# Swiss Data Science Center (SDSC)

# 2025 Call for Collaborative Research Data Science Projects

# **Data Form** (5 pages max.)

Use the tables in this document to provide details about the datasets that will be used in the project. Each table should represent a specific dataset or coherent part of it. Each table should be filled with specific details that are useful to understand the volume, the type and the different characteristics for each dataset. In particular, whenever possible, specify the number of observations, the number of variables / features / dimensions, the number of annotations / labels, etc. Provide a name for each dataset and answer the questions provided in the templates. You are welcome to refer to these tables from the proposal text, by using the corresponding dataset name.

Note that most of the required data should be available (and annotated if necessary) prior to the beginning of the project, given that the project lasts only two years. However, if the goal of the project is to develop methods to help acquire and / or annotate data, these are obviously not needed. Nonetheless, it is important to describe the data to be acquired. In cases where DTUAs or ethical committee approval is required, accessibility by April 2027 is acceptable, provided that supporting documents demonstrating the feasibility of this timeline are submitted

We provide templated tables for common types of data (data array, images, time series, video, text, georeferenced data, etc.), but you are welcome to add custom tables for particular data types and add/remove specific points in each table as needed. You have the possibility of submitting a sample of your data and metadata on CMT for illustrative purposes (treated as confidential and for internal review only).

In particular, to the extent possible, please:

(1) list separately different data types, even if there are correspondences between the datasets (e.g., pictures of plants & genetic markers for the same plants).

(2) list separately datasets composed of data of the same type if the observations in each dataset correspond to different acquisition protocols (E.g., two databases of spectroscopy data acquired under different conditions).

Please indicate for each data sets if there are particular procedures or security requirement to access and work with these datasets and specify them.

|  |  |  |
| --- | --- | --- |
| Dataset 1 *(replace with dataset name)* | Data type | IMAGE |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* What relevant information do the images contain?
* What was the protocol used to acquire the data?
* In which part of the project is this data needed? (*short answer*)
* Number of images:
* Image size: *(in pixels by pixels)*
* What is the size of the objects of interest compared to the resolution of the image?
* What preprocessing is already done?
* What further (pre-)processing will be needed? (by the SDSC? by the partners?)
* Is the data corrupted (noise, blur, etc.)? How significant are the biases and uncertainties on the images and on the labels?
 |
| Additional comments or remarks |

|  |  |  |
| --- | --- | --- |
| Dataset 2 *(replace with dataset name)* | Data type | TIME SERIES |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* What phenomenon do these time series capture?
* What was the protocol used to acquire the data?
* In which part of the project is this data needed? (*short answer*)
* If the time series is multivariate, how many dimensions does it have and to what do they correspond?
* Number of times series:
* Lengths of time series: *(specify a range)*
* Is the sampling frequency high or low with respect to the signal of interest?
* What preprocessing is already done?
* What further preprocessing will be needed? (by the SDSC? by the partners?)
* What kinds of data corruption present a challenge (noise, etc.)? How significant are the biases and uncertainties on the measurements and on the labels?
* What is the proportion of missing data? Is the data missing at random, or is there a specific reason for missing entries?
* Is the data expected to change over time (e.g., drift), over repetitions of the same experiments or over a change of sensors?
 |
| Additional comments or remarks |

|  |  |  |
| --- | --- | --- |
| Dataset 3 *(replace with dataset name)* | Data type | DATA ARRAY |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* In which part of the project is this data needed? (*short answer*)
* How many measurements/features/entries are there per record/observation?
* How many records/observations are there?
* What preprocessing is already done?
* What further preprocessing will be needed? (by the SDSC? by the partners?)
* What kinds of data corruption present a challenge (noise, etc.)? How significant are the biases and uncertainties on the measurements and on the labels?
* What is the proportion of missing data? Is the data missing at random, or is there a specific reason for missing entries?
 |
| Additional comments or remarks |

|  |  |  |
| --- | --- | --- |
| Dataset 4 *(replace with dataset name)* | Data type | TEXT CORPUS |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* What is the nature of the corpus?
* What was the protocol used to acquire the data?
* In what language(s) is the corpus?
* How are the texts used in the project?
* In which part of the project is this data needed? (*short answer*)
* How many texts are there, and what are their lengths?
* Is the text already extracted (OCR) and in digital form?
* What preprocessing is already done?
* What further preprocessing will be needed? (by the SDSC? by the partners?)
* What kinds of data corruption present a challenge? Does the corpus present some biases for the task considered? Are the labels uncertain?
 |
| Additional comments or remarks |

|  |  |  |
| --- | --- | --- |
| Dataset 5 *(replace with dataset name)* | Data type | GEOREFERENCED DATA |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* What is the nature of the measurements (e.g., geolocated time series, geolocated images, geolocated physical model outputs)?
* What was the protocol used to acquire the data?
* In which part of the project is this data needed? (*short answer*)
* What is the spatial resolution/number of measurements in space?
* Are the temporal measurements regularly spaced?
* How many temporal measurements are there/what is the temporal frequency of the data?
* Is the sampling frequency high or low with respect to the signal of interest?
* What preprocessing is already done?
* What further preprocessing will be needed? (by the SDSC? by the partners?)
* What kinds of data corruption present a challenge (noise, etc.)? How significant are the biases and uncertainties on the measurements and on the labels?
* What is the proportion of missing data? Is the data missing at random, or is there a specific reason for missing entries?
* Is the data expected to change over time (e.g., drift), over repetitions of the same experiments or over a change of sensors?
 |
| Additional comments or remarks |

|  |  |  |
| --- | --- | --- |
| Dataset 6 *(replace with dataset name)* | Data type | VIDEO / 4D DATA |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* What relevant information do the videos contain?
* What was the protocol used to acquire the data?
* In which part of the project is this data needed? (*short answer*)
* Frame rate:
* Range of durations of the videos:
* Frame size (in pixels):
* Number of videos:
* What is the resolution in space and time relative to the objects of interest?
* What preprocessing will already be done?
* What further preprocessing will be needed? (by the SDSC? by the partners?)
* What kinds of data corruption present a challenge (noise, motion blur, etc.)? How significant are the biases and uncertainties on the data and on the labels?
* Is the data expected to change over time (e.g., drift), over repetitions of the same experiments or over a change of sensors?
 |
| Additional comments or remarks |

|  |  |  |
| --- | --- | --- |
| Dataset 7 *(replace with dataset name)* | Data type  | OTHER / COMPLEX / STRUCTURED DATA |
| * Is this dataset fully acquired and readily usable? Security requirements, if any?
* Is the data going to be updated during the project?
* Does the dataset still need to be annotated by the partners? If it does, how much work is required?
* Is the data openly accessible and/or will it be made openly accessible?
* What does the data represent?
* What is the structure of the data?
* What was the protocol used to acquire the data?
* In which part of the project is this data needed?
* What preprocessing will already be done?
* What further preprocessing will be needed? (by the SDSC? by the partners?)
* What is the proportion of missing data? Is the data missing at random, or is there a specific reason for missing entries?
* What kinds of data corruption present a challenge (noise, motion blur, etc.)? How significant are the biases and uncertainties on the labels?
* Is the data expected to change over time (e.g., drift), over repetitions of the same experiments or over a change of sensors?
 |
| Additional comments or remarks |