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

Authors: Juliana Menger, Marcus Guderle

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

Table of content

1. Create a dataset

The "Create" field, which is used to generate a dataset, is located in the second position in the top menu. (Figure 1):
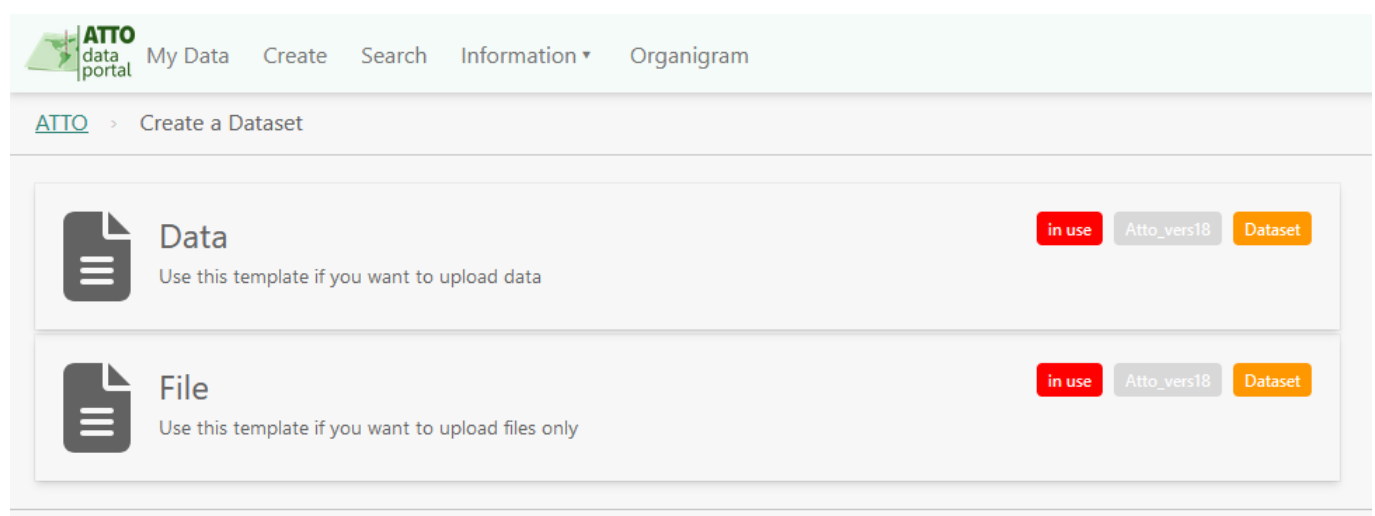

Figure 1: Overview of "Create" menu for creating a dataset

Click on "Create" to open a page with two templates to choose from. If primary data are to be uploaded as a structured dataset, the "Data" template needs to be selected. If files are to be uploaded to the data portal, select the "File" template.

It is advisable to always upload tabular data as a structured dataset, as this offers several advantages for the end user. Among other things, the variable and the respective values can be viewed. Furthermore, when downloading the dataset, information about the respective data structure, the metadata, if already published, the citation as a BibTex file and the metadata as an .xml file are also provided.

In addition, the primary data can be selected and the respective selection can be downloaded using the download function in the "Data" tap.

After selecting the respective template, you will be asked for the title and a description of the dataset (Figure 2). Both fields must be filled in to create the new dataset. The title of the respective data set should contain clear information about the content of the dataset and the year in which the primary data was collected. The description is the same as the abstract of the dataset. Both information is then automatically transferred to the metadata form and can be revised again at a later stage.
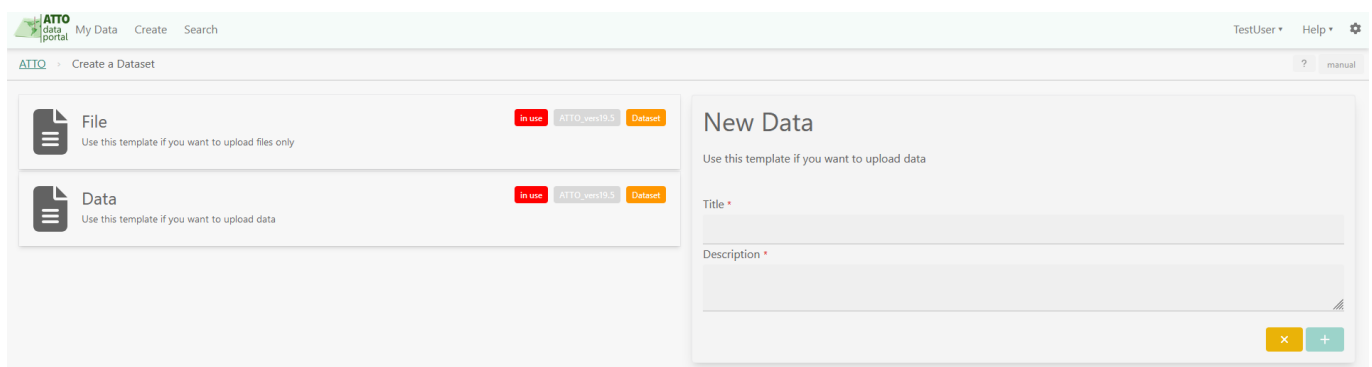

Figure 2: Input mask for title and description for the dataset to be created.

Once both fields have been completed, the dataset can be generated by clicking on the green plus sign located in the bottom right corner. Afterwards, the general editing page of the dataset appears (Figure 3). From here, there are various ways to continue working:

- Select an existing data structure
- Upload data
- Edit the metadata

Figure 3: Editing page of a newly generated dataset.

2. Select an existing 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. It is possible to create data structures for tabular (structured) data such as Excel tables, CSV-files, text-files etc..

In the editing mode of the dataset, it can be created in the drop-down menu "Data Structure" before an upload. Furthermore, it can be created from an uploaded file. These options should only be used by the data management team. We recommend that you select an existing data structure. Please send a request for a data structure including the header of your dataset with a clear description and the unit of each variable to the data management team (attodbm@bgc-jena.mpg.de). After the data management team has created the data structure, you will be informed of its name so that you can select it from the drop-down menu (Figure 4).
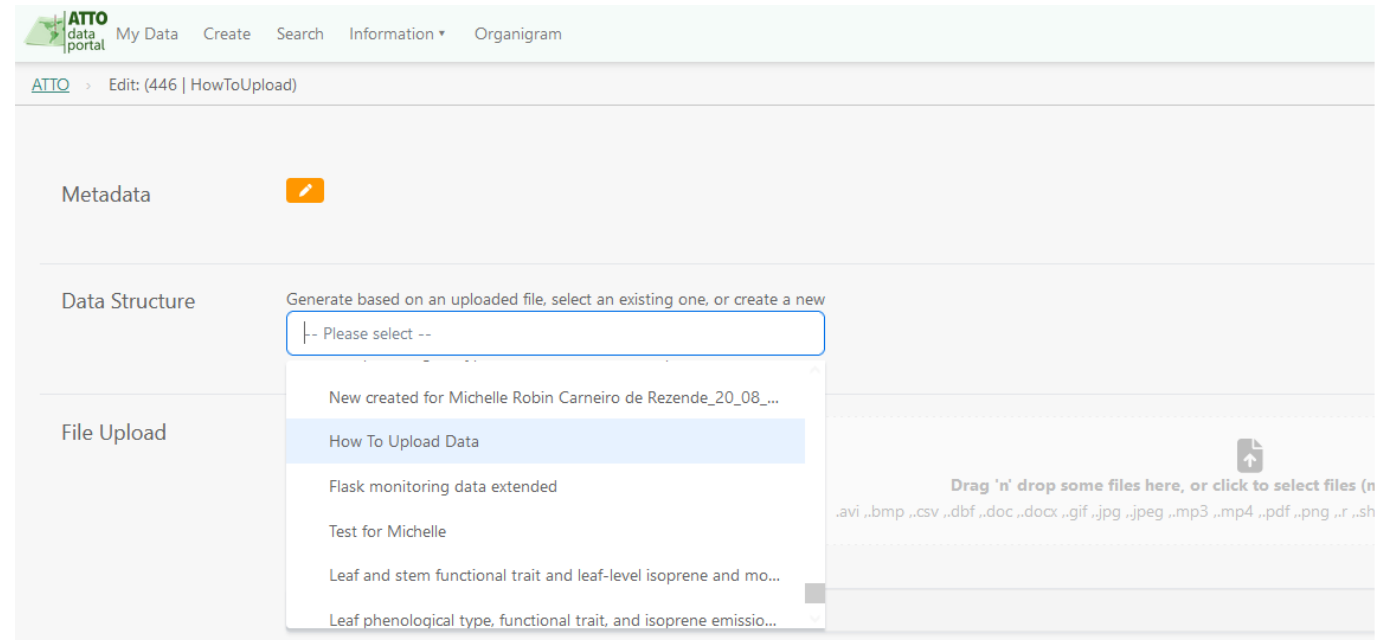
Figure 4: Drop-down menu for selecting respectively generating a data structure.

3 Upload data

3.1 Upload primary data

After selecting the respective data structure for the dataset to be uploaded, you can import the primary data from a .csv, .xls, .txt or other data types in the next step. To do this, please use the "File Upload" section on the general editing page. Here, the file can either be selected from a directory by clicking into the field or simply dragg and dropp the respective file into the field (Figure 5).
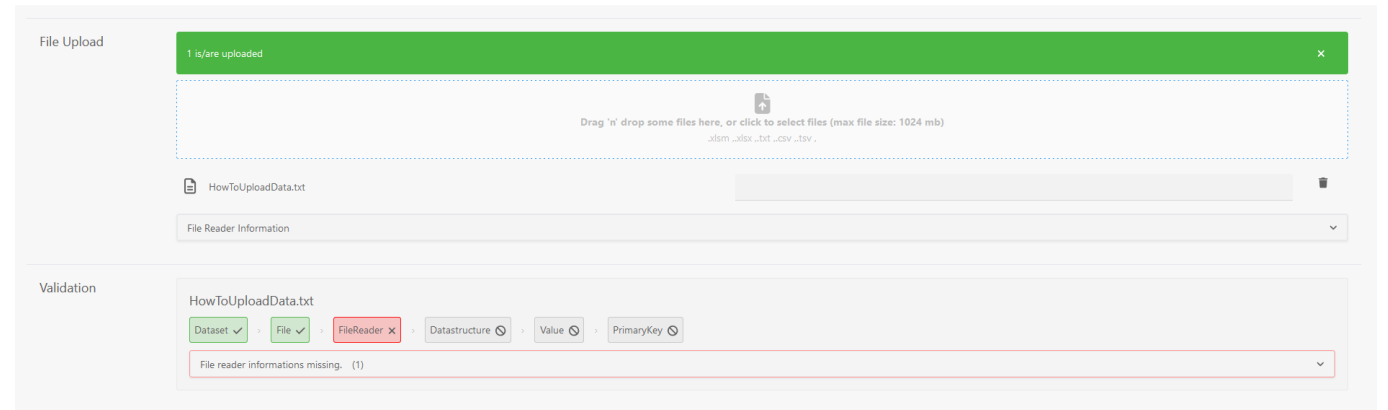
Figure 5: File upload and Validation menu.

Figure 5 also shows the validation section. The green taps show the steps successfully completed for the upload. The taps highlighted in red show what still needs to be processed.

In this example, the assignment of the file reader information such as the separator between the respective data columns, the decimal separator, the definition of text in the file and the assignment of where variables, units and data can be found in the file is still missing. To activate this, click on "File Reader Information" in the "File Upload" section. The corresponding file for which the assignments are to be made can then be selected. This is the file that was previously added using drag'n'drop, for example (Figure 6).

Figure 6: Start to modify file reader information.
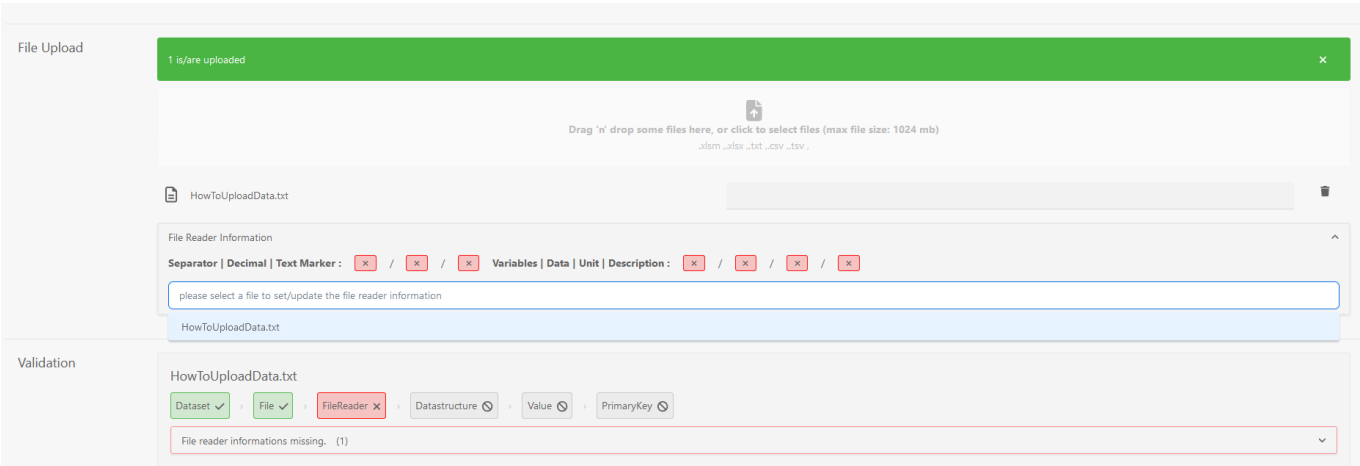
The interface opens where the file reader information can be edited (Figure 7). In the top bar, the column separator, the decimal separator, the text marker and the encoding can be changed. If one of the respective definitions is changed, you can see what has changed in the primary data display below. Normally, the system recognizes the respective definitions automatically.
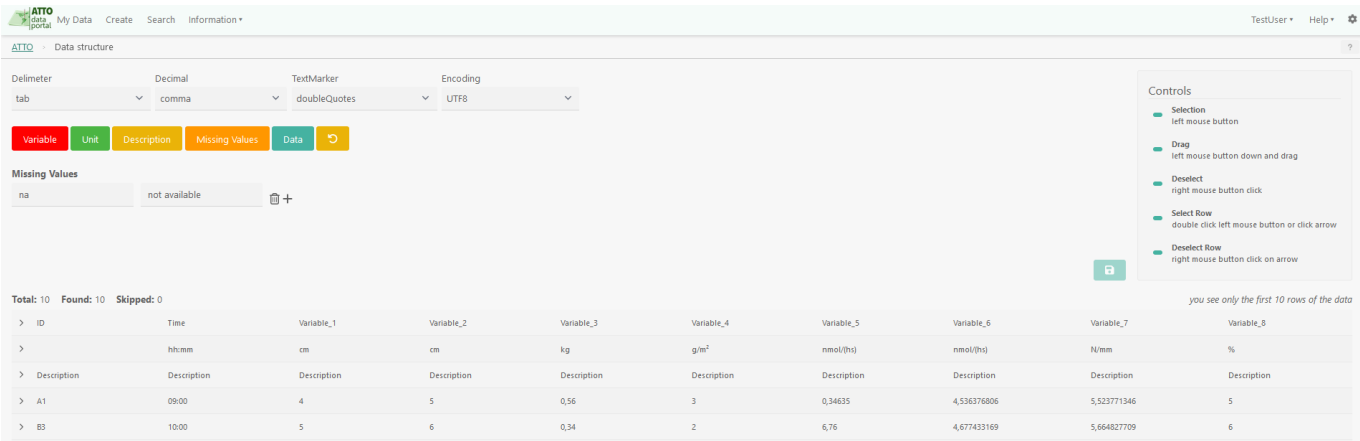

Figure 7: Modify file reader information.

In the second bar from the top you can see the color coding of the individual elements that the primary data should contain: Variables, Unit, Description of variables, Missing values and Data (Figure 8). The next step is to define the respective lines in the primary data. In our example, the variables are defined in the first row in the table below. To do this, click on the ">" sign in the left-hand column and the row will be highlighted in dark gray. Then click on the red button with "Variable" at the top and the first row in the table will turn red. Do the same for the rows in which the units and the variable description can be found. To define the values of the data, all you need to do is select the first row of values and confirm by clicking on the green "Data" button. All other lines with data values are automatically recognized by the system. The missing values can still be defined in the bar. Some use "na" or "NA" or "-9999". Please enter this here accordingly.

Once all elements have been defined, they can be saved by clicking on the green save button in the right-hand area above the data table. You will then be redirected back to the editing page.
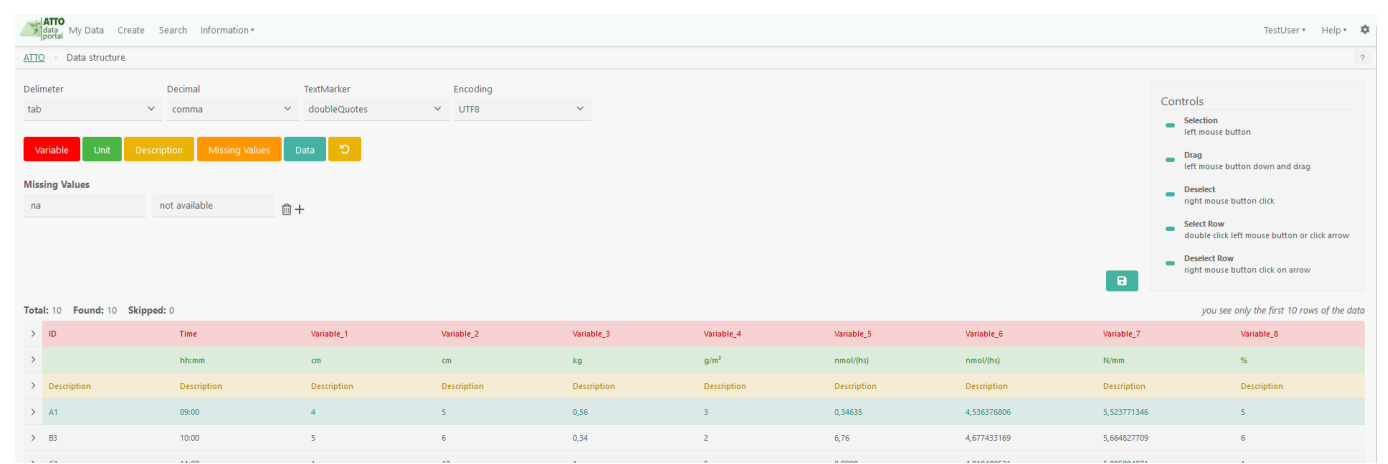
Figure 8: Define Variables, units, description and data in the uploaded primary data file.

On the editing page in the "Validation" area, the taps should now all be green (Figure 9). If this is not the case, you will need to correct this at the appropriate point.

Figure 9: Finished validation of file upload process.

If everything is now green, the dataset can be finally uploaded in the "Submit" area using the green Submit button (Figur 9). You will then be asked again for confirmation, which can be done by clicking on "Confirm" (Figure 10).
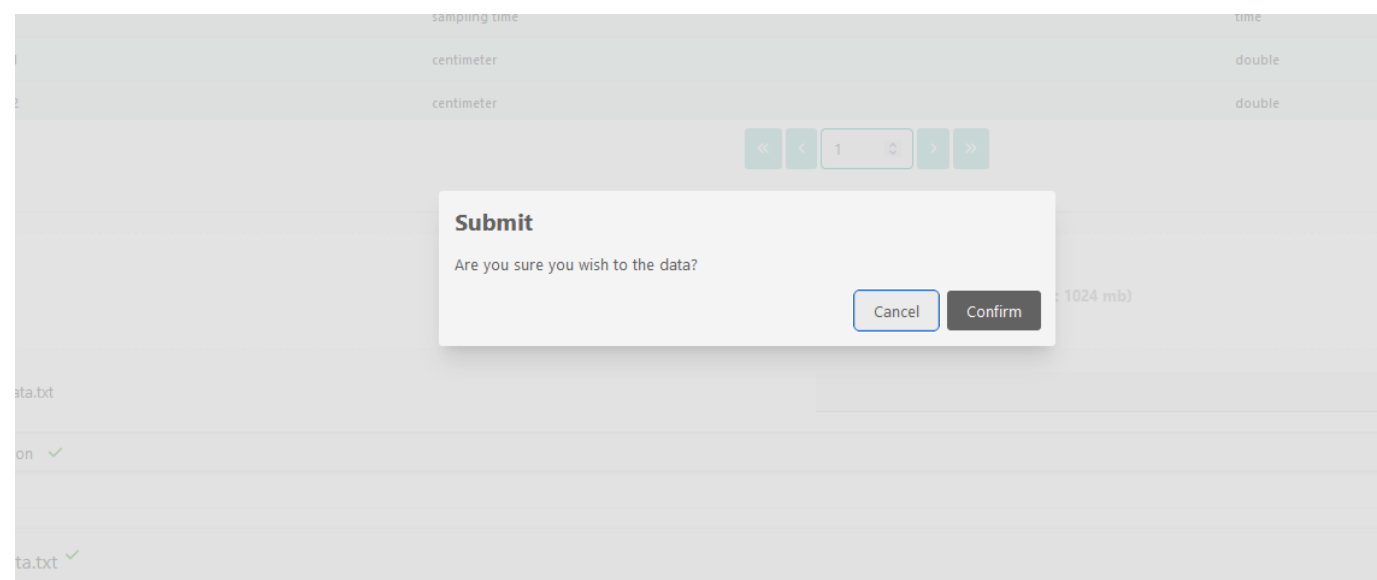
Figure 10: Submission of primary data.
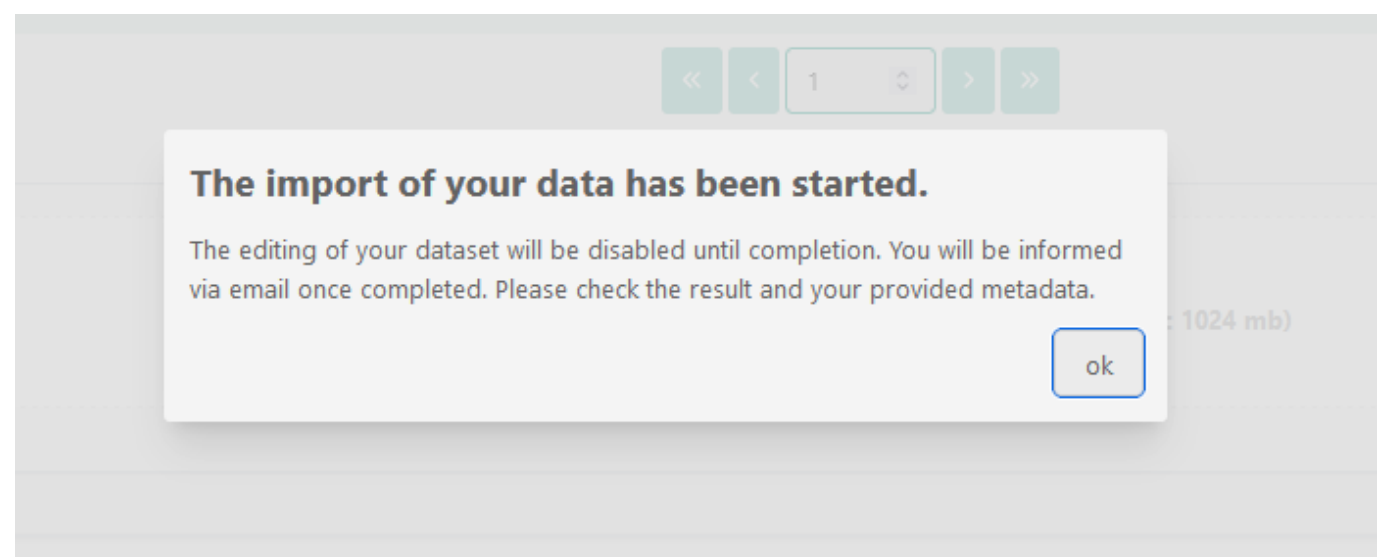
Figure 11: Confirmation window of the data import.

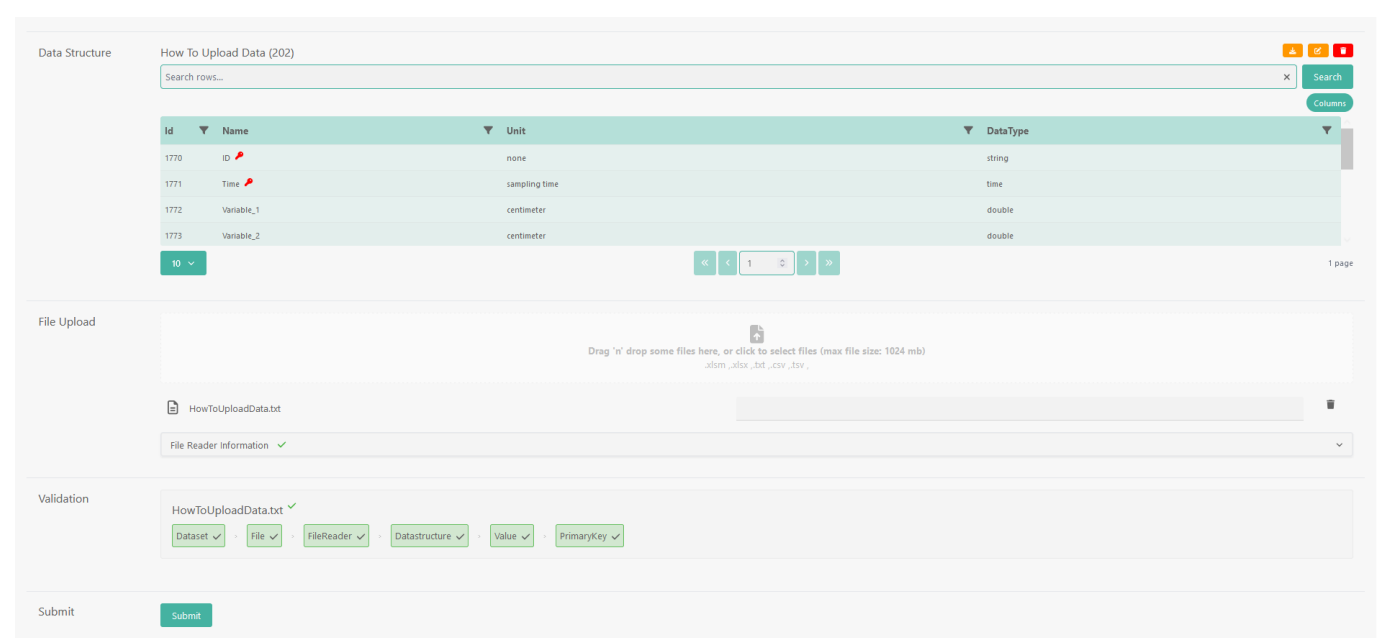## 3.2 Add attachments to datasets

In certain cases it is helpful and advisable to add additional information (descriptions as word document, text file, excel file) or a source code or images to the respective dataset. This can be done on the editing page in the lower section in the "Attachments" area (Figure 12). The respective file or several files can be attached by clicking in the field or via drag'n'drop. Corresponding instructions can also be found in the "Attachments" field.

Figure 12: "Attachments"-section in the general editing page of a dataset

Once the file has been added, the corresponding file name is displayed in the left-hand area. To the right of this is a free text field to add a description of the respective attachment (Figure 13). If a file is successfully added as an attachment, it appears in the overview. Each individual attachm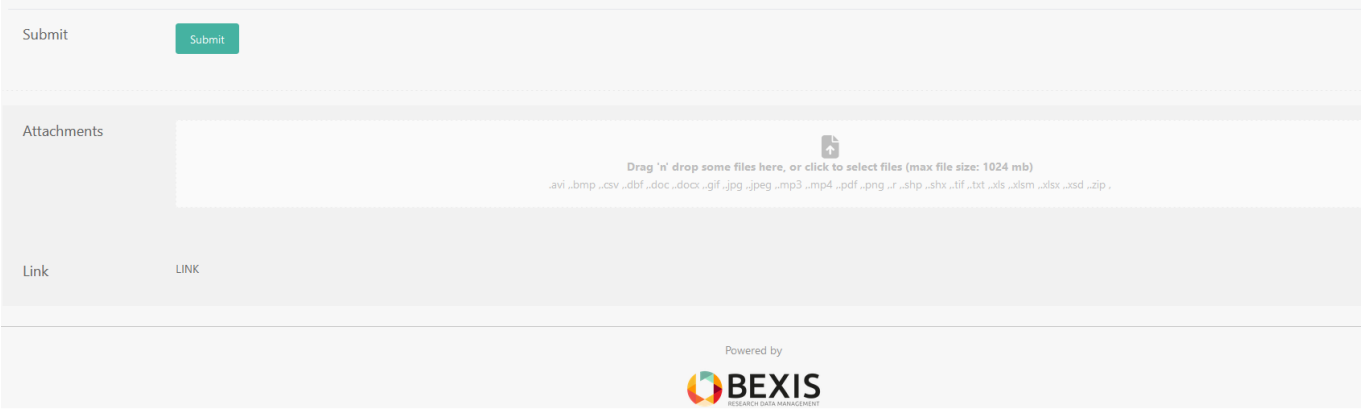ent is assigned an ID. The file name, file type and file size are also listed (Figure 14). While moving the mouse pointer over the information icon, the previously entered description of the attachment is displayed. You can also add more than one attachment to one dataset.
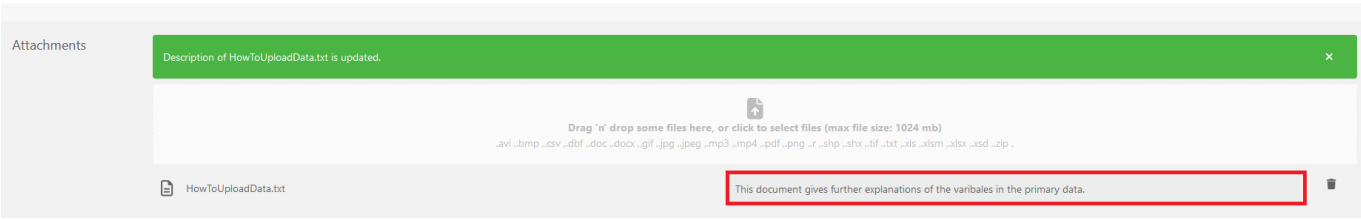
Figure 13: "Attachments"-tab with uploaded file

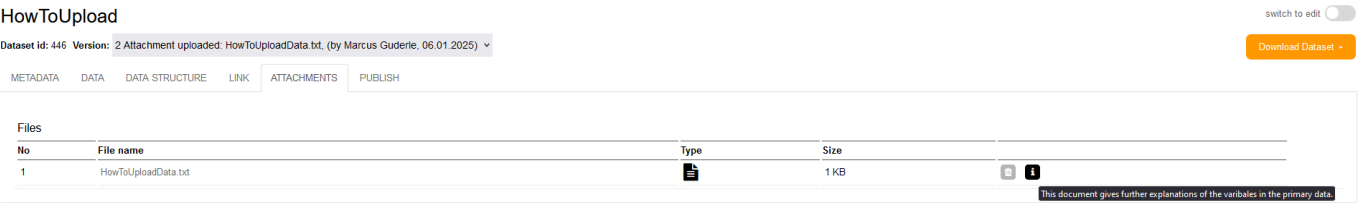*Figure 14: "Attachments"-section with description field*

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

4 Metadata form

The following section introduces the updated metadata schema for datasets following the DataCite schema.

To access the metadata form, click on the small orange pencil icon right next to "Metadata" at the top of the general editing page (Figure 15).

Figure 15: General editing page directing to the metadata form

The metadata structure contains seven main sections (Figure 16):

- Identification: Basic information on the data like citation, title, abstract, identifier URL etc.
- Data Accessibility: Information if the dataset was used for student theses and the type of license under which the dataset is published.
- Funding: Information about the funder of the project and the respective award number and title.
- Creators: Information about the contact person and the authors of the dataset. The contact persons is the person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource. Authors are the main researchers involved in producing the data, or the authors of the dataset, in priority order.
- Temporal Coverage: Information about the start and end date of the respective data collection.
- Geographical Coverage: Spatial region or named place where the data were gathered.
- Methods: Information about field sampling and data preparation as well as identifiers of other related resources, such as articles and other datasets.

Figure 16: Metadata form with seven main sections

Some fields are a free-text field, which have to be filled in by the data creator. Others have drop-down menus or an autocomplete function which 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) and Datacite.

Table 1 in the Appendix gives an overview and explanation of each variable of the metadata form. You also get this information by the information icon to the right of the respective fields. There is also an information icon at the top right of the metadata form. When this is clicked, the information for each field is displayed (Figure 16 & 17). All fields marked with a red asterisks * are mandatory for a valid metadata form.

Figure 17: Information text for each metadata field

After you complete the form, you can validate (see Figure 18) the entries and the system will give you further information in case any entries are wrong or whether more entries are required (Figure 19).

Figure 18: Validation button at the bottom left of the metadata form

Figure 19: Validation information if metadata are NOT valid

When you completed and validated the metadata form, please click "Save" (bottom right of the metadata form, see Figure 18) in order to save the entries. If there are metadata entries missing, the window shown in Figure 19 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@bgc-jena.mpg.de) for help.

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
Identification*		
Citation string	Full citation string of the dataset. To be filled by data manager.	Free text entry
Title*	Title of the dataset well describing the data (what, where, when). E.g.: Atmospheric boundary layer height dataset (2014-2020) measured at the ATTO site	Free text entry
Abstract*	A brief overview of the resource to be uploaded to the ATTO data portal. It should include basic information that summarizes the dataset, including background, kind of data collected, study site and timeframe.	Free text entry
Identifier URL	**Applies only for data sets, which will be published via MPDL/DataCite.** The uniform resource locator of the dataset. To be filled by data manager. e.g. 10.17871/atto_xxxx	Free text entry
Identifier type	**Applies only for data sets, which will be published via MPDL/DataCite.** A registered DOI for public datasets or a URL for internal datasets. To be filled by data manager.	Free text entry
Publisher	The name of the entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the resource. **In the case of the ATTO project, Publisher will be the Max Planck Institute for Biogeochemistry, Jena due to the requirements of the MPDL/DataCite.**	Automatically filled in field
Publisher identifier	Uniquely identifies the publisher, according to Research Organization Registry.	Automatically filled in field
Contributor	The institution or person responsible for collecting, managing, distributing, 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/DataCite.**	Automatically filled in field
Contributor type	The type of contributor of the resource.	Automatically filled in field
Contributor identifier	Uniquely identifies an individual or legal entity, according to Research Organization Registry (ROR).	Automatically filled in field
Publication Year	**Applies only for datasets, which will be published via MPDL/DataCite.** The year when the data was or will be made publicly available. Format: **YYYY**	Free text entry
Resource Type	The general type of the resource. E.g. Audiovisual, Collection, Dataset etc.	Dropdown menu

Variable name	Description	Type of text box
Version	The version number of the resource. To be filled by data manager.	Free text entry
Keywords	Please provide at least 3 keywords. Use a new field for each keyword, which can be added by clicking on the plus sign (E.g., flux, CO2, tower).	Free text entry
Data Accessibility*		
Data are part of student thesis*	Data used for student theses may remain under embargo for up to 6 months after thesis defense. Please refer to ATTO Data Policy (https://www.attodata.org/ato/home/Download?fileName=ATTO_DATA%20POLICY_20180917.pdf).	Dropdown menu
License	ATTO Data are published under licence CC BY 4.0, which allows datasets to be shared and adapted since appropriate credit is given and changes are indicated. Please check license URI for more details.	Automatically filled in field
License URI	The uniform resource identifier of the license.	Automatically filled in field
Funding*		
Funder*	Funder of the project, please use standard name given by the Research Organization Registry (ror.org). E.g.: Federal Ministry of Education and Research; Fundação de Amparo à Pesquisa do Estado de São Paulo.	Free text entry
Funder identifier*	URI of the funder as given by the Research Organization Registry (ror.org). E.g.: https://ror.org/04pz7b180, https://ror.org/02ddkpn78	Free text entry
Award number	The process number assigned by the funder to a sponsored grant.	Free text entry
Award title	The title of the funded project.	Free text entry
Creators*		
Contact Person*	Person with knowledge of how to access, troubleshoot, or otherwise field issues related to the resource.	
Fullname*	Fullname of the contact person/author.	Free text entry/Autocomplete list
ORCID*	ORCID is a free, unique, persistent identifier (PID) for individuals to use as they engage in research, scholarship, and innovation activities. E.g.: https://orcid.org/0000-0001-5727-2427	Free text entry/Autocomplete list

Variable name	Description	Type of text box
Affiliation*	Use the standard name of your affiliation as given by the Research Organization Registry (ror.org). E.g.: Max Planck Institute for Biogeochemistry; Instituto Nacional de Pesquisas da Amazônia.	Free text entry/Autocomplete list
Affiliation ROR*	Uniquely identifies the research institution, according to the Research Organization Registry (ror.org). E.g.: https://ror.org/051yxp643; https://ror.org/01xe86309.	Free text entry/Autocomplete list
Further Affiliation	In case the user/author has more than one affiliation.	
Affiliation	Use the standard name of your affiliation as given by the Research Organization Registry (ror.org). E.g.: Max Planck Institute for Biogeochemistry; Instituto Nacional de Pesquisas da Amazônia.	Free text entry
Affiliation ROR	Uniquely identifies the research institution, according to the Research Organization Registry (ror.org). E.g.: https://ror.org/051yxp643; https://ror.org/01xe86309.	Free text entry
Author*	The main researchers involved in producing the data, or the authors of the dataset, in priority order.	
Fullname*	Fullname of the contact person/author.	Free text entry/Autocomplete list
ORCID*	ORCID is a free, unique, persistent identifier (PID) for individuals to use as they engage in research, scholarship, and innovation activities. E.g.: https://orcid.org/0000-0001-5727-2427	Free text entry/Autocomplete list
Affiliation*	Use the standard name of your affiliation as given by the Research Organization Registry (ror.org). E.g.: Max Planck Institute for Biogeochemistry; Instituto Nacional de Pesquisas da Amazônia.	Free text entry/Autocomplete list
Affiliation ROR*	Uniquely identifies the research institution, according to the Research Organization Registry (ror.org). E.g.: https://ror.org/051yxp643; https://ror.org/01xe86309.	Free text entry/Autocomplete list
Further Affiliation	In case the user/author has more than one affiliation.	
Affiliation	Use the standard name of your affiliation as given by the Research Organization Registry (ror.org). E.g.: Max Planck Institute for Biogeochemistry; Instituto Nacional de Pesquisas da Amazônia.	Free text entry

Variable name	Description	Type of text box
Affiliation ROR	Uniquely identifies the research institution, according to the Research Organization Registry (ror.org). E.g.: https://ror.org/051yxp643; https://ror.org/01xe86309.	Free text entry
Temporal Coverage*		
Start Date*	Start date of data collection.	Calendar
End Date	End date of data collection. Only fill if end date is known.	Calendar
Temporal Resolution	The difference in time between two successive samples. E.g.: 1 min, monthly.	Free text entry
Geographical Coverage*		
Location*	Spatial region or named place where the data were gathered.	Dropdown menu
Data collection Layer*	Please inform whether data are collected below or aboveground.	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
Other location	Please provide location name if not found in the location list	Free text entry
Latitude	Latitude of "Other location" in decimals.	Free text entry
Longitude	Longitude of "Other location" in decimals.	Free text entry
Coordinate reference system	Please provide the coordinate reference system in which geographical coordinates of "Other" location" were taken.	Dropdown menu
Methods		
Field data collection*	Please provide a description of the field sampling procedures used for data collection. The content provided here would be similar to a description of sampling procedures found in the methods section of a journal article.	Free text entry
Data preparation and processing	Please describe procedures for further preparation and /or processing of samples before analysis.	Free text entry
Related resources		

Variable name	Description	Type of text box
Related resource	Please provide identifiers of other related resources, such as articles and other datasets	Free text entry
Software	Information on the software used for dataset creation, quality check and assurance.	
Application		
Name	Name of the software used for dataset creation, quality check and assurance.	Free text entry
Version	The version number of the resource. To be filled by data manager.	Free text entry
Availability		
Company name	Name of the Software developer.	Free text entry
URL	URL of the Software developer.	Free text entry
Additional Information		
Remarks	Add comments or notes about the dataset.	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.