CropBooster-P
Deliverable 6.1
Title: Data Management Plan

Start date of the project: **November 1st, 2018** / Duration: 36 **months**
Planned delivery date: M6 (April 2019)
Actual submission date: April 12th 2019
Work package: WP6 / Task: 6.1

Work package leader: WR   Deliverable leader: WR

Version: 1

Date of version: April 2019

Dissemination type	ORDP
Dissemination level	Public

2

Abstract

The CropBooster-P Data Management Plan outlines the data management strategies that will be implemented throughout the project research data lifecycle. In particular, it describes (i) the type, format and volume of the generated data, (ii) the metadata and documentation provided to make it findable, interoperable and reusable, (iii) the long-term preservation plan, (iv) how data will be shared and licensed for re-use, (v) the resources that need to be allocated to data management, (vi) data storage and back up policies during the active phase of the project, and (vii) the handling of personal data.

Table of contents   3

1. Summary

The CropBooster-P Data Management Plan outlines the data management strategies that will be implemented throughout the project research data lifecycle. In particular, it describes (i) the type, format and volume of the generated data, (ii) the metadata and documentation provided to make it findable, interoperable and reusable, (iii) the long-term preservation plan, (iv) how data will be shared and licensed for re-use, (v) the resources that need to be allocated to data management, (vi) data storage and back up policies during the active phase of the project, and (vii) the handling of personal data.

The CropBooster-P project is a Coordination and Support Action which in its nature is not a research project. Consequently, only a limited amount of data will be generated that could be considered as research data and thus should be covered by a dedicated management plan. Currently, the only foreseen activities that will generate research data are a series of workshops that will be carried out in the framework of Work Package 2; Assessing Economic, Social and Environmental Impact. In these workshops, specific opinions and attitudes of stakeholders towards societal developments regarding agriculture, food and bioeconomy will be collected. The current version of the Data Management Plan thus is geared only towards managing the specific data types generated during these workshops.

As stipulated in the Guidelines on FAIR Data Management in Horizon 2020, this Data Management Plan will be updated when important changes to the project occur, like the production of research data in any of the other CropBooster-P Work Packages. In addition this Data Management Plan will be updated as part of the periodic reviews and at the end of the project.

2. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?
- The project aims to collate opinions of stakeholders and citizens on the development of innovations in plant production; and specific approaches to achieve that aim. To do so, workshops and a citizen jury will be conducted. We will combine this with systematic review of the literature, where secondary datasets may be collated.

What types and formats of data will the project generate/collect?
- The data will be audio recordings (mp3 or FLAC format) and/or transcripts of stakeholder and citizen workshops and juries (pdf format). These might be amended with written notes of participants if relevant and available.

Will you re-use any existing data and how?
- Existing unpublished data is unlikely to be used.
- Secondary data and conclusions available from scientific publications and databases will be used.

What is the origin of the data?
- New data generated will be from self-conducted workshops and citizen juries.

What is the expected size of the data?
- Several hours of audio recordings, several dozens of pages of transcript. It is expected that the storage will not exceed 10GB.

To whom might it be useful ('data utility')?
- As this kind of data is very contextualised in its raw form this data will be almost exclusively relevant to CropboosterP project partners.

3. FAIR data

3.1 Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- No. As it is almost impossible to fully anonymise this type of data given the provided expert information (i.e. the heart of the data) may already suffice to identify individuals, it is not suitable to be made discoverable and identifiable. Also as highlighted in the data utility question above, the data produced from stakeholder workshops is highly contextualised, and additional documentation to provide context would be insufficient to allow interpretation making this data of limited value to others beyond the project.

What naming conventions do you follow?
- Social science interview / qualitative data conventions

Will search keywords be provided that optimize possibilities for re-use?
- No

Do you provide clear version numbers?
- Yes

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.
- Metadata outlining the date, place and language of data collection, as well as protocols will be provided.

3.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

How will the data be made accessible (e.g. by deposition in a repository)?
- Data will be deposited in a (closed) repository. Lancaster University data will be stored on the Lancaster University Research Data Repository (via Pure), Wageningen University will be stored in a registered repository after consultation with Wageningen Data management support (via Pure).

What methods or software tools are needed to access the data?
- Standard word/audio processing software

Is documentation about the software needed to access the data included?
- No

Is it possible to include the relevant software (e.g. in open source code)?
- No

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

- Closed university repository. Due to the expected impossibility to meaningfully anonymise the data, open access cannot be granted. In addition, as raw data of this type, due to its highly specific context, is of limited use for re-use, the benefits for providing open access are extremely limited.

Have you explored appropriate arrangements with the identified repository?

- Yes, we have sought guidance and advice from Lancaster University's Research Data Management Team.

If there are restrictions on use, how will access be provided?

- Data will be restricted to consortium participants due to reasons stated in 2.1. Access will be controlled through the repository system.

Is there a need for a data access committee?

- No

Are there well described conditions for access (i.e. a machine readable license)?

- No

How will the identity of the person accessing the data be ascertained?

- The repository is a closed password protected server. Requests for access will be on a case-by-case basis, where the identity of the person accessing the data is ascertained by the manager of the repository.

3.3 Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

- Raw data is verbatim transcription of interview audio. This should be accessible to all who have access to generic word processing software.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

- Raw data is verbatim transcription of participants, making this non applicable. Editorial comments in transcription where needed are done in standard format [square brackets]

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

- Not applicable

3.4 Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?
- Please see answers in 2.1 and 2.2. Data will be used internally in consortium only.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Not applicable

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
- Not applicable, see above.

How long is it intended that the data remains re-usable?
- At least 10 years after publication of the original papers

Are data quality assurance processes described?
- Yes

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

- Data will be anonymised / pseudonomised before storage as much as possible. During processing and analysis data will be stored at secured university server space (not local). Data held on Lancaster systems are stored in a resilient storage infrastructure which is dual homed in the Lancaster University data centres (on site). There are multiple levels of redundancy built into these storage arrays – snapshots and backups are automated and taken regularly.

- Long term storage location only grants access to a few, named people (less than 5) and university SSO (single sign-on) credentials will be required to gain access

Is the data safely stored in certified repositories for long term preservation and curation?

- Lancaster University data will be stored on the Lancaster University Research Data Repository (via Pure).
- Wageningen University will be stored in a registered repository after consultation with Wageningen Data management support (via Pure).

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

- All legal and ethical issues regarding data sharing are described Deliverable D 7.1, Ethics requirements for studies involving human participants in CropBooster-P and in Deliverable D7.2, Ethics requirements for studies involving human participants in CropBooster-P.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

- Long term preservations yes. Data sharing no (see above)