Standard Operating Procedure

Number: **003**

Name: **DR instructions**

Title: Instructions for creating a Dataset Report (DR) hosted on the EU Pollinator Hub

Category: Data management

Status: Approved

Version: 1.0

Date: 2023-06-30

Author: Michael Rubinigg

This document is intended for the regulation of processes within the EU Pollinator Hub. It is published under the CC BY 4.0 licence. You may distribute, transform, adapt and build upon the material in any medium or format, including for commercial purposes, as long as you give credit to the author.

Table of contents

1 Document history 3

1.1 Current Release 3

1.2 Change history 3

2 Abbreviations 4

3 Introduction 5

4 Validity 6

4.1 Area 6

4.2 Time 6

4.3 Related Documents 6

4.4 Related Tasks 6

5 Procedure 7

5.1 Drafting 7

1 Document history

1.1 Current Release

1.1.1 Drafted

Submitted on 2023-06-05 by Michael Rubinigg, quality manager.

1.1.2 Reviewed

Reviewed on 2023-06-30 by Noa Simòn Delso, general manager.

1.1.3 Approved

Approved on 2023-06-30 by Michael Rubinigg, quality manager.

1.2 Change history

n.a.

2 Abbreviations

EU	European Union
EUPH	EU Pollinator Hub
n.a.	not available
UID	Unique Identifier
SOP	Standard Operating Procedure

3 Introduction

The process of accepting, preparing and integrating data on the EU Pollinator Hub is documented in a Dataset Report (DR). The information contained in the DR ensures full traceability of the process of dataset preparation and integration on the EU Pollinator Hub (EUPH). The present SOP contains the instructions for automated or manual drafting or editing this report.

4 Validity

4.1 Area

The of validity of this document is limited to the drafting of Dataset Reports as part of the process of data preparation – including data profiling and data cleansing – and its integration – including standardisation – on the EUPH from the moment a dataset enters the EUPH to the moment it is integrated into the EUPH data model.

4.2 Time

The period of validity of this SOP begins on 2023-06-30 and ends with its revocation.

4.3 Related Documents

Document type	Document ID	Title
SOP	001	SOP instructions
SOP	004	Organisation chart
SOP	006	Dataset preparation
SOP	007	Data profiling
SOP	008	Data cleansing
SOP	018	Data exploration
SOP	019	Data visualisation

4.4 Related Tasks

Persons who have been assigned one of the following tasks, as defined in SOP-004, assume a responsibility within this SOP:

- Quality management

5 Procedure

5.1 Drafting

5.1.1 General rules

5.1.1.1 Editing

- The person responsible for quality management provides a template in an appropriate format, which shall be used for drafting a Dataset Report.

5.1.1.2 Text

- In the present SOP, text to be included as such in the Dataset Report to be drafted is written in ***bold italic***, information intended as placeholder is written between [square brackets].
- Text in the Dataset Report shall appear as follows:
  - General: Font type Calibri, full justified; line spacing 1,5 lines; alignment justified
  - Body text: Font weight normal; font size 11 pt
  - Headings: Font weight normal; font size 14 pt (first level), 13 pt (second level), 12 pt (third level), 11 pt (fourth level). First level headings start after a page break.
  - Code of programming languages or SQL queries shall be written:
    - in `Courier New`, 10 pt within body text;
    - in `Courier New`, 8 pt; line spacing 1,0 lines in any other case;
  - The name of databases, tables of databases, columns of tables, descriptors of columns, directories, and files shall be written in *italic*.
  - The reported content of fields in a database shall be enclosed in braces, separated by comma. Example: {1, one, NULL}
  - The symbol to be used as decimal separator is the baseline dot.

5.1.1.3 Dates

- Dates are given in the following format, including a leading zero: [YYYY-MM-DD] where Y stands for year, M for month and D for day.

5.1.1.4 Document header

The document header contains information on the institution by which the dataset has been issued.

- Even page numbers: Name of the institution (***EU Pollinator Hub***), left aligned.
- Odd page numbers, except for cover page, left aligned: ***DR*** [UID of the dataset] [dataset name]***, Version*** [version number]

- Odd page numbers, including cover page, right aligned: Logo of the institution.

5.1.2 Cover page

- The cover page contains a summary of the most important control information of the dataset required to configure management and compliance standards:
  - ***Dataset Report***, font weight normal; font size 20 pt; center alignment
  - ***UID:*** The UID of the dataset written in **bold letters**.
  - ***Name:*** A clear, concise name of the dataset, containing no more than 30 characters, including blank spaces, written in **bold letters**.
  - ***Title:*** A long title of the dataset, clearly describing the scope of the dataset containing no more than 256 characters including blank spaces.
  - ***Status:*** One of the following status information: ***Draft*** | ***Approved*** | ***Expired***.
  - ***Version****:* Version number.
  - ***Date****:* Date of release of the current version.
  - ***Author****:* Name of the author(s) of the current release of the Dataset Report.
  - Disclosure rules: Text containing the explicit statement by whom the document is intended to be used as well as to whom, by whom and for which purpose the document may be disclosed.

5.1.3 Table of contents

- A Dataset Report may include a section ***Table of Contents*** at two levels with the page number.

5.1.4 Document history

- Section ***1 Document history*** contains a complete list with the following information on previous versions and/or releases of the Dataset Report in decreasing chronological order.
  - ***Version****:* Version number.
  - ***Date****:* Date of release of the current version.

5.1.5 Release

- Section ***1.1 Release*** contains the following technical information in the current release:
  - ***Version*** [version number].
  - ***released on*** [date of release].
  - ***Written by*** [full name of all authors separated by comma]**.**
  - ***Reviewed by*** [full name of all reviewers separated by comma].
  - ***Approved by*** [name of person responsible for the release of reports].

5.1.6 Revision

- Section ***1.2 Revision*** contains a tabular summary of the revisions made within each release:
  - ***No***: Identifier of the table row (incrementing integer).
  - ***Date***: Date of the finalisation of a revision.
  - ***Description***: Short description of the reviewed item before and after the revision.
  - ***Reason***: Reason for revision.

5.1.7 Abbreviations

- Section ***2 Abbreviations*** contains a list of abbreviations used in the document in increasing alphabetical order and the description.
  - Abbreviation
  - Description

5.1.8 Executive Summary

- Section ***3 Executive Summary*** contains a concise summary of the following items:
  - ***Data overview:*** Concise information (less than 200 words) about: *e.g.* provider of the data (who?); origin of the data and the subject area covered by the dataset (what?); geographical origin of the data (where?); time period covered by the dataset (when?).
  - ***Data value:*** Concise information (less than 200 words) about: *e.g.* reason for collecting the data; possible benefits resulting from the collection of the data (why?).
  - ***Data description:*** Concise description (less than 200 words) of the dataset metrics: *e.g.* number of tables; size of the datasets in bytes.
  - ***Data application:*** A concise summary (less than 200 words) of possible applications for which the data could be used, scientific or technical questions or other issues that could be addressed (for what?).

5.1.9 Introduction

Section ***4 Introduction*** contains a detailed discussion of the specific reasons for which the dataset has been collected and possible benefits resulting from the collection of the dataset, detailed information on the provider of the dataset, its geographical origin, the time period and subject area covered by the dataset.

5.1.10 Material and Methods

- Section ***5 Material and Methods*** contains the material and methods used to prepare, clean, validate, and analyse the data. If available, the name of material (equipment, software

applications) must comprise the name of the producer and model or version number. If available, references to methods should be provided.

5.1.10.1 Data Acquisition

- Section ***5.1 Data Acquisition*** contains:
  - A description of or a reference to material and methods (provide references if possible) used to acquire the data in the dataset.
  - A table with the names of all files containing data and/or metadata that were made available by a data provider.
    - ***No***: Identifier of the row (incrementing integer).
    - ***File***: Name of the file.
    - ***D***: **YES** if the file contains data, **NO** if it contains no data.
    - ***M***: ***YES*** if the file contains metadata, **NO** if the file contains no metadata.
    - ***Arrival***: Date and time of arrival of the file at the EU Pollinator Hub.
    - ***Provider***: Acronym or, if not available, name of the data provider.

5.1.10.2 Data Preparation

- Section ***5.2 Data Preparation*** contains a detailed description of material and methods used to prepare the dataset.

5.1.10.3 Data Validation

- Section ***5.3 Data Validation*** contains a detailed description of material and methods used to validate the data in the dataset for the purpose of data quality assessment.

5.1.10.4 Data Analysis

- Section ***5.4 Data Analysis*** contains a detailed description of material and methods used to analyse the data in the dataset for the purpose of data quality assessment.

5.1.11 Dataset description

5.1.11.1 First level section

- First level section ***6 Data description***.

5.1.11.2 Second level section containing metadata of the dataset

- Second level section ***6.1 Dataset*** contains a tabular summary of the standardised metadata associated with the dataset.
  - A tabular summary of all tables contained in the dataset.

- ***UID***: Unique Identifier of the table.
- ***Table***: Name of the table.
- ***Description***: Description of the table.

- A tabular summary containing the name of the following parameters (***Parameter***) and their content (***Content***) of the standardised metadata of the dataset:
  - ***UID***: Unique Identifier of the dataset on the EUPH.
  - ***Name***: Name of the dataset.
  - ***Title***: Title of the dataset.
  - ***Licence***: Licensing type of all the data contained in the dataset. If individual licences are applied for the tables, write ***See tables***.
  - ***Published***: Date and time at which the dataset becomes available on the EUPH. If no date has been set write **No date set**.
  - ***Description***: Detailed description of the dataset using less than 100 words and containing one or more of the following information: What does the data measure or track? Who provided the data? Which period does it cover? For whom is the data relevant?
- A tabular summary containing the name of the following parameters (***Parameter***) and their content (***Content***) of the standardised metadata of each data provider:
  - ***UID***: Unique identifier of the data provider.
  - ***Name***: Full name of the data provider.
  - ***URL***: URL of the data provider's website.
  - ***Acronym***: Acronym of the data provider.
  - ***Description***: A detailed description of the data provider.

5.1.11.3 Second level section containing metadata of the tables

- Second level section ***6.2 Tables*** contains one or more third level sections with information on the tables associated to this dataset.

5.1.11.4 Third level section [table name]

- All third level sections contain descriptive information on the tables from a given data provider described in the parent section:
  - A tabular summary containing the name of the following parameters (***Parameter***) and their content (***Content***) of the standardised metadata of each table:
    - ***UID***: Unique Identifier of the data used as part of the URL on the EUPH.
    - ***Name***: Name of the table.
    - ***Licence***: Licence type of the data contained in the table.

- ***Description***: A concise description of the table using less than 100 words and containing one or more of the following information: Which information does the table contain? Which subject area, period and geographic area does it cover?

- Name of the table.
- Detailed description of the information contained in the table.
  - Detail description of columns.
- Detailed metrics of the data contained in the table:
  - Groups, categories.
  - Appropriate measures of size (*e.g.* number of records, % of total records, bytes).
  - Appropriate measures of central tendency (*e.g.* arithmetic mean, median).
  - Appropriate measures of dispersion and/or variability (*e.g.* range, standard deviation).
  - Anomalies (*e.g.* biases).
- Contextual information on:
  - Geographical origin of the data contained in the table.
  - Subject area and time period covered by the data contained in the table.
- Detailed results from the process of data exploration, as defined by SOP-018.
- Detailed results from the process of data visualisation, as defined by SOP-019.

5.1.11.5 Metadata

- Fourth level section ***Metadata*** contains:
  - An unordered list of potential links between columns of the same table, of different tables of the same dataset or of different tables of different datasets as follows: ***Column*** [*column*] ***is joined to*** [*dataset*].[*table*].[*column*]
  - A tabular summary of the structural analysis of the table described in the parent section:
    - ***Name***: Name of the column.
    - ***Description***: Concise description of the column as provided by the data provider.
    - ***Data type***: Data type in which values are provided.
    - ***Descriptor***: Name of the EUPH Descriptor.
    - ***Unit***: Name of the Unit in which the values are provided.

5.1.11.6 Descriptive measures

- Fourth level section ***Descriptive measures*** contains a tabular summary of the content analysis of the table described in the parent section.
  - ***Name***: Name of the column.
  - ***Length***: Range of length of characters

- ***Mean***: Arithmetic mean of all (numeric) values in the column. N/A if not applicable.
- ***Min***: Lowest value in the column.
- ***Max***: Highest values in the column.
- ***Total***: Total number of records.
- ***Missing***: Number and percentage (between brackets, rounded to one decimal place) of values in a column containing NULL.
- ***Zero***: Number and percentage (between brackets, rounded to one decimal place) of values in a column containing the value 0.
- ***Blank***: Number and percentage (between brackets, rounded to one decimal place) of values in a column containing exclusively one or more blank characters.
- ***Distinct***: Number and percentage (between brackets, rounded to one decimal place) of distinct values in a column, including NULL, Zero and blank.

5.1.11.7 Changes made to preparatory file

- Fourth level section **Changes made to preparatory file** contains an ordered list of all changes made in the preparatory file described in the parent section according to SOP-006.

5.1.11.8 Changes made to data

- Fourth level section **Changes made to data** contains an ordered list of all changes made in the data during data preparation according to SOP-008.

5.1.11.9 Unresolved issues

- Fourth level section **Unresolved issues** contains an ordered list of all unresolved issues in the table of the parent section requiring intervention by the data provider according to SOP-007 and SOP-018.

5.1.12 References

- Section ***7 References*** contains an ordered list of referenced documents with the exception of other quality assets.
  - *[Id]*: Identifier of the reference.
  - *[List of authors]*
    - In the case of author(s) being physical persons ([*first letter of Name*].[*first letter of Middle Names*]. [*Surname*]) repeat for all authors.
    - In the case of author(s) being juridical persons provide name of institution.
  - Publication date *[Year-Month-Day-Hour-Minute]*

- *[Title]*
- *[Publication]*
- *[List of editors]*
    - In the case of author(s) being physical persons ([first letter of Name].[first letter of Middle Names]. [surname]) repeat for all authors.
    - In the case of author(s) being juridical persons provide name of institution.
- *[Edition]* ***edition***.
- *[Volume]*, *[Issue]*, *[Pages]*
- *[Publisher]*, *[Town], [Country]*
- In case of websites *[Url],* ***Accessed on*** [Date and time of access].

5.1.13 Annexes

- If applicable, section ***8 Annexes*** may contain additional information, *e.g.* queries, programming code, *etc.* used to process data, which are not described in existing work instructions.