Standard Operating Procedure

Number: **007**

Name: **Data profiling**

Title: Instructions for profiling of datasets 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 Responsibilities 7

5.2 Descriptive analysis of datasets 7

6 References 9

7 Annexes 10

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
SOP	Standard Operating Procedure
WI	Work Instruction

3 Introduction

Data entering the EU Pollinator Hub or data running through a revision process is subjected to a profiling process, which includes a descriptive statistical analysis of data values and an exploration of relationships between data elements both within the dataset and across datasets. The results of this profiling process are used to produce a Dataset Report.

4 Validity

4.1 Area

The of validity of this document is limited to the descriptive analysis of datasets entering the BeeHub Operational Platform or having passed a revision process.

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
WI	005	pMA Profiling
WI	023	Data profiling

4.4 Related Tasks

- 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 person responsible for data preparation, as defined by SOP-004, assures that all individuals who carry out activities covered by this SOP are familiar with the processes described in the SOP.

5.2 Descriptive analysis of datasets

5.2.1 Structural analysis

5.2.1.1 Performing and reporting the structural analysis using phpMyAdmin

- Structural parameters of a column are assessed according to WI-005.
- All previous steps of section 5.2.1.1 are repeated for all columns in a table.
- All previous steps of section 5.2.1.1 are repeated for all tables of a dataset.

5.2.1.2 Performing and reporting the structural analysis using the EUPH Profiling Tool

- Structural parameters of a column are assessed according to WI-023.
- All previous steps of section 5.2.1.2 are repeated for all columns in a table.
- All previous steps of section 5.2.1.2 are repeated for all tables of a dataset.

5.2.2 Content analysis

5.2.2.1 Performing ads reporting of the content analysis using phpMyAdmin

- The content of a column is assessed according to WI-005.
- All previous steps of section 5.2.2.1 are repeated for all columns in a table.
- All previous steps of section 5.2.2.1 are repeated for all tables of a dataset.

5.2.2.2 Performing ads reporting of the content analysis using the EUPH Profiling Tool

- The content of a column is assessed according to WI-023.
- All previous steps of section 5.2.2.2 are repeated for all columns in a table.
- All previous steps of section 5.2.2.2 are repeated for all tables of a dataset.

5.2.3 Identification of duplicate records

5.2.3.1 Assessment and reporting of the content of a file using phpMyAdmin

- The content of a table is analysed according to WI-005.
- All previous steps of section 5.2.3.1 are repeated for all columns in a table.

- All previous steps of section 5.2.3.1 are repeated for all tables of a dataset.

6 References

There are no sources in the current document.

7 Annexes

There are no annexes in the current document.