" button, the metadata form will be saved and you have the possibility to edit the form at a later point of time – even after primary data are already uploaded.

Note: After a dataset is published with an assigned DOI, do not change any part of the dataset. Please contact the data management team (attodbm@mpi-mail.mpg.de) for help.

Figure 7: Warning if metadata information is missing

3.3 Upload primary data

After the metadata form is saved, you will be automatically directed to the top of the page where you will find a link to the Primary Data and thus the Data Upload Wizard (see Figure 8 & 9). If you want to upload your data as tabular data (.xlsm, .txt, .csv) now, please click on the link. The term "Tabular data" is used for all datasets where the internal structure of the data is "known" to the system. For example, in a data table the header, which defines the columns (i.e. variables) is the structure of the data. Before uploading/importing data to the system the data structure needs to be created by the data management team (see also paragraph 2 - Defining a data structure).
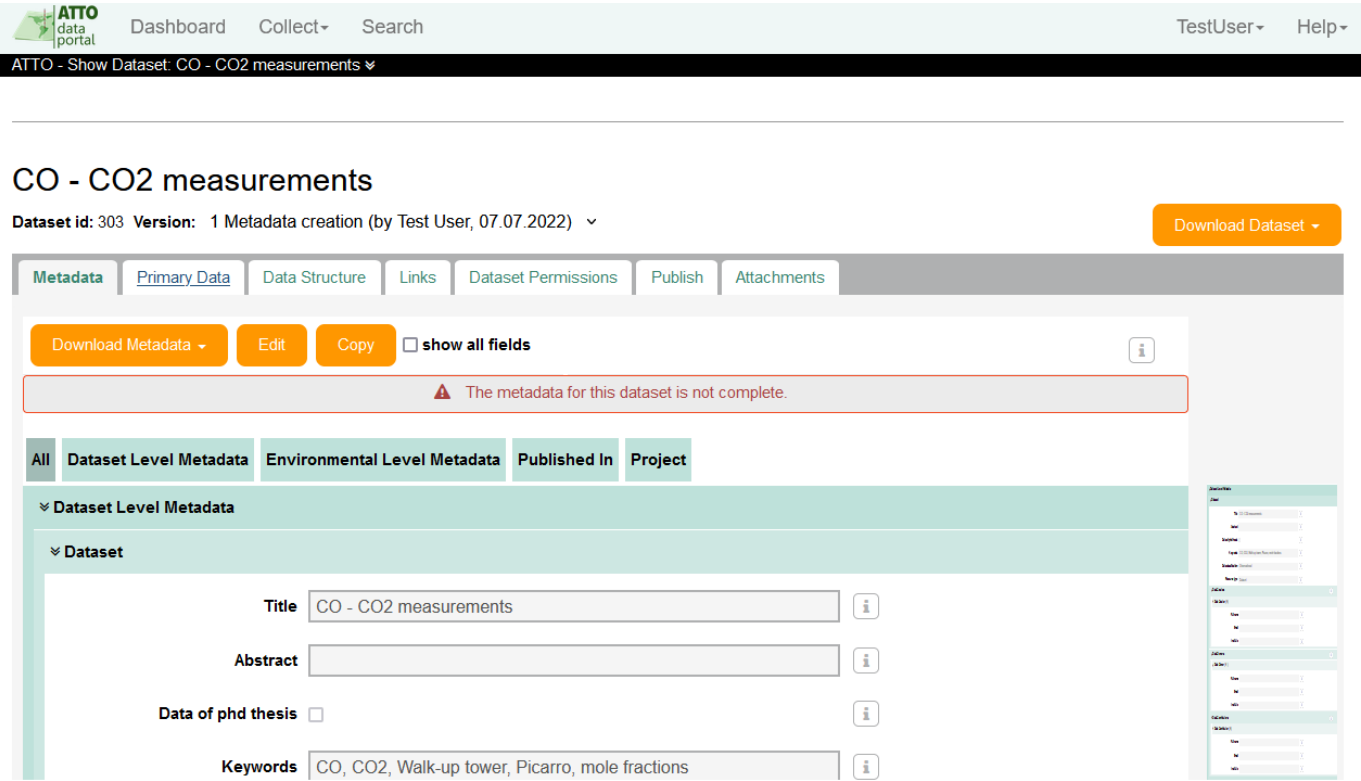
Figure 8: Redirection to the link of the Primary Data
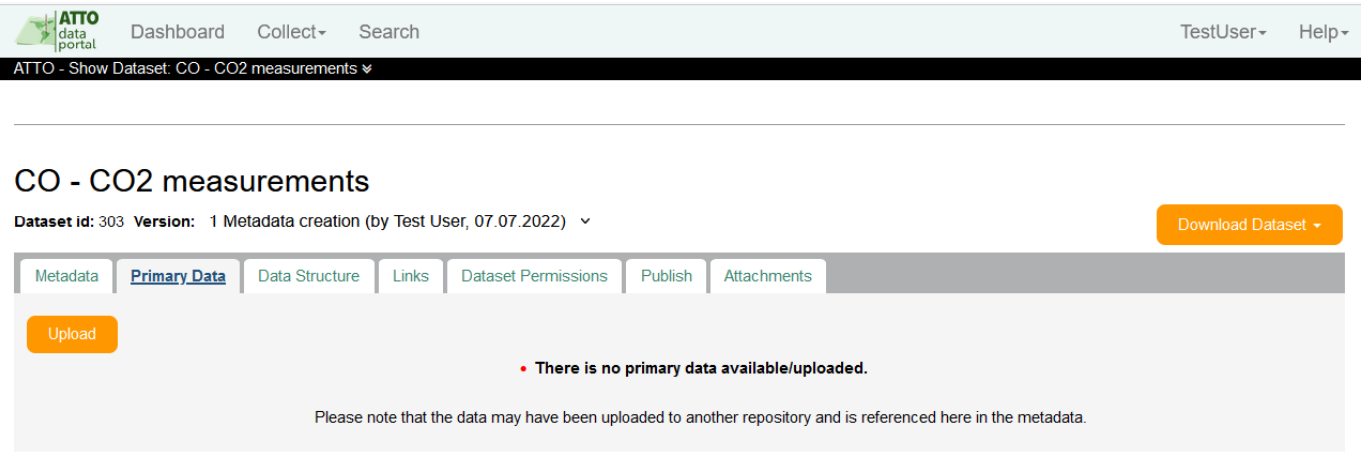
Figure 9: Data Upload Wizard

First, please select an existing file containing your data. You can either select a file from your local computer or a file that has been uploaded to the server prior to starting the Data Upload Wizard (see Figure 10). The second option is designed for files larger than 4 MB that may take several minutes to transfer. The wizard supports file formats of Microsoft Excel (.xlsm) or ASCII (.txt, .csv). Once a file was successfully selected, click the "Next" button and proceed to the next step.

Figure 10: Data Upload Wizard - file selection

For all ASCII files you need to provide information on the formatting and file structure (Figure 11).

- Separator: Please choose which separator is used to separate data values from each other in the ASCII file.
- Decimal: Please choose the punctuation, which is used for decimal values. This is depending on your language settings of your computer.
- TextMarker: Please choose how text is marked in your data file.
- Orientation: Please choose if your data is oriented column-wise or row-wise.
- Offset: Define whether your dataset contain empty rows or columns on top or to the left before the header and the actual data values start.
- Variables: Please define the row/column where the header starts. Usually, datasets contain of a header, which defines variable names, types etc.
- Data: Please specify the row/column where the primary data starts.

More information can also be found in the online Help via "Data Collection".

Figure 11: Data Upload Wizard - file information

Next, the created data set must be specified once again, to which the data to be uploaded should be assigned.
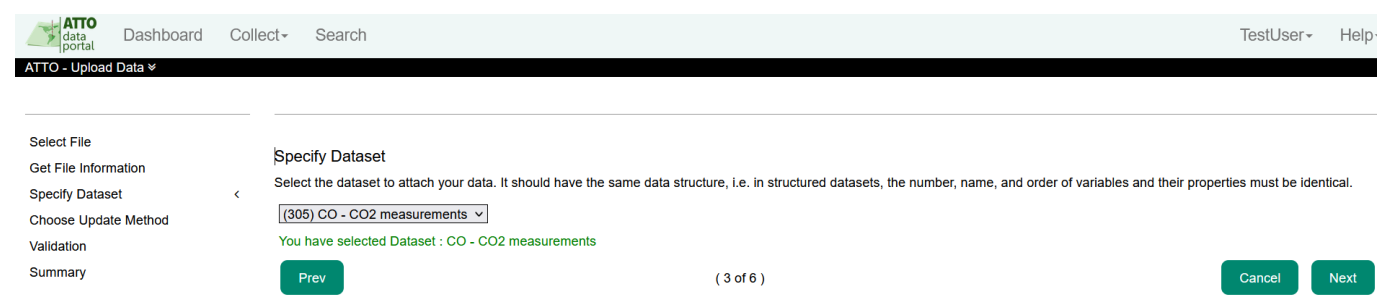
Figure 12: Data Upload Wizard - dataset information

After defining the required information, click "Next" and you are asked to validate your dataset (Figure 13). In this step, the system checks whether the header information and the number of variables matches with the predefined data structure. If this is the case, you will see the green stroke "Validated!!" on the top of the page. Please click "Next" to get a summary of your created dataset and to finalize the upload process by clicking "Finish".
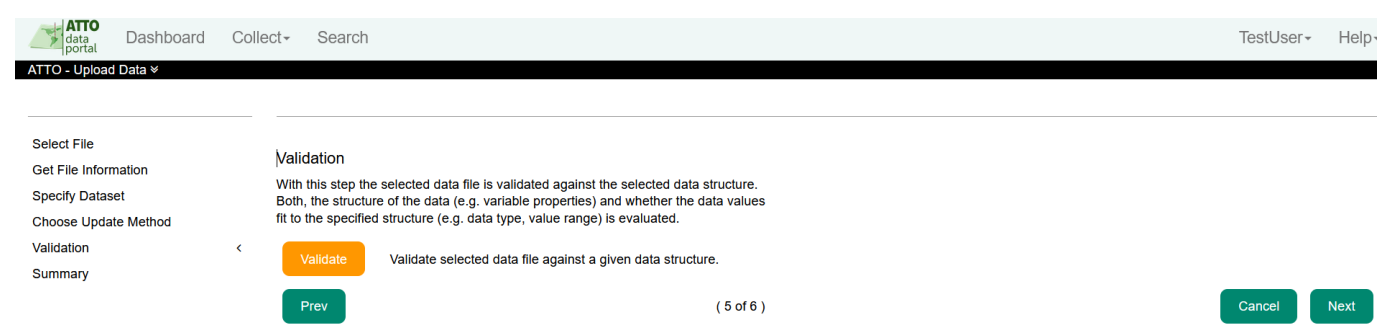
Figure 13: Data Upload Wizard - Validation

In case you want to upload your primary data at a later point of time and assign them to your prior defined metadata, you can use the "Upload" option via the "Collect" tab (Figure 1). Here you can chose as well, whether you want to upload tabular data or files like images or other data associated to your dataset (.avi, .bmp, .csv, .dbf, .doc, .docx, .gif, .jpg, .jpeg, .mp3, .mp4, .pdf, .png?raw=true, .shp, .shx, .tif, .txt, .xls, .xlsm, .xlsx, .xsd, .zip). The uploading process is similar to the application of the Upload Data Wizard described above.

3.4 Import primary data

Using the Import Data Wizard is another possibility to create a dataset including metadata in one workflow. Please click "Import Data" via the "Collect" tab (Figure 1). After you selected a file from your local computer or a prior uploaded file, you have to choose the "ATTO" metadata schema. The title is by default the name of your file, but can be changed in the textbox "Title" (Figure 14).
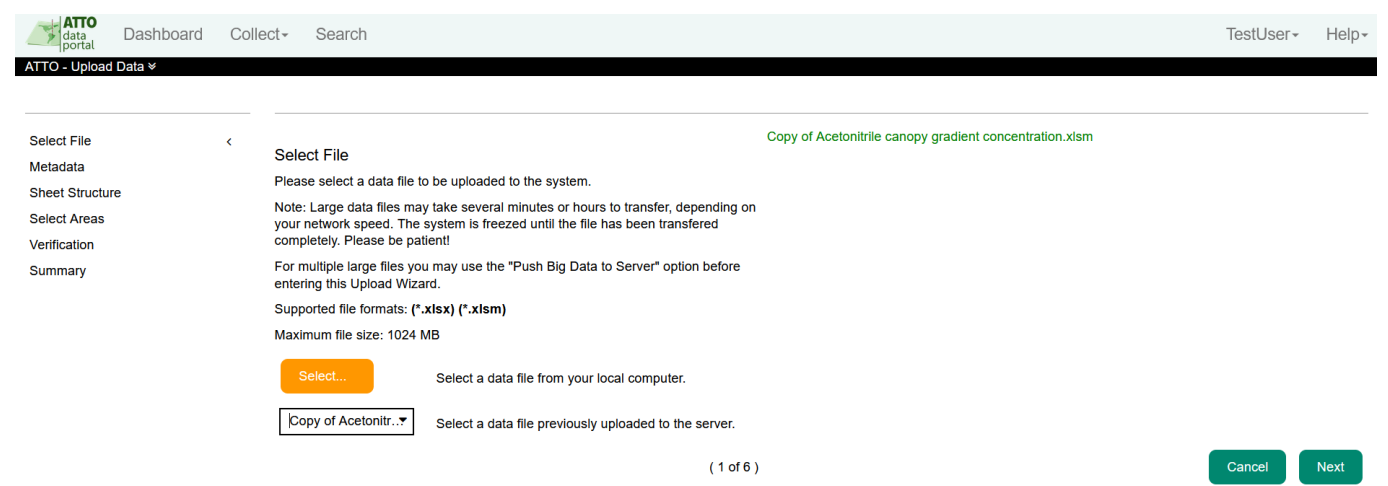
Figure 13: Import Data Wizard
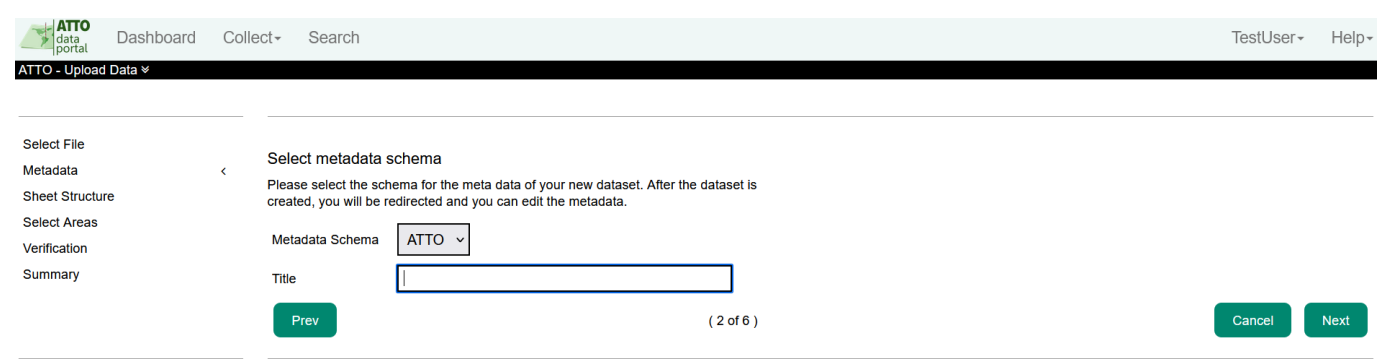
Figure 14: Define metadata schema & title

You will be directed to a table, which represents your selected file. Here you can define which parts of the file contain the variables/header and which part are the primary data. Please select the row with variables/header and click the "Header" button (see Figure 15). The selected part will be highlighted in red. In order to specify the data, select multiple rows which contain the primary data and click the "Data" button (see Figure 15). If you want to skip some parts of your data you can just mark the rows above and below the respective parts. With the "Expand Selection" button you can expand the last selection of your data to the last row of the file. This is recommended for large datasets
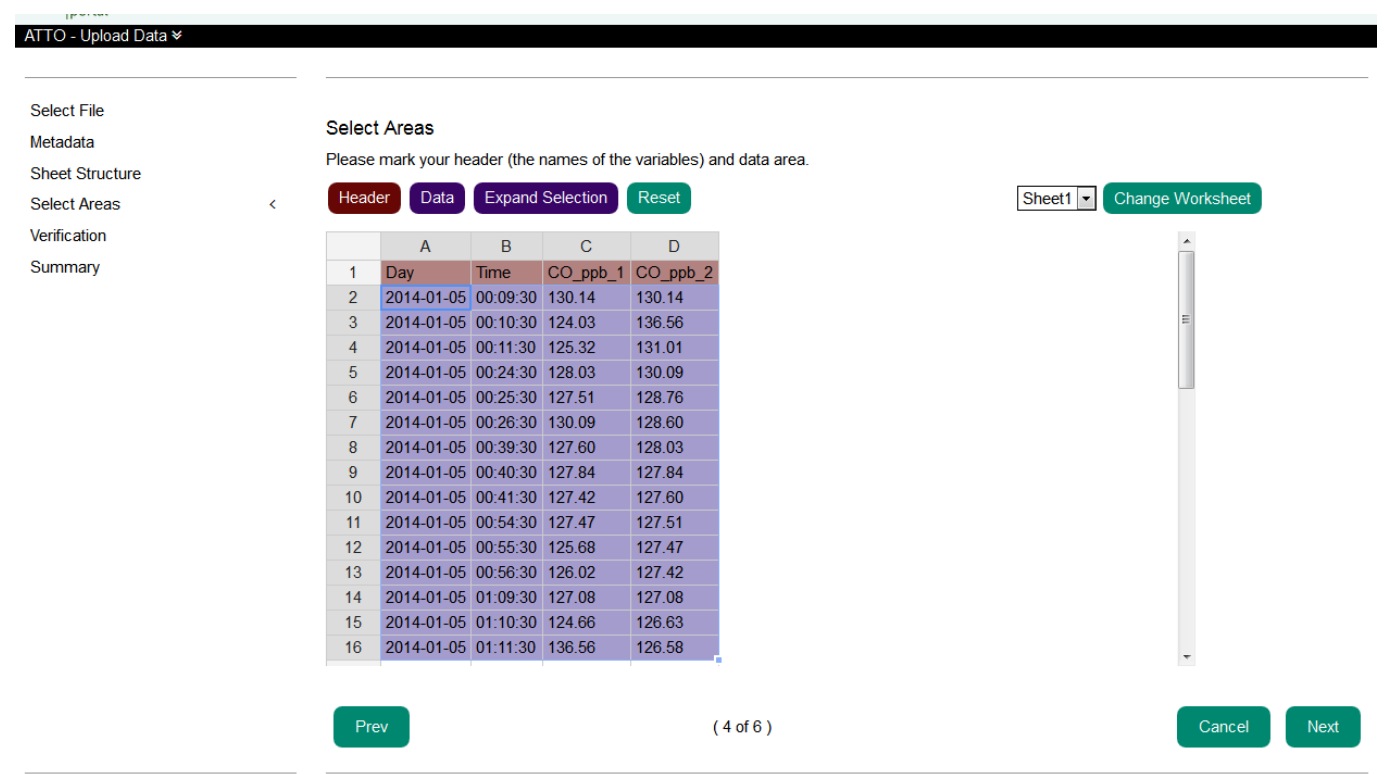
Figure 15: Select areas in the Import Data Wizard

You also have the option to import data from another worksheet of your file with the "Change Worksheet" button on the right side. Please note that only data from one sheet can be uploaded at a time.

When you finished the selection, click the "Next" button to proceed to the verification step, where you have to define the units and datatypes of the variables in the dataset you want to upload (Figure 16). A table with a dropdown menu shows you the variables in the dataset and provides you suggestions for the respective definition of units and datatypes based on other already uploaded datasets. While choosing one of the suggestions, the respective unit and datatype are automatically filled in. You also can define these attributes by yourself by entering free text. With the "Validate" button (see Figure 16) you can verify whether the selected datatypes are suitable for the dataset or if you have to adjust a particular specification.

With a click on the "Next" button you can proceed to a summary of your upload process, which provides an overview of your dataset. Please click the "Finish" button and the data structure and the dataset will be created. In the next step, you will be redirected to the new dataset and you can fill in the metadata form as described in section 3.2 Metadata.

Figure 16: Verification process in the Import Data Wizard

3.5 Push big file

In case you want to upload a big file, you can use the "Push Big File" option via the "Collect" tab. Datasets with the following formats are supported: .avi, .bmp, .csv, .dbf, .doc, .docx, .gif, .jpg, .jpeg, .mp3, .mp4, .pdf, .png?raw=true, .shp, .shx, .tif, .txt, .xls, .xlsm, .xlsx, .xsd, .zip.

Figure 17: Push Big File

Each registered user can upload files to a personal folder where files are stored temporary. An overview of these files is given on the left side of the Push Big File Wizard (see Figure 17). You can delete these files by clicking on the small bin next to the file name. For uploading a dataset, please click on the "Select" button to select a file from your local computer and then click "Push" .

3.6 Add data or update dataset

The system allows to add data to an already existing dataset or to update datasets in case the processing standard was updated. The first case might mainly apply for continuous measurements of time series like temperature profiles or radiation etc. and allows extending the already uploaded time series with the fresh data.

Please click on the "Collect2 tab in the main menu (see Figure 1) and choose "Upload Data". In the Upload Data Wizard you first have to choose whether your data are tabular data or a file.
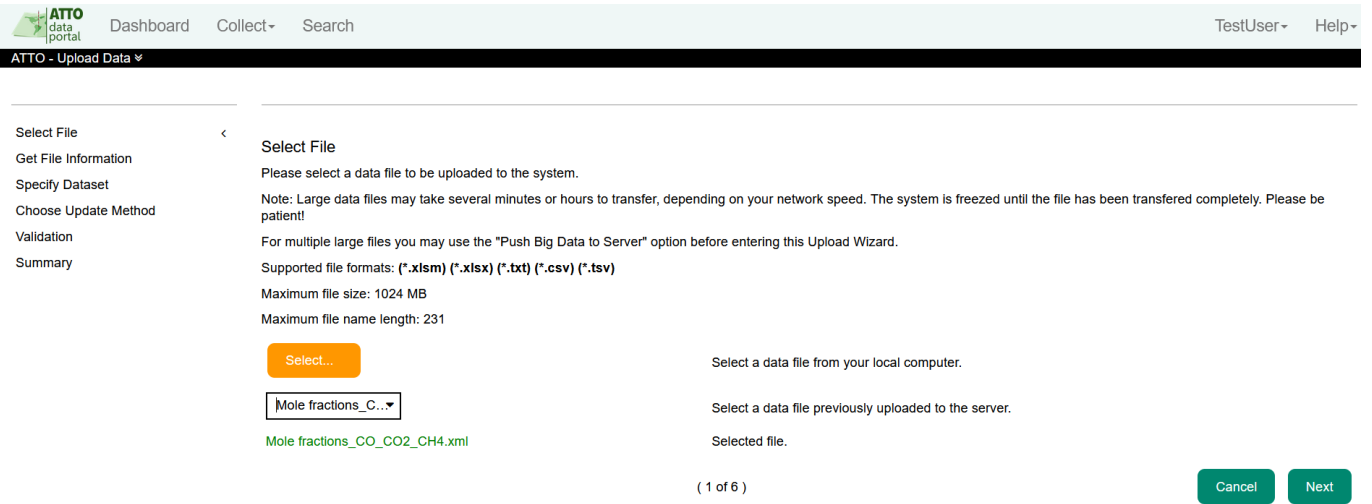
Figure 18: Upload Data Wizard

The file selection is same as described in section 3.3 Upload primary data. After you selected your data/file either from your local computer (or from the server in case it was uploaded before) the file information have to be defined (see section 3.3 Figure 11). In the next step, please specify the dataset to which the uploaded data should be added. A list of all datasets (dataset-ID and title) is shown in a dropdown menu (Figure 19).
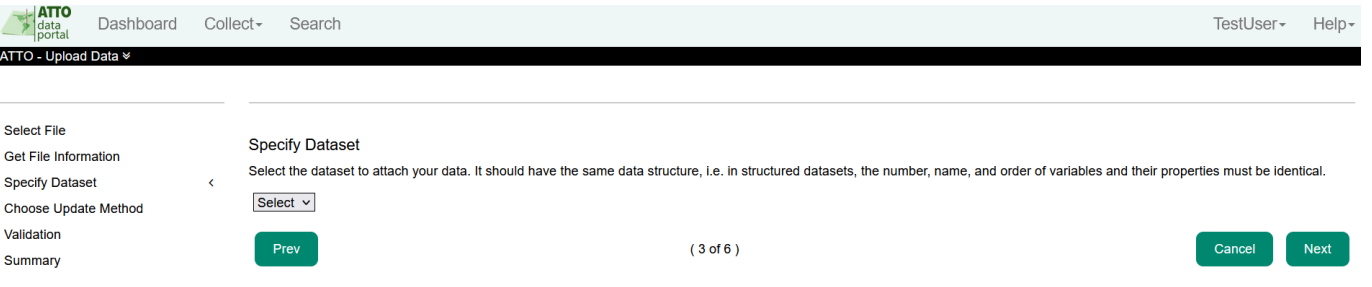
Figure 19: Upload Data Wizard - specify dataset

Click next and define the primary keys of the dataset to be uploaded. Please check the selected variables before clicking next. In case all primary keys are defined accordingly, they are displayed in green in the upper right corner of the webpage.

Click "Next" and you are asked to validate your dataset. In this step, the system checks whether the header information and the number of variables matches with the predefined data structure. If this is the case, you

will see the green stroke "Validated!!" on the top of the page. Please click "Next" to get a summary of your created dataset and to finalize the upload process by clicking "Finish".

> Note: A dataset ID and a dataset version ID are assigned to the uploaded datasets. During each change of the metadata and/or the primary data, the dataset version ID is updated while the dataset ID is always the same. However, all dataset versions are stored on the server. Furthermore, while downloading data, the dataset version ID is taken into account. Please consider the change of datasets before publishing the dataset with assigned DOI since the link to the DOI has to be updated as well after this process. Further information regarding data publication can be found in a separate document.

3.7 Add attachments to datasets

The Attachments-tab (Figure 20) allows users, who have upload rights to dataset, to attach files such as descriptions as word document, text file, excel file, source code, images, etc. to the respective dataset. Other users can only see or download these attachments.

In order to add attachments to datasets, a brief description of the attachment should first be given (Figure 21). Then you can select a file from the local hard disk by clicking the "Select" button and upload it to the database by clicking the "Push" button (Figure 21).
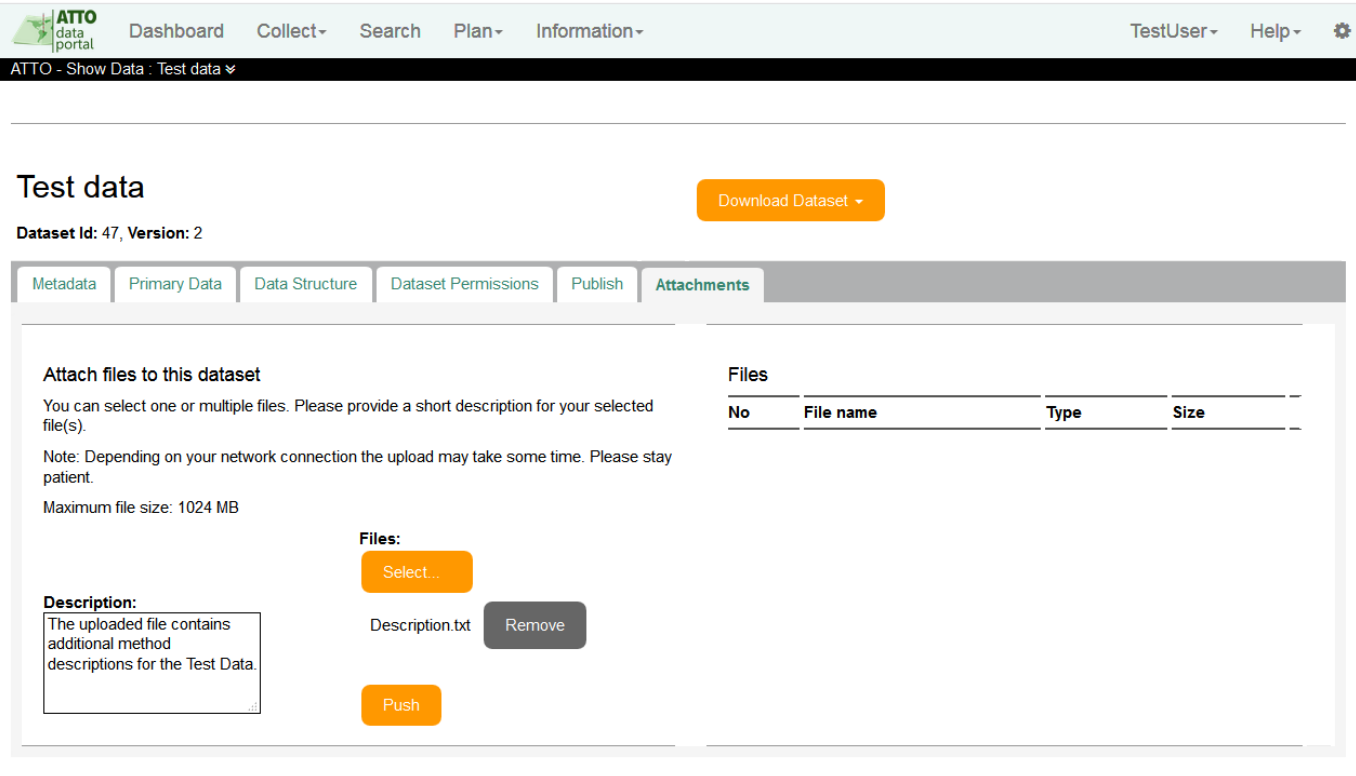
Figure 20: "Attachments"-tab in the dataset view

If a file is successfully added as an attachment, it appears in the overview. Each individual attachment is assigned an ID. The file name, file type and file size are also listed (Figure 21). While moving the mouse pointer over the information icon, the previously entered description of the attachment is displayed. You can also add more than one attachment to one dataset.
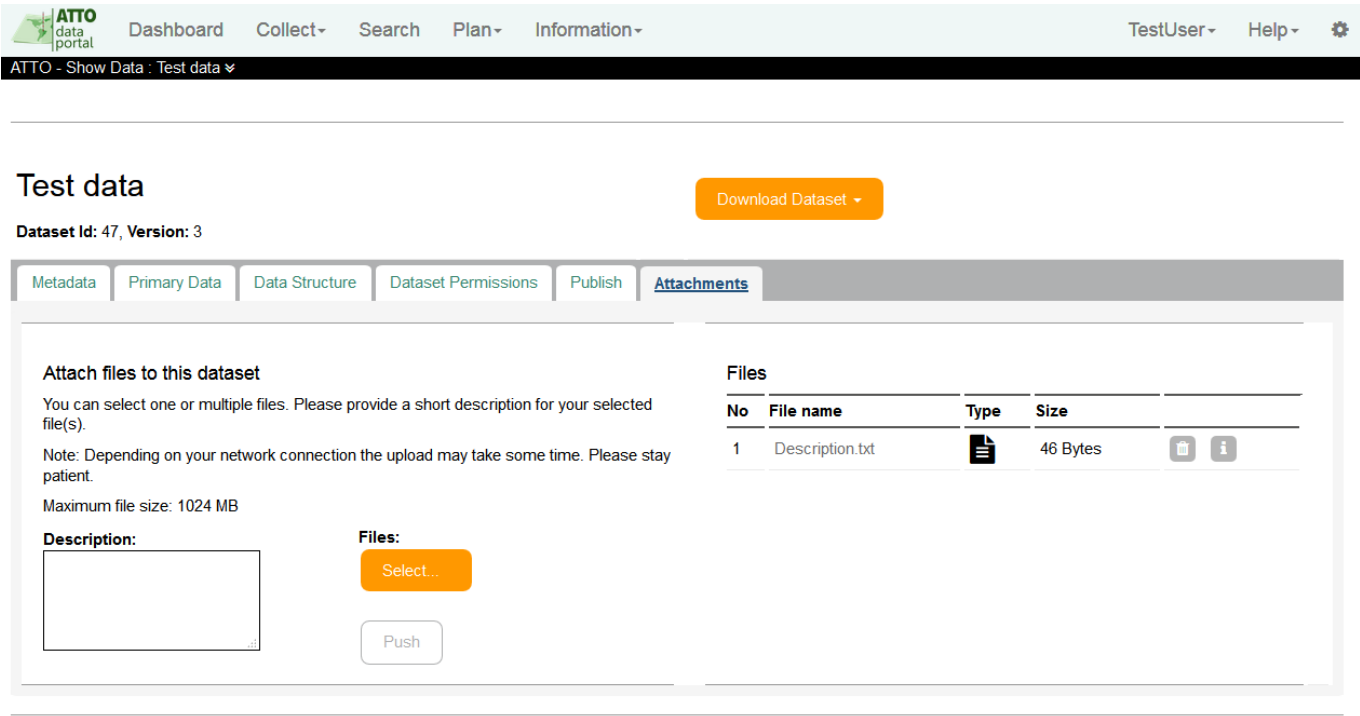
Figure 21: "Attachments"-tab with uploaded file

The attachments of a data set can be downloaded by clicking on the respective file name.

Appendix

Table 1: Overview of metadata variables. These information are also available by hovering the mouse pointer over the respective variable name left side of the text boxes and the green header of the subsections. All fields marked with an asterisks * are mandatory for a valid metadata form.

Variable name	Description	Type of text box
Dataset Level Metadata		
Title*	Title of the dataset to be uploaded to the ATTO data portal.	Free text entry

Variable name	Description	Type of text box
Abstract*	A brief overview of the resource to be uploaded to the ATTO data portal. The abstract should include basic information that summarizes the dataset. Example: Dataset provides information about water temperature collected in Lake Candia.	Free text entry
Data Of Phd Thesis*	Was the dataset collected as part of a doctoral thesis? If so, data users are obliged to consider the special regulations in the data policy regarding PhD data.	Checkbox
Keywords	Provide a set of related keywords describing the content of the dataset. Please separate the keywords with commas (e.g. flux, CO2, tower).	Free text entry
Identifier Type	**Applies only for data sets, which will be published via MPDL.** A persistent identifier that identifies a dataset, assigned by the MPDL for this dataset. Currently, only DOI is possible.	Dropdown menu
Identifier Value	**Applies only for data sets, which will be published via MPDL.** Please insert the DOI assigned by the MPDL. e.g. 10.17871/atto_xxxx	Free text entry
Publisher	**Applies only for datasets, which will be published via MPDL.** The name of the entity that holds, archives, publishes the dataset. In the case of datasets, "publish" is understood to mean making the data available to the community of researchers. **In the case of the ATTO project, Publisher will be the Max Planck Institute for Biogeochemistry, Jena due to the requirements of the MPDL.**	Dropdown menu
Contributor	**Applies only for datasets, which will be published via MPDL.** The institution responsible for collecting, managing, distributing, publishing or otherwise contributing to the development of the resource. **Contributor will be the Instituto Nacional de Pesquisas da Amazônia, Manaus due to the requirements of the MPDL.**	Dropdown menu
Publication Year	**Applies only for datasets, which will be published via MPDL.** Year when the data is made publicly available. If an embargo period has been in effect, use the date when the embargo period ends. Format: **YYYY**	Free text entry
Data Classification	Principal classification of the data to be uploaded to the ATTO data portal. E.g. Observational, Experimental, Simulation, Derived, Metadata, Other	Dropdown menu
Resource Type	The type of the dataset to be uploaded to the ATTO data portal. E.g. Audiovisual, Collection	Dropdown menu
Format	Technical format of the original resource. Use file extension or MIME type where possible. E.g. jpg, png?raw=true, xlsx, csv, dat etc.	Free text entry

Variable name	Description	Type of text box
Version	Version number of the resource/dataset. If the primary resource has changed the version number increases. Register a new identifier for a major version change. Individual stewards need to determine which are major vs. minor versions.	Free text entry
Data Creator(s)*	The main researchers involved working on the data, uploading the data to the data portal or the authors of the publication in priority order. May be a corporate/ institutional or personal name.	One or more persons can be selected via the + button on the right side.
Data Owner(s)*	Person(s) who own this dataset. Mostly these are the PIs of the project.	One or more persons can be selected via the + button on the right side.
Data Contributor(s)*	Person(s) contributed to the data or/and helped to collect the data.	One or more persons can be selected via the + button on the right side.
Full name*	Fullname of the Data Creator, Data Onwer and Data Contributor respectively (e.g. John Doe). Name format: Given Family	Free text entry
Email*	Email of the Data Creator, Data Onwer and Data Contributor respectively (e.g. john@doe.com).	Free text entry
Institute*	Affiliation of the Data Creator, Data Onwer and Data Contributor respectively (e.g. Max Planck Institute for Biogeochemistry).	Free text entry
Street	Street of the respective contact address.	Free text entry
Zipcode	Postal code of the contact address.	Free text entry
City	City of the contact address.	Free text entry
Country	Country of the contact address.	Dropdown menu

Variable name	Description	Type of text box
Dataset Method Specification*	Provides a documentation of the procedures followed to produce any object in the dataset to be uploaded to the ATTO data portal. These shall include information about procedure steps, software used within individual steps and source data.	Free text entry
Dataset Sampling Specification*	Allows for a text-based/human readable description of the actual sampling procedures used within the dataset collection. This element shall include information about dataset lineage.	Free text entry
Dataset Sampling Equipment Info*	Provides information about any instruments used in the data collection or quality control and quality assurance. The description should include vendor, model number, optional equipment, etc.	Free text entry
Data Processing Level*	Processing level of the data to be uploaded to the ATTO data portal. E.g. Level 1B for unit processed data according to the NASA data processing level standards or Level 2A for processed and aggregated data according to the AmeriFlux.	Free text entry
Data Processing Level Standards*	Processing level standards used, e.g. NASA or AmeriFlux and European Fluxes networks.	Free text entry
Data Collection Layer*	Information if the data were collected above ground, below ground or in water.	Dropdown menu
Layer Start	Start height/depth of layer in cm.	Free text entry
Layer End	End height/depth of layer in cm	Free text entry
Start Date*	Start date of the data/time series.	Calendar
End Date	End date of the data/time series.	Calendar
Temporal Resolution	Information about the temporal resolution of the data/time series e.g. 1 min, monthly.	Free text entry
Number Of Parameters	Information about the number of parameters in the dataset.	Numbers
Location*	Information about the location at the ATTO site where measurements were taken. Please use the common name, which is also given in the general map provided by the coordinators (e.g. Tall Tower, Walk-Up Tower, Triangular Mast, River, Tree(s), Soil, Other).	Dropdown menu
Location Other	In case "Other" is selected in the field "Location", please enter here the name and/or coordinates of the location (e.g. Tall Tower (S 02 08.756 W 059 00.335)).	Free text entry

Variable name	Description	Type of text box
Abbreviation(s)*	Abbreviations used in the dataset to be uploaded to the ATTO data portal (e.g. m for meters, t for tonnes).	Free text entry; More than one variable can be selected via the + button on the right side.
Description	Long-form of the abbreviations used in the dataset to be uploaded to the ATTO data portal (e.g. meters, tonnes).	Free text entry; More than one variable can be selected via the + button on the right side.
Data Quality*	Quality assurance - has the dataset undergone any quality checks? This is important for scientists interested in using the data for further analysis.	Multiple fields
Check Performed*	Has the dataset undergone any quality checks? If yes, please click checkbox.	Checkbox
Issues*	Please mention the inconsistencies or errors present in the dataset to be uploaded to the ATTO data portal.	Free text entry
Curation Info	Information about data curation.	Free text entry
Plot Info	In case the dataset is based on a plot experiment, please insert additional information here.	Checkbox on the right side in the header.
Total Plots	Total number of plots the dataset is based on.	Numbers
Plot(s) *		
Plot Size	Plotsize in sqm.	Free text entry
Number of Plots	Number of plots of particular size on which dataset is based.	Numbers
Sub Plots	In case the experiment contains subplots, please insert additional information here.	Checkbox on the right side in the header.

Variable name	Description	Type of text box
Subplot Size	Subplotsize in sqm.	Free text entry
Number of Subplots	Number of subplots of particular size on which dataset is based.	Numbers
Software*	Information on the software used for dataset creation, quality check and assurance.	
Application(s) Name*	Name of the software used for dataset creation, quality check and assurance.	Free text entry; More than one variable can be selected via the + button on the right side.
Version*	If applicable, version of the software.	Free text entry
Minor	An optional minor version number (e.g. '2' in 1.2)	Free text entry
Modifier	Unconstrained text specifying status & optional number, e.g. 'beta', 'alpha', 'internal'. If missing, release status is assumed.	Free text entry
Date Issued	Date of software version release.	Calendar
Availability*		
Company Name	Name of the Software developer.	Free text entry
URL	URL of the Software developer.	Free text entry
Additional Info		Checkbox on the right side in the header.
Related Dataset	Provide information if any additional related datasets where used to create this dataset, if applicable.	Free text entry
Dataset Level Additional Metadata	Provide additional Metadata for the dataset when required and not listed above.	Free text entry
Environmental Level Metadata		

Variable name	Description	Type of text box
Spatial Coverage*	Provide information on the spatial coverage of the dataset to be uploaded to the ATTO data portal, e.g. Local; Regional; National; Continental; Global.	Dropdown menu
Georeference Source	Source used for georeferencing such as gazetter, online resource url and name, gps type etc.	Free text entry
Georeference Remarks	Remarks related to georeferencing.	Free text entry
GeoLocation(s)*		More than one variable can be selected via the + button on the right side.
Geo Location Name*	Provide detailed information about the location at the ATTO site where measurements were taken.	Free text entry
Latitude*	Latitude of the location in decimals.	Free text entry
Longitude*	Longitude of the location in decimals.	Free text entry
Coordinate Reference System*	WGS 84; UTM Zone 38 North ETRS89 etc.	Dropdown menu
Precision	How precise is the location Lat/Long reading?	Free text entry
Altitude	Altitude of the location in metres.	Free text entry
Depth	If aquatic then depth of the location.	Free text entry
Slope	Angle of the slope.	Free text entry
Aspect	Direction of the slope.	Dropdown menu
Habitat	Habitat info of the location.	Free text entry

Variable name	Description	Type of text box
Remarks		Checkbox on the right side in the header.
Environmental Level Additional Metadata	Additional environmental metadata that you think is not covered in this schema.	Free text entry
Published In*		
DOI	If data are already published in a repository or in the context of a paper, please insert here the respective DOI. e.g. 10.17871/atto_xxxx	Free text entry; More than one variable can be selected via the + button on the right side.
Project*		
Project Title	Name of the project for which the dataset is collected. A project can have many datasets with different dataset titles.	Free text entry
Funder(s)*	Information about the funder(s) of the Project e.g. BMBF, INPA.	More than one variable can be selected via the + button on the right side.
Funder Identifier	Uniquely identifies a funding entity, according to various types.	Free text entry
Award Number	The code assigned by the funder to a sponsored award (grant).	Free text entry
Award Title	The human readable title of the award (grant).	Free text entry

> **Note:** The order of the metadata variables in this table are according to the order in the metadata form on the ATTO data portal.