Open Research Data Pilot and project Data Management Plan

Deliverable 11.2

20 December 2018

Mark Brown (RHUL)

PoshBee

Pan-european assessment, monitoring, and mitigation of stressors on the health of bees

Prepared under contract from the European Commission

Grant agreement No. 773921
EU Horizon 2020 Research and Innovation action

Project acronym:	**PoshBee**
Project full title:	**Pan-european assessment, monitoring, and mitigation of stressors on the health of bee**
Start of the project:	June 2018
Duration:	60 months
Project coordinator:	Professor Mark Brown
	Royal Holloway and Bedford New College www.poshbee.eu

Deliverable title:	Open Research Data Pilot and project Data Management Plan
REPORTED IN THE DoA]	
Deliverable n°:	11.2
Nature of the deliverable:	ORDP: Open Research Data Pilot
Dissemination level:	Public

WP responsible:	WP11
Lead beneficiary:	Pensoft

Citation:	Brown, M.J.F. (2018). *Open Research Data Pilot and project Data Management Plan.* Deliverable D11.2 EU Horizon 2020 PoshBee Project, Grant agreement No. 773921.

Due date of deliverable:	Month n° 6
Actual submission date:	Month n° 7

Deliverable status:

Version	Status	Date	Author(s)
1.0	Final/Draft	20 December 2018	Mark Brown
			RHUL

The content of this deliverable does not necessarily reflect the official opinions of the European Commission or other institutions of the European Union.

Table of contents

_____

POSHBEE: DATA MANAGEMENT PLAN

1. Open access statement and data sharing

In the context of the Horizon 2020 programme, the European Commission published a document titled "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020". The document clearly describes the need that led to the mandate for open access to scientific publications, research data and their associated metadata that have been produced under funding from the Horizon 2020 programme. At the same time, the document states the European Commission's view on this aspect "information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full".

In this context, this document provides the plan for the management of research outcomes (and more specifically, the research publications and data) that will be produced during the PoshBee project lifetime, as well as those that will be collected from the PoshBee partners for the respective use cases. It aims to ensure that the research activities of the project are compliant with the H2020 Open Access policy and the recommendations of the Open Research Data pilot, where applicable. In this context, the project's Data Management Plan (DMP) described in this document outlines how research data will be collected, processed or generated both within the project and how these data will be curated and preserved during and after the project.

1.1. What is the Open Research Data Pilot?

Open data is data that are free to use, reuse, and redistribute. The Open Research Data Pilot aims to make the research data generated by selected Horizon 2020 projects open. It will be carefully monitored and used to inform future EC policy. As a Horizon 2020 project participating in the pilot, PoshBee will:

- Develop (and keep up-to-date) a Data Management Plan (DMP).
- Deposit our data in a research data repository.
- Make sure third parties can freely access, mine, exploit, reproduce and disseminate it (where applicable and not in conflict with any IPR considerations).
- Make clear what tools will be needed to use the raw data to validate research results (or provide the tools themselves).

The pilot applies to (1) the data (and metadata) needed to validate results in scientific publications, and (2) other curated and/or raw data (and metadata) that is specified within this DMP.

1.2. Where to store the data after the project?

A data repository is a digital archive collecting and displaying datasets and their metadata. Many data repositories also accept publications, and allow linking between publications and their underlying data. Data generated within PoshBee will be hosted on disciplinary or institutional repositories, as appropriate (some publishers require specific repositories, whilst others only require that the data are deposited), with the requirement that these repositories provide free and open access to the data deposited.

PoshBee will collect a range of data types, both generated by the project and obtained and collated from existing external sources. Data collection templates will be created within and, in collaboration,

across WPs (depending upon the aspect of the project). The templates will ensure that collected data is interoperable both within PoshBee's context and in the context of outside data aggregators and archives. Data types will include: experimental data, including measures of bee health, chemical data, genetic data, proteomics data, landscape data, and survey data.

In data publishing and dissemination policies, PoshBee will follow the basic postulates of the **Open Knowledge/Data Definition, and the Panton Principles for Open Data in Science**. PoshBee will strengthen and develop various Open Source databases, providing access to the data sets underlying the published research. Database Rights will be generated with the collation of data collections, addition of new data and new ways of displaying and archiving datasets. The partners will, to the greatest extent possible, manage the Database Rights with the aim of using, to the maximum possible extent, copyright licenses that enable the free distribution of data. One such license is the **Open Data Commons Attribution License (ODC-By)**, which allows users to freely share, modify, and use the published data (bases) provided that the data authors are acknowledged (cited in academic articles or acknowledged when used for other purposes). Data that are deemed to be commercially sensitive will be clearly identified and treated in accordance with the relevant IPR considerations. Authors should explicitly inform the Project Coordinator and Project Management Group, if they want to publish data associated with a journal article under a license that is different from the Open Data Commons Attribution License (ODC-By).

In cases where data used in a publication have been previously deposited or published elsewhere under a license different from the above, the author should explicitly mention that in the text of the manuscript, cite the respective license and link to it. Some of the data, on the discretion of the data collectors and in coordination with the Project Management Group, may also be published under the Creative Commons CC0 (also cited as "CC-Zero" or "CC-zero") and the Open Data Commons Public Domain Dedication and License (ODC-PDDL). According to the CC0 license, *"the person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighbouring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission."*

Publication of data under a non-attribution waiver such as CC0 avoids potential problems of "attribution stacking" when data from several sources are aggregated for re-use, particularly if this re-use is undertaken automatically. In such cases, while there is no legal requirement to provide attribution to the data creators, the norms of academic citation best practice for fair use still apply, and those who re-use the data should reference the data source, as they would reference other research articles. The Attribution-ShareAlike Open Data Commons Open Database License (OdbL) is NOT recommended for use for POSHBEE data, although it may be used as an exception in particular cases. The OdbL license assumes that "*If one publicly uses any adapted version of the database, or works produced from an adapted database, he or she must also offer that adapted database under the OdbL.*"

The following methodology and reporting steps are recommended to partners managing data:

DATASET CONTENT, PROVENANCE AND VALUE

What type of data have been collected or created?

Data can be derived from one or more datasets that relate to each use case. The DMP will explain the background (retaining provenance) of the described dataset. Imported data can then be combined,

_____

processed and analysed, generating additional data. A description of the operations leading to these data should also be included.

What is the value for other researchers?

Candidates for reuse are identified in respective user-driven requirements and use cases deliverables.

STANDARDS AND METADATA

Which data standards will the data conform to?

The consortium will strive to comply with or reuse existing standards whenever possible. While original data sources may conform to different formats and standards, data processed by POSHBEE analyses will likely have been transformed into formats complying with a set of well-known standards.

What documentation and metadata will accompany the data?

In addition to the data collection activities, POSHBEE will also generate its own valuable data assets in terms of metadata that will improve the description, interlinking, normalization, unification, and quality assessment of the collected datasets.

DATA ACCESS AND SHARING

Which data are open, re-usable and what licenses are applicable?

It is envisaged that most of the datasets resulting out of project activities will be of an open nature, i.e., data which are freely accessible and protected by minimally restrictive or unrestricted licenses. However, some data may also be obtained via private access, while other data generated in the project may have IPR sensitivity. In the former case, the consortium will ensure that any imported data conforms to existing or indicated licenses. In particular, the attachment of the Open Data Commons Open Database License (ODbL) to open datasets could be adopted, promoting the three core requirements of attribution, share-alike, and the retention of its open nature. Additional usage and sharing restrictions on the dataset will be defined through additional licences or modifications of existing alternatives. Justifications for restrictions to dataset access or re-use should be explained clearly in data reports.

How will open data be accessible and how will such access be maintained?

Data access will vary depending on storage location. Measures will be taken to enable third parties to freely access, re-use, analyse, exploit and disseminate the data (bound by the license specifications). Different access procedures will be implemented, enabling the export of an entire dataset as well as the provision of a querying interface for the retrieval of relevant subsets. Access mechanisms will also be supported as much as possible by metadata enabling search engines and other automated processes to access the data using standard Web mechanisms.

Which privacy protocols are implemented?

In the case that a dataset contains sensitive corporate or personal data, privacy protocols will be established and followed throughout the aggregation, processing and publishing stages. The anonymisation of personal information should precede the processing stage, especially if additional

_____

data pre-processing measures need to be taken to safeguard individuals or groups. If the data processing results still produce sensitive data, access controls will be enforced and described.

DATA ARCHIVING, MAINTENANCE AND PRESERVATION

Where will each dataset be physically stored?

Depending on the nature of the data, a dataset might eventually be moved to an external repository, e.g. the European Open Data Portal. Data generated via other means can have additional hosting arrangements.

What physical resources are required to carry out the plan?

During the pilot project phase, hosting, persistence and access will be managed by the project partners' infrastructure. Partners with the most suitable hosting and processing capabilities have been identified early in the project lifetime

What are the physical security protection features?

Once a dataset is published and access enabled, security will be addressed to ensure that the data cannot be tampered with and its veracity can be guaranteed.

How will each dataset be preserved to ensure long-term value?

Since the majority of data integrated and generated within the POSHBEE infrastructure will abide by the Linked Data principles, the consortium will follow the best practices for supporting the life cycle of Linked Data. This includes its curation, repair and evolution, thus also increasing the likelihood that machine-readable structured datasets (and associated metadata) resulting out of project efforts can also be of long-term use for third parties.

Who is responsible for delivering the plan?

Consortium members collecting and re-using data will be tasked to follow the data management steps and make sure all aspects are considered when producing reports and products across all WPs.

2. Data publishing

Data publishing is an act of making data available on the Internet, so that they can be downloaded, analysed, re-used and cited by people and organisations other than the creators of the data. This can be achieved in various ways. In the broadest sense, any upload of a dataset onto a freely accessible website could be regarded as "data publishing". There are, however, several issues to be considered during the process of data publication, including:

- Data hosting, long-term preservation and archiving
- Documentation and metadata
- Citation and credit to the data authors
- Licenses for publishing and re-use
- Data interoperability standards
- Format of published data
- Software used for creation and retrieval

_____

- Dissemination of published data

PoshBee data may be published in Pensoft's Open Access Journal Research Ideas and Outcomes (RIO) under a project collection such as eg. EUBON project: https://riojournal.com/collection/2/
The journal accepts various types of unconventional research outputs such as: data management plans, software descriptions, project reports, Grant proposals, etc.

3. Conclusion

PoshBee has developed a comprehensive DMP, as described above. This will be shared and communicated within the PoshBee consortium, to ensure that all beneficiaries are enabled to comply with the DMP throughout the lifetime of PoshBee, and beyond. Where necessary, as the field of open access evolves, we will update our DMP and guidance to reflect advances in open access for data.