



FAIR data guidelines for pig research

Deliverable 3.3

Start date of the project: March 1st, 2021
Duration: 60 months
Deliverable Title: FAIR data guidelines for pig research
Deliverable Number: D3.3
Deliverable Lead: Rob Lokers
Related Work package: WP3
Author(s): Rob Lokers, Hendrik Boogaard, Catherine Hurtaud, Nina Melzer, Catherine Larzul
Contributor(s): Claudia Kasper, Sarah Fischer
Communication level: PU
Due date: M24
Actual submission date: M29
Revision: V1

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

Contents

1 General introduction

2 Open Science

- 2.1 *Definition of Open Science*
- 2.2 *What to do and why?*
- 2.3 *Useful links on Open Science*

3 FAIR principles

- 3.1 *Introduction*
- 3.2 *Open Data*
- 3.3 *FAIR principles and Fair Data sharing*
 - 3.3.1 *Reasons to comply with FAIR*
 - 3.3.2 *Making Data Findable*
 - 3.3.3 *Making Data Accessible*
 - 3.3.4 *Making Data Interoperable*
 - 3.3.5 *Making Data Re-usable*
- 3.4 *Some myths and misunderstandings around FAIR*
- 3.5 *Useful links on FAIR data sharing*

4 Metadata and data standardization

- 4.1 *Definition of metadata*
- 4.2 *Common metadata*
 - 4.2.1 *Introduction to Dublin Core (DC)*
 - 4.2.2 *Highly recommended elements of the Dublin Core*
 - 4.2.3 *Recommended elements of DC*
 - 4.2.4 *Restrictions and limitations of DC*
 - 4.2.5 *Example of a dataset published in Zenodo*
- 4.3 *Data and metadata standardization*
 - 4.3.1 *Introduction*
 - 4.3.2 *Additional standardisation*
 - 4.3.3 *Short insight into the ABCD standard ("Access to Biological Collections Data")*
- 4.4 *Useful links on meta-/data and standardization*

5 Ontologies

- 5.1 *Introduction*
- 5.2 *ATOL ontology*
- 5.3 *Examples of annotation of data in publications*
- 5.4 *Annotation of data: how to proceed?*
- 5.5 *Useful links on ontologies*

6 Data curation

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

6.1 *Introduction*

6.2 *Organisation and storing data*

6.2.1 Folder and files

6.2.2 Relational databases

6.3 *Data quality*

6.3.1 General aspects

6.3.2 Data plausibility examples - introduction

6.3.3 Data plausibility examples - formatting problems

6.3.4 Data plausibility examples - checking individual variables

6.3.5 Data plausibility examples - checking dependent variables

6.4 *Version and backup*

6.5 *Useful links on data curation*

7 **Data publication**

7.1 *How to publish?*

7.2 *The data paper*

7.3 *Choosing a repository*

7.4 *Metadata repository*

7.5 *Useful links on data repository*

8 **Data Management Plan**

8.1 *Introduction*

8.2 *Writing a Data Management Plan (DMP)*

8.3 *DMP and the FAIR principles*

8.4 *Useful links*

Annex I - Example of an ABCD data standard in pig research

Mandatory elements

Core elements for storing experimental data

Abbreviations

ABCD	= Access to Biological Collections Data Schema
AHOL	= Animal Health Ontology for Livestock
ATOL	= Animal Trait Ontology for Livestock
BPS	= BioCase Provider Software
CHEBI	= Chemical Entities of Biological Interest
CSV	= Comma Separated Values (file format)
DC	= Dublin Core
DMP	= Data Management Plan
DCC	= Digital Curation Centre
EOL	= Environment Ontology for Livestock
FAIR	= Findable, Accessible, Interoperable and Re-usable
FBN	= Research Institute for Farm Animal Biology
INRAE	= National Research Institute for Agriculture, Food and Environment
ICT	= Information Communication Technology
IPR	= Intellectual Property Rights
LPT	= Livestock Product Trait Ontology
NWO	= Dutch Research Council
ODI	= Open Data Institute
ORCID	= Open Researcher and Contributor Identifier
SOP	= Standard Operating Procedure
TDWG	= Biodiversity Information Standards
VTO	= Vertebrate Trait Ontology

1 General introduction

Data Science is at the heart of a lot of research activities, also for pig research. Consequently, the role of data and “taking care of data” is becoming increasingly important, as it has become an indispensable asset. The amounts of available data are growing exponentially, and to make that data most valuable for the research community, it is key to ensure that relevant data is available, can be found, and be (re)used. This is also becoming an important factor in research, as many funders emphasise the importance of high-quality data and data management, and making data re-usable for the wider community. To ensure the re-usability of the data, they are often checked against the FAIR (Findable, Accessible, Interoperable and Re-usable) criteria (see chapter 3).

The PIGWEB project has performed surveys to explore the current landscape of pig research when it comes to data and data management.

Some relevant outcomes of the landscape survey¹ in pig research were:

- Collaboration is key to pig researchers’ work.
- Many researchers are involved in tasks related to data processing.
- In general, researchers are relatively comfortable with sharing data and data being re-used. They are mostly positive about data sharing and see the benefits of data re-use.
- Some researchers never re-use data, and most of them only re-use their own data.
- Successful data re-use is achieved in about half of the attempts.
- Researchers are not very familiar with the FAIR principles and FAIR policies and think they generally do not deliver FAIR data. They feel they need help with (FAIR) data sharing.
- In general, researchers see many barriers for data sharing, like lack of time, lack of budget, lack of knowledge, and lack of rewards for data sharing.
- Researchers feel they get too little credit for data, where citation and co-authorship would be good incentives.

This reflects a couple of aspects around (FAIR) data sharing and re-use. First, researchers seem to see the value of sharing and re-using data but are practically hindered by a lack of knowledge and resources. Secondly, the incentives to share data seem to be insufficient. These might be the main causes of the currently low data sharing and re-use adoption. At the same time, there might be some misunderstanding regarding the current opportunities and incentives.

The FAIR data guidelines for pig research in this Deliverable introduce the FAIR principles and the requirements for delivering FAIR data, and the various aspects regarding data management and curation that are relevant for efficient data sharing and re-use. The objective is to provide knowledge and introduce good practices and tools that can support the adoption of FAIR data practices by the broader community’s adoption of FAIR data practices. Moreover, it attempts to lower some of the barriers to data sharing and re-use by discussing some observed misunderstandings and interpretations and clarifying some often less well-known opportunities and incentives.

This Deliverable starts with an introduction to Open Science and the FAIR principles, explaining the motivation behind the FAIR data movement and how it relates to the broader process of working with data. The various steps of data curation, the handling of data, from data collection to data publication and re-use, are presented. Some key aspects in this process are discussed in more detail, specifically

¹ <https://www.pigweb.eu/deliverablemilestones>.

how data can be harmonized using common standards, formats, semantics etc., and how data can (should) be published so they can be easily re-used. A separate section focuses on data management plans (see chapter 8). A data management plan describing how data can be handled in a project, which is a mandatory deliverable for more and more research projects. In the various chapters, several user cases from the pig research domain are used to illustrate how FAIR data and data management aspects can be applied practically in research.

2 Open Science

Open Science is a fundamental concept on how to perform research that is increasingly adopted by the global scientific community. It is closely linked to FAIR, where the FAIR principles are a good way to implement some of the key aspects of Open Science. Therefore, to understand the broader idea behind FAIR principles and why making data FAIR is useful, we will first take a closer look at Open Science.

2.1 Definition of Open Science

There is no single, unique definition of Open Science. However, looking at the various definitions available (Figure 2-1), we can clearly see the relevant aspects.

Open Science
The movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of society , amateur or professional (source: wikipedia.org)
Open science encompasses unhindered access to scientific articles, access to data from public research, and collaborative research enabled by ICT tools and incentives (source: OECD)
The practice of science in such a way that others can collaborate and contribute , where research data, lab notes and other research processes are freely available, under terms that enable re-use, redistribution and reproduction of the research and its underlying data and methods (source: fosteropenscience.eu)
Open Science is the movement that aims at more open and collaborative research practices in which publications, data, software, and other types of academic output are shared at the earliest possible stage and made available for re-use (NWO , NL)

Figure 2-1 Four definitions of Open Science

Open Science is about performing research in such a way that the results are as broadly accessible and understandable as possible. Results, in this case, should be seen in a broad sense and it is not only about scientific publications and data. It is also about good, understandable descriptions of that data, how it was generated and processed, and about the software and algorithms used for that (see Figure 2-2).

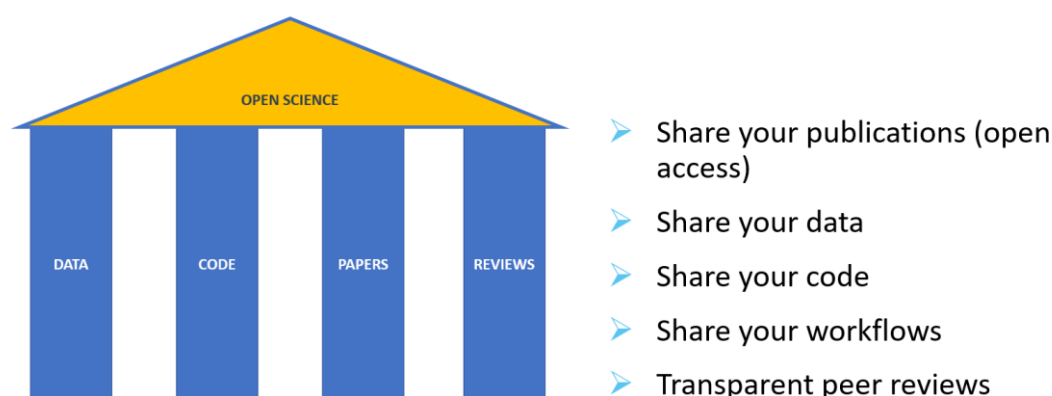


Figure 2-2 The main pillars of Open Science

Basically, it is about producing understandable results and sharing them broadly and as openly as possible. The overall objective is that fellow researchers and other stakeholders can re-use the

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

scientific output. The idea is that this will lead to collaboration and co-development, broader contributions to the scientific process, and ultimately to more innovations and more value and impact for society.

2.2 What to do and why?

Open Science is basically about sharing research work in such a way that others can easily understand and re-use it. Sharing scientific output transparently is relevant for others (and for the creator) to be able to integrate the data into other research. It also makes collaborating and co-developing with others and extending scientific networks easier. Open Science aims to make research more democratic and inclusive, for example, because stakeholders that otherwise might not be able affording to pay for data can now use it. Moreover, it makes full reviewing of scientific work possible, so that it can be verified and validated so that, in the end, Open Science contributes to reducing cases of fraud and misconduct.

All these factors contribute to enhancing the impact of research in the broader scientific community and society, ultimately leading to increased innovation in a more efficient manner. This is also why more and more funders require that results from their funded projects are made openly available and that researchers adopt Open Science and FAIR practices as a fundament of research.

The main reasons for promoting and adopting Open Science are:

- It helps to maximise the impact of your research.
- It provides the foundation for others to build upon.
- It supports the validation and reproducibility of scientific work.
- It reduces cases of academic misconduct.
- It supports a levelled playing field.
- It responds to funder requirements.

In the next chapter the FAIR principles are explained. They are an important element of Open Science; understanding and applying them in research is a big step towards practicing it.

2.3 Useful links on Open Science

- Foster Open Science ([here](#))
- Wikipedia ([here](#))

3 FAIR principles

3.1 Introduction

Data sharing and data re-use are important objectives of Open Science. Obviously, sharing data is only useful if others can work with that data and re-use it in a good and efficient way in their research. There are a few aspects then that are relevant with respect to data:

- To know that the data exists, others should be able to discover it.
- To be able to start working with the data, others should be able to get the data.
- Additionally, for others to decide if they can use the data, they should also be able to “easily” understand the data, its background, how it was generated, processed etc.
- To start working with the data, it is important that it is in a form where it can easily be processed and combined with other data, e.g., using analytical tools.

These aspects are the core of the FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**e-usable) principles and FAIR data, and they will be further explained in this chapter.

It is good to realize that data can be classified in different categories, which plays a role in how that data and the derived data will (and can) be dealt with when it comes to data sharing.

Open data - Data that anyone can access, use, and share. Data must be licensed to make clear that anyone can use the data in any way they want, including transforming, combining, and sharing it with others, even for commercial purposes (the Open Data Institute).

Shared data - Shared data may be made widely accessible but could have some restricting conditions, such as non-commercial re-use or re-use with attribution. It is important to note that not all shared data has to be available to anyone.

Closed data - If researchers are dealing with sensitive data (e.g., personal or commercially data), it may not be possible to share the data.

3.2 Open Data

A term that is often used in connection with data sharing is Open Data. Open Data is related to FAIR data, but it is not the same (see Figure 3-1).

“Open data is data that anyone can access, use and share”

Figure 3-1 Definition of Open Data (the Open Data Institute)

The definition of Open Data provided by the Open Data Institute is simple and clear. Anyone should be able to get the data, work with it, and reshare it with others. Besides, Open Data should have a licence attached that says that it is Open Data. An Open Data license might also indicate that users must credit the publishers (attributions) and that people using the data and combining it with other data should again publish the results as Open Data (share alike).

The ODI also mentions that good Open Data has the following characteristics:

- The data can be linked, so that it can be easily shared and talked about.
- The data is available in a standard, structured format, so that it can be easily processed.
- The data has guaranteed availability and consistency over time, so that others can rely on it.
- The data is traceable, through any processing, right back to its origin, so others can work out whether to trust it.

Obviously, publishing data as Open Data is very useful. But it is also reflecting an “ideal situation” where you are free to decide that your research data will be openly available to others. In practice, there are many situations where this will not be the case. Data that you use or produce might be sensitive. Also, source data might already have a license attached that does not permit derived data to be published as Open Data. So, referring to the types of data mentioned in the introduction, one might be dealing with shared or even closed data in many cases. Even if such scenarios seem to contradict the idea of Open Data, the characteristics of Open Data, and in particular the FAIR principles, are useful to use to facilitate and allow your own or internal organisation’s re-use of the data.

3.3 FAIR principles and Fair Data sharing

The FAIR principles stand for: **F**indable, **A**ccessible, **I**nteroperable and **R**e-usable.

Findable	The data should be discoverable.
Accessible	The data should be available and obtainable (if needed with authentication and authorisation)
Interoperable	The data should be parseable and integratable with other data (e.g. for analysis and processing)
Re-usable	The data should be well-described, allowing the most comprehensive re-use possible and the least cumbersome integration

The following sections shortly explain FAIR, why to adopt and how integrate these FAIR principles. Many of the terms used will be further explained in the next chapters of this guidance.

3.3.1 Reasons to comply with FAIR

There are many good reasons to make data FAIR. It all boils down to making research more efficient and impactful, by making it easier to find, get, understand, and use data generated by others.

A lot of research is done worldwide, resulting in a wealth of potentially valuable data. Rather than constantly reinventing the wheel by generating similar data over and over again, we can make use of the work of fellow researchers. This is one of the aims of the FAIR principles. When data is made FAIR, re-using it in other research becomes much easier. This requires, among other things, that data is formatted so it can be easily used with data science tools and that it is well documented, so its background and the options for re-use can be easily understood. And, of course, the data should be available for download online so that it can be easily found and downloaded.

It is often overlooked that publishing FAIR data allows you to get credit for your work. First, the simple fact that others can more easily get and re-use your data will increase the chance that it gets cited. Moreover, connecting a suitable license to your data, requiring attribution, ensures your work gets appropriate credits.

Another reason to work on FAIR data is that many funders of research require you to do so. They are generally looking for ways to make their investments in research more effective and to increase the societal impact of the research they fund. Consequently, many research organizations have made FAIR part of their strategies and integrated it in their data policies. In most cases, this requirement is enforced in the form of a mandatory research data management plan (DMP) (see chapter 8).

Finally, it is becoming a common practice that scientific publishers and journals require the underlying data along with the publications. Again, one of the reasons is to make it easier to understand and re-use (parts of) the published research. But it is also a way to increase the transparency and traceability of research, so results can be reproduced and verified if needed.

3.3.2 Making Data Findable

Users looking for specific data for their research or to underpin decision making will usually try to find data using search capabilities on the Internet. Chances that they will find a dataset will increase when the following is implemented.

- The dataset is published in a data repository or data catalogue. See also chapter 7.
- The dataset has metadata (data about the data) attached. It is this metadata that is usually published and made searchable through the data catalogues. Thus, sufficient and high-quality metadata will increase the chance that people will find the data.
- The dataset can be identified and accessed by means of a standard identification mechanism. Digital Object Identifiers (DOIs) are commonly used to identify and create a link to a dataset. They are essentially unique URLs that lead to the dataset and remain linked with the data for its entire lifetime.

3.3.3 Making Data Accessible

In principle, once a dataset has been found by the user, it should be available and obtainable (usually meaning downloadable) for re-use. Even if the access to the dataset itself may be restricted, it is important that at least the metadata of the dataset is available. Firstly, the existence of the dataset is then documented for all, which increases the transparency of research. Secondly, the license information provided as part of the metadata could clarify the conditions for obtaining and re-using the data. Additionally, the available metadata provides a possibility to obtain further information about the dataset or even contact the creator directly and possibly circumvent restrictions and negotiate access.

- Data repositories and catalogues will provide the option to download the dataset if allowed.
- It is important to select a data repository that offers long-term storage of the dataset, linking it to a DOI to reference and identify it. This ensures that the data will not “get lost” over time and delegates the responsibility to ensure that. See also chapter 7.

3.3.4 Making Data Interoperable

An important aspect of making data FAIR is data interoperability. It essentially means that data has to be syntactically parseable² and semantically³ understandable. This will allow easy data exchange and re-use between researchers, institutions, organisations, or countries. It also allows that data can be easily automatically processed or combined with other data.

- Make your data available as structured data in non-proprietary formats. As an example: structuring data in an Excel-file is already better than a scan of a table from a document, again CSV as a non-proprietary format is more interoperable than Excel).
- Document your data by providing metadata according to a recognized metadata standard.
- If possible, use common taxonomies or ontologies to tag and describe your data.

3.3.5 Making Data Re-usable

The final step is making data re-usable. In fact, a lot of this is already accomplished if the data is made Findable, Accessible, and Interoperable. These last steps are especially important to clarify for re-users how they can or cannot use the data. Many of the potential misinterpretations and misuses of a dataset can be prevented by:

- Describing the dataset further, for example by providing information about its provenance. It should be clear which data and/or instruments were used to generate the data, what the processing steps were, and which technical and use restrictions that might result in. Also, think about describing other relevant information that cannot be included in the dataset's metadata, e.g. a good description of the dataset attributes. See also chapters 4, 5 and 6.
- Attaching a license to the data (e.g. through its metadata) so that people know about the restrictions for re-use. The least restrictive license allows for the widest re-use, as licences are not only binding for external users, but also for the data creator's re-use of the data.

3.4 Some myths and misunderstandings around FAIR

In discussions on FAIR data, there can be some confusion on the advantages and disadvantages of FAIR data. Researchers tend to see several risks and disadvantages associated with publishing FAIR data for re-use. Some of the most well-known “myths” and misunderstandings are discussed here, using some often-heard statements.

Statement: “Being FAIR” means that I give up control over my data.”

This is definitely not the case. There are often valid reasons to restrict access to data and there are several ways to do that. Reasons could be that data contains personal information, is competitive or sensitive. In these cases, one can, or even has to restrict access to the data. It is, however, good practice to publish the dataset's metadata. In this way, others can find the data and learn that it exists. The

² Syntactic interoperability defines the way in which data services will be invoked (is also related to schematic interoperability which defines the structure (application schema, data model) in which the data will be offered by a service e.g., GML, JSON etc.

³ Semantic interoperability ensures that the content of the schema (the data itself) can be understood by humans or machines.

restrictions can indicate which (part of) the data may be obtained and who to contact for further information.

Statement: “Others might “misuse” my data”.

First, it is good to realise that misuse of data in most cases is not intentional but caused by others not fully understanding the data and its background and the consequences of using it for a specific purpose. This risk can be decreased by good data documentation, as described in the previous sections. Make sure that it is clear to others why and how the data was derived, which choices were made, and how it should (or should not) be re-used.

Another good way to ensure that others use your data only as intended is to attach a suitable license. A license explains the conditions under which data can be re-used. Many standard licenses (e.g. [Creative Commons](#) or [Open Data Commons](#)) offer good options to ensure that you are cited, protect your IPR (Intellectual Property Rights) and restrict re-use (e.g. for commercial purposes).

Statement: “I don’t benefit from data sharing.”

This is an often-heard misunderstanding about (FAIR) data sharing. As already before, it is possible to link a license to a dataset that requires that others re-using a dataset provide credit the originator of the data (e.g., through a citation or appropriate references). Besides, more and more data journals are established that allow data sets to be published based on a scientific (peer) review process, with a DOI, allowing others to properly cite the dataset.

3.5 Useful links on FAIR data sharing

In the next sections, guidelines to explain some of the relevant steps of FAIR data sharing are given. Where possible, relevant resources, case studies and examples from the pig research domain will be used. For generic information related to the topics in this chapter, the following links may be consulted:

- FAIR data on Wikipedia ([here](#))
- FAIR resources ([here](#))
- FAIR self-assessment ([here](#))
- OpenAIRE guide to FAIR for researchers ([here](#))
- Open Data ([here](#))
- European Open Science Cloud – EOSC portal ([here](#))

4 Metadata and data standardization

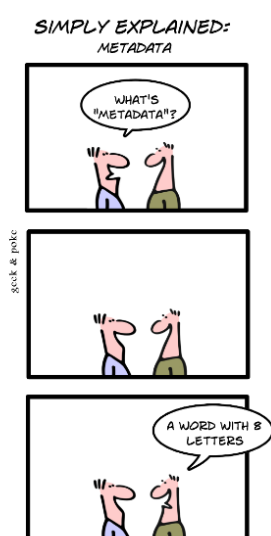
In the pig research community, a large variety of pig phenotype data has been collected, but there is a lack of standards. Systematic documentation of research data is key for making data publishable, discoverable, citable, and re-usable. By doing so you comply with the FAIR data principles (see chapter 3). This will not only be of great benefit to your peers but also to yourself. It will make your research more efficient. Think of the ease to find and re-use data, minimizing the risk of errors, improving quality and so on. Publishing your metadata (including or excluding the research data) will improve the visibility of your work and acknowledges the agencies funding your research. Finally, it contributes to responsible and transparent animal experimentation as the pig community is informed about past and on-going research and, if possible, peers can re-use and build on previous work.

In this section, we first briefly define metadata and introduce different levels of metadata: 1) common aspects of the dataset and 2) a more detailed description of the data in a standardized manner (for more information see [here](#)).

4.1 Definition of metadata

There are several definitions of metadata: “data about data” (e.g. [Wikipedia](#)), “a description of the data” (e.g. [atlan.com](#)) and “information on a thing”. The Digital Curation Centre ([DCC](#)) defines metadata as a subset of documentation information that uses standardized terms and is presented in a structured way.

Let us think of a simple example describing the metadata of a book and a video. Metadata of both items includes common items such as the author, title, and date of publication. Metadata also has items specific to the data type. For instance, a book has a number of pages, while a video has a certain duration. The following video link gives a good introduction to the concept of metadata: [Metadata MOOC 1-1: Introduction](#).



Source: Geek and Poke (17/4/2010)

4.2 Common metadata

To correctly use and reference a dataset, different types of common metadata should be documented and provided: “Descriptive”, “Technical”, and “Access and Rights”. Usually, data are collected and processed in the context of a study or project. So, the common metadata will give context to the study or project, including references to the funding agencies. This is similar to publishing a study in a peer-reviewed journal.

4.2.1 Introduction to Dublin Core (DC)

We advise publishing a dataset (metadata and, if possible, the data itself) in a trusted repository that supports a recognized/common metadata scheme, such as [Zenodo](#), which is compliant with the DataCite metadata schema ([Zenodo](#)) (see chapter 7). Usually, metadata schemas, like [DataCite](#) are based on the Dublin Core Metadata Initiative ([DCMI](#)).

While sharing data might not be possible, at least you can publish the common metadata of the study or project. By doing so, peers learn about your work, can reference it, or even start to collaborate with you. Often data search engines are used for this (e.g. [Dataset Search](#), [B2FIND](#), [Zenodo](#); see chapter 7).

The original [Dublin Core elements](#) contain 15 simple core “elements” created in 1995. Different properties in the form of “terms”, “classes” and “vocabulary” were added since. Also, other elements were added to address the metadata types “Provenance” and “Preservation” in addition to the types “Descriptive”, “Technical” and “Access and Rights”. The Dublin Core does not require all elements to be filled in. However, we highly recommend filling in as many elements as possible to provide a basic coverage.

The Dublin Core (DC) provides more terms and properties that could be used (e.g., media type, specific terms about time periods) but here we restrict them to the most common ones. When annotating a metadata file, check all the elements at DCMI [namespace/elements/1.1/](#). A simple way to capture your common metadata is using the Dublin Core Metadata Generator ([here](#)) or the following link: [metadataetc.org](#).

4.2.2 Highly recommended elements of the Dublin Core

For a basic coverage of your data, the following list of the simple Dublin Core elements is recommended (see Table 4-1 for a detailed definition and example):

- Creator (Who)
- Contributor (Who)
- Title (What)
- Description (What)
- Date (When)
- Coverage (Where)
- Rights (Access)

Table 4-1 Overview of highly recommended Dublin Core elements with example entries (which do not reflect a real dataset)

Element	Example	Remark ⁴
Creator	FBN Dummerstorf	A creator can be an institution or a real person (name, Orcid id). <i>"An entity primarily responsible for making the resource."</i>
Contributor	FBN Dummerstorf	It could be the same as for the Creator but also differ. <i>"An entity responsible for making contributions to the resource."</i>
Title	Pig weights example	Title for the whole data set <i>"A name given to the resource."</i>
Description	Data includes four weight measurements	Description of the data <i>"An account of the resource."</i>
Date	2005-05-01	Date of last modification ⁵ <i>"A point or period of time associated with an event in the lifecycle of the resource."</i>
Coverage	54.005546, 12.232895 1998-12-09 till 2004-11-11	Coverage can be used for multiple things. They are here considering, e.g. the location in the form of coordinates ⁶ and the time frame where the data is collected. <i>"The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant."</i>
Rights⁷	CC-BY4.0 (link)	Widespread licences are the Creative Commons (About CC Licenses - Creative Commons) or the Open Data Commons (Home — Open Data Commons: legal tools for open data). Instead of just writing the license, a link could also be provided. For more information see chapter 7. <i>"Information about rights held in and over the resource."</i>

4.2.3 Recommended elements of DC

After describing the essential data details, it is advisable to include further explanations. To ensure a comprehensive metadata description, it is recommended to complete the following elements (Table 4-2).

Table 4-2 Overview of recommended Dublin Core elements with example entries (which do not reflect a real dataset)

Element	Example	Remark ⁴
Subject	Piglets	What subject is in the dataset? It could be given the animal or even more specific information. <i>"The topic of the resource."</i>
Type	Dataset	For what kind of data is the provided metadata file? Multiple possible terms ⁸ :

⁴ Remarks in italic and between quotes are retrieved from the original definition at: Caverlee, J., Mitra, P., Laarsgard, M. (2009). Dublin Core. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_894

⁵ Format: ISO8601 (ISO 8601 - Wikipedia)

⁶ This would also require the used co-ordinate system in this case WGS84 having latitude and longitude decimal degrees (e.g. alternatively it could also be defined in degrees, minutes and seconds)

⁷ The rights element of DC covers this setup while it can be extended by more elements (e.g. referenced by, license, references).

⁸ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-7>

Element	Example	Remark ⁴
		<ul style="list-style-type: none"> - Collection - Dataset - Event - Image - InteractiveResource - MovingImage - PhysicalObject - Service - Software - Sound - StillImage - Text <i>"The nature or genre of the resource."</i>
Format	Txt	What format does the data have? Could be different types (e.g., Common MIME types ⁹) <i>"The file format, physical medium, or dimensions of the resource."</i>
Language	Ger	Provides in which language the data and metadata are presented. Potentially is recommended to use the ISO 639 Standard ¹⁰ <i>"A language of the resource."</i>
Relation	https://doi.org/10.1038/sdata.2016.18	Recommended practice is to identify the related resource by means of a URI. If this is not possible or feasible, a string conforming to a formal identification system may be provided. <i>"A related resource."</i>
Identifier		Demonstrates, if available, an unambiguous reference to the resource within a given context. Recommended practice is to identify the resource by means of a string conforming to an identification system. <i>"An unambiguous reference to the resource within a given context."</i>
Publisher	FBN Dummerstorf	An entity is responsible for making the resource available. <i>"An entity responsible for making the resource available."</i>
Source		A related resource from which the described resource is derived. The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system. For example, the corresponding farm (e.g. breeding) could be deposited as the source in animal sciences. <i>"A related resource from which the described resource is derived."</i>

4.2.4 Restrictions and limitations of DC

Since Dublin Core is a linear metadata standard, the defined elements should not be used multiple times (e.g., multiple creators in separate creator XML-tags). This is not recommended because by multiple options for the same tag, processing tools or repositories may not be able to work with the multiple occurrences and just take the first element in the first XML tag as creator. Nevertheless, some DC creator tools, like Dublin Core Metadata Generator ([here](#)) provide this option. Be aware that this could lead to problems.

⁹ https://developer.mozilla.org/en-US/docs/Web/HTTP/Basics_of_HTTP/MIME_types/Common_types

¹⁰ https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

Even though it is possible to add own elements (e.g., pig breed, that contain further information), this is not recommended since this can either result in the standard not being recognised as such, or the information may simply be ignored in the mostly automatic further processing.

Dublin Core does not provide a specific field for storing funding information or grant identifiers. Other metadata schemas like datacite or crossref provide specific terms for this purpose. Due to no specific assigned field, multiple fields could be used to add grant numbers in DC (e.g., identifier, relation, included in description). For instance, Zenodo does not provide an extra element for grant numbers. Here, we recommend adding this information within the description or relation.

4.2.5 Example of a dataset published in Zenodo

Figure 4-1 illustrates common metadata items for a dataset named “Pigs feeding behaviours from two different farms, including behaviours during a tail biting event” (see [link](#)).

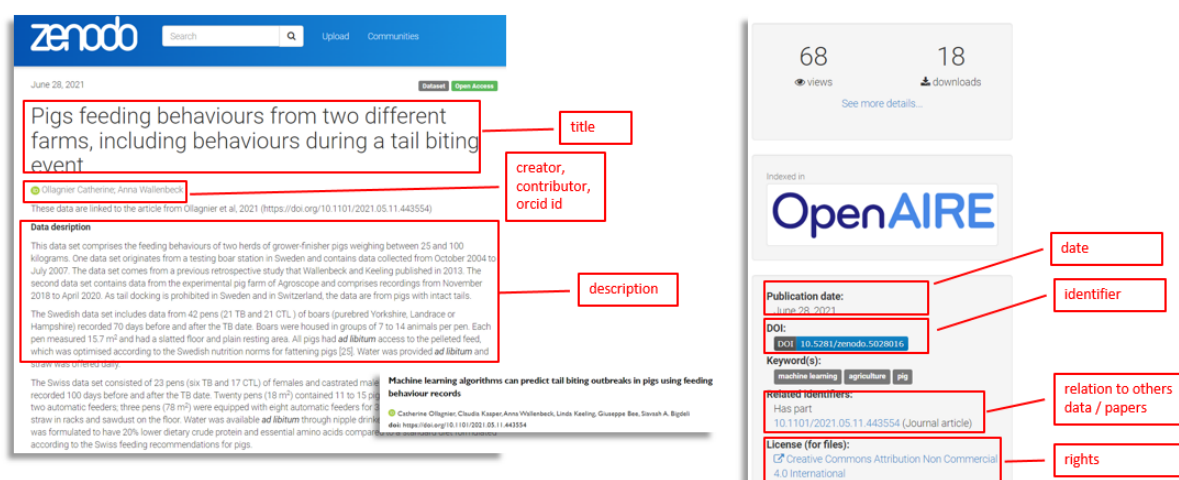


Figure 4-1 Example of common metadata of a dataset published in the Zenodo repository

4.3 Data and metadata standardization

4.3.1 Introduction

Publishing common metadata is a good start, but finding, understanding, and re-using data will still be difficult. Through the common metadata, you learn about the dataset, its context and offer opportunities to re-use (part of) the published dataset. To efficiently find, understand, and re-use the data, either by yourself or peers, data must be completely clear in terms of definition, units, used coding/classification and provenance (i.e., how was the data collected, processed, and updated?). If there is any doubt or room for interpretation, there is a risk of overlooking data when searching for data and the re-use of data may occur in a different or wrong way. See also the following link of CESSDA on data management ([here](#)).

For example, an outsider could question the term “pig pen” (see Figure 4-2). With pig pen somebody could refer to a fictional character in the comic strip Peanuts by Charles M. Schulz.

pig pen?



a fictional character in the comic strip Peanuts by Charles M. Schulz

Figure 4-2 What is a pig pen? An example to emphasize the need explicit description of the object

This example illustrates why it is necessary to use standardized vocabular. In the case of “pig pen” it is encouraged to use the standard Ontology term EOL0001902 from the Environment Ontology of Livestock (EOL) (see Figure 4-3; for more see chapter 5).

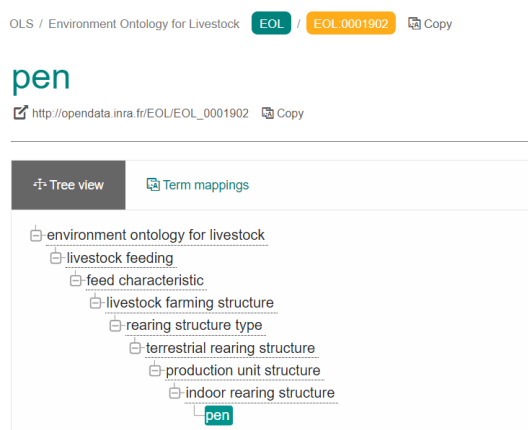


Figure 4-3 Pen definition in Environment Ontology for Livestock (EOL)

4.3.2 Additional standardisation

Further standardized terms are desired to improve the interoperability of the dataset, to ease the dataset search, and to provide a machine-readable documentation of the dataset, which facilitates its re-use. This could be achieved by switching from metadata standards to more complex formal data standards (e.g., ABCD). In contrast to pure metadata files, those standards allow describing the data, its structure, and partly its metadata in a standardized and machine-readable format (e.g., xml). Metadata information could also be included, but the extent highly depends on the chosen data standard and could still require an additional metadata file.

The main advantage of formal data standards (e.g., “Access to Biological Collections Data”; ABCD) is the more detailed description of some metadata or structural information, including:

- Specification of data types (e.g., in the case of several different types, such as observations, interviews, images, questionnaires, in a dataset)
- Size information
- Definitions (e.g., variables, names, indicators)
- Location information (e.g., coordinates)

PIGWEB

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004770

- Experimental setups (e.g., standard operating procedures (SOP))
- Machines/instruments used
- People involved
- Processing information (e.g., workflows or scripts)
- Quality checks performed
- Data annotation (e.g., taxonomic determination, ontology)

However, a balance needs to be found between what is of interest for the community to offer in a formal standardized way, the complexity of the data set, and the required effort to do this formal standardization. So, you would document your data as complete as possible, preferably in a formal standardized way, including:

- data structure: data type (e.g., observations, interviews, images, questionnaires), file type, format, naming convention size
- definitions: variable descriptors and, if possible, additionally use ontology terms (see chapter 5)
- units and classification: explanation of the used units and classification schemes used
- information on data acquisition: instruments (e.g., type, calibration), hardware and software, protocols (SOP), sampling strategies, population, units, data collectors, date of data collection, geographical coverage
- information on data processing and cleaning: describe the processing procedures, the data quality checks and classification of data (e.g., taxonomic classification). This includes multiple versions of the data and the corresponding scripts, which could be managed, for example by git, and point out missing or incorrect values or where data were anonymised or modified

4.3.3 Short insight into the ABCD standard (“Access to Biological Collections Data”)

Within the [biology](#) domain, the Access to Biological Collections Data ([ABCD](#)) Schema¹¹ is one of the available standards. The standard was developed between 2001 and 2006 with the aim to harmonise the exchange of biological collection data using the XML format. In contrast to DC, ABCD provides an enormous number of additional terms. Due to its hierarchical structure (see Figure 4-4), it is highly flexible.

¹¹ Access to Biological Collections Data Task Group. 2005. Access to Biological Collection Data (ABCD). Biodiversity Information Standards (TDWG) <http://www.tdwg.org/standards/115>.

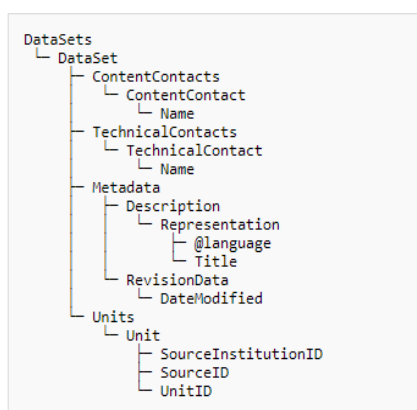


Figure 4-4 A small sub-set of the ABCD schema, which is required to be assigned for a valid ABCD file.

ABCD's origin in the biodiversity community lead to the specification of, for example, botanical terms included in this standard. Such specific expressions made the ABCD standard being very extensive. However, since most of the terms are specific to the field of application, they do not arise for use in pig research and can be omitted. The remaining terms are mostly kept generic so that they can be adapted for different fields of application.

To simplify the use of ABCD, the BioCase (Biological Collection Access Service for Europe) Provider Software exists, even though its application for pig research is currently under review. Additionally, ABCD is compatible with other existing standards, such as DC or DarwinCore, and is used as interface for the GBIF repository (see chapter 7). Besides the BioCase software, we are also testing the usability of ABCD, and the resulting recommendations are work in progress. To give you a first overview, the example structure is shown in Figure 4-4 while in Figure 4-5 for this structure an ABCD-xml file is illustrated.

```

<?xml version='1.0' encoding='UTF-8'?>
<DataSets xmlns='http://www.tdwg.org/schemas/abcd/2.06'>
  <DataSet>
    <ContentContacts>
      <ContentContact>
        <Name>John Doe</Name>
        <Email>doe@test.de</Email>
      </ContentContact>
    </ContentContacts>
    <TechnicalContacts>
      <TechnicalContact>
        <Name>Jane Austen</Name>
        <Email>austen@example.org</Email>
      </TechnicalContact>
    </TechnicalContacts>
    <Metadata>
      <Description>
        <Representation language='eng'>
          <Title>Pig weights collection</Title>
        </Representation>
      </Description>
      <RevisionData>
        <DateModified>2005-05-01T00:00:00</DateModified>
      </RevisionData>
    </Metadata>
    <Units>
      <Unit>
        <SourceInstitutionID>FBN</SourceInstitutionID>
        <SourceID>Sus scrofa</SourceID>
        <UnitID>0123</UnitID>
      </Unit>
    </Units>
  </DataSet>
</DataSets>
  
```

Figure 4-5 An example of a minimal ABCD xml file, filled with random examples.

An insight of the complexity of the standard is given in Table A1 of Annex I, where an example for some pig research data relevant ABCD terms is listed.

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

We recommend using the ratified version 2.06, by the Biodiversity Information Standards (TDWG), instead of the newer version 3.0. Also, feel free to contact the authors of the Deliverable about occurring problems so that we could collect them and consider their solution in preparation for ABCD guidelines for pig research.

4.4 Useful links on meta-/data and standardization

- Introduction to metadata ([here](#))
- Dublin Core ([here](#) and [here](#))
- Dublin Core Metadata Generator ([here](#))
- Metadataetc.org ([here](#))
- Example in Zenodo ([here](#))
- Data management - CESSDA ([here](#))
- Access to Biological Collection Data (ABCD) ([here](#))
- BioCASE Provider Software to prepare ABCD metadata ([here](#))

5 Ontologies

5.1 Introduction

Describing the set of phenotypic characters or traits of phenotypes of interest in a homogeneous and, if possible, unambiguous way is one of the challenges of life sciences. This goal requires that the phenotypic characters are accurately defined, standardized, measured, and referenced (Hocquette *et al.* 2012). Among the standardization tools at our disposal, ontologies appear relevant because they permit to integrate heterogeneous data from different sources.

What is an ontology? Bard and Rhee (2004) define ontologies as a “formal way to represent knowledge in describing the concepts both by their meaning and the relationships between them”. In practice, ontologies consist mainly of classes (or “concepts” or “types”), relations (or properties), and sometimes rules of reasoning. Classes and properties are used to describe the data via their ID.

An ontology is a formal, explicit description of concepts that address a defined field of organized information. It makes concepts readable for machines by describing both the concept (or “classes”) meaning and their relationship (or “properties”) to each other.

Knowledge basis is built on standardized and harmonized concepts. There is a hierarchical structure with main branches and subbranches (or parent and child traits) according to the research area (e.g., nutrition, welfare) chosen to describe the knowledge. The language is shared between partners within a project. An ontology is essential to answer FAIR principles. See Figure 5-1 and Figure 5-2 for some examples.

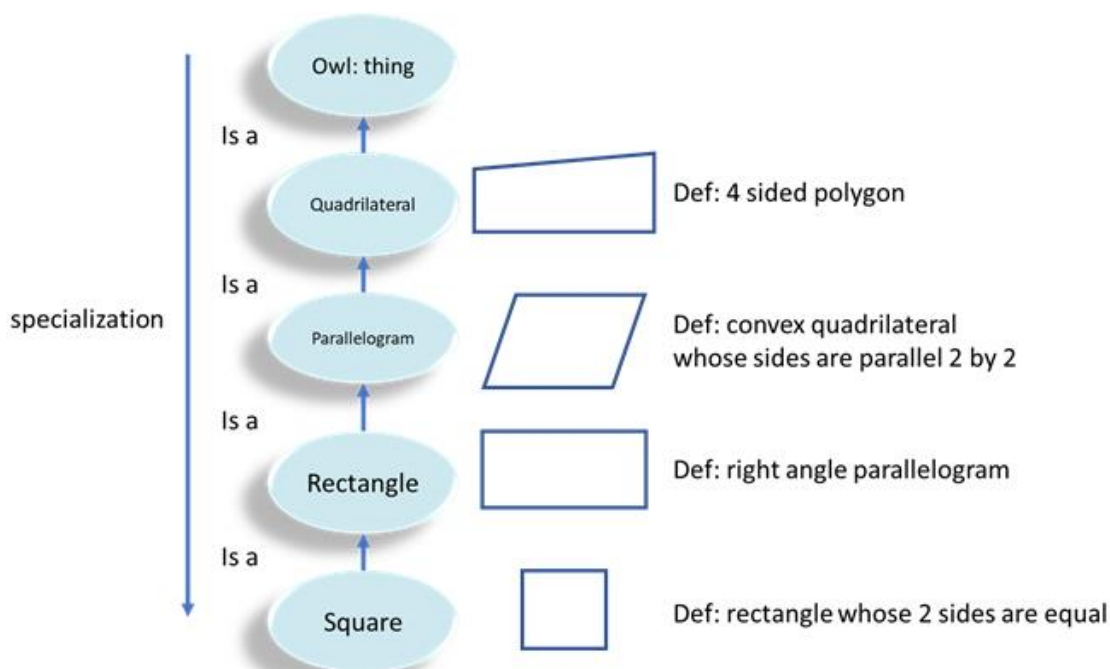


Figure 5-1 An example of ontologies and its hierarchical structure: a “square” belonging to a rectangle, which belongs to a parallelogram etc.

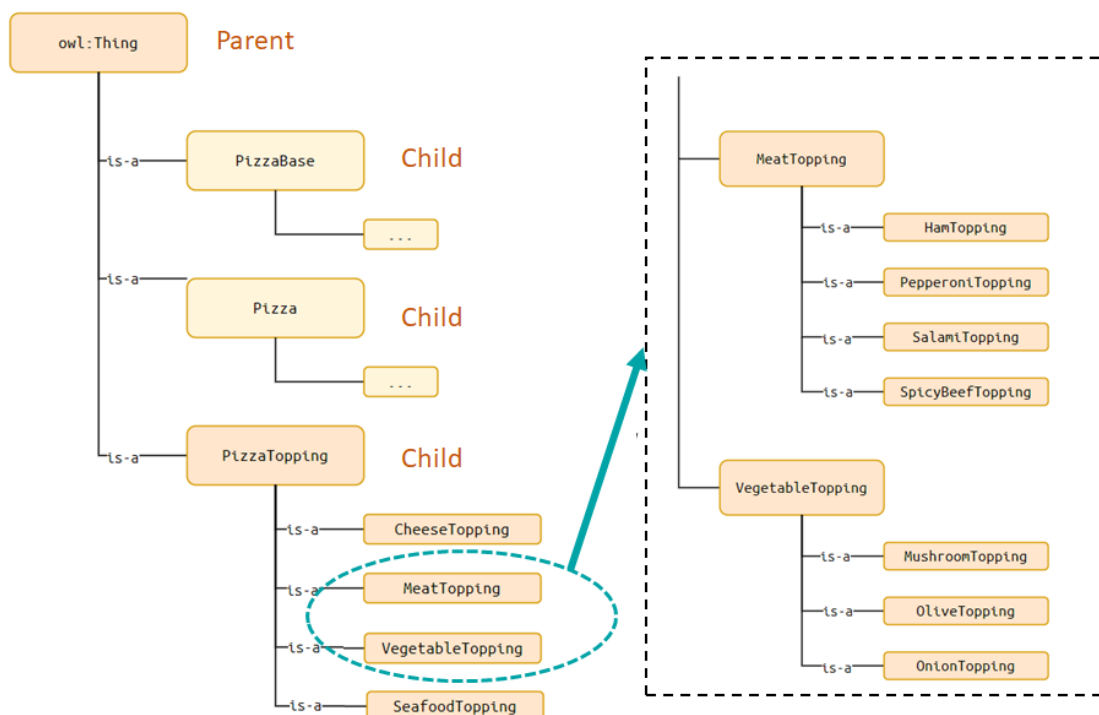


Figure 5-2 An example of ontologies and its hierarchical structure: pizza elements and their children such as different toppings

5.2 ATOL ontology

The Animal Trait Ontology for Livestock (ATOL) proposes a common language between zootechnicians, physiologists and geneticists, facilitates collaborative projects between disciplines and/or animal models, and facilitates information exchange by using referenced traits in publications and databases.

ATOL aims to implement a multi-species ontology shared by the international scientific, teaching, and technical animal science community for experimental data annotation while having a language usable by software (e.g., database management, semantic analysis, modelling). ATOL experts are mainly from INRAE, but also from some European organisations (resulting from the AquaExcel and SmartCow projects).

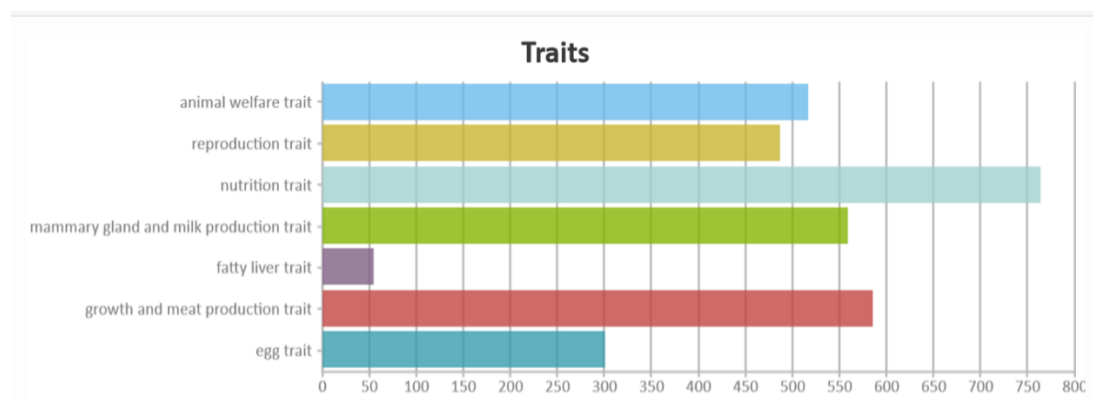


Figure 5-3 Traits in the main branches of ATOL

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

The ATOL's main branches concern: animal welfare, reproduction, nutrition, mammary gland and milk production, fatty liver, growth and meat production and egg. Each branch contains between 50 and 750 traits (see Figure 5-3).

Each trait has multiple attributes: general information concerning "Identity", "Name", "Definition", "Source", synonyms as "Exact synonyms" and "Related synonyms", measurement methods and species. The relationship between the different traits is termed "is a" (see Figure 5-4).

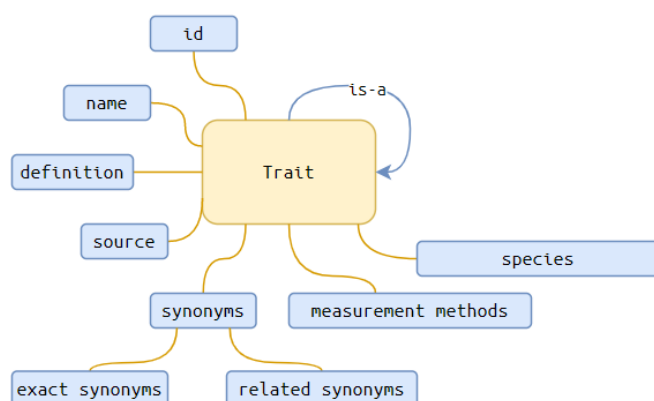


Figure 5-4 Attributes of a trait in ATOL

In addition to the hierarchy, the ATOL website allows consulting a range of information related to the concept, its origin, to facilitate its use (see Figure 5-5). These are:

- i) an identifier ATOL, supplemented by the reference of the initial identifier in VT (VTO is an ontology by J. Reecy and C. Parks from Iowa State University), if necessary (e.g., for "investigative behaviour trait", ATOL:0000844). The source of the concept (INRAE or another ontology such as "Iowa State University Curator") is associated to the identifier.
- ii) a name (here: "investigative behaviour", "investigation"), which corresponds to the most frequent use and any synonyms which may be exact or close according to the degree of functional or semantic similarity.
- iii) a definition whose form follows a standardized framework (for example: "any measurable or observable characteristic related to the behaviour devoted to investigate the environment (physical or social), expressed by motor activities such as sniffing, pecking, scratching, licking, biting, looking at").
- iv) a validation of the suitability of the trait for different species (e.g., "present" for all the mammals).
- v) if available, links to sites providing information on the phenotypic trait (e.g., publications, candidate genes, RNA, databases).
- vi) if available, known phenotypes associated with the ATOL character.
- vii) if available, the methods to measure this trait, with links to databases on the procedures of measurement.

The screenshot displays the INRAE ATOL ontology interface. On the left, a tree view shows the hierarchy: **psychoneurophysiological state trait** > **behaviour trait** > **motor activity trait** > **investigative behaviour**. A blue arrow points to 'investigative behaviour'. The main panel shows the details for 'investigative behaviour trait' (ATOL_0000844). The 'Informations' tab is active, showing the following details:

Name	investigative behaviour trait
Nom	caractère de comportement d'investigation
Definition (en)	any measurable or observable characteristic related to the behaviour devoted to investigate the environment (physical or social), expressed by motor activities such as sniffing, pecking, scratching, licking, biting, looking at
Definition (fr)	toute caractéristique mesurable ou observable associée au comportement dédié à l'investigation de l'environnement (physique ou social), exprimé par des activités motrices comme le reniflage, le picage, le grattage, le léchage, le mordillement, le regard
Source	INRAE
Link	No results
Comments	No results

Handwritten notes in blue ink include: 'Species specificity' with an arrow pointing to the 'Definition' field, and 'video, reference article methods' with an arrow pointing to the 'Link' field. Below the table, it says 'Ontology, Book of ref adapted from X,...'. The top navigation bar shows 'Ontologies - ATOL - EOL - AHOL' and 'Releases'.

The second screenshot shows the same trait with different tabs: 'Synonyms', 'Exact/related', and 'Species'. The 'Synonyms' tab is active, showing:

- Exact synonyms:** Investigative behavior
- Related synonyms:** exploratory behaviour

The 'Species' tab is also active, showing a list of species: Birds, Fish, Mammals, Cattle present, Goat present, Horse present, Mouse present, Pig present, Rabbit present, Sheep present. The 'Measurement methods' tab shows 'No results'.

Figure 5-5 Attributes of a trait in ATOL







There may be links with other ontologies such as:

- **EOL: Environment Ontology for Livestock:** An ontology that describes elements related to the livestock environment.

- environment ontology for livestock
 - livestock farming environment
 - livestock farming structure
 - livestock farming system
 - livestock feeding

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

- **AHOL: Animal Health Ontology for Livestock:** an ontology that describes **health issues** such as diseases, symptoms, and involved pathogens. Work on AHOL is in progress, but health traits are findable in ATOL
 - ▾  disease
 -  chronic disease
 -  communicable disease
 -  genetic disease
 -  infectious disease
 -  non communicable disease
- **CHEBI: Chemical Entities of Biological Interest** (*Developed and maintained by EMBL-EBI*): A structured **classification of chemical compounds** of biological relevance.
- **VTO: Vertebrate Trait Ontology** (*Developed and maintained by Iowa State University*): Controlled vocabulary for the description of traits (measurable or observable characteristics) pertaining to the **morphology, physiology, or development of vertebrate organisms**.
- **LPT: Livestock Product Trait Ontology** (*Developed and maintained by the Iowa State University*): Controlled vocabulary for the description of traits (measurable or observable characteristics) pertaining to **products produced by or obtained from the body of an agricultural animal or bird maintained for use and profit**.

5.3 Examples of annotation of data in publications

Example 1: Hurtaud *et al.*, 2023 in Animal Open Space:

Table 5

Plasma metabolites and hormone concentrations based on non-restricted (NON RESTR) and restricted (RESTR) feeding treatments for dairy cows.

Characteristic	Ontology ¹ id	Feeding		SEM	P-value	
		NON RESTR	RESTR		Group	Feeding
Acetate, mmol/L	/	1.16	1.04	0.044	0.138	0.077
Non-esterified fatty acids, μmol/L	VT:0001553	147	449	34.1	0.610	<0.001
β-hydroxybutyrate, μmol/L	VT:0010996	609	647	34.0	0.066	0.227
Glucose, mg/L	ATOL_0000097	734	705	5.4	0.220	<0.001
Lactose, mg/L	/	51.2	44.9	2.49	0.408	0.066
Triglycerides, mg/L	VT:0002644	71.2	92.7	2.98	0.895	<0.001
Cholesterol, mg/L	VT:0000180	1 724.8	1 866.5	57.06	0.751	0.001
Urea, mg/L	VT:0005265	198	224	6.5	0.347	<0.001
Insulin, μU/mL	VT:0001560	8.045	6.719	0.497	0.103	0.042
Insulin-like growth factor 1, ng/mL	ATOL_0000990	180.49	158.25	4.966	0.054	<0.001
Prolactin, ng/mL	ATOL_0001699	12.58	9.38	6.27	0.046	<0.001

¹ Traits in reference to ontologies: ATOL (Animal Trait Ontology for Livestock, <https://www.atol-ontology.com/en/erter-2/>) and VT (Vertebrate Trait ontology, <https://biportal.bioontology.org/ontologies/VT/?p=summary>).

Example 2: supplementary table using ontology:

Description focused on the specific species of the study

generic concept

Supplementary Table S1 Posture and behavioural activities items of gestating sows recorded in continuous and scan sampling

Traits ¹	ATOL ref ²	Description of the measurement in the experimental study
Continuous sampling ¹		
Agonistic behaviour	ATOL_0000902	Sow initiating aggressive acts such as threat, push, bites, head knock (initiator) , or sow exhibiting subordinate response as avoidance, flight in response to aggressor (receiver)
Physical investigation	ATOL_0000845	Sow exhibiting investigation expressed by motor activities such as sniffing, licking, scratching, rooting acts, performed on the physical substrate of the pen (wall, floor) or available substrate (piece of wood, chain, straw)
Scan sampling ²		
Posture	ATOL_0000370	position of the limbs or carriage of the body
Lying	ATOL_0000837	Sow with body in contact with the ground or substrate, maintaining a recumbent or ventral position
Standing	ATOL_0000835	Sow in upright position on extended legs
Sitting	ATOL_0000836	Sow with posterior of the body trunk in contact with the ground and supports most of the body weight
Behavioural activity		
Lying in feeding stall	ATOL_0000816	Sow with limited or no movement of the body in lying posture, within the feeding stall
	ATOL_0000837	
Physical investigation	ATOL_0000845	Sow exhibiting investigation expressed by motor activities such as sniffing, licking, scratching, rooting acts, performed on the physical substrate of the pen (wall, floor) or available substrate (piece of wood, chain, straw)
Social interactions	ATOL_0000899	Sow exhibiting acts directed towards and influenced by conspecifics of the social group, perceived by the receiver as negative (aggression or flight) or positive (approach, non aggressive reaction)
Other		any behavioural activity other than the activities traits described above

¹ Behavioural activity was recorded in continuous way during 2 hours, with identification of the initiator and the receiver within the agonistic interactions, and the substrate investigated by sows, pen (walls and floor), chain, piece of wood or straw

² each pig or the number of pigs was scored for category of traits: posture and behavioural activity. Within each category, traits were mutually exclusive.

³ traits adapted from reference to the ontology ATOL: <http://www.atol-ontology.com/index.php/en/>

Resting
Immobility + lying

Example 3: Reference index for publication in scientific journal:



animal

Divergent selection for residual feed intake in group-housed growing pigs: characteristics of physical and behavioural activity according to line and sex

M. C. Meunier-Salaün^{1,2,†}, C. Guérin^{1,2}, Y. Billon³, P. Sellier⁴, J. Noblet^{1,2} and H. Gilbert^{4,5}

¹INRA, UMRI1348 PEGASE, F-35590 Saint-Gilles, France; ²Agrocampus, UMRI1348 PEGASE, F-35590 Saint-Gilles, France; ³INRA, UE1372 GenESI, F-17700 Surgères, France; ⁴INRA, UMRI1313 GABI, F-78350 Jouy-en-Josas, France; ⁵INRA, UMRI1388 GenPhySE, F-31326 Castanet-Tolosan, France

(Received 10 October 2013; Accepted 10 June 2014)

Animal, page 1 of 9 © The Animal Consortium 2014
doi:10.1017/S1751731114001839

Supplementary Table S1 Posture and behavioural activities items recorded in the video analysis¹.

Traits ²	ATOL ref ³	Description of the measurement
Posture		
Lying	ATOL:0000837	body in contact with the ground or substrate, maintaining a lateral or ventral position
Standing	ATOL:0000835	upright position on extended legs
Sitting	ATOL:0000836	posterior of the body trunk in contact with the ground and supports most of the body weight
Behavioural activity		
Immobility		body with lack of motor activity in standing or sitting, except eye, ears or small head movements
Resting		body with lack motor activity in lying down, except eye, ears or small head movements
Feeding	ATOL:0000363	presence at the feeder with access door to feed opened
Drinking	ATOL:0000361	presence at the drinker with head above or within the water bowl
Investigation	ATOL:0000844	investigation expressed by motor activities such as sniffing, licking, scratching, rooting acts, performed on physical or social substrate
Pen	ATOL:0000845	investigation of the pen
Social	ATOL:0000846	investigation of a conspecific
Novel objet ⁴		investigation of the unfamiliar objects introduced during the motivation test
Agonistic ⁵	ATOL:0000902	social interaction encompassing the actions of both the instigator and the receptor, in the ESF area or in the other part of the pen
Mobility		standing and moving from place to place, head up
Other		any behavioural activity other than the activities traits described above

¹ Video recordings were carried out at 17 wks (24h-analysis) and 18 wks of age (motivation test, 9 hours). A 5-min instantaneous scan sampling technique was utilized.

² each pig was scored for category of traits: posture and behavioural activity. Within each category traits were mutually exclusive.

³ traits adapted from reference to the ontology ATOL: <http://www.atol-ontology.com/index.php/en/>

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

5.4 Annotation of data: how to proceed?

For example, a data table to be annotated using 3 ontologies.

Item	signification
pdsabt	weight of the pig at the end of the experiment (in kg)
gmqeng	growth rate between the beginning and the end of the food experiment (in grams/day)
consoj	average daily feed consumption (in grams)
IC	feed efficiency (daily feed intake/growth rate)
pds foie	liver weight (in grams)
pds reins	kidney weight (in grams)
pds pannes	weight of perirenal (or perivisceral) adipose tissue (in grams)
pds carcch	carcass weight immediately after slaughter (hot) (in kg)
moyELD	thickness of dorsal subcutaneous adipose tissue at the G2 anatomic site (in mm)
tractus vide	weight of the empty digestive tract (in grams)
pds digesta	weight of digestive contents at slaughter (in grams)
carcas froide	weight of the carcass after 36 hours of soaking at 4°C (cold) (in kg)
pds pannes froides	weight of perirenal (or perivisceral) adipose tissue after 36 hours of soaking at 4°C (in grams)
pds jambon	weight of the ham after 36 hours of cooling at 4°C (in grams)
pds des os du jambon	Weight of bones of the HAM (g)

The ontologies:

ATOL: <https://www.atol-ontology.com/en/erter-2/>

VT on bioportal: <https://bioportal.bioontology.org/ontologies/VT/?p=classes&conceptid=root>

LPT on bioportal: <https://bioportal.bioontology.org/ontologies/LPT/?p=classes&conceptid=root>

For example, **pdsabt** as “weight of the pig at the end of the experiment (in kg)”. In ATOL, the trait “body weight” that matches the trait that was measured.

The screenshot shows the ATOL ontology interface. On the left, a search bar contains the text 'weight'. Below it, a tree view shows the hierarchy of traits, with 'body weight - ATOL_0000351' selected. On the right, the details for this trait are displayed. The trait is named 'body weight' and 'masse corporelle'. Its definition in English is 'any measurable characteristic related to the body weight of an organism'. The source is 'VTO:CP "Cari Park, Iowa State University Curator"'. There are no results for the link or comments.

For “**pds pannes froides**” as the “weight of perirenal adipose tissue...”, the trait is not available in ATOL, but is found in VT.

PIGWEB

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004770

Jump to:

- vertebrate trait
 - organ system trait
 - alimentary system trait
 - circulatory system trait
 - connective tissue trait
 - connective tissue development trait
 - connective tissue morphology trait
 - adipose morphology trait
 - adipocyte morphology trait
 - adipose amount
 - adipose distribution trait
 - adipose molecular composition trait
 - brown adipose morphology trait
 - fat pad morphology trait
 - abdominal fat pad morphology trait
 - epididymal fat pad morphology trait
 - femoral fat pad morphology trait
 - gonadal fat pad morphology trait
 - inguinal fat pad morphology trait
 - interscapular fat pad morphology trait
 - mammary fat pad morphology trait
 - mesenteric fat pad morphology trait
 - parametrial fat pad morphology trait
 - pericardial fat pad morphology trait
 - renal fat pad morphology trait
 - renal fat pad mass**
 - retroperitoneal fat pad morphology trait
 - subscapular fat pad morphology trait
 - total fat pad mass
 - uterine fat pad morphology trait
 - white adipose morphology trait

Details Visualization Notes (0) Class Mappings (4)

| | |
|-------------------|---|
| Preferred Name | renal fat pad mass |
| Synonyms | |
| Definitions | The amount of matter in the encapsulated adipose tissue associated with the kidney. |
| ID | http://purl.obolibrary.org/obo/VT_0010429 |
| created_by | caripark |
| creation_date | 2012-10-10T11:59:52Z |
| definition | The amount of matter in the encapsulated adipose tissue associated with the kidney. |
| has_obo_namespace | Trait.ontology |
| id | VT:0010429 |
| label | renal fat pad mass |
| notation | VT:0010429 |
| prefLabel | renal fat pad mass |
| treeView | renal fat pad morphology trait |
| subClassOf | renal fat pad morphology trait |

The same holds for “pds des os du jambon” as “the weight of the bones of the ham”, which cannot be found in in ATOL or in VT, but is available in LPT.

Jump to:

- livestock product trait
 - colostrum trait
 - dressed carcass trait
 - dressed carcass composition trait
 - dressed carcass size trait
 - dressed carcass subdivision trait
 - belly trait
 - breast trait (poultry)
 - drumstick trait
 - ham trait
 - ham composition trait
 - ham bone weight**
 - ham bone-to-muscle ratio
 - ham fat percentage
 - ham fat thickness
 - ham fat weight
 - ham muscle percentage
 - ham muscle weight
 - ham size trait
 - jowl weight
 - leg trait
 - loin trait
 - neck weight
 - rib cut trait
 - shoulder trait
 - thigh trait (poultry)
 - wing trait
 - egg trait
 - meat trait
 - milk trait
 - wool trait

Details Visualization Notes (0) Class Mappings (0)

| | |
|-------------------|---|
| Preferred Name | ham bone weight |
| Synonyms | |
| Definitions | Any measurable or observable characteristic related to the relative heaviness of the bone in the ham. |
| ID | http://purl.obolibrary.org/obo/LPT_1000719 |
| created_by | caripark |
| creation_date | 2009-05-06T09:06:13Z |
| definition | Any measurable or observable characteristic related to the relative heaviness of the bone in the ham. |
| has_obo_namespace | Product.ontology |
| id | LPT:1000719 |
| in_subset | http://purl.obolibrary.org/obo/TEMP#slim_pig |
| label | ham bone weight |
| notation | LPT:1000719 |
| prefLabel | ham bone weight |
| treeView | ham composition trait |
| subClassOf | ham composition trait |

It results in the following annotated table:

| Item | ATOL id | VT id | LPT id | Signification |
|-----------|------------------------------|-------|--------|---|
| pdsabt | ATOL_0000351 | | | weight of the pig at the end of the experiment (in kg) |
| gmqeng | ATOL_0002175 | | | growth rate between the beginning and the end of the food experiment (in grams/day) |
| consoj | ATOL_0005508 | | | average daily feed consumption (in grams) |
| IC | ATOL_0001580 | | | feed efficiency (daily feed intake/growth rate) |
| pds foie | ATOL_0000459 | | | liver weight (in grams) |
| pds reins | ATOL_0005578 | | | kidney weight (in grams) |

PIGWEB

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004770

| | | | | |
|----------------------|------------------------------|----------------------------|-----------------------------|--|
| pds pannes | | VT:0010429 | | weight of perirenal (or perivisceral) adipose tissue (in grams) |
| pds carcch | ATOL_0001057 | | | carcass weight immediately after slaughter (hot) (in kg) |
| moyELD | ATOL_0001517 | | | thickness of dorsal subcutaneous adipose tissue at the G2 anatomic site (in mm) |
| tractus vide | ATOL_0005122 | | | weight of the empty digestive tract (in grams) |
| pds digesta | ATOL_0002256 | | | weight of digestive contents at slaughter (in grams) |
| carcas froide | ATOL_0001057 | | | weight of the carcass after 36 hours of soaking at 4°C (cold) (in kg) |
| pds pannes froides | ATOL_0000552 | | | weight of perirenal (or perivisceral) adipose tissue after 36 hours of soaking at 4°C (in grams) |
| pds jambon | VT:0010449 | | LPT:1000563 | weight of the ham after 36 hours of cooling at 4°C (in grams) |
| pds des os du jambon | | | LPT:1000719 | Weight of bones of the HAM (g) |

5.5 Useful links on ontologies

- ATOL ontologies (normal link [here](#), actual link [here](#))
- EOL ontologies (link [here](#))
- CHEBI ontology (link [here](#))
- VT ontology (link [here](#))
- LPT ontology (link [here](#))
- Excel plugin RightField for linking to ontology ([here](#))

6 Data curation

6.1 Introduction

Recorded data must be stored as raw data together with its structure and metadata file. For good scientific practice, further processing has to be realized without changing the raw data (reproducible). Therefore, the whole data lifecycle works on an image of the raw data.

Good data preparation, such as structuring, formatting, checking, and correcting data takes time but it is well worth the effort as it greatly increases the efficiency and proper use in later stages of a project. Research organizations, especially the larger ones, may have specialized staff (e.g., data stewards) and an infrastructure to organize and manage data during the different stages of the data life cycle (Figure 6-1).

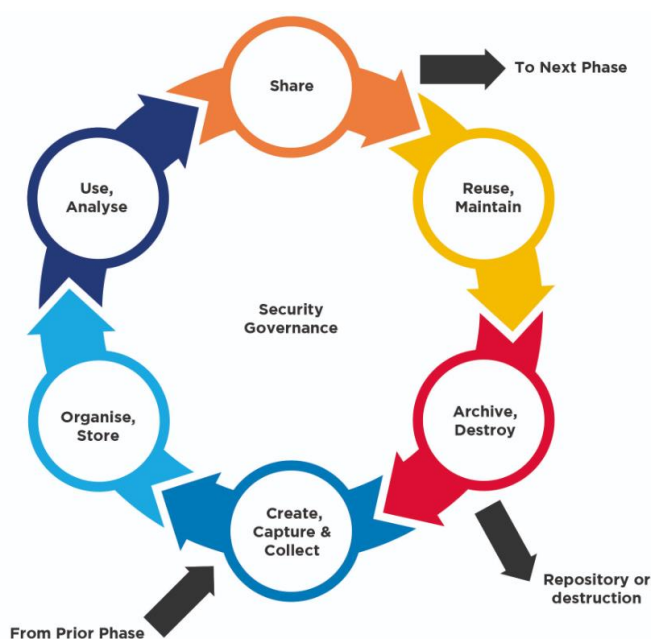


Figure 6-1 Data Life Cycle (source: Infrastructure Data Management Framework (IDMF), Data.NSW New South Wales, Australia)

6.2 Organisation and storing data

6.2.1 Folder and files

The folder structure depends on the plan and organisation of the study. Think of the hierarchy and decide whether a deep or shallow hierarchy is preferable. In the case of independent data collections, it is better to create a separate folder for each collection. Regarding folder names, it is advised to use logical, short keywords, underscores, or dashes/hyphens without spaces, dots, or special characters.

Data are preferably stored in open data formats (e.g., csv), avoiding proprietary formats (e.g., Excel). File names must be logical and usually include items such as date (yyyy-mm-dd), version, content-related key words, project number and author/creator. As with folder names, use underscores or dashes/hyphens and do not use spaces, dots, or special characters.

The naming convention and folder structure should be explained in a readme or codebook.

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

6.2.2 Relational databases

Data can also be stored in a relational database (open solutions like [MySQL](#), [SQLite](#) etc.). It is important to define a logical data model¹² that avoids data redundancy and build a database following certain specifications (e.g., character, integer, double) and constraints (e.g., plausible data ranges, checks on unique entries) to enable data consistency and contribute to higher data quality.

The naming of tables and columns should be logical, clear, avoiding too long names and duplications. Again, the used convention should be explained in a readme or codebook.

As an example from the PIGWEB community, IRTA uses a MySQL database to manage data about growth and fattening control (i.e., feed intake, body weight, backfat thickness, loin depth). Data is sent daily from the feeder station server to the MySQL database server. The process is automated using scripts to minimize human errors and secure a repeatable and controlled process. Their database aids to preserve data integrity and structure.

6.3 Data quality

6.3.1 General aspects

Data quality covers many different aspects. Data occurs in various formats, quality, varying time periods and various levels of how the collected data are stored. Data may include a variety of errors, such as:

- Human mistakes (e.g., transmission failure, typing errors, copy & paste).
- Technical equipment data problems (e.g., switching between summer and winter time).
- Used software (e.g., software can change format. A common problem is the date format (e.g., Excel)).

For data cleaning and to improve reproducibility, we recommend documenting every change. For documentation purposes, either use notebooks in script languages like Python or R, or simple documents like plain txt files. To keep track of changes, data can be either deleted or adjusted using versioning (see section 6.4) or the affected data can be flagged.

Prediction of possible mistakes or errors with the data is difficult, but there are some common problems which are listed here:

- *Always double-check handwritten information entered in a computer*
- Check the format and eventually changing the format (e.g., use a standardized format, use same separator)
- Check type, and/or set variable type.
- Check for data plausibly for each variable:
 - Missing or empty values
 - Extreme/impossible values using graphical representation and descriptive statistics
- Check dependent variables on consistency (e.g., the weight of different parts of pig cannot exceed the total weight of pig or the weaning date cannot be before the birth date)
- No “blind trust” in retrieved data:
 - Was data curation performed?

¹² https://en.wikipedia.org/wiki/Relational_model

- Make sure that all desired quality checks are performed

We recommend using descriptive statistics (e.g., minimum, maximum, mean) as well as graphical representations (e.g., boxplot, scatterplot, bar plot) for checks.

6.3.2 Data plausibility examples - introduction

The following examples are only for demonstration purposes and are taken from a piglet dataset. The presented examples cover some of the potential problems that can occur in pig datasets. The examples and errors should reveal the logic of plausibility checks and is not an exhaustive list. The plausibility checks and the extent to which they should be performed are highly dependent on the setup of the data and/or planned analyses. The software used to adapt these examples could also vary from simple editing tools for small datasets to complex programming languages.

6.3.3 Data plausibility examples - formatting problems

Here three common issues are presented.

First example: the dataset was stored in a *.txt file. For unknown reasons, two weights were wrongly merged in the file. This might have happened when the data were produced, generated and/or stored. It always occurred when the weaning weight was larger than 9.99 kg, so that the life weight at 21 days and the weaning weight were wrongly merged. As a result, the corresponding dates appears incorrectly under weaning weight instead of under farrowing date (Figure 6-2).

How this can be revealed? It can be captured by checking for empty or missing values and then checking the revealed rows if this is plausible. Or, one could check the length of the characters within each column and find discrepancies.

| Animal_ID | Father | Mother | Indicator_(1_still_ born-0_live_born) | Farrowing _year | Gestation _period | Suckling _period | No_stillborns _per_litter | No_live_borns _per_litter | No_losses _rearing | Litter _No | Sex_ piglet | Birth_ weight | Weight_14th _life_day | Weight_21st _life_day | Weaning _weight | Farrowing _date |
|------------|--------|--------|---------------------------------------|-----------------|-------------------|------------------|---------------------------|---------------------------|--------------------|------------|-------------|---------------|-----------------------|-----------------------|-----------------|-----------------|
| 872/2/2/16 | Jagger | 872 | 0 | 2004 | 114 | 29 | 1 | 10 | 0 | 2 | 2 | 1.68 | 5 | 7.04 | 8.74 | 14.4.04 |
| 872/2/2/17 | Jagger | 872 | 0 | 2004 | 114 | 29 | 1 | 10 | 0 | 2 | 2 | 1.66 | 4.62 | 5.72 | 7.3 | 14.4.04 |
| 872/3/1/18 | Zambo7 | 872 | 0 | 2004 | 113 | 30 | 0 | 15 | 0 | 3 | 1 | 1.4 | 5.32 | 7.3010.02 | 28.9.04 | |
| 872/3/1/19 | Zambo7 | 872 | 0 | 2004 | 113 | 30 | 0 | 15 | 0 | 3 | 1 | 1.3 | 4.8 | 6.88 | 9.58 | 28.9.04 |
| 872/3/1/20 | Zambo7 | 872 | 0 | 2004 | 113 | 30 | 0 | 15 | 0 | 3 | 1 | 1.64 | 5.56 | 7.9811.00 | 28.9.04 | |
| 872/3/1/21 | Zambo7 | 872 | 0 | 2004 | 113 | 30 | 0 | 15 | 0 | 3 | 1 | 1.48 | 5.3 | 7.7810.90 | 28.9.04 | |

Figure 6-2 Example of wrong formatting resulting in merged terms, shown in blue.

Second example: the farrowing date format in the dataset is not standardized (i.e., by including or excluding a “0” before single-digit days and/or months, see Figure 6-3), which create problems when descriptive statistics will be performed. It is highly recommended to standardize the date using ISO format ([ISO 8601 – Wikipedia](https://en.wikipedia.org/wiki/ISO_8601)) and specify the column type as date format.

How this can be revealed? One could check the length of characters of each entry within the column to check if it has the same length.

| Animal_ID | Father | Mother | Indicator_(1_still_ born-0_live_born) | Farrowing _year | Gestation _period | Suckling _period | No_stillborns _per_litter | No_live_borns _per_litter | No_losses _rearing | Litter _No | Sex_ piglet | Birth_ weight | Weight_14th _life_day | Weight_21st _life_day | Weaning _weight | Farrowing _date |
|------------|--------|--------|---------------------------------------|-----------------|-------------------|------------------|---------------------------|---------------------------|--------------------|------------|-------------|---------------|-----------------------|-----------------------|-----------------|-----------------|
| 878/1/2/7 | Eddi | 878 | 0 | 2004 | 115 | 27 | 0 | 8 | 2 | 1 | 2 | 1.1 | 3.62 | 5.62 | 7.58 | 1.1.04 |
| 878/1/2/8 | Eddi | 878 | 0 | 2004 | 115 | 27 | 0 | 8 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 1.1.04 |
| 878/2/1/10 | Grafik | 878 | 0 | 2004 | 114 | 30 | 0 | 12 | 0 | 2 | 1 | 1.9 | 6.02 | 7.38 | 9.4 | 26.5.04 |
| 878/2/1/11 | Grafik | 878 | 0 | 2004 | 114 | 30 | 0 | 12 | 0 | 2 | 1 | 2 | 5.82 | 7.22 | 9.06 | 26.5.04 |

Figure 6-3 Example of different formatted date cells in the farrowing date column, shown in blue.

Third example: in the dataset we observed “0” values in the weight columns (see Figure 6-4). At this point it is not clear why, and possible reasons include:

- real value on the scale showed “0”

PIGWEB

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004770

- incorrect rounding of, for example, 0.4 kg or in combination with an error in the manual data entry (e.g., slipped decimal point of 0.24 kg instead of 2.4 kg)
- weighing did not occur (e.g., due to piglet loss)

Without this information the interpretation of the data is very difficult. It is not recommended to use a “0” for empty or missing values, because it can affect the subsequent analyses.

The error can be revealed if the value is not in the plausible range.

| Animal_ID | Father | Mother | Indicator_(1_still_born_0_live_born) | Farrowing_year | Gestation_period | Suckling_period | No_stillborns_per_litter | No_live_borns_per_litter | No_losses_rearing | Litter_No | Sex_piglet | Birth_weight | Weight_14th_life_day | Weight_21st_life_day | Weaning_weight | Farrowing_date |
|-------------|--------|--------|--------------------------------------|----------------|------------------|-----------------|--------------------------|--------------------------|-------------------|-----------|------------|--------------|----------------------|----------------------|----------------|----------------|
| 797/2/1/22 | Jagger | 797 | 0 | 2004 | 115 | 28 | 2 | 12 | 0 | 2 | 1 | 1.28 | 3.94 | 5.52 | 6.86 | 22.2.04 |
| 797/2/1/23 | Jagger | 797 | 0 | 2004 | 115 | 28 | 2 | 12 | 0 | 2 | 1 | 0.84 | 3.38 | 4.92 | 6.28 | 22.2.04 |
| 797/2/2/0/1 | Jagger | 797 | 1 | 2004 | 115 | 28 | 2 | 12 | 0 | 2 | 2 | 0.92 | 0 | 0 | 0 | 22.2.04 |
| 797/2/2/0/2 | Jagger | 797 | 1 | 2004 | 115 | 28 | 2 | 12 | 0 | 2 | 2 | 1.02 | 0 | 0 | 0 | 22.2.04 |
| 797/2/2/24 | Jagger | 797 | 0 | 2004 | 115 | 28 | 2 | 12 | 0 | 2 | 2 | 1.36 | 4.02 | 5.52 | 6.4 | 22.2.04 |

Figure 6-4 Example of potential formatting problem, when encountering zero value which are not specified, shown in blue.

6.3.4 Data plausibility examples - checking individual variables

First example: check unique variables. For instance, we checked that the animal ID is unique in the dataset. How this can be revealed? Check that each animal ID is unique and compare the number of unique IDs against the row numbers in the dataset.

Second example: check indicators or categorical variables. In our case, we can check if the indicator for still-born and live-born piglets match the other information. How this can be revealed? Check that the unique values in the variables have the same length as well as the same unique entries as with the provided indicator set.

Third example: check that the values for the given variable lie in a reasonable range. For instance, time periods that are inherently limited (e.g., as gestation length) so that only a specific time can apply. For gestation length, we expected that the values are ranges between 105-125 days. How this could be revealed? This can be obtained via graphical representation of the variable as well as using descriptive statistics (e.g., minimum, maximum, mean). Figure 6-5 detecting one outlier.

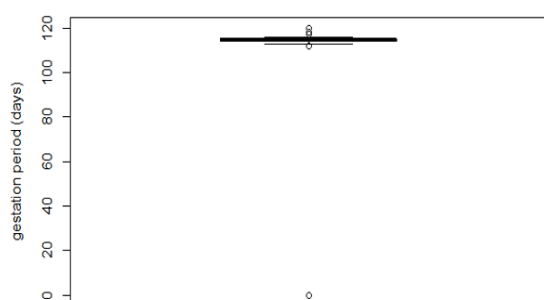


Figure 6-5 Graphical presentation to check if the gestation period lies in a reasonable range using a boxplot.

6.3.5 Data plausibility examples - checking dependent variables

These checks depend on the dataset and we provide here three examples.

First example: check that the year is identical in the two variables “farrowing_year” and “farrowing_day” (see Figure 6-6).

PIGWEB

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004770

How this can be revealed? By extracting the year from the whole farrowing date and compare it with the farrowing year. Further, we observed mismatches and tried to reveal the reason. In this example the given farrowing_year is probably the weaning_year because the affected litters all had their farrowing dates in December (earliest 5th of December) and as farrowing_year the year after was recorded. This lines up with the normalised suckling period of 28 days (at least this is the case for the piglets of this barn). However, it would need to be verified again if this assumption is correct.

| Animal_ID | Father | Mother | Indicator_(1_
still_born-
0_live_born) | Farrowing
_year | Gestation
_period | Suckling
_period | No_still_
borns_p
er_litter | No_live_
borns_p
er_litter | No_losses
_rearing | Litter
No | Sex
piglet | Birth_
weight | Weight_14
th_life_day | Weight_21
st_life_day | Weaning
weight | Farrowing
date |
|-----------|--------|--------|--|--------------------|----------------------|---------------------|-----------------------------------|----------------------------------|-----------------------|---------------|----------------|------------------|--------------------------|--------------------------|--------------------|--------------------|
| 108/1/1/1 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 1 | 1.44 | 5.66 | 8.18 | 10.23 | 1999-12-22 |
| 108/1/1/2 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 1 | 1.68 | 5.82 | 8.22 | 10.28 | 1999-12-22 |
| 108/1/1/3 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 1 | 1.56 | 5.40 | 8.46 | 10.52 | 1999-12-22 |
| 108/1/1/4 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 1 | 1.36 | 4.38 | 6.32 | 7.70 | 1999-12-22 |
| 108/1/1/5 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 1 | 1.58 | 5.70 | 8.00 | 10.09 | 1999-12-22 |
| 108/1/2/6 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 2 | 1.66 | 5.50 | 7.36 | 9.16 | 1999-12-22 |
| 108/1/2/7 | Bennet | 108 | 0 | 2000 | 114 | 28 | 0 | 7 | 0 | 1 | 2 | 1.48 | 5.14 | 5.58 | 9.43 | 1999-12-22 |

Figure 6-6 Example of problem between two dependent variables - farrowing year and farrowing date where years exemplarily differ, considered columns are marked blue.

Second example: combinations of columns, for instance, check that the piglets from a litter have the same mother and father (see Figure 6-7).

How this can be revealed? In this case, we created two new variables. The first variable combines the information “Mother-Litter_No” and the second variable combines the information “Mother-Litter_No-Father”. Then, we compare the number of the unique entries of the new variable 1 against the unique entries of the new variable 2. If the number does not match, we can check where the discrepancies occur.

| Animal_ID | Father | Mother | Indicator_(1_
still_born-
0_live_born) | Farrowing
_year | Gestation
_period | Suckling
_period | No_still_
borns_p
er_litter | No_live_
borns_p
er_litter | No_losses
_rearing | Litter
No | Sex
piglet | Birth_
weight | Weight_14
th_life_day | Weight_21
st_life_day | Weaning
weight | Farrowing
date |
|------------|--------|--------|--|--------------------|----------------------|---------------------|-----------------------------------|----------------------------------|-----------------------|---------------|----------------|------------------|--------------------------|--------------------------|--------------------|--------------------|
| 124/2/2/24 | Konrad | 124 | 0 | 2000 | 114 | 28 | 0 | 14 | 4 | 2 | 2 | 1.28 | 4.04 | 5.72 | 6.58 | 2000-08-09 |
| 124/2/2/25 | Konrad | 124 | 0 | 2000 | 114 | 28 | 0 | 14 | 4 | 2 | 2 | 1.06 | 2.54 | 0.00 | 0.00 | 2000-08-09 |
| 124/2/2/26 | Konrad | 124 | 0 | 2000 | 114 | 28 | 0 | 14 | 4 | 2 | 2 | 1.24 | 2.48 | 0.00 | 0.00 | 2000-08-09 |
| 124/2/2/27 | Konrad | 124 | 0 | 2000 | 114 | 28 | 0 | 14 | 4 | 2 | 2 | 1.04 | 0.00 | 0.00 | 0.00 | 2000-08-09 |
| 226 | Kontal | 124 | 0 | 2000 | 114 | 28 | 0 | 14 | 4 | 2 | 2 | 1.76 | 4.68 | 6.12 | 8.21 | 2000-08-09 |
| 229 | Kontal | 124 | 0 | 2000 | 114 | 28 | 0 | 14 | 4 | 2 | 2 | 1.40 | 4.62 | 7.00 | 8.34 | 2000-08-09 |

Figure 6-7 Example to check if a litter has no unique father, considered columns are marked blue.

Third example: check if the sum of the given number of still-born piglets and number of losses are matching with the “0” weaning weight information (as stated in section 6.3.3, we assume that “0” were wrongly used for missing values) per litter.

How this can be revealed? A newly created variable “Mother-Litter_No-Father” can be used to filter each litter for this information. This variable was created in the previous example and is required to be performed before this check to assure that wrong mother- father – litter_number associations are already excluded as cause in this case. As the information of number of still-born piglets and number of losses are redundant for each piglet in each litter. After verifying this, simply the first entry of the litter can be used. Then calculate the sum of both variables and compare against the number of how many zeroes are contained in the corresponding weaning weight.

| Animal_ID | Father | Mother | Indicator_(1_
still_born-
0_live_born) | Farrowing
_year | Gestation
_period | Suckling
_period | No_still_
borns_p
er_litter | No_live_
borns_p
er_litter | No_losses
rearing | Litter
No | Sex
piglet | Birth_
weight | Weight_14
th_life_day | Weight_21
st_life_day | Weaning
_weight | Farrowing
date |
|---------------------|---------|--------|--|--------------------|----------------------|---------------------|-----------------------------------|----------------------------------|----------------------|---------------|----------------|------------------|--------------------------|--------------------------|--------------------|-------------------|
| 179/5/1/32 | Edmundo | 179 | 0 | 2002 | 115 | 28 | 0 | 5 | 2 | 5 | 1 | 2.32 | 7.06 | 9.76 | 11.58 | 2002-07-04 |
| 179/5/1/33 | Edmundo | 179 | 0 | 2002 | 115 | 28 | 0 | 5 | 2 | 5 | 1 | 1.86 | 5.02 | 7.32 | 9.54 | 2002-07-04 |
| 179/5/1/34 | Edmundo | 179 | 0 | 2002 | 115 | 28 | 0 | 5 | 2 | 5 | 1 | 1.56 | 5.12 | 7.28 | 9.20 | 2002-07-04 |
| 179/5/2/35 | Edmundo | 179 | 0 | 2002 | 115 | 28 | 0 | 5 | 2 | 5 | 2 | 2.18 | 0.00 | 0.00 | 0.00 | 2002-07-04 |
| 179/5/2/36 | Edmundo | 179 | 0 | 2002 | 115 | 28 | 0 | 5 | 2 | 5 | 2 | 2.08 | 3.50 | 4.42 | 5.18 | 2002-07-04 |
| No. of zero rows =1 | | | | | | | | | | | | | | | | |

Figure 6-8 Example on how to compare the number of zero weights piglets with number of losses or still-born piglets does not add up in the blue marked cells.

6.4 Version and backup

As data is curated, new versions of the data set will emerge. This includes raw data, processed data, quality checked data. Therefore, applying logical versioning is needed.

It helps to:

- Keep track of changes
- Access specific versions
- Increase transparency (easier to follow work progress)
- Properly implement the provenance of your data (who, what, when)

In simple projects, versioning can be done via filenames and associated version control table (see Figure 6-9).

| File Name: | | VisionScreenResults_00_05 | |
|----------------|---------------|--|----------|
| Description: | | Results data of 120 VS Tests Essex nurseries | |
| Maintained By: | | Sally Watsley | |
| Last Modified: | | 06/03/2018 | |
| Based on: | | VisionScreenDatabaseDesign_02_00 | |
| Version | Responsible | Notes | Amended |
| 00_01 | Sally Watsley | Test results 1-120 entered | 05/02/18 |
| 00_02 | Steve Knight | Checked by SK | 14/02/18 |
| 00_03 | Vani Yussu | Checked by VY, independent from SK | 16/02/18 |
| 00_04 | Karin Mills | Version 00_02 and 00_03 merged by KM | 06/03/18 |

Figure 6-9 Version control via filenames and version control table

In larger and collaborative projects, data management is preferably done via a (institutional or public) data repository that supports versioning (see chapter 7). Alternatively, data files and versioning could be managed via git¹³, although this is mainly developed for managing software code.

In case versioning was not done or went wrong, software is available to compare files, such as [Beyond Compare](#) or [WinMerge](#).

Of course, data loss must be avoided by applying a proper back-up strategy. Unique data is more critical than copies of secondary data as the latter can be reproduced, provided the processing was

¹³ <https://en.wikipedia.org/wiki/Git>

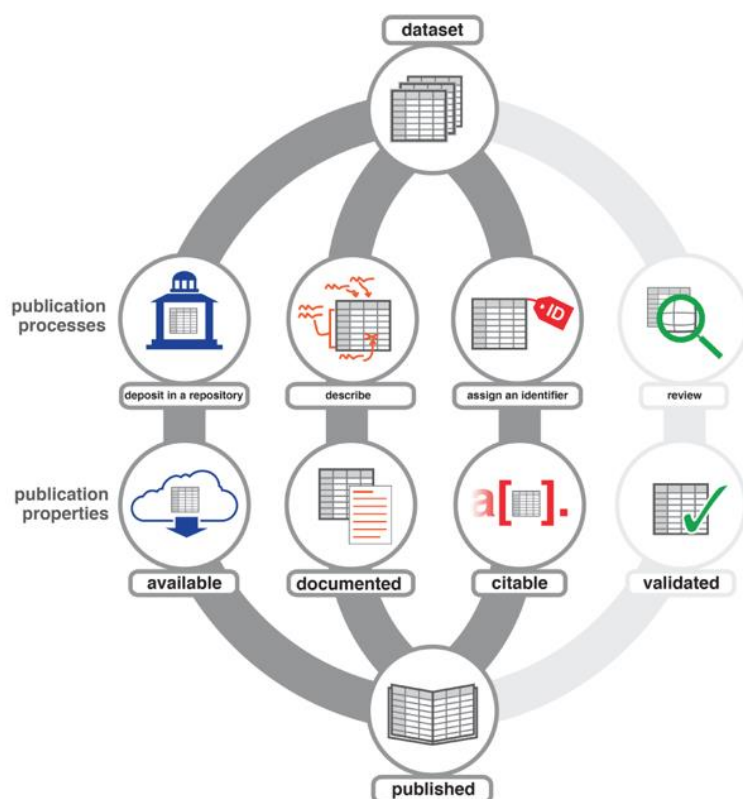
documented in scripts and/or writing. The back-up schedule and method depend on the importance and frequency of changes. We recommend using fully managed file services with automated back-up offered by the IT services of your organization.

6.5 Useful links on data curation

- Mantra training material ([here](#))
- Data file structure -CESSDA ([here](#))
- File naming and folder structure – CESSDA ([here](#))

7 Data publication

Several issues must be considered for data publication. Sharing means that you give others access to your data. This access can be given to a limited number of people or without restriction to a large audience. Partial information (e.g., only metadata) or full access to data can be granted. The access can be publicized, and it is possible to provide information to potential users in very different ways. The physical location of data can be internal or external using either private cloud or ins. In general, to be published, datasets are typically deposited in a repository to make them available, documented to support reproduction and re-use, and assigned an identifier to facilitate citation.



Kratz J and Strasser C. Data publication consensus and controversies [version 3]. *F1000Research* 2014, 3:94 (doi: 10.12688/f1000research.3979.3)

Figure 7-1 Pathways from data to published datasets

7.1 How to publish?

There are several ways to publish datasets with documentation. A basic approach is to integrate data in a published article. In that case, all information on the data is provided in the article. However, the data will still be difficult to find independently of the article and in a format with little or no re-usability. Providing data as supplementary material to an article offers easy access to data with fewer size or format constraints, but data are still difficult to find.

The solution promoted here is to use a repository. There are numerous repositories known and recognized by the scientific community, with little or no size limit. Datasets are provided with a digital identifier. They may be linked to a published article, be amended if necessary, and can be organised in collections. They can be accessed through an article or directly harvested through repository search

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

engines. Metadata is varying informative depending on the repository. If datasets are not associated with an article, extra documentation must be provided to ensure easy data re-use. Repositories dedicated to specific communities should be the first choice when they exist, because dissemination towards the specific community is more efficient.

7.2 The data paper

In addition to publishing data in a topic-specific repository, increased visibility can also be achieved with a data paper. This can also benefit from peer-review, ensure the quality of data and documentation, and increase authors' recognition for their work. Several editors promote open data policy and offer the possibility to publish data and data papers with a peer-review processing ([Scientific data](#), [BMC research notes](#), [Data in Brief](#), [F1000Research data notes](#), [GigaScience data note](#), [Plos One Databases](#), [Animal Open space](#), [elife Tools and Resources](#), [Open data journal for Agricultural Research](#), [BMC journals](#) all publish databases articles).

Depending on the journal and its requirements, the content and size of the data paper may vary from a summary to a comprehensive article. Datasets will either be deposited in a data repository, integrated in the paper, or supplied as supplementary files. The type of license applied to the datasets must not be overlooked. The distribution license applied to data papers (i.e., to the article itself) is generally the Creative Commons CC-BY (attribution requirement, see Figure 7-2). On the other hand, the type of distribution license applied to datasets depends on their location: either on the website of the journal or publisher or in a data warehouse. Some journals and data repositories use a CC0 license by default (i.e., without attribution). Other journals and data repositories accept a delay (embargo) before releasing the data, use licenses that exclude commercial use of the data or generally restrict its re-use.

7.3 Choosing a repository

Putting data in a data repository can provide numerous advantages to sustain FAIR principles with physical infrastructures, archiving policy (with long term storage and availability) and can help structuring files or datasets with versioning. For data discovery, datasets receive a persistent identifier, and through direct access or API, data can be searched, found, and retrieved. The identifier is often a Digital Object Identifier (DOI) that remains the same even if the location of the dataset changes or if a new version of the dataset is provided. When assigned, a DOI is not retrievable. The repository provides a data citation, including a title and an authors' list to simplify data citation. Finally, the repository can provide tools to manage data access rights. In that case, licences define the terms of use when releasing data into the public domain (with or without an embargo). It is possible to use existing licenses (e.g., Creative Common, see Figure 7-2) that establish the rules for re-use. In some cases, specific terms can also be defined.



Figure 7-2 Understand the particularities of Creative Commons licenses

The choice of a repository is dictated by rules. The first step is to check if subject- or domain-specific repositories exist, which is the best option to target your peers. There are, at the time of writing, no specific repositories for pig data, and institutional or national data repositories are good options as they can provide easy access support (e.g., 4TU.ResearchData, DANS-EASY, data.gouv.fr). There are also generalist and multidisciplinary data repositories. In that case, it is recommended to use well-known repositories (e.g., Zenodo, b2share, b2find), maintained by known entities to guarantee the application of FAIR principles (see also Table 7-1 for some other repositories).

Table 7-1 Overview of some repositories taken from WUR <https://library.wur.nl/repositoryfinder>

| Repository | Discipline | Associated journal(s) or publisher(s) |
|---|--|--|
| Pangaea | Earth & Environmental Science | No partnerships or integrations known, but recommended as the standard repository in the discipline by various publishers. |
| GBIF/NLBIF | Biology, Biodiversity | None known, but recommended by publishers including PLOS and Springer Nature. |
| NCBI : Genbank | Biology, Genetics | No partnerships known, but the use of Genbank is encouraged by many publishers. Examples are PLOS, Springer Nature and Elsevier. |
| EMBL-EBI : ArrayExpress, ENA, BioStudies, PRIDE, BioModels, IntAct, MetaboLights | Biology, Genetics, Bioinformatics | EMBL-EBI repositories are often recommended by publishers. Examples are PLOS, Springer Nature and Elsevier. |
| Dryad | Multidisciplinary (focus on life sciences) | Hundreds of journals offer integrated data submission with Dryad: browse the list. |
| Harvard Dataverse | Multidisciplinary (focus on social sciences) | Various publishers recommend Harvard Dataverse, and some journals have set up their own Dataverse. |
| Mendeley Data | Multidisciplinary | Integrated into the workflow of Elsevier journals. |
| DataverseNL | Multidisciplinary | None known |

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

| Repository | Discipline | Associated journal(s) or publisher(s) |
|--------------------------|-------------------|--|
| Figshare | Multidisciplinary | Many publishers have a partnership with Figshare, including Springer Nature, PLOS, and Wiley |

A global registry of research data repositories that covers research data repositories from different academic disciplines can be useful tool (e.g., [re3data](#), [FAIRsharing](#), [b2find](#)). Additionally, [CoreTrustSeal](#) offers core level certification to any interested data repository.

7.4 Metadata repository

For the PIGWEB community, a central metadata [repository](#) has been created under the data.gouv.fr repository. It will be used to collect metadata from all PIGWEB datasets stored in open repositories and provide an easy access to them.

7.5 Useful links on data repository

See previous sections for links.

8 Data Management Plan

8.1 Introduction

Data play an important role in research. Research often uses, generates, processes, and publishes data as an output. We might even state that “data are the crown jewels of research”. Consequently, data management, taking good care of data during the whole research process, is a key aspect of performing science.

Of course, you can try to find out what to do with your data “on the job”, but as with any key activity, it makes more sense to think ahead and know how you are going to handle the data before you start. Therefore, it is useful to document how your data management will be implemented in a data management plan (DMP). It will support the management of your data by providing a structured plan, pre-planned quality control checks, and an overview for data re-use, helping others to understand and reproduce what you have done. Due to the growing awareness of the importance of data management, many funders now require a DMP to be delivered in the early stages of research projects and maintained throughout the process. With its strong focus on data management, the DMP helps to make the data FAIR at an early stage, which is also in line with good scientific practice.

DMPs are not only relevant for researchers. They are also an important asset for others in a research organisation, such as data managers, privacy officers, and information security officers, because a DMP also addresses the safe handling of sensitive data (e.g., privacy-sensitive data, legal data, business-related confidential data). Your organisation probably requires data handling to comply with certain laws, policies, and guidelines, and a DMP can be used to determine whether the intended handling of data is well organized or can be improved.

When you (need to) adopt the FAIR principles as part of how you handle your research data, the DMP will describe how you will implement that as part of your research.

8.2 Writing a Data Management Plan (DMP)

There are a lot of aspects that are relevant when describing how you are going to manage data in a research project. Usually, a DMP describes (at least) the following aspects:

- Organisational context
- Description of the research project
- The roles of involved persons
- The data and software used to work with data
- How short-term storage is arranged during the project
- How the data is structured
- How data is documented, and metadata is added
- How aspects like sensitive data, data sharing, data ownership and access to data are dealt with
- How data is published
- How long-term storage is arranged

This is quite a long list, and it might seem complex to cover all these topics in a good way. The good news is that there are many templates available with a clear structure, providing instructions on how to fill in the details. Many funders provide templates for the DMP that they require as part of their

projects. As an example, the European Commission offers specific templates and instructions for e.g., the Horizon 2020 and Horizon Europe programs. You can probably find some other examples from national and institutional programmes. Besides, it is a good idea to inform yourself if your own institute offers any templates.

Still, it can be challenging to complete all sections of a DMP. At the same time, you are not the first one to write a DMP. Example DMPs are usually available that can inspire you on how to create your own, although sometimes finding them can be a challenge. Probably the most useful and powerful tool that can help you to get started with your DMP is [DMPonline](#), an online tool developed by the Digital Curation Centre (DCC) in the UK. The tool supports researchers by offering access to DMP templates, accompanied by good practice guidelines.

DMPonline offers a range of generic and commonly used templates, some of which follow the requirements of specific funders. There are, for instance, templates specifically for the EU Horizon 2020 and Horizon Europe research programs. Additionally, templates of many regional and national funders, universities and research institutions are available.

Many research organisations have arranged institutional access to the DMPonline web tool, so you can log in with your organisational credentials and access the most relevant templates for your organisation. The tool guides you through the steps of creating a DMP using the specific template that you have selected. It also supports a review process by allowing you to share and request feedback from fellow researchers or supervisors. At any stage, you can also download the full plan in a specific format (e.g., PDF, Word, CSV, or plain text).

A valuable feature of DMPonline is that it allows you to access DMPs of other researchers. Everyone can publish a created DMP to make it accessible to others using the tool. This provides you with a wealth of example plans, possibly also dealing with your domain, or covering the same types of data that your research handles.

8.3 DMP and the FAIR principles

Some templates and instructions might explicitly refer to FAIR as a guideline to hold on to. In any case, it is useful to consider how you will ensure that your data is Findable, Accessible, Interoperable and Re-usable, if any conditions impose restrictions and how to deal with them. **Error! Reference source not found.** shows which aspects of the FAIR principles should somehow come back in a DMP.

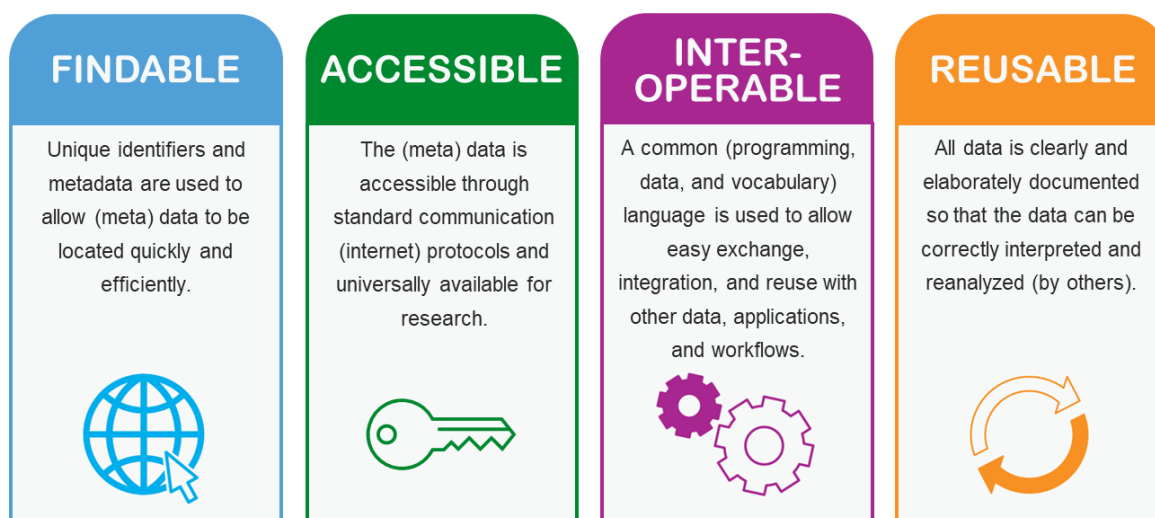


Figure 8-1 How the FAIR principles influence a DMP

8.4 Useful links

- DMPonline ([here](#)) (on-line writing of DMP)
- DMPonline manual ([here](#))
- Fictional pig research DMP ([here](#))
- DMP template repository ([here](#)) (in French)

Annex I - Example of an ABCD data standard in pig research

Mandatory elements

The mandatory elements listed here must be filled out to be a valid ABCD file. Other elements are required depending on which platform you want to share the data.

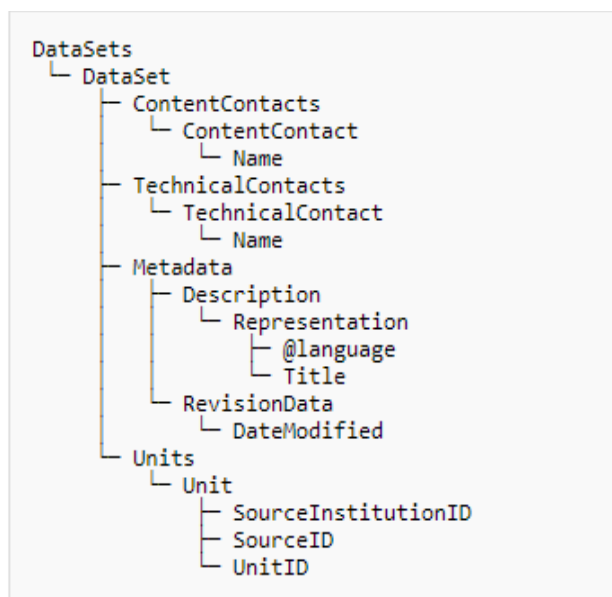


Figure A1: Hierarchical structure of the mandatory concepts to be a valid ABCD file (copied from https://wiki.bgbm.org/bps/index.php/Main_Page)

As illustrated in Figure A1, the metadata applies to the DataSet and provides information for each unit. Considering livestock data, a unit could be seen as a single animal or a group of animals, e.g. when they are not easily separable, like fish in a tank.

Core elements for storing experimental data

Of the ABCD elements, the MeasurementsOrFacts class is best suited for the structured description of the experiments or observations carried out. Table A1 splits the MeasurementOrFacts class into sub-elements, describes its definition in the remark and shows an example.

Table A1: ABCD elements about an experiment or an observation stored in MeasurementOrFact.

| Group | Element | Example | Remark |
|---------------------------|---------------------|------------------|--|
| MeasurementOrFactAtomised | MeasuredBy | Employee A | Attribution of the measurement to a Person |
| | MeasurementDateTime | | Date (and time) the measurement was taken |
| | Duration | 1998-2000 | Duration of measurement in time. |
| | Method | Gestation period | The method used to make a measurement. |
| | Parameter | days | Describes the type of measurement or fact, such as width, abundance, |

PIGWEB

| Group | Element | Example | Remark |
|-------------------|--|---|--|
| | | | circumference, temperature etc. |
| | AppliesTo | days of gestation of the mother sow | Depending on the use of the type, this can further specify the actual part measured. For example, a temperature measurement may be a surface, air or sub-surface measurement. Possible to provide here technical information, for instance, a scale for weighting. |
| | LowerValue | 0 | Lower or only value or fact text. |
| | UpperValue | 150 (just an approximated upper value) | Upper value where there is a range. |
| | UnitOfMeasurement | days | Unit of measurement. |
| | Accuracy | 24h | Statement of the accuracy of measurement |
| | MeasurementOrFactReference/ TitleCitation | ATOL ontology term for "gestation length" | Reference (publication) where this measurement was taken from. |
| | MeasurementOrFactReference/ CitationDetail | | Specific page, figure or illustration number(s) within the reference. |
| | MeasurementOrFactReference/ URI | URI | URI to Reference. This can be a known methodology or an institutional internal Standard Operating Procedure (SOP). Beside the ATOL ontologies also Livestock product trait ontology (here) or Vertebrate Trait Ontology (here) can be used. |
| | IsQuantitative | TRUE | Flag indicating if the value represents the numerical result of a quantitative measurement (TRUE) or a descriptor with the textual or categorical result (FALSE). |
| MeasurementOrFact | MeasurementOrFactText | Gestation period | Free text alternative to atomised version. |

PIGWEB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004770

| Group | Element | Example | Remark |
|-------|---------------------------------|---------|----------------------|
| | MeasurementOrFactText/ language | eng | Language of the Text |

There are two mutually exclusive elements for the time measurement in the MeasurementOrFact. Either the duration (Duration) or the concrete date and time (MeasurementDateTime) for the measurement is described.