Standard Operating Procedure

Number: **006**

Name: **Dataset preparation**

Title: Instructions for the preparation of datasets for integration into 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 Responsibilities 7
5.2 Arrival of a dataset 7
5.3 Dataset Report 8
5.4 Preparation of a dataset 9
6 References 10
7 Annexes 11

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

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

3 Introduction

This document describes the processing of a dataset from the delivery by the supplier of the dataset until its integration into the EU Pollinator Hub.

4 Validity

4.1 Area

The of validity of this document is limited to the processing of a dataset from the delivery by the supplier of the dataset until its integration into the EU Pollinator Hub.

4.2 Time

The period of validity of this SOP begins on 2022-10-05 and ends with its revocation.

4.3 Related Documents

Document type	Number	Name
SOP	001	SOP instructions
SOP	003	DR instructions
SOP	004	Organisation chart
WI	002	Raw data preparation
WI	003	pMA Setup
WI	004	pMA Data Import

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:

- Data preparation

5 Procedure

5.1 Responsibilities

- The person responsible for data preparation, as defined by SOP-004, or a person appointed by him/her for this purpose, is responsible for the management of the processes defined in this SOP.
- The task of carrying out certain parts of this process can be transferred to another person.

5.2 Arrival of a dataset

5.2.1 Acceptance of a dataset

- This section describes the procedure of acceptance of a dataset from a data provider.

5.2.1.1 Creation of a directory for new dataset

- For new datasets a new directory is created in the *Data* directory. The name of the directory consists of a unique consecutive integer the name of the dataset.
- Within the newly created directory, the following subdirectories are created:
  - ***01 Raw Data***
  - ***02 Cleaned Data***
  - ***03 Processed Data***
  - ***04 Documents***

5.2.1.2 Archiving

- All delivered files are saved in directory *01 Raw Data* of the directory created for the dataset.

5.2.1.3 Check of integrity and completeness

- The delivered files are tested for the presence of malware.
- The delivered files are opened with a suitable software and the readability is checked.
- Face validity is checked by visually inspecting the content of the file.

5.2.1.4 Take-over confirmation to the supplier of the dataset

- The supplier of the dataset receives a written confirmation on the arrival of the files via e-mail.
- The supplier is informed of any anomaly encountered during check of integrity and completeness as described in section 5.2.1.3.

5.3 Dataset Report

5.3.1 Starting a Dataset Report

5.3.1.1 Creating a file

- The template provided for drafting a Dataset Report is saved in directory *02 Cleaned Data* of the directory created for the dataset.
- The name of the file consists of ***DR***, followed by an underscore, followed by the dataset block and the version block of the UID used for the dataset directory (5.2.1.1), followed by an underscore, followed by ***V***, followed by the version number of the dataset report.

5.3.2 Recording the origin of datasets

5.3.2.1 Documentation of raw data files

- The following information relating to raw data is recorded in section *5.1 Data Acquisition* of the Dataset Report.
  - A complete list of names of all files containing data and/or metadata submitted by the data provider.
  - The date of arrival of the dataset.
  - An accurate and traceable description of the source from which the datasets were obtained.

5.4 Preparation of a dataset

5.4.1 Evaluation of a dataset

5.4.1.1 Structural evaluation

- Based on the existing structure of a dataset, it is decided whether it can be integrated as delivered or whether the dataset has to be prepared manually.

5.4.2 Manual preparation of a dataset

5.4.2.1 Preparation of transfer files

- Files for the transfer of raw data to the data preparation and profiling tool are prepared according to WI-002.
- All changes to data in the preparatory file, as defined by WI-002, which are necessary to assure a complete and accurate import of data into a data preparation and profiling tool, are documented in the appropriate paragraph of the Dataset Report:
  - The name of the column where a change was applied.
  - At least the number of records per column in which data had to be modified or completed.
  - An accurate description of changes that were applied.
  - A reason for the change.
- All unresolved issues encountered during the procedure are documented in the appropriate paragraph of the Dataset Report.

5.4.2.2 Preparation of raw data

- Data preparation and profiling tools are configured according to WI-003.
- All issues encountered during this procedure are documented in the appropriate paragraph of the Dataset Report.

5.4.2.3 Import

- Data is imported into the data preparation and profiling tool according to WI-004.
- All issues encountered during this procedure are documented in the appropriate paragraph of the Dataset Report.

6 References

There are no sources in the current document.

7 Annexes

There are no annexes in the current document.