PLAID
Data Management Plan

D1.1 PLAID Data Management Plan

Author: Claire Hardy, The James Hutton Institute
Project Coordinator: Lee-Ann Sutherland, The James Hutton Institute

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727388 (PLAID).

Contents

Introduction

A well devised Data Management Plan is essential in designing and producing a high quality research project. Data management refers to all aspects of creating/collecting, storing, delivering, maintaining, and archiving/preserving data. Data that has been generated often has a longer lifespan than the research project that produces the data. Researchers may continue to work on data after funding has ceased, follow-up projects may analyse or add to the data, and data may be re-used by other researchers. The Data Management Plan therefore needs to go beyond the life of the project. The plan is a living document that must react, evolve and develop, to respond to changes in the data and or project.

FAIR data management
In general terms, research data should be 'FAIR', that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation solution.

More information about FAIR:

FAIR data principles (FORCE11 discussion forum)
FAIR principles (article in Nature)

Managing the data effectively during the project lifecycle is very important for the success of any project.
The benefits of managing your data include:

•Meeting EU grant requirements.
•Ensuring integrity and reproducibility.
•Increasing efficiency.
•Ensuring data and records are accurate, complete, authentic and reliable.
•Enhancing data security and minimising the risk of data loss.
•Preventing duplication of effort by enabling others to use your data.

http://datalib.edina.ac.uk/mantra/datamanagementplans/

1. Data summary

1.1 Purpose of Data collection

The purpose of the data generated in the PLAID project is to enable the development of best practice guidelines and indicators to enable successful on farm demonstration leading to peer-to-peer learning.
The initial research data generated will enable a conceptual framework to be developed, which will be further developed to produce the analytical framework. The analytical framework will allow the characteristics and themes to be developed therefore allowing the inventory to be collected in a useful and meaningful manner.

The on-line georeferenced data base will allow farmers to locate appropriate demonstration locations to enable new innovations to be observed on commercial demonstration farms. PLAID will also enable the communication of knowledge between European regions through the development of virtual (on-line) demonstration activities on commercial farms.

The subsequent data generated will be used to draw up guidelines which will be used to develop easily accessible targeted decision-support tools. The tools will combine checklists of questions to ask and useful templates for establishing a demonstration event. Multiply versions of the tool will be available for different user groups, for example a farmer will need different questions to a funder when it comes to calculating the budget for a demonstration event on a farm. A range of these tools will be produced in consultation with a wide range of stakeholders. They will be made available in hard copy, electronically or useable directly through the PLAID project web-site using clickable forms.

1.2 Types and Formats of data to be generated

Type and format of the data collected during this project will be varied and are summarized below in the table showing deliverables and the data to be generated.

Table of Data types to be generated during the PLAID project

Deliverable Number	Short description	Data type	Data format
D2.2/2.2	Framework and typology initial and revised	Document	Word files
D2.3	Information notes on farm demonstration activity	Document	Word files
D3.1	Data collection template to allow inventory entry	Document	Word and excel files
D3.2	Information on national demonstration activities	Poster and document	Word (Text, images)
D3.3	Searchable geo-referenced data base	Post codes farm data, web-sites	Excel or other database
D3.4	Summary report on characteristics and looking at country differences	Document	Word files, figures, graphs
D4.1	14 virtual demonstration farms	Consent forms, meeting notes, contact details, videos	Word files, website addresses, Mpeg4
D4.2	28 Demonstration videos	Videos	Mpeg4
D4.3	Good practice guidelines for virtual demonstration	Information notes	Word documents, (Text, images)
D5.1/T5.1	Framework, feedback interviews, surveys	Documents: Interviews, surveys, notes	Word, Mp3, answers
D5.1/T5.3	24 case studies	Documents: Interviews, reports,	Word, Mp3, text
D5.2	Summary report on case studies	Document	Word files, graphs, figures, images
D5.3	Best practice materials	Information notes	Word, Text, images
D6.1	Validated decision-support tools	Information notes	Web-tool, Text, images
D6.2	Policy recommendation summary	Report	Text, graphs, figures, images
D6.3	Policy and governance briefs	Document	Word files
D7.1	Initial dissemination plan	Document	Word files
D7.2	Midterm dissemination report – summarizing activities in first reporting period	Document	Word files

D7.3	Final dissemination report – summarizing activities in second reporting period	Document	Word files
D7.5	Midterm practice Abstracts -10 abstracts in EIP Agri format based on research undertaken in the first reporting period	Document	Word files
D7.6	Final practice Abstracts - 10 abstracts in EIP Agri format based on research undertaken in the second reporting period	Document	Word files
D8.1	Ethics requirement consent form template	Document	Word files

In addition to the deliverables listed here general activities of the PLAID consortium will generate:

- press articles in partners' media or other media related to agriculture, publications
- internet posts through social media
- videos on PLAID YouTube channel (like PLAID presentation video)
- an email list, used for PLAID's outputs dissemination, which will be confidential
- PowerPoint presentations or posters from events / fairs / conferences

For geo-referenced material the format used will be GPS Latitude longitude referenced to WGS84. The World Geodetic System (WGS84) is the reference coordinate system used by the Global Positioning System. It comprises of a reference ellipsoid, a standard coordinate system, altitude data and a geoid. Similar to the North American Datum of 1983 (NAD83), it uses the Earth's centre mass as the coordinate origin.
http://gisgeography.com/geodetic-datums-nad27-nad83-wgs84/

This is the standard but needs to be specified to ensure all geo-references from all over Europe are compatible from the outset.

This list of data produced will be reviewed and updated periodically to ensure all data sets are included in the data management plan.

1.3 Data reuse

PLAID aims to reuse existing datasets already produced in previous research projects that have looked into European research into farming systems, agricultural innovation and social learning processes. PLAID will draw on the outputs of 11 preceding projects, in addition to the EIP Agri Focus groups (e.g. New Entrants to Farming, Arable Organic Crops) in which consortium members have participated. The existing data is available in the format noted below.

Acronym	How PLAID will draw on project outputs	Data Format
PRO AKIS FP7	The PRO AKIS inventory of advisory systems in the EU 27 will be used to identify points of contact for completing the demonstration farm inventory. Lessons learned from compiling the PRO AKIS data base will be applied (WP3).	PDF files
Solinsa FP7	Concepts relating to learning in informal networks will be integrated into the PLAID conceptual framework (WP2)	Word files
FarmPath FP7	Concepts relating to innovation conception and spread will be integrated into the PLAID conceptual framework (WP2)	Word files
TESS FP7	Lessons learned about on-line mapping of initiatives will be applied. (WP3)	Oral discussions.
Valerie FP7	Lessons learned about increasing up-take of research in practice will be integrated into WP3 conceptual framework.	PDF files
AgriSpin H2020	Source of data and contacts for the PLAID inventory (WP3)	e-book PDF files

WINETWORK ERA-NET	Source of data and contacts for the PLAID inventory (WP3)	PDF files
SUFISA H2020	Input to analysis of finance and governance in WP6	PDF files
RETHINK RURAGRI ERA-NET	Input to conceptual framework (WP2)	PDF files
SMART RURAL ERA NET	Experience of training rural farmers in the use of mobile applications will inform methods in WP4.	PDF files
FLINT FP7	Experience developing indicators will inform WP6.	MP4, PDF

1.4 Data Origin

Data origin will be reported in each deliverable. The use of existing data or the generation of new data will be clearly identified in each deliverable produced. The ethical clearance of all data both new and existing will be addressed by each work package leader and will be clearly reported upon. For further ethical information see section 8.0.

1.5 Expected Size of the Data

Most data sets will be in word documents ranging from 20kb to 300kb. Video format: mp4 or mov data size: 300 GB (30 videos estimated 10 GB for the raw material for each video produced). This however will change once uploaded to YouTube. This section will be reviewed and updated regularly.

2 FAIR Data

2.1 Making data findable

2.1.1 Digital Object Identifier (DOI)

A **Digital Object Identifier** or DOI is a persistent identifier or handle used to uniquely identify objects. The DOI is bound to the metadata describing the object, such as a URL, indicating where the object can be found. Thus, by being actionable and interoperable, a DOI differs from identifiers such as ISBNs and ISRCs which aim only to uniquely identify their referents.

The DOI for a document remains fixed over the lifetime of the document, whereas its location and other metadata may change. Referring to an online document by its DOI provides a more stable link than simply using its URL, because if its URL changes, the publisher only needs to update the metadata for the DOI to link to the new URL.

To monitor embargo periods, sometimes encountered when self-archiving (3.11 Green access – see later), the publication date and embargo period must be stated. The DOI will identify the publication and enable a link to the authoritative version of the publication.

2.1.2 Metadata

An identifier such as a DOI is of no value without some related **metadata** describing what it is that is being identified. The DOI requires the metadata to have two parts: “first, the DOI standard mandates a particular minimum set of metadata (the "Kernel" metadata) to describe the referent of a DOI name, supported by an XML Schema; secondly, to promote interoperability and assist RA's in the creation of their own schemas the IDF provides a Data Dictionary or ontology of all terms used in the Kernel, and other terms registered by Registration Agencies, and supports a mapping tool called the Vocabulary Mapping Framework.” (DOI Handbook – Data model)

https://www.doi.org/doi_handbook/4_Data_Model.html

Bibliographic metadata identifies the deposited publication; standard format dictates the following should be included

- the terms *["European Union (EU)" & "Horizon 2020"]["Euratom" & Euratom research & training programme 2014-2018"]*
- the name of the action, acronym and grant number
- the publication date, the length of the embargo period (if applicable) and a
- persistent identifier.

The bibliographic metadata makes it easier to find publications, ensures EU funding is acknowledged and helps monitor publications originating from H2020 funding enabling statistics to be produced and the impact of the programme to be assessed.

2.1.3 Findable

All final versions of deliverables once uploaded to the EC SyGMA system will be available with limited access on the PLAID website intranet site.
Final versions of deliverables marked as suitable for public dissemination, as indicated in the PLAID Grant Agreement, once uploaded to EC SyGMA system and having agreement from the principle author and Project Coordinator will be marked with a ISBN, DOI and necessary metadata. A DOI Registration Agency (see doi.org for relevant resources) will be selected and their guidelines followed as required. The ISBN identifiers will be registered with those available to the James Hutton Institute, and documented as required.
These procedures will ensure data is findable and searchable when stored in the public domain.

2.2 Naming convention - Project Documentation

How files are organized and allocated names has a substantial impact on the traceability of those files subsequently and the ability to determine their content. Files names therefore need to be allocated consistently and should be provided with a descriptive name so when organizing files it is obvious where to find specific data and what the files contain. It is essential to set up a clear directory structure that includes useful information to clearly identify the document. Consider what is important about the document and what you will require to identify the file at a later date. Files that will be shared or submitted should use the following convention:-

- Consortium. PLAID or FarmDemo (i.e. joint collaborative work)
- WP number: The work package the document is referring to
- Partner short name (see below).
- Staff initial.
- Version number

2.2.1 Midterm and Final reports

Midterm and final reports will use a similar naming convention as follows:-

- Consortium. PLAID or FarmDemo (i.e. joint collaborative work)
- Partner short name (see below).
- Staff initial.
- Version number

2.2.2 Deliverables

For final version of deliverables it is proposed that the following naming convention is used
PLAID><WP number><Deliverable number><Deliverable Title><Principle Author><date>
Where date is DDMMYY
E.g. PLAID_WP1_D1.2_Website_Claire Hardy_310317

2.2.3 Working documents

To prevent confusion and so all partners can track documents, making the data findable, both in terms of time and to attribute ownership, PLAID consortium partners will adopt the following naming convention. With the consortium name referring to either PLAID or FarmDemo (joint work) and partner abbreviations as listed HUT, NAAS, ACTA or IDELE, VINI, ISP, BSC, ASC, INTIA, LEAF, WUR, CRR, Delphy, FiBL

Consortium name><WP number><partner abbreviation><staff initials of principle author><version number<>short title
E.g. 1. Elzen - Plaid-AgriDemo framework memo
would become FarmDemo_WP2_WUR_BE_1.0_ framework
2. PLAID_WP2_T2.1_BSC_01022017
would become PLAID_WP2_BSC_Talis_1.0_successful demos

2.2.4 Minutes

When writing minutes for intermediate meetings it is important to identify either where, it is face to face, or skype, a date, work package, short title and version

Location or skype><DDMMYY<>Work package number<>version
e.g. skype_030317_WP2_Minutes3

Minutes for main meetings should identify consortium, location, date, short title
PLAID<>location>< DDMMYY<>short title
e.g. PLAID_Ghent_110117_1pre

2.3 Outline the approach for clear versioning

2.3.1 Data versioning - Definition

Versioning refers to saving new copies of files, so previous versions can be referred to, if necessary to allow changes to be tracked, it is practically useful when working on joint documents then the changes can be attributable as well.

2.3.2 Numbering versions

When creating new versions of your files, record what changes are being made to the files and give the new files a unique name.

- Include a version number, e.g. "1," "2,"
- New documents or datasets with no existing major version, numbered 0.0.
- Major versions with significant changes in status, numbered 1.0, 2.0, 3.0 etc.
- Minor Versions based working edits and changes, numbed 1.1, 1.2, 1.3, etc.

A unique version number should be assigned to each version of a document depending on whether the changes are significant (major) or not (minor) and allocate the new number accordingly. Lead authors determine the significance of the changes made and then allocate major and minor version numbering. The Project Coordinator will designate the final version by assigning a major version number to a document, and renaming the final version for uploading and public dissemination.
Naming conventions and versioning follow the approach and best practices as indicated by Stanford University Libraries, UK Data Archive.

3.0 Making Data Openly Accessible

What is open access? It is the ability to access on-line scientific data both peer-reviewed research articles and research data, both the raw and processed data generated and utilized in the scientific publications. As stipulated in the guidelines on Open access to scientific publications and research data in H2020 (version 3.1 August 2016) 'access includes not only the basic elements – the right to read, download, and print – but also the right to copy distribute, search, link, trawl and mine'.
PLAID will ensure open access as set out in the Grant Agreement specifically detailed in Article 29.2.

3.1 Open Access Methods

3.1.1 'Green' open access or Self-archiving

Green open access is when the author (or representative) deposits the published article (or peer reviewed final manuscript) in an archive recognized on-line repository (see below). This should occur at the same time or shortly after publication, although some publishers request an embargo period.

3.1.2 'Gold' open access or open access publishing

Gold open access occurs when an article is published in open access mode, the payment of publication costs is shifted away from the subscribers and instead the author makes a one off payment termed the Article Processing Charge (APC).

Even where gold access has been decided upon an electronic machine-readable copy (for a definition please see guidelines on Open access to scientific publications and research data in H2020 version 3.1 August 2016) must still be deposited in an on-line repository as soon as possible, to ensure that the article is preserved in the long term (repository being an on-line archive). Institutes repositories, subject based and central based repositories are all suitable see notes on the Open Access Infrastructure for Research in Europe (OpenAIRE) for guides on determining the most suitable repository to be used.

3.1.3 Open access as agreed by PLAID

'PLAID has agreed to encourage gold access publishing. €3000 has been allocated for this purpose. Owing to the relatively short duration of the PLAID and the time required for peer-review, two peer reviewed outputs are expected during the project, with several more published after the project concludes. Funding for gold open access will be sought from other sources after the project concludes (e.g. partner's institutional budgets), and green open access (self-archiving) undertaken if funding is not found.' PLAID Grant Agreement – associated document page 28.

3.1.4 Open Access Infrastructure for Research in Europe (OpenAIRE)

The original OpenAIRE project was started in 2009, this was followed by the OpenAIREPlus project in 2011. Then the OpenAIRE2020 project which is a research and innovation EU project that started in 01/01/2015 involving 50 partners to promote open access, improve discoverability and reusability of both research publications and data. It assists in monitoring H2020 research outputs.

OpenAIRE gives access to over 19,145,958 publications, 45,418 datasets, and 6,212 repositories. It is the recommended entry point for researchers to determine what repository to choose. It also offers support services for researchers, such as the National Open Access Desks. Other useful listings of repositories are:

- Registry of Open Access Repositories (ROAR)
- Directory of Open Access Repositories (OpenDOAR)

https://www.openaire.eu/

3.1.5 Open Research Data Pilot (ORD Pilot)

PLAID Work Package 1Task1.2 states 'PLAID will volunteer for the pilot action on Open Access to Research Data'. Article 29.3 in the Grant Agreement sets out the terms for open access to research data.

The ORD Pilot covers 2 types of data

- the 'underlying data' (the data needed to validate the results presented in scientific publications), including the associated metadata (i.e. metadata describing the research data deposited), as soon as possible.
- any other data (for instance curated data not directly attributable to a publication, or raw data), including the associated metadata, as specified and within the deadlines laid down in the DMP – that is, according to the individual judgement by each project/grantee.

To meet the ORD Pilot requirements PLAID must ensure two requirements have been met:

1. The data described above must be deposited in a repository, an on-line research data archive, subject, institutional or centralized based (see 3.14 for repository listings). The Open Access Infrastructure for Research in Europe (OpenAIRE) provides information and support on linking publications to underlying research data. Some repositories like Zenodo (an OpenAIRE and CERN collaboration), allows researchers to deposit both publications and data, while providing tools to link them. Zenodo and some other repositories as well as many

academic publishers also facilitate linking publications and underlying data through persistent identifiers and data citations.

2. Projects must take measures to enable third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) the research data.
   The most effective way of doing this is to attach Creative Commons Licences (see 5.2.1) to the data deposited.

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 Version 3.1 25/08/16

Zenodo will be the PLAID project designated repository for data storage, to ensure compliance with the ORD pilot.

4.0 Making Data Interoperable

4.1 Vocabularies and keywords

Controlled vocabularies are a standardized set of pre-defined terms used to describe a database, often referred to as subjects, subject headings, subject terms, descriptors, or index terms. Whilst keywords are often taken from a standard Thesaurus' which can be used to attribute keywords to dataset and publications, making them interoperable. See below for clarification.

Controlled Vocabulary	**Keywords**
Pre-defined, higher order terminology	Natural (conversational) language
Only searches the subject or descriptor field in a database record	Searches all fields in a database record including title, abstract, text, etc.
Results usually relevant to your topic	Results may or may not be relevant to your topic
Narrows your search providing fewer results	Broadens your search providing more results
Less Flexible - searches only available, pre-defined subject headings	Flexible - searches newer or distinctive terms and acronyms with no defined subject heading
Examples: Canines Skin Neoplasms Hypertension Insomnia	**Examples:** Dogs Skin Cancer High Blood Pressure Sleeplessness

http://mesacc.libguides.com/c.php?g=612321&p=4251654

Using them combined will help increase the interoperability of the data.

4.1.1 The US Library of Congress (Library of Congress, 2017)

The US Library of Congress provides an international standard vocabulary for use in bibliographic records. It provides controlled vocabulary words and phrases used to index content. The thesaurus contains preferred and variant terms. Searches for terms will produce an existing label and identifier, and variants of broader or narrower terms, and related terms. Specific identifiers will be embedded in the metadata for datasets and public deliverables leading to them being more effectively located by internet search engines, and documented in international libraries.

The US Library of Congress, https://www.loc.gov/

4.1.2 HASSET

Humanities and Social Science Electronic Thesaurus is a subject thesaurus that has been developed by the UK Data Archive over more than 25 years. HASSET was initially based on the UNESCO Thesaurus (ISBN: 92-3-101469-2). HASSET is the basis for the multilingual European Language Social Science Thesaurus (ELSST) which is currently translated into eleven languages and is used to aid retrieval in the CESSDA Catalogue. HASSET may be used by other organisations, for example to index materials,

provided it is for non-commercial purposes and appropriate acknowledgement is given. A license form allowing you to use or adapt HASSET in your own organisation can be requested by contacting the UK Data Service; please address your query to the Thesaurus Team. The HASSET thesaurus can be used to apply keywords to social science documents before archiving helping to make them Interoperable and findable.
HASSETT, http://www.data-archive.ac.uk/find/hasset-thesaurus

4.1.3 ELSST

European Language Social Science Thesaurus is a broad-based, multilingual thesaurus for the social sciences.
It is currently available in 12 languages: Danish, Czech, English, Finnish, French, German, Greek, Lithuanian, Norwegian, Romanian, Spanish, and Swedish.
ELSST was originally developed in 2000 as part of the EU-funded LIMBER project. It has been enhanced and extended through additional funding from the ESRC, the University of Essex, and through subsequent EU grants. ELSST is used for searching the CESSDA data catalogue and thus facilitates access to data resources across Europe, independent of domain, resource, language or vocabulary. https://elsst.ukdataservice.ac.uk/elsst-guide

5.0 Increase Data Re-use

5.1 Copyright

Is the legal right that gives the creator of a piece of work the exclusive right for its use and distribution, usually for a limited period of time. The EU encourages authors to retain their copyright but grant licences to publishers. Within the EU, the act of compiling a database attracts copyright insofar as the compiler has exercised intellectual judgement in selecting or arranging the data.
Copyright is an intellectual property right assigned automatically to the creator that prevents unauthorised copying and publishing of an original work. Copyright applies to research data and plays a role when creating, sharing and re-using data. Most research outputs such as spreadsheets, publications, reports and computer programs fall under literary work and are therefore protected by copyright. Facts, however, cannot be copyrighted.
For data collected via interviews that are recorded and/or transcribed, the researcher holds the copyright of recordings and transcripts but each speaker is an author of his or her recorded words in the interview (Padfield, T. (2010). Copyright for archivists and records managers, 4th ed., London: Facet Publishing).

Type of work	Copyright duration
Literary and artistic works	70 years from the end of the year of the death of creator
Sound recordings	50 years from date of creation
Typographical arrangements	25 years from date of publication
Crown Copyright	50 years from date of publication or 125 years from date of creation

5.2 Licences

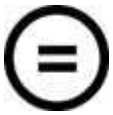
A licence is a legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held. Licences typically grant permissions on condition that certain terms are met. There are different types of licences that allow different levels of copyright freedom, the traditional approach of all rights reserved is not always appropriate, especially not for open access EU funded projects, however the author's rights should still be protected and equally the owner of the data. So a balance should be sought to allow open access within out breaching ethical requirements.

5.2.1 Creative Commons CC

Creative Commons CC is a global movement to encourage the distribution and sharing of any work that has been created. Creative Commons licences are free to everyone; they are available online and have been legally and linguistically adapted for use in over fifty countries. They make finding resources on the web, that you can legitimately use and re-use much easier, it provides an easy way to manage copyright enabling the sharing of materials.
Creative Commons doesn't replace copyright or ignore it and it doesn't mean 'copyright cancelled'; it builds on copyright. Some rights are reserved. Creative Commons licences are available from a free and open website which automatically generates a 'button', made up from a set of 'icons', and a link that you can fix to your work.
The licences are made up by selecting four basic conditions each represented by an 'icon':

BY means 'attribute' - if you use or re-use the work you must acknowledge the creator - name and a web link if available is usual.

NC means 'non-commercial' - you can use or re-use the work for personal or education use but not for a commercial project.



ND means 'non-derivative' - you can use or re-use the work but only as it is - you mustn't edit, mash-up or manipulate it to make something new.

SA means 'share alike' - if you have adapted someone's work - that is, created a 'derivative work' - and the SA icon has been used by the original creator it means you are obliged to release your new work under the same licence.

The creator chooses the conditions they require for their licence and the CC licence page generates a 'button':

CC-BY-NC means: Creative Commons licence; please acknowledge the creator. You are free to use, re-use, modify or develop the materials; BUT you must ask permission to use the material for commercial purposes

CC-BY-NC-SA means: you can use or re-use the work provided you acknowledge the creator, but don't use it for commercial purposes AND you must retain the CC licence on your use of the materials - You must 'share alike'

CC-BY-NC-ND this is the similar to above, but ND means 'no-derivatives'; i.e. you can't mash-up the music, manipulate the photograph or whatever - without asking permission.

Creative Commons licences have 3 layers. The top layer can be recognised by search engines, 'machine readable' and 'invisible' to the user, the second layer is called 'human readable', it has the icons that you can see but if you click on the icon it shows a description, the third layer, details the 'legal code' .
https://creativecommons.org/

5.2.2 Open Data Commons

Open Data commons was set up in 2007. The first licence produced was The Open Data Commons Attribution Licence (ODC-By). It allows licensees to copy, distribute and use the database, to produce works from it and to modify, transform and build upon it for any purpose. If the content is generated from the data, that content should include or accompany a notice explaining that the database was used in its creation. http://www.dcc.ac.uk/resources/how-guides/license-research-data

5.2.3 Open Government Licence (OGL)

The Open Government Licence (OGL) was released as part of the UK Government Licensing Framework in September 2010; version 2 was released in June 2013. It is intended for UK public sector and government resources although it does not state that it can't be used by licensors outside of the UK, its wording implies this, it is however very similar to CC BY.
http://opendatacommons.org/licenses/by/{version}

5.2.4 End User Licence

Users sign an End User Licence which has contractual force in law, in which they agree to certain conditions, such as not to disseminate any identifying or confidential information on individuals, households or organisations; not to attempt to identify any individuals; not to use the data to attempt to obtain information relating specifically to an identifiable individual; lastly not to share the data with unregistered users. This type of licence is often used for software but is useful for some survey and qualitative data for example data collected for the inventory i.e. GIS data.

6.0 Quality Assurance of Data

6.1 Quality Assurance of collected data

Data quality assurance will follow good practice standard procedures to allow data collection and analysis relevant to each task in the individual work packages. Academic project partners are all experienced in the collection and publishing of peer-reviewed outputs which are based on the data collected.

6.2 Quality Assurance of Deliverables

Task leaders will ensure the quality of the data produced, draft copies of deliverables will be passed to the Work Package leaders for assessment and finally to the PLAID Project manager. The data will be checked at each point to ensure the quality has been maintained. The Project manager will ensure final copies of the deliverables are passed to the Project Coordinator for the final check before uploading to the EC SyGMA.

6.3 Quality Assurance of documents for open access

Documents, both peer reviewed publications and the underlying data that are required to be deposited in open access repositories, will be prepared with the necessary: vocabulary, keywords, DOI, metadata, ISBN and licence before archiving. (see relevant sections).

6.4 Quality Assurance of Inventory

The procedure for checking quality assurance of the Data for the Inventory will be discussed at the June meeting 2017.

7.0 Data Security

7.1 General

All partners have back-up systems for their organisational data (see Annex II).

All final outputs will be stored on the James Hutton Institute servers; the Data Management complies with the requirements of its funders, insurers and adheres to the James Hutton Institute IT Services Back-up Policy. Final copies of all documents, relating to project management and reporting, are held by the James Hutton Institute in restricted access, shared folders. Project Coordinator and Project Manager have access to these files. The folders are backed-up daily (see below). Project management files which may be used by staff working remotely from the office will be used on laptops which have encrypted drives, requiring passwords to unlock, as well as passwords for logging in to the relevant operating systems (PC or Mac).

7.1.1 Backup Policy (James Hutton)

The backup system covers all central servers, all backups are fully automated and the status is monitored on a daily basis. The software used is Symantec NetBackup 7.6 which runs on Dell PowerEdge R720XD servers. There is currently one server per site which has local disk and tape storage attached. The backup system uses deduplication technology which stores only changed data instead of all data (applies only to disk). Daily there are incremental backups (6 month retention), weekly (6 month retention) and monthly (life of tape) there are full backups.

7.1.2 Storage (James Hutton)

Backups are stored on disk and tape. The backups stored on disk are only readable to the backup software and with the appropriate catalog information. Tapes are handled correctly to limit damage and are stored in fireproof safes that only limited members of staff have access to. Backup data transmitted across sites is encrypted. The live data, backup system with local disk, and fireproof safes storing tapes are all sited in different locations. A copy of critical backup data is also stored on the sister site.

8.0 Ethical Requirements

8.1 Ethics of Data Management

- Every researcher is expected to maintain high ethical standards. Ethical guidelines for research involving people are typically issued by professional bodies, host institutions and funding organisations. The key principles of research ethics that have a bearing on sharing or archiving confidential research data are: a duty of confidentiality towards informants and participants.
- a duty to protect participants from harm, by not disclosing sensitive information.
- a duty to treat participants as intelligent beings, able to make their own decisions on how the information they provide can be used, shared and made public (through informed consent).
- a duty to inform participants how information and data obtained will be used, processed, shared, disposed of, prior to obtaining consent.
- a duty to wider society to make available resources produced by researchers with public funds (data sharing required by research funders).

Relevant laws include the Data Protection Act 1998 and other legislation.
Data protection legislation applies only to personal data or sensitive personal data and not to all research data collection from participants or to anonymised data.
http://www.data-archive.ac.uk/

8.2 Research Ethics Policy for Human Participants

The James Hutton Institute has a Research Ethics committee that has a responsibility to protect the rights of human participants involved in research projects and to ensure the safety of the researchers.

There are four key criteria that the James Hutton Institute expects research proposals involving human participants to meet:

1. Research should be designed, reviewed and undertaken to ensure integrity, quality and transparency.
2. Research participants and staff should be informed fully about the purpose, methods and intended possible uses of the research, what their participation in the research entails and what risks, if any, are involved.
3. The confidentiality of information supplied by research participants and the anonymity of respondents must be respected.
4. The independence of research must be clear, and any conflicts of interest or partiality must be explicit.

The correct ethical form will be completed and submitted to the Ethics Committee 2 months prior to work commencing, this will allow the proposed work to be reviewed and gain approval.

The primary role of the committee is to protect the dignity, rights and welfare of those acting as human participants in research studies and of researchers. It will also consider the consequences of the proposed research for others affected directly by it or those who might benefit or suffer from its outcomes in the future. (The Hutton Research Ethics Policy for Human Participants 31 July 2013)

8.2.1 Procedure for Ethic consent forms

Once approval has been received the templates for consent for data gathering will be circulated to all participants. When gathering data for PLAID the consent form must be filled out completely and accurately ensuring at all times a clear signature has been obtained authorizing the gathering, use and storage of data. This form must be securely stored, along with a digital back up and kept with the data at all times.

9.0 Acknowledgements

This Data Management Plan has followed the Template Horizon 2020 Data Management Plan Template and has taken inspiration from both David Miller (James Hutton) and the SIMRA Data Management Plan.

10.0 Bibliography

http://datalib.edina.ac.uk/mantra/datamanagementplans/

http://gisgeography.com/geodetic-datums-nad27-nad83-wgs84/

DOI Handbook – Data model https://www.doi.org/doi_handbook/4_Data_Model.html

Stanford University Libraries, http://library.stanford.edu/

UK Data Archive. http://www.data-archive.ac.uk/

OpenAIRE https://www.openaire.eu/

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 Version 3.1 25/08/16

Vocabularies, http://mesacc.libguides.com/c.php?g=612321&p=4251654

The US Library of Congress, https://www.loc.gov/

HASSETT, http://www.data-archive.ac.uk/find/hasset-thesaurus

ELSSETT, https://elsst.ukdataservice.ac.uk/elsst-guide

Padfield, T. (2010). Copyright for archivists and records managers, 4th ed., London: Facet Publishing.

Creative Commons, https://creativecommons.org/

Open Data Commons Attribution Licence, http://www.dcc.ac.uk/resources/how-guides/license-research-data

Open Government Licence, http://opendatacommons.org/licenses/by/{version}

Data protection legislation http://www.data-archive.ac.uk/

Ethics requirements, The Hutton Research Ethics Policy for Human Participants 31 July 2013

ANNEX I

Data Management Plan Summary

The Data Management Plan (DMP) will be reviewed and updated at regular intervals. It is a living document and will therefore evolve as the project progresses.
The following are sections that should be regarded as high importance.

- **Data origin** (1.4) must be reported in each deliverable and in working documents where appropriate
- It is important to adhere to guidelines to make **data findable** (2.1) in particular by adding digital object identifiers, metadata and ISBN identifiers to published documents (including related data)
- **Naming conventions and versioning** must be used and adhered to, again to help in making data findable (2.2 & 2.3)
- The PLAID project has agreed in the Grant Agreement (Article 29.2) to ensure **open access** to all scientific publications and research data. 'Green' open access (3.1.1) as a minimum, 'Gold' open access (3.1.2) where possible
- The PLAID project will also ensure published documents and occupying data are deposited in an on-line repository (Zenodo has been selected for the PLAID project) following the guidelines for the **Open Research Data Pilot** (3.1.5)
- **Vocabulary** - US Library of Congress standard vocabulary will be used during preparation for archiving (4.1)
- **Keywords** – HASSET/ELSST keywords will be used as standards (4.1.2 & 4.1.3)
- Copyright – **Creative Commons licences** will be used where appropriate (5.2.1), with end user licences used on some data
- **Quality Assurance** – will follow good practice standard procedures to allow data collection and analysis relevant to each task in the individual work packages (6.0)
- **Data storage and security** – see Annex II for details of data storage of partner organisations. The James Hutton data storage is detailed (7.1.1) as all the final outputs generated by the PLAID project will be stored at this facility ensuring the security and longevity.
- **Ethics** – standard templates will be used when data is collected to protect the dignity, rights and welfare of those acting as human participants in research studies and of researchers (8.2)

ANNEX II

PLAID Consortium Partner Backup Policies

Partner	Backup Policy
HUT	See main text
NAAS	Weekly there is back up on our data base information system on NAAS IBM server and parallel on HD of IBM server i.e. we have two independent places for back up of our data base. For internet page- a daily back up on server of company which supports our site.
ACTA and Idele	All computers are connected to the server. Weekly there is a back-up of the whole server. Every night there is a back-up of what was modified since the previous week-end. Data saved on employees' own computers (desk or local folders) are saved every 15 minutes on another server.
VINIDEA	Files are stored in a central server situated in our office. This server has 2 internal HDDs. Information are stored in the 1st HDD and the 2nd HDD works as a mirror. All data (documents and email messages) are also synchronized on our computers for off line working. Once a week there is a full backup on an external unit, with daily incremental copies during the night. There are 2 external units alternating every Monday.
ISP	Data on ISP servers (Sharepoint, Outlook, …) have a back up on a daily basis, with a rotation of 1 month
BSC	The Baltic Studies Centre implements the following data management principles and procedures: • Research data (statistical materials, interview transcripts, analytical documents, research reports, scientific papers, etc.) are stored on BSC Research Repository. • The data are stored for at least five years after the original date of collection or publication. • Sensitive data (data containing personal information, interview transcripts, personalised research documentation, etc.) are destroyed after five years from the date of their original collection.
ASC	Employees have notebooks which are not automatically backed up. However each employee has an assigned space on NAS. Space on the NAS is mapped as network drive and access is possible only with username and password. From a remote location, the NAS can be accessed through OpenVPN and therefore can gain secure access to the folder. Drives in NAS are in RAID 1 configuration to provide redundancy or in RAID 10 for those who need redundancy and performance. We use Synology NAS solution. We use Backup Strategy 3-2-1, one copy on employee computer, one on NAS and one offsite for very important data (thru Synology Hyper Backup). All employees also get USB flash drive and free access to DVD-R/RW, CD-R/RW, ... to make extra copies if they need one.
INTIA	The backup service in the servers of Nasertic, Government of Navarra and Public Companies under corporate standard. The software used to make copies is ArcServe UDP, all types of files are backed up as long as they are not in use and the results of the backup jobs are monitored from Monday

	to Friday and issues are issued to solve the problems. Further information is available.
LEAF	1. All LEAF PC's are networked through a secure login and password to an MS Windows 7 secure server with a dedicated firewall. 2. The server is third party administered, managed and monitored for security and safety 24/7. 3. All LEAF ICT systems are supported by external ICT service providers 4. The server filing system has security permissions administered through the Operations Manager to limit and control access to data and files. 5. The files are backed up automatically and remotely on a daily basis and a full server backup is conducted weekly to a secure cloud server. 6. An additional 'disaster recovery' back up of the server's operating system and office PCs is conducted weekly. 7. An additional internal back up is conducted quarterly to a portable HDD which is retained in a fire proof safe. 8. Laptops are not directly backed up but staff save files to the server as required. Additionally all staff have access to secure MS Office one drive for temporary storage of files online 9. Back up data is stored for 6 years Data on LEAF websites I retained on third party secure servers that are backed up to alternate servers daily.
WUR	The security of WUR information systems, conducted by the FB-IT department, is based on an appropriate set of measures, according to the Information Security Code (BS 7799, ISO / IEC 17799). The measures have been taken with the aim of ensuring that availability, integrity and confidentiality function at a level that fits an institution for research and education. The information security is reviewed annually by the auditor. **Specific measures** • Physical security; The IT infrastructure (network, servers and Internet access) is located in areas protected from fire, burglary, power failure and the like. All files are stored in two geographically separated WUR data centers. In case of loss of one full data center, data loss will take up to 10 minutes. Additional measures have been taken to restore files when two full data centers are lost. Files that are more recent than two weeks old can be lost. • Security of information systems and infrastructure; At all levels in the information chain, measures to protect data and systems against hacking (unauthorized access), viruses and the like have been taken. Periodic security scans take place within the framework of the Technical Security Audit Plan; Backup en restore of stored data and systems. • Password Policy; Information systems and data are protected with the default username / password combination. • Backup Policy; The backup system covers all central servers, all backups are fully automated and the status is monitored on a daily basis. Daily there are incremental backups (6 month retention), weekly (6 month retention) and monthly there are full backups. It is possible to restore deleted or overwritten files up to one year back.
CRR	All staff save work on to individual computers with daily backups taken to the Norges teknisk-naturvitenskapelige universitet (NTNU) server. Sent and received E-mails using the @bygdeforskning address are also automatically backed up on the NTNU server.

DELPHY	Backup is made every night of all data on network disks by the Delphy ICT Manager
FiBL	Storage of PLAID project data at FiBL - The FiBL-server which has a back-up every day. - All video footage is stored on a separate external hard disk, with regular backup to a second hard disk. - The edited videos are stored on YouTube and at a third external hard disc.