D7.2

Data Management Plan

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

Document info

Project number	760891 – CHIC
Funding scheme	Collaborative Project
Work programme	H2020-NMBP-BIOTEC-07-2017: New Plant Breeding Techniques (NPBT) in molecular farming: Multipurpose crops for industrial bioproducts
Deliverable number	D7.2
Deliverable title	Data management plan
Dissemination level	Public
Due date	30.6.2018
Actual submission date	15.7.2018
Project start date	January 1st, 2018
Duration	54 months
Work package concerned	WP7
Concerned work package leader	Macarena Sanz
Type	ORDP: Open Research Data Pilot
Author	Katarina Cankar, Macarena Sanz, Armin Spoek, Dirk Bosch
Contributor	P1-WR, P17-IDC, P10 TU-GRAZ
Reviewers	Chic Project Management Team (PMT), all Partners

Document history

Date	Author	Action	Status
01-07-2018	Katarina Cankar, Dirk Bosch. In close consultation with Diego Orzaez and Asun Fernandez (NEWCOTIANA)	1st draft	Finished
06-07-2018	Katarina Cankar, Dirk Bosch, Macarena Sanz Armin Spoek. In close consultation with Diego Orzaez and Asun Fernandez	2nd draft	Sent to partners for comments
11-07-2018	Katarina Cankar, Dirk Bosch	3rd draft	Remarks from partners
12-07-2018	Katarina Cankar, Dirk Bosch	final	Submitted to EC

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

1 Summary

This document outlines the data management strategies that will be implemented throughout the CHIC 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.

As stipulated in the Guidelines on FAIR Data Management in Horizon 2020, this DMP will be updated when important changes to the project occur and, at least, as part of the periodic reviews and at the end of the project.

The data management plans of the CHIC project and of the NEWCOTIANA project (grant agreement 760331) have been generated in close collaboration between these two projects. Both projects will generate datasets on technical performance, safety assessment, socio-economic and stakeholder interactions related to the use of NPBT for the development of multipurpose crops for molecular farming. Aligning and standardising the data management between these projects will facilitate data reuse and data interoperability. In addition, no reporting and metadata standards are currently available for NPBTs. The CHIC and the NEWCOTIANA projects will together contribute to the development of reporting requirements for datasets related to NPBTs.

2 CHIC Data Summary

CHIC aims to develop new chicory varieties with improved dietary fiber characteristics and improved terpene composition. Additionally, we will address the self-incompatibility which hampers the breeding efforts for this crop. This goal will be achieved by new plant breeding techniques NPBTs. More precisely in CHIC we will develop and apply gene editing approaches all based on the CRISP/Cas technology. We will use stable agrobacterium-mediated gene editing, transient gene editing techniques and the application of ribonucleoproteins to edit the genome DNA in the chicory protoplasts.

In the CHIC project data to assess the technological performance of these different methods will be collected. Additionally, the data related to the risk assessment of different NPBT techniques used, such as the off-target effects, will be generated. In improved chicory lines data about the dietary fiber and terpene composition and bioactivity will be evaluated. Economic feasibility and socio-economic impact of the newly produced chicory varieties will be evaluated. The data generated will contribute to evidence-based informed decisions on the legal status and regulation of NPBT crops.

To accomplish this a series of datasets will be generated:

- Improved genome assembly of *C. Intybus*
- RNAseq data on *C. intybus*
- gRNA inventory and gene editing efficiencies

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

- genetic part and construct designs
- Dietary fiber characterisation
- Terpene characterisation
- Data on regulatory networks for secondary metabolite biosynthesis in *C. intybus*
- Bioactivity data for dietary fiber / terpenes
- Phenotypic and agricultural parameters of newly developed *C. intybus* varieties
- Safety assessment, including untargeted effects, of different NPBT applications
- Socio-economic impacts
- Broader societal impacts
- Stakeholder views

Table 1 provides a list of research data categories that will be produced in CHIC and the expected data volume for each of them.

CHIC will re-use the constructs, RNAseq data and available genome data as well as established protocols for tissue-culture cultivation, protoplast transformation and regeneration of chicory that is available at different partners to maximize the use of resources.

Stakeholder views on commercial cultivation and use of GE chicory will be collected in order to clarify possible hurdles and facilitating factors for chicory innovation using GE techniques. Stakeholder views will be collected in the course of document reviews, interviews, questionnaires, workshops and focus groups. Data will be gathered as audio recordings, transcripts, interviews, and workshop notes. Only data gathered in the course of the CHIC project will be used

These data will not only serve to meet the objectives of the current project, but will also be useful for stakeholders including the scientific community, plant breeders, farmers, industry, legislators and regulators, and the general public.

Thus, the scientific community will benefit from the development of the NPBT techniques for chicory. The improved gene editing and knowledge on the off-target effects can be applied broader for gene editing of (the asteraceous) crops. The project will create added value for chicory farmers, by providing improved dietary fibre yield and quality and terpene yields. Additionally, the work on chicory incompatibility and the development of NPBTs will benefit chicory breeders. Finally, the generated data on the utility, efficiency and safety of NPBTS as well as the generated communication materials will help EU and National legislators and regulators and the general public make informed decisions on the regulation and public acceptance of NPBTs.

To enhance the usability of the data, open or otherwise widely-used file formats will be the preferred option for data collection (see Table 1). Formats that are open and/or in widespread use stand the best change to be readable in the future; on the contrary, proprietary formats used only by a particular software are prone to becoming obsolete. In those cases in which the laboratory instrument used to perform the measurement outputs the data in an instrument specific proprietary format, a converted version of the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

output file to an open data format will be shared together with the original file thus fostering data interoperability.

Table 1. CHIC foreseen data types, size and selected file formats.

Research data	
Data Type	File Format
Genome sequence data (raw and processed data)	bam, fastq
DNA parts and constructs	genbank, fasta plain text, ASCII (.txt), Ab1 (.ab1),
qPCR data	Raw data, comma-separated values (.csv), text (tab delimited) (*.txt)
RNA-Seq data (raw and processed data)	.ffn
Metabolomics data (raw and processed data)	mzML, netCDF (.cdf), comma-separated values (.csv), text (tab delimited) (*.txt)
Images (e.g. microscopy, immunoblots)	TIFF (.tiff), png (.png), jpeg (.jpg)
Tabular data (e.g. ELISA tests, metabolite yield, purity and functionality)	comma-separated values (.csv), text (tab delimited) (*.txt), MS excel (.xlsx)
Plant phenotypic data (contained and field conditions)	text (tab delimited) (*.txt), comma-separated values (.csv), MS excel (.xlsx)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

Plant genotypic descriptions	text (tab delimited) (*.txt), comma-separated values (.csv), MS excel (.xlsx)
Stakeholder views : audio recordings, transcripts, interview and workshop notes, questionnaires	audio recordings (mp3), MS Word (.docx), MS excel (.xlsx), comma-separated values (.csv).
Standard operating procedures, protocols	pdf (.pdf), MS word (.docx)
Scientific publications	pdf (.pdf), MS word (.docx)
Project reports	pdf (.pdf), MS word (.docx)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

2. FAIR Data

2.1. Making data findable, including provisions for metadata

The provision of adequate metadata (a description of the key attributes and properties of each dataset) is fundamental to enable the finding, understanding and reusability of the data, as well as the validation of research results. Descriptive metadata in particular, aims to provide searchable information that makes data discovery and identification possible. CHIC will adopt the DataCite Metadata Schema, one of the broadest cross-domain standards available, as the basis for dataset description. The minimum set of descriptors established for a CHIC dataset include:

- Type: a description of the resource.

Recommended best practice: use of a controlled vocabulary such as the DCMI Type Vocabulary).

- Identifier: a unique string that identifies a resource. Provided by repository where the dataset is stored.

Preferred option: digital object identifier (DOI); also accepted URL, URN, Handle, PURL, ARK.

- Publication date: date when the data was or will be made publicly available.

Format: YYYY-MM-DD

- Title: a name by which a resource is known (free text).

- Authors: the main researcher(s) involved in producing the data, or the authors of the publication, in priority order and affiliation. Recommended inclusion of a name identifier (e.g. ORCID)

Personal name format: family, given. Affiliation format: free text

- Description: additional information that does not fit in any of the other categories. Example: publication abstract.

Format: open.

- Version: the version number of the resource.

Format: track major_version.minor_version. Examples: 1.0, 2.1

- Language: primary language of the resource

- Rights: information about rights held in and over the resource

Values: openAccess, embargoedAccess, restrictedAccess, closedAccess.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

- Licence: information about the type of licence applying to the dataset
- Contributors: institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.

This property must also be used to allow unique and persistent identification of the funder. Values: European Commission (EU), H2020, Research and Innovation action, CHIC, Grant Agreement Number 760891 .

- Subject: subject, keywords, classification code, or key phrase describing the resource (free text).

Additionally, metadata elements and documentation providing specific information about the data collection processes, methodology, data analysis procedures, variable definitions, or relationships between the different files of a dataset will be compiled to ensure data interpretability and reusability. These metadata elements will be covered in section 2.3. The relevant metadata categories mentioned above will also be applied for data related to stakeholder interactions.

2.2. Making data openly accessible

CHIC project results will be made openly accessible provided that open publication does not interfere with the obligation to protect and exploit the results or the protection of personal data.

Regarding protection of results, to ensure that dissemination of the CHIC research outputs does not jeopardize their exploitation potential, project results will be subject to evaluation prior to any dissemination activity. CHIC IPR management and dissemination strategies are described in document D7.1 – PEDR. Results approved for dissemination will be made accessible through a variety of channels including project webpage (www.chicproject.com) social-media, scientific conferences, scientific publications in peer-reviewed journals, and data repositories, among others.

Regarding the protection of personal data, stakeholder views will be either audio recorded or documented in writing as interview or workshop notes. A restricted access policy will be implemented for stakeholder consultation data in order to insure confidentiality of personal data. These raw data will be only be handled and analysed by the teams conducting the respective research tasks. Summaries of stakeholder views will be presented in project reports which will be made publicly available on the project website and in open-access repositories. In these reports stakeholder views will be presented in a pseudonymised way. No reference will be made to individual stakeholder representatives or individual stakeholder organisations.

Being part of the Open Research Data Pilot (ORDP), the CHIC consortium is committed to provide Open Access (free-of-charge access) to all scientific publications and associated research data. The Open Access policy implementation is described in D7.1 – PEDR.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

Open Access (OA) to CHIC peer reviewed scientific publications will be mostly granted through "Gold" OA, although "Green" OA will be also be considered if "Gold" OA is not provided by the selected journal. Final versions of articles accepted for publication and their associated metadata (see section 2.1 and below) will be deposited in Zenodo, an interdisciplinary open data repository service created through the European Commission's OpenAIRE project and hosted at CERN, and will be made openly accessible at the time of publication ("Gold" OA) or with a maximum of 6 months embargo (for "Green" OA). Zenodo is compliant with the FAIR principles: it assigns a DOI to each deposited object, supports DOI versioning, is compliant with the DataCite Metadata Schema, is searchable, provides clear and flexible licensing, and provides secure back-up (see section 4).

In addition to the scientific publication, OA will also be provided to the research data required to validate the published results. Although Zenodo allows the deposit of data as well as publications, the use of discipline-specific repositories is often a more convenient option since (i) they have been developed to cover the subject specific needs and (ii) being widely used by the community, facilitate integration with other datasets. At present time, several discipline-specific repositories are under consideration for the deposit of CHIC datasets. These include:

- Metabolights: a metabolomics cross-platform and cross-species repository maintained by the European Bioinformatics Institute (EMBL-EBI). Metabolights supports the Core Information for Metabolomics Reporting (CIMR) metadata standard and submission of datasets follows the ISA-Tab format, a general purpose framework with which to collect and communicate complex metadata used by a growing number a repositories and publishers.

- Gene Expression Omnibus (GEO), Sequence Read Archive (SRA): two public data repositories at the US National Center for Biotechnology Information (NCBI) suitable for the deposit of RNA-Seq data (GEO) and high-throughput sequencing data (SRA) which are compliant with the Minimum Information about a high-throughput SEQuencing Experiment (MINSEQE) standard.

All data deposited in a discipline-specific repository will also have a record in Zenodo for the associated publication with a link to the externally deposited data files. Additionally, Zenodo will be the repository of choice for those data types for which a disciplinary repository is not available. The deposited dataset will include all the information needed to interpret and re-use the data following reporting standards when available (see section 2.3). These will include: publication file, raw and processed data files (in open or widely used formats), detailed protocols with information on instruments and settings used, a codebook for the variables used, and a readme file describing the files that compose the dataset and the relation between them.

As already mentioned, open or widely used file formats that can be accessed with open software (or software that is in widespread use) will be the preferred option for data collection. When the use of proprietary formats is necessary, the name and version of the software used to generate the file will be indicated in a readme.txt file included in the dataset.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

All data deposited in a repository will be made openly accessible under no access restrictions other than the embargo period for "Green" OA publications mentioned above.

2.3. Making data interoperable

Promoting data exchange and integration to its full potential requires the use of standardised data formats, metadata elements, and ontologies that ensure the reusability of the underlying data. As discussed in section 1, open or otherwise widely used file formats will be used to collect and share the data derived from CHIC research activities, thus facilitating data retrieval and analysis by other users.

With regard to metadata, likewise discipline-specific repositories, discipline specific metadata schemes broadly accepted by the scientific community should be the preferred alternative since they have been developed to cover subject specific needs. Accordingly, disciplinary repositories often show compliance with such specific metadata standards in combination with (recommended) controlled vocabularies. Metadata standards and ontologies that will be used to document datasets generated within the CHIC project include:

- Core Information for Metabolomics Reporting (CIMR) (metabolomics data)
- Minimum Information about a high-throughput SEQuencing Experiment (MINSEQE) (RNA-Seq and genome sequence data)
- Minimum Information about a Plant Phenotyping Experiment (MIAPPE) (plant phenotypic data)
- Minimum Information about a Proteomics Experiment (MIAPE) (protein mass spectrometry data)
- Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) (qPCR data)
- Data Documentation Initiative (DDI) (survey data)
- Plant Ontology
- Gene ontology
- OBI ontology
- NCBI taxonomy

There is currently no available reporting standard for CRISPR experiment metadata. To cover this need, the NIST Genome Editing Consortium works on the development of suggested minimal information reporting for public studies and the generation of a common lexicon for genome editing. CHIC will follow the progress of the Genome Editing Consortium on the development of a standard CRISPR metadata.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

At the same time, NEWCOTIANA and CHIC partners have initiated a common dialogue to define the metadata elements that should be collected for each genome editing experiment in order to facilitate sharing, validation, and interpretability of the results. A first draft metadata checklist (see Annex 1) covering the whole genome editing workflow has been assembled as a result of this work. This draft will continue to be refined in future working discussions involving both projects.

2.4. Increase data re-use (through clarifying licences)

Re-use is one of the pillars of FAIR data. Data re-use increases the impact and visibility of research, maximises transparency and accountability, promotes the improvement and validation of research methods, stimulates innovation through new data uses, and saves resources avoiding unnecessary replications. For data to be re-usable it should be in an open or widely-used file format, well described with rich metadata that meet domain-relevant community standards, and released under a clear data usage licence. The way CHIC will approach the first two points has already been discussed in section 1 (file formats) and sections 2.1 and 2.3 (metadata). Regarding licensing, as a default standard CHIC will share scientific publications and the associated research data under a Creative Commons Attribution Licence CC-BY whenever possible. CC-BY does not impose any restriction on access and reuse of the data; it allows users to copy, distribute, transmit, adapt and make commercial use of the data with the sole condition that the creator is appropriately credited. Most data repositories as well as most open access and hybrid publishers support the use of CC-BY licence.

Data quality check is the responsibility of the partners involved in the generation of the dataset and will be supported by a peer-review process at publication. Should errors be detected in already published data, these will be corrected and adequately documented in a new version of the dataset.

4. Allocation of resources

Adequate data management is an integral part of good research practice and as such it concerns every person involved in the research process. All CHIC partners have agreed to the general guidelines set up in this DMP and it is the responsibility of the group leaders to ensure that they are known and implemented by all members of their research group. For each dataset, the partner that generates the data is accountable for registering and storing all data and metadata according to the guidelines of this DMP, applying adequate back up policies, and sharing all public data through the selected open access repository. The project coordinator is in addition responsible for the maintenance of the project

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

website and the Sharepoint hosting service (see Section 4) for the sharing and storing of CHIC main documents during the active phase of the project.

As indicated in section 2.2, "Gold" OA publication will be chosen as the preferred publication option. Article processing charges for OA publishing were budgeted at the proposal stage and will be covered by the main partner of the publication out of their allocated funds. The estimated costs of applying open access publication is 2500€. It is not possible at this stage to determine the number of publications that will be produced. Resources for data storage and back up during the active phase of the project will be provided by the respective partner's institutions (costs included in standard indirect costs). No direct costs for data sharing and long term preservation are anticipated given that all the considered data repositories are free of charge.

5. Data security

All CHIC partners have adequate storage capability and back up policies at their respective institutions that guarantee the safe storage of the generated research data during the active phase of the project. Additionally, a variety of platforms are being used for internal data sharing, which also serve to the purpose of backup storage. All project documents (grant and consortium agreements, deliverables, meeting minutes, project reports and presentations, scientific manuscripts) are stored in a shared folder in Sharepoint, an Wageningen University and Research hosted sharing platform administered by WR that supports control access back-up and file version control.

Sustainable long-term preservation of the data beyond project completion is guaranteed by the use of trustworthy repositories such as Zenodo. Zenodo accessibility principles guarantee that deposited data and metadata will be retained for the lifetime of the repository, which is currently the lifetime of the host laboratory CERN, with an experimental programme defined for the next 20 years at least. Data files and metadata are backed up nightly and replicated into multiple copies in the online system ensuring file preservation. Finally, to preserve data authenticity and integrity, all files are stored along with a MD5 checksum of the file content and are regularly checked against their checksums to assure that file content remains constant.

Audio recordings and written notes of stakeholder views, as well as internal reports will be stored on password-protected servers only accessible for the partner conducting the research tasks.

6. Ethical aspects

In order to comply with the EU General Data Protection Regulation (GDPR) stakeholders participating in the project will be informed about the purpose, method, storage, processing, and publication of personal stakeholder data and data containing stakeholder views and asked for their permission. Stakeholder data collected in the course of the CHIC project will not include any sensitive personal data in the meaning of the GDPR. In publicly available reports stakeholder views will be presented in a anonymized way.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

There are no other ethical or legal issues relevant to data sharing of stakeholder data nor on the ethics Deliverables.

7. Annexes

Annex1: Draft metadata scheme for gene editing experiments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 760891

Annex 1. Draft metadata scheme for gene editing experiments.

Metadata Entry	Metadata Type
Experiment ID	unique identifier; format (NWC_ExpXXXX)
Partner	string
Transformation Method	{Agrobacterium, Biolistic plasmid, Biolistic RNP, Electroporation plasmid, Electroporation RNP, Other}
Nuclease	{Cas9, Cpf1, TAL, other}
Vector name (if applicable)	string
Cloning method (if applicable)	{Type II, Type IIS, GB, Moclo, Gibson}
Reference (Method)	string
Multiplexing	{Yes, No}
Cleavage method	{tRNA, other}
Number gRNAs	number
gRNAs name (optional)	string
Protospacer sequence	string (nucleotide sequence)
Protospacer lenght	number
PAM	string
Targeting strand	{Positive, Negative}
Target gene	string (Gene ID indicating version or coordinates +/- 500bp from PAM)
Target sequence	string (nucleotide sequence +/- 500 bp fromPAM)
Species	string
Guide design software (version)	string
On-target score	number
On-target score prediction algorithm	string
Off-target score	number
Off-target score prediction algorithm	string
Vector sequence	string
Construct sequence	string
Vector sequence link (benchling)	string
Transformation protocol	string
On target efficiency analysis (method description)	string
On target efficiency	{A: percentage of bialelic; B: percentage of heterozygous; C: percentage of chimeras}
Off target analysis (method description)	string
Number Off targets	number
Name Mutant(s) selected from the experiment	string
Mutated sequence selected	string (FASTA format target region +/- 500bp from PAM)
Number Off targets in selected mutant	number
Mutated off-target sequence in selected mutant	string (sequence)
Phenotype	string