Contents

[Introduction 1](#_Toc184031282)

[1. Data Summary 1](#_Toc184031283)

[2 FAIR data 4](#_Toc184031284)

[2.1. Making data findable, including provisions for metadata 4](#_Toc184031285)

[2.2. Making data accessible 5](#_Toc184031286)

[2.3. Making data interoperable 7](#_Toc184031287)

[2.4. Increase data re-use 7](#_Toc184031288)

[3. Ethics 8](#_Toc184031289)

**D3.2.1.2** **SFGIN Data Management Plan (DMP) V0.1 2nd December 2024**

# Introduction

The aim of this document is to provide guidelines on principles guiding the data management in the **Soft Fruit Genetic Improvement Network (SFGIN)** and what data will be stored.

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

The DMP will ensure that any data generated by the **SFGIN** will be findable, accessible, interoperable and reusable (FAIR) and operate to open science and data principles wherever and as much as possible.

# 1. Data Summary

The SFGIN will provide a platform for the genetic improvement of the major soft fruit crops in the UK, a space for knowledge exchange and training to deliver the knowledge into industry outputs. The project will enable the soft-fruit sector to generate improvements in sustainability, resilience and resource-use efficiency that fit into the UK Government’s Net Zero Strategy, which intends to halve net UK emissions in around a decade and eliminate them by 2050, while at the same time maintaining quality and nutritional standards as well as economic production. Research and development towards these goals through a soft-fruit GIN will provide vital knowledge, tools and germplasm resources to enable the soft-fruit industry to meet these goals and deliver significant benefits towards the public good.

Over the past two decades, UK soft-fruit breeding programs have focussed on supplying the industry with new high quality, high yielding varieties that meet the demands of the retail customers. To this end the key traits that have been prioritised in new varieties are large fruit size, high yields, shelf-life, attractive berries and good flavour. There is an inherent trade-off between many of the essential traits within soft-fruit and yield and quality attributes have been favoured over other attributes. This strategy has led to the development of many successful new varieties and has contributed to significant increases in average fruit size, as well as in yields. However, this has only been possible with significant inputs, increasing costs, and there has been no focus on increasing resilience of soft-fruit crops. Whilst soft-fruit purchasing is not monopolised in the UK, there are a limited number of very large retail customers, who are themselves being challenged to be more competitive on price. This means attributes such as nutritional quality, sustainable production and resource-use efficiency are of significantly lower priority in the production cycle, where the key to profitability is viewed as delivery of consistent yields and quality at a low production cost. It is therefore extremely timely for the formation of a pre-competitive soft-fruit GIN focussed on resilience, sustainability, nutritional and efficiency traits.

Recently external pressures have created significant challenges for the UK soft-fruit industry, presenting a real risk of a sharp decline in the sustainability for UK growers and a real risk of market failure. These pressures include rising costs in production, reduction in chemical controls, emergence of new pest/pathogens, reduction in availability of primary resources (water and nitrogen), lack of availability of sustainable substrates, and climate change (higher temperatures) all of which put pressure on production of all soft-fruit crops grown in the UK through reduced yields, reduced quality and coupled with rising costs across the sector and seasonal labour availability, ultimately result in a lack of profitability. The proposed soft-fruit GIN will promote research aimed at crop improvement in key traits that will serve the public good and promote sustainability and resilience within the sector, with an emphasis on transfer of knowledge, tools and resources between academics and industrial stakeholders on crop genetics and resilience, benefiting stakeholders across the breeding and production chains.

The proposed research within the soft-fruit GIN will provide solutions that allow the cutting of emissions from the input intense, soft-fruit sector and enable the Defra food strategy. It will achieve this by helping to increase availability of sustainably grown UK fruit, supporting increased consumption whilst maximising land-use efficiency, and by creating a shift in the UK food culture through the provision of cost-effective, high-quality UK-grown soft-fruit across as long a season as is possible to achieve sustainability.

**Q. Will the SFGIN re-use any existing data and what will it be re-used for?**

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 the **SFGIN**.

It is our intention to ensure that all data is made freely available to **SFGIN** partners, the UK soft fruit industry and the wider research community.

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

The **SFGIN** 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 SFGIN?**

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 the **SFGIN** are well described and annotated with metadata using open standards as far as possible.

**Q. What is the expected size of the data that SFGIN will 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 uncompressed.

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

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 this data be useful outside the SFGIN project?**

The data will primarily benefit the **SFGIN** 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 the **SFGIN** Germinate database (<https://germinate.hutton.ac.uk/sfgin>).

# 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 the **SFGIN** 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. Datasets will be assigned DOIs or PUIDs wherever possible.

**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.**

The **SFGIN** 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  the **soft fruit GIN** data  platform (https://germinate.hutton.ac.uk/sfgin) 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.* 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 SFGIN data management platform based on Germinate which is developing its DOI capabilities.

***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 the **soft fruit GIN** 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 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. In the case data is made public either through Germinate or alternative repositories, completely anonymous access will be provided and we do not track IP addresses.

***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. 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 the **SFGIN** 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 the **SFGIN**. We will use publicly available open-source and well-documented certified software, and any software or computer code created under the **SFGIN** 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 be followed to make our data interoperable to allow data exchange and re-use within and across disciplines?**

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 the **SFGIN** 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.)?**

All documentation will be provided through the Germinate GIN website (https://germinate.hutton.ac.uk/sfgin) or through the GitHub GIN repository (<https://github.com/sfgin>).

**Will data be made freely available in the public domain to permit the widest re-use possible and will the data created under this GIN be licences using standard reuse licences?**

Yes, our data, where possible, 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. There may be instances where data has a commercial aspect and, in these cases, while the data will be submitted to our central database it may not be publicly available.

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

Project data will be hosted in Germinate (https://germinate.hutton.ac.uk/sfgin) after the project end and plain text copies of data held, backed up and archived both by JHI and NIAB as per local archive procedures for a period of 10 years.

# 3. Ethics

**Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?**

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.

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

Personal data may be collected for dissemination and communication activities using specific methods and procedures developed by the **SFGIN** partners (JHI & NIAB) will adhere to current GDPR laws and include a liaison phase with data protection and governance officers at either site to ensure compliance.

The current information governance officers are:

**The James Hutton Institute**

Evangelia Apostolakopoulou Email: [Evangelia.Apostolakopoulou@hutton.ac.uk](mailto:Evangelia.Apostolakopoulou@hutton.ac.uk)

**NIAB**

TBC