Standard Operating Procedure

Number: **018**

Name: **Data exploration**

Title: Procedure for data exploration on data hosted on the EU Pollinator Hub

Category: Data management

Status: Draft

Version: 1.0

Date: 2024-02-07

Authors: 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 Automated and semi-automated exploration 7
5.3 Manual data exploration 8
6 References 12
7 Annexes 13

1 Document history

1.1 Current Release

1.1.1 Drafted

Submitted on 2023-06-29 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 2024-02-07 by Michael Rubinigg, quality manager.

1.2 Change history

n.a.

2 Abbreviations

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

3 Introduction

The current document describes the process of data exploration with the aim to understand the characteristics of data provided to the EU Pollinator Hub, to describe these characteristics in the Dataset Report and to report inconsistencies found during this process.

4 Validity

4.1 Area

The of validity of this document is limited to the process of data exploration with the aim to understand the characteristics of data provided to the EU Pollinator Hub.

4.2 Time

The period of validity of this SOP begins on 2024-02-07 and ends with its revocation.

4.3 Related Documents

Document type	Number	Name
SOP	001	SOP instructions
SOP	004	Organisation chart
SOP	007	Data profiling
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:

- Data exploration

5 Procedure

5.1 Responsibilities

- The person responsible for data exploration, as defined by SOP-004. Is responsible for all processes described in this SOP.
- The task of carrying out certain parts of this process can be transferred to another person.

5.2 Automated and semi-automated exploration

5.2.1 Data profiling

5.2.1.1 Autmated and semi-autoimated data profiling

- The results from data profiling, described in SOP-007 or obtained in the data profiling tool of the EU Pollinator Hub, are considered in this procedure.

5.2.1.2 Reporting

- Results from this analysis must be reported in the Dataset Report as described in SOP-003.

5.2.2 Data visualisation

5.2.2.1 Autmated and semi-autoimated data visualisation

5.2.2.2 Reporting

- The results from data profiling, described in SOP-019, or obtained from the data visualisation tool of the EU Pollinator Hub are considered in this procedure.
- Results from this analysis may be reported in the Dataset Report as described in SOP-003.

5.3 Manual data exploration

5.3.1 Identity resolution

5.3.1.1 Identification of distinct data elements referring to same real-word identity

- Columns are analysed for distinct data elements referring to the same real-world entity using appropriate data management software.
- Results from this analysis are reported in the Dataset Report according to SOP-003 in an appropriate way.
- Deviations from the expected results are reported in the Dataset Report according to SOP-003. The following parameters are tracked:
  - Name of the column where the issue was detected.
  - Number of records per column that are affected by the issue.
  - Queries used to execute this process.
- All previous steps of section 5.3.1.1 are repeated for all columns in a table.
- All previous steps of section 5.3.1.1 are repeated for all tables of a dataset.

5.3.2 Linkage

5.3.2.1 Linkage of columns with datasets hosted on the EUPH

- Relevant columns of a table (referred to as base table) are linked to relevant columns of other tables (referred to as target table) in order to verify the occurrence of all data elements of the target table in the base table using appropriate data management software.
- Results from this analysis are reported in the Dataset Report according to SOP-003 in an appropriate way.
- Deviations from the expected results are reported in the Dataset Report according to SOP-003. The following parameters are reported:
  - Name of the column where the issue was detected.
  - Number of records per column that are affected by the issue.
  - Queries used to execute this process.
- All previous steps of section 5.3.2.1 are repeated for all columns in a table.
- All previous steps of section 5.3.2.1 are repeated for all tables of a dataset.

5.3.3 Data drill-down

5.3.3.1 Identification of distinct data elements referring to same real-word identity

- Data in tables is linked, grouped and filtered in various dimensions with the aim to find inconsistencies in the data using appropriate data management software.
- Results from this analysis are reported in the Dataset Report according to SOP-003 in an appropriate way.
- Deviations from the expected results are reported in the Dataset Report according to SOP-003. The following parameters are tracked:
  - Name of the column where the issue was detected.
  - Number of records per column that are affected by the issue.
  - Queries used to execute this process.
- All previous steps of section 5.3.3.1 are repeated for all tables of a dataset.

6 References

There are no sources in the current document.

7 Annexes