Disclaimer:

This deliverable has been submitted to the European Commission but has not yet been formally accepted. For this reason, its content may be subject to changes following the review process.

Funded by the
European Union

D9.1 Data Management Plan

(Version 0.2, 28th December 2023)

THE JAMES HUTTON INSTITUTE (JHI)

Deliverable description

DELIVERABLE: D9.1 Data Management Plan
WORK PACKAGE: WP9. Project management
AUTHOR(S): Paul Shaw The James Hutton Institute (JHI)
DUE DATE: 31/12/2023
ACTUAL SUBMISSION DATE: 28/12/2023
DISSEMINATION LEVEL PU: Public (must be available on the website)
GRANT AGREEMENT No: 101082091
PROJECT STARTING DATE: 01/07/2023
PROJECT DURATION: 60 months

Quality of information - Disclaimer according to the Art. 17.3 of GA
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or EUROPEAN RESEARCH EXECUTIVE AGENCY (REA). Neither the European Union nor the granting authority can be held responsible for them.

REVISION HISTORY			
Version	Date	Modified by	Comments
0.1	12/12/2023	JHI	Initial submission of Data Management Plan (DMP)
0.2	28/12/2023	UMIL	Second Draft
0.3			
0.4			
0.5			

Table of Contents

Glossary of terms

API	Application Programming Interface
BrAPI	Breeding Application Programming Interface
CC-BY	Creative Commons Attribution Licence
CC-REL	Creative Commons Rights Expression Language
DDBJ	DNA Data Bank of Japan
DMP	Data Management Plan
DOI	Data Object Identifier
dPCR	Digital Polymerase Chain Reaction
EBI	EMBL European Bioinformatics Institute
FAO	Food and Agriculture Organization of the United Nations
GB	Gigabyte
INSDC	International Nucleotide Sequence Database Collaboration
FAIR	Findable, Accessible, Interoperable and Reusable
MIAPPE	Minimum Information About a Plant Phenotyping Experiment
RDM	Research Data Management
SNP	Single Nucleotide Polymorphism
TB	Terabyte
WGS	Whole Genome Sequencing

Introduction

The aim of this document is to provide guidelines on principles guiding the data management in BEST-CROP and what data will be stored by using the responses to the EU questionnaire on Data Management Plan (DMP) as a DMP document.

This DMP presents how data will be handled during and after the project. It will be updated/its validity checked during the duration of BEST-CROP at 6 monthly intervals starting with the submission of the draft data management plan at month 6 then updated at month 12.

The DMP will ensure that any data generated by BEST-CROP will be findable, accessible, interoperable and reusable (FAIR).

1. Data Summary

There is a need for a ground-breaking technology to boost crop yield (both grains and biomass) and its processing into materials of economic interests. Novel crops with enhanced photosynthesis and assimilation of green-house gasses, such as carbon dioxide (CO_2) and ozone (O_3), and tailored straw suitable for industrial manufacturing will be the foundation of this radical change. We are an alliance of European plant breeding companies, straw processing companies and academic plant scientists aiming to use the major advances in photosynthetic knowledge to improve barley yield and to exploit the variability of barley straw quality and composition. We will capitalize on very promising strategies to improve the photosynthetic properties and ozone assimilation of barley: i) tuning leaf chlorophyll content and modifying canopy architecture; ii) increasing the kinetics of photosynthetic responses to changes in irradiance; iii), introducing photorespiration bypasses; iv) modulating stomatal opening, thus increasing the rate of CO_2 fixation and O_3 assimilation. Beside the higher yield, the resulting barley straw will be tailored to: i) increase straw protein content to make it suitable as an alternative feed production source; ii) control cellulose/lignin contents and lignin properties to develop construction panels and straw reinforced polymer composites. To do so, we aim to exploit barley natural- and induced-genetic variability as well as gene editing and transgenic engineering. Based on precedent, we expect that improving our targeted traits will result in increases in above ground total biomass production by

15-20% without modification of the harvest index, and there will be added benefits in sustainability via better resource-use efficiency of water and nitrogen. A public dialogue will be established to ensure stakeholder engagement and explore the acceptability of a range of technologies as potential routes to crop improvement and climate change mitigation.

Q. Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

The project builds on existing data sets. Each project partner may provide historical datasets or background information such as characterisation and evaluation data on germplasm that will be used in BEST-CROP.

It is our intention to ensure that all BEST-CROP data is made freely available to BEST-CROP partners and the wider research community.

Q. What types and formats of data will the project generate or re-use?

The BEST-CROP project will collect and/or generate the following types of raw data: phenotypic, genotypic, germplasm characterisation and evaluation, other NGS data, metabolomic and other forms of transcriptomic data. In addition, both historical data provided by project partners and derived data from the original raw data sets will be collected, used, stored and archived. This is important, as different analytical pipelines or techniques may yield different results or include ad-hoc data analysis components. Specific care will be taken to document and archive these resources including the analytical pipelines or computer code to ensure complete reproducibility.

Q. What is the purpose of the data generation or re-use and its relation to the objectives of the project?

BEST-CROP will advance the state of the art by addressing the following specific high-level objectives:

1. Boost the growth of the circular bioeconomy: tailoring of barley straw for efficient transformation into high-value bio-based compounds and materials

that replace products currently obtained from high-polluting industrial sectors with high dependency on non-renewable energy sources.

2. Mitigate ozone air pollution extremes during drought by providing a strategy of air phytoremediation through the modulation of stomatal conductivity for ozone without a negative effect on drought tolerance and yield.
3. Address the global food security crisis by delivering highly productive barley lines. Furthermore, barley represents an optimal model species for other cereals with a view to project medium-term replicability.

Data collection, integration and visualization through a standardized data management process is necessary to ensure that we understand not only the underlying biological principles, but also be informed about the provenance of data. It is therefore necessary to ensure that the data generated under BEST-CROP are well described and annotated with metadata using open standards where appropriate.

Q. What is the expected size of the data that you intend to generate or re-use?

We expect to generate raw data (sequence, image, and evaluation data) in the range of 5TB of data. The size of the derived data will be around 20 GB.

Q. What is the origin/provenance of the data, either generated or re-used?

Historical and background datasets will be provided by consortium partners, and we will do our best to ensure that these datasets are well described and annotated where possible.

Any data obtained from public repositories will be annotated and described with reference and links maintained to the original source.

Data of different types or representing different domains will be generated using unique approaches. For example:

- WGS (Whole Genome Sequencing) data will be generated by long or short read platforms.
- SNP (Single Nucleotide Polymorphism) data will be generated using dPCR assays to identify lines with alleles of interest from pools of DNA.

- Image data will be generated by equipment such as cameras, scanners, and microscopes combined with software. Original images which contain metadata such as exif photo information will be archived.
- Genomic data will be created from sequencing data, which will be processed to identify genes, regulatory elements, transposable elements, and physical markers such as SNPs and structural variants.
- Genetic data will be generated targeting crosses and breeding experiments and will include recombination frequencies and crossover event that position genetic markers and quantitative trait loci that can be associated with physical genomic markers/variants.
- Targeted assays data (e.g., glucose and fructose concentrations or production/utilisation rates) will be generated using specific equipment and methods that are fully documented in the laboratory notebook.
- Model data will be generated by using software simulations. The complete workflow, which includes the environment, runtime, parameters, and results, will be documented and archived.
- Computer code will be produced by programmers.
- Excel data will be generated by data analysts by using MS Office or open-source software.
- The origin and assembly of cloned DNA will include (a) source of original vector sequence with Add gene reference where available, and source of insert DNA (e.g., amplification by PCR from a given sample, or obtained from existing library), (b) cloning strategy (e.g., restriction endonuclease digests/ligation, PCR, TOPO cloning, Gibson assembly, LR recombination), and (c) verified DNA data sequence of final recombinant vector.
- Phenotypic data will be generated using phenotyping platforms and corresponding ontologies, including number/size of organs such as leaves, flowers, buds etc., size of whole plant, stem/root architecture (number of lateral branches/roots etc), organ structures/morphologies, quantitative metrics such as colour, turgor, health/nutrition indicators, among others.

Q. To whom might it be useful ('data utility'), outside your project?

The data will primarily benefit the BEST-CROP research and industrial partners but will also be made available to selected stakeholders closely involved in the project, and then the wider scientific community. The data will be disseminated through our BEST-CROP Germinate database (https://ics.hutton.ac.uk/germinate-bestcrop).

2. FAIR data

2.1. Making data findable, including provisions for metadata

Q. Will data be identified by a persistent identifier?

All data sets will receive unique identifiers on submission to Germinate and they will be annotated with necessary metadata. All germplasm used in BEST-CROP will be assigned unique identifiers on submission to Germinate and any germplasm subsequently submitted to genebanks assigned a digital object identifier (DOI) as per the FAO International Treaty on Plant Genetic Resources for Food and Agriculture guidelines (https://glis.fao.org/glis/linkdir/channel?c=docs). Any germplasm obtained from genebanks will have their DOIs preserved throughout this work to ensure clear links back to the original germplasm source.

Q. Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

BEST-CROP will rely on established and developing community standards such as Dublin Core (https://en.wikipedia.org/wiki/Dublin_Core; ISO 15836) to broadly describe digital resources plus additional recommendations necessary in plant science including Minimum Information About a Plant Phenotyping Experiment (MIAPPE; https://www.miappe.org) which will help facilitate reusability by other researchers. Our database Germinate is moving towards MIAPPE compliance and offers BrAPI (https://brapi.org) compliance for data exchange with other compliant databases where appropriate.

Q. Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Keywords about experiments will be included as well as an abstract/description about the datasets we will generate. Germinate also offers full text-searching capabilities and with BrAPI compliance will allow for detailed searching of resources and future integration with other BrAPI tools.

Q. Will metadata be offered in such a way that it can be harvested and indexed?

To maintain data integrity and to be able to accurately and reproducibly re-analyse data, datasets will be assigned version numbers (e.g., raw data must not be changed and will not get a version number and considered immutable). Data variables will be allocated standard names. For example, genes, proteins and metabolites will be named according to approved nomenclature and conventions. These will also be linked to functional ontologies where possible. Datasets will also be named in a meaningful way to ensure clarity and easy readability by humans. Plant names will include traditional names, binomials, and all strain/cultivar/subspecies/variety identifiers all stored and searchable from within Germinate.

Germinate offers standardised links which can be harvested and indexed providing links to resources that can be used in publications, reports and to link from other websites.

2.2. Making data accessible

Q. Will the data be deposited in a trusted repository?

Data will be made available via the BEST-CROP data platform (https://ics.hutton.ac.uk/germinate-bestcrop) using a web-based, user-friendly front end that allows data searching, visualization and download in standard formats. Data will be either stored directly here or links made to appropriate specialized repositories such as INSDC (GenBank, EBI, DDBJ) for nucleotide sequence data, PIR/UniProt/SWISS-PROT for proteins, PDB for protein structures, GEO for transcriptomics, PRIDE for proteomics data, and METLIN for metabolomics data. For unstructured and less standardized data (e.g., experimental phenotypic

measurements), these will be annotated with metadata, stored in Germinate and if applicable assigned a digital object identifier (DOI).

Q. Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Germinate will be further developed as part of BEST-CROP and the team responsible for it at the James Hutton Institute are involved in this work. For the specialised repositories listed above the submission is for free and therefore prior arrangements are not necessary.

Q. Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data will be made available via the BEST-CROP data management platform based on Germinate which is developing its DOI capabilities beyond that of storing a DOI with a datasets metadata.

Q. Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

By default, all data sets from BEST-CROP will be shared with the community and made openly available under an unrestrictive CC-BY licence (https://creativecommons.org/licenses/by/4.0). However, before the data are released publicly, consortium partners will be provided with an opportunity to check for potential IP (according to the consortium agreement and background IP rights).

Q. If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be published as soon as possible to guarantee reusability. IP issues will be checked before publication. All consortium partners will be encouraged to

make data available before publication, openly and/or under pre-publication agreements if required. Datasets submitted to Germinate will be available to all consortium partners privately until agreed that the data can be made public.

Q. Will the data be accessible through a free and standardized access protocol?

Yes. Germinate offers web-based access both through browser based or API interfaces including through the BrAPI plant breeding standard.

Q. If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

There are no restrictions, beyond the aforementioned IP checks, which are in line with e.g. European open data policies.

Datasets will be made available using temporary restricted access to consortium partners through Germinate until checks undertaken and the data published publicly through the same platform.

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

If the data is not yet finalised or under IP checks, the data will be hosted and username and password will be required which will only be available to verified consortium partners (see also our GDPR statement). In the case data is made public either through Germinate or alternative repositories, completely anonymous access will be provided.

Q. Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

We do not think that this is required but this will be revisited when our DMP is reviewed.

Q. Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Yes, where possible, e.g., CC REL (https://wiki.creativecommons.org/wiki/CC_REL) will be used for data where we can and will soon be supported by Germinate.

Q. How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

We will ensure that data from BEST-CROP is available for a period of at least 3 years past the end of the project. All data held will be downloadable in either plain text or other open formats. Data submitted to public repositories would be subject to their local data storage policies. Data held at the James Hutton Institute will be archived as per UKRI guidelines for a period of at least 10 years.

Q. Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

No specialized software will be needed to access the data, usually just a modern browser and standard text editing software (either commercial or open source). Access will be possible through web interfaces. For data processing after obtaining raw data, typical open-source software can be used and will be the default position for BEST-CROP. We will use publicly available open-source and well-documented certified software and any software or computer code created under BEST-CROP will be released through our GitHub repository using an unrestrictive Apache 2 licence.

2.3. Making data interoperable

Q. What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

As noted above, we foresee using minimal standards such as MIAPPE. The minimal information standards will allow the potential for future integration of data from BEST-CROP across projects, and its reuse according to established and tested protocols. Specialized repositories will be used for common data types. Whenever possible, data will be stored in common and openly defined formats (usually plain

text .txt format) including all the necessary metadata to interpret and analyse data in a biological context. By default, no proprietary formats will be used. However, Microsoft Excel files (according to ISO/IEC 29500-1:2016) will probably be used as intermediates by the consortium members due to familiarity with the software and ease of use for basic data handling and quality checking. In addition, text files might be edited in text processor files, but will be shared as pdf.

Q. 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? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Common and open ontologies will be used where appropriate.

Q. Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

The references to other data will be made in the form of DOI. Links to related work can be assigned to datasets held within our data repository.

2.4. Increase data re-use

Q. How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

The documentation will be provided in the form of ISA (Investigation Study Assay) and CWL (Common Workflow Language).

Q. Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Yes, our data will be made freely available in the public domain to permit the widest re-use possible. Open licenses, such as Creative Commons (CC), will be used whenever possible and as described in the Grant Agreement.

Q. Will the data produced in the project be useable by third parties, in particular after the end of the project?

There will be no restrictions once the data is made public.

Q. Will the provenance of the data be thoroughly documented using the appropriate standards? Describe all relevant data quality assurance processes.

BEST-CROP data will be checked and curated by using pre agreed data collection protocols, personnel training, data cleaning, data analysis, and quality control.

3. Other research outputs

Software code used for analysis of data generated under BEST-CROP will be made available through the project website and held in the project GitHub repository. All code will be released under unrestrictive open-source licencing (Apache 2) and made available freely to the community.

Germinate which will be used as the data repository for BEST-CROP data is open source and available for download from https://germinateplatform.github.io/get-germinate. All development of Germinate funded by BEST-CROP will be release under the same licence terms and freely available.

4. Allocation of resources

Q. What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

BEST-CROP will bear the costs of data curation, and data maintenance/security before transfer to public repositories. Subsequent costs are then borne by the operators of these repositories. All data stored in Germinate will be available for a minimum period of 3 years past the end of the project.

Q. How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

The cost born by the BEST-CROP project are covered by the project funding.

Q. Who will be responsible for data management in your project?

The responsible person will be Dr Paul Shaw, Information and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland. Paul.Shaw@hutton.ac.uk.

Q. How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

The data officer or undefined will ultimately decides on the strategy to preserve data that are not submitted to end-point subject area repositories when the project ends. This will be in line with EU guidelines, institute policies, and data sharing based on EU and international standards.

Current UKRI guidelines are that data will be archived for a minimum of 10 years.

5. Data security

Q. What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Online platforms will be protected by vulnerability scanning, two-factor authorization and daily automatic backups allowing immediate recovery. All partners holding confidential project data to use secure platforms with automatic backups and offsite secure copies.

Q. Will the data be safely stored in trusted repositories for long term preservation and curation?

Wherever there are certified repositories, these will be used as end-point repositories. Transcriptomics data and gene sequence data will be also made available upon publication via the standards ENA/SRA, metabolite data in e.g. Metabolights (and/or Nationwide repositories like the German NFDI the French INRAe), Proteomics data in e.g. Pride/Proteomexchange. In addition, the national resource will maintain safekeeping of data also after the project ends.

6. Ethics

Q. Are there, or could there be, any ethics 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).

Currently, we do not anticipate ethical or legal issues with data sharing. In terms of ethics, since this is plant data, there is no need for an ethics committee, however, diligence for plant resource benefit sharing is considered (see Nagoya protocol) and we have standard material transfer agreements in place for germplasm exchange between partners.

Q. Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

The only personal data that will potentially be stored is the submitter name and affiliation in the metadata for data. In addition, personal data will be collected for dissemination and communication activities using specific methods and procedures developed by the BEST-CROP partners to adhere to GDPR.

7. Other issues

Q. Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

BEST-CROP will use common Research Data Management (RDM) tools such as GitHub for managing all code developed under BEST-CROP. All code will be released under an Apache 2 licence and available without restriction to the community.

8. Expected Datasets

Will be updated throughout duration of project.

Partner	Year	Description of experiment	No of lines	No of traits	Reps/Sites etc.	Images Y/N (and numbers)	Month data to be uploaded	Task
JHI	2023	Phenotypic screening Laureate M2 mutants	4800	7		Y (at least 24)	October 2023	
UPOL	2023-	Gene edited barley lines	30 lines 10 plasmids	10		Y	Throughout project duration	
ITAL BIOTEC	2023-	Database of input and output flows for LCA WP6.1						
ITAL BIOTEC	2023-	Database of project stakeholders (Includes personal information: email addresses) WP6.2						
ITAL BIOTEC	2023-	Process flow for TEA (WP6.2)						
ITAL BIOTEC	2023-	Different contact database for project newsletter, events organized in the context of the project, stakeholders, other projects for networking (WP8) -> data may be managed by an external platform that we use for distributing the newsletter (usually Mailchimp) or collecting registrations to the events (usually EventBrite).						
ITAL BIOTEC	2023-	Results of a socio-economic study (mainly results of a survey to be shared with consumers)						
UDUS	2023-	Generation of transgenic and gene edited barley lines	5-10 per construct	10		Y	Troughout project duration	
CREA	2024	Phenotypic screening winter barleys for straw N content	120	1-3	2 years, 2 reps/year, 2 tissues (8 measurement in total for each line)	N	October 2024	Task 2.1
CREA	2025	Results from GWAS for straw related traits and allele mining for candidate geens		1-3		N	June 2025	Task 2.2
CREA	2023-2028	Pyramiding of natural, chemical-induced and GE/GM-derived mutations	4000	>100 SNPs for foreground and background selection		N	Throughout project duration	Task 3.1 and 3.2
CREA	2023-2028	Field phenotyping of mutants, germplasm sets and advanced breeding lines	100	10	4 sites	N	Throughout project duration	Task 2.1 and Task 4.2
IMT	2025-2028	Results on the morphology, thermal/fire and mechanical properties of barley stem fragments	20-30			Y	Throughout project duration	Task 4.1
IMT	2025-2028	Results on the processing parameters to develop barley-based panels and composites	20-30			Y	Throughout project duration	Task 5.3
IMT	2025-2028	Results on the microstructure, mechanical and thermal/fire properties of barley-based panels and composites	20-30			Y	Throughout project duration	Task 5.3
MOGU	2025-2028	Use of mycelium strains as binders to create panels with improved acoustic / insulation / fire performance				Y	M56	Task 5.2
UMIL	2023	Forward Genetic Screening of the *Chlorina* mutant collection provided by Prof. Hansson to characterise new pale-green lines with increased/wt-like photosynthetic performance	8	5-8		Y	Beginning 2024	Task 1.1
UOD	2023-	Phenotypic analsis of straw lignin content	20-30		2 treatments, 2 reps per treatment		early 2024	