Technical and Scientific Description
Appendix 6. Data Management Plan

Version 7.1
16/10/2020

Cover page, from left to right, and top to bottom:
Enclosed growth facility at Højbakkegård, Denmark (© University of Copenhagen)
Inside the Gembloux Ecotron, Belgium (© University of Liège)
Preparing a macrocosm at the Montpellier Ecotron, France (© CNRS)
The Planaqua CEREEP aquacosm facility in Saint-Pierre-Lès-Nemours, France (© ENS – CNRS)
The O3HP open-air manipulation platform at the Observatoire de Haute-Provence, France (© CNRS)
The open-air FACE experiment from Risö fields, Denmark (© Technical University of Denmark)

Data Management Plan

1. FRAMEWORK AND SCOPE

1.1. Reference documents

(a) Statutes of AnaEE-ERIC
(b) AnaEE Scientific and Technical Document
(c) User Access policy document

1.2. Administrative information

Initiative	**AnaEE-ERIC** Analysis and Experimentation on Ecosystems
Responsible	AnaEE
Contact details	
Reference number/ID	

1.3. Purpose of the Data Management Plan

In the current context of open science, and consequently the sharing of data from scientific research, the AnaEE infrastructure aims to offer all the tools and services enabling the scientific community, but also experts from the industry, NGOs, policy makers, or the general public, to access and reuse the data produced during the projects it hosts. The diversity of scientific themes addressed during research projects leads us to handle a wide variety of data related to the physical, chemical and biological aspects of an ecosystem. This is why it is essential to take data management into account throughout the data life cycle, from the design of experiments to the sharing of data in a sustainable manner. The different actors of the infrastructure as well as the project leaders will be involved in the data management process in order to provide quality data and metadata according to FAIR principles.
This Data Management Plan (DMP) is therefore intended to be the reference document specifying good practices to be implemented at each stage of the data life cycle. A data management plan is, of course, a living document that will have to adapt to changes in ecological research but also to technical developments relating to the data. This is why we will try to describe how it contributes to the continuous improvement of the production of FAIR datasets by implementing quality control throughout the process of producing and then sharing the data.

1.4. Research framework, activities and objectives of data collection

AnaEE is a distributed research infrastructure of experimental platforms all across Europe. (Figure 1) The platforms may either be open-air or enclosed; their specificities is that all platforms are designed to perform manipulation of the ecosystem in order to simulate the stresses applied to ecosystems by climate changes and anthropogenic activities, and test different evolution scenarios. In addition, AnaEE features analytical platforms that are able to perform deep analysis, and modelling platforms to allow the users to interpret the data within the framework of various theoretical models.

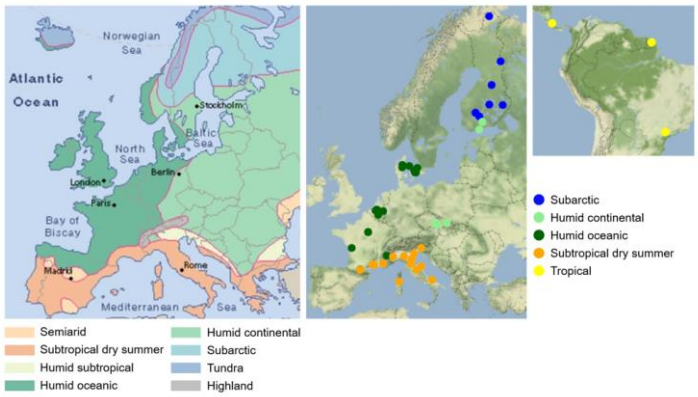

Figure 1: Climatic zones of Europe (left panel), and the geographic coverage of the AnaEE platforms in founding countries (middle and right panels). Note that AnaEE includes 3 tropical platforms in central and South America.

The ERIC will coordinate the activities of AnaEE (Figure 2). It will evaluate and optimize the proposals. A technology centre (TC) helps the platforms to maintain a high standard in quality, and foster new technological solutions. The Interface and Synthesis Centre (ISC) will make studies to integrate the results and exchange with the society. The Data and Modelling Center (DMC) will make the data available with the appropriate standards, and provide models for the interpretation. The Central Hub (CH) performs the central administrative tasks. It will also be the central portal for the access to all AnaEE services, and for the proposals from users.

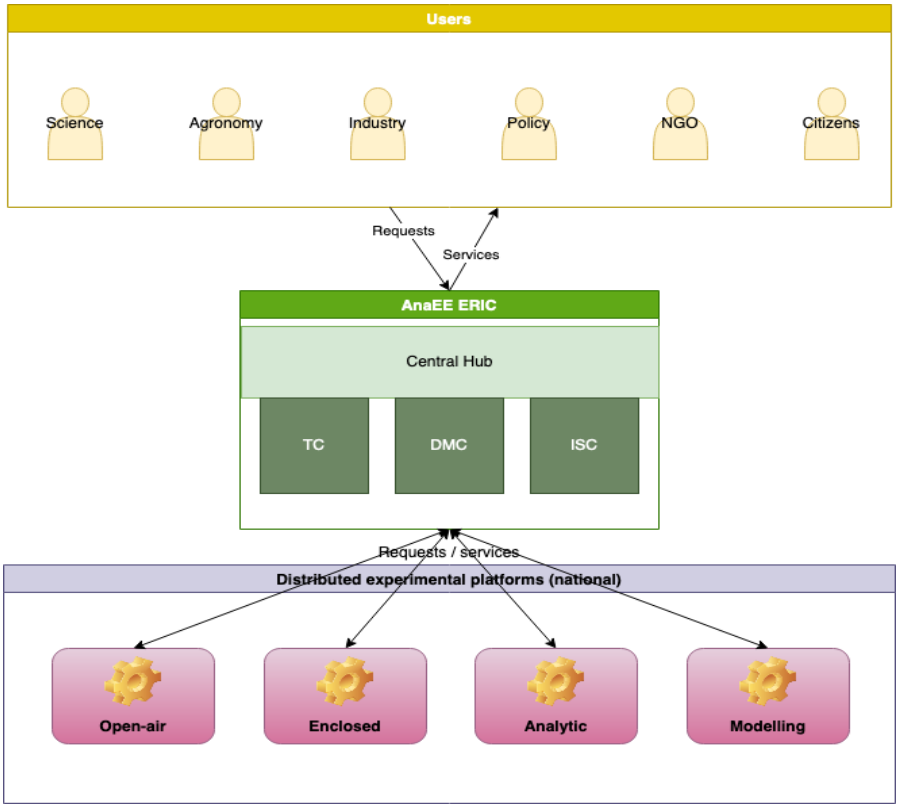

Figure 2: The organization of AnaEE. The ERIC will be the interface with the users, with an integrated offer of services, and receiving requests. Thanks to the Services Level Agreements, the requests will be analysed and sent to the appropriate distributed platforms, taking into account their offer of services.

The access to the services provided by AnaEE is documented in the appropriate document and in the user guide.

The review procedure and project life is a 2 step procedures that can be described as follows (Figure 3):

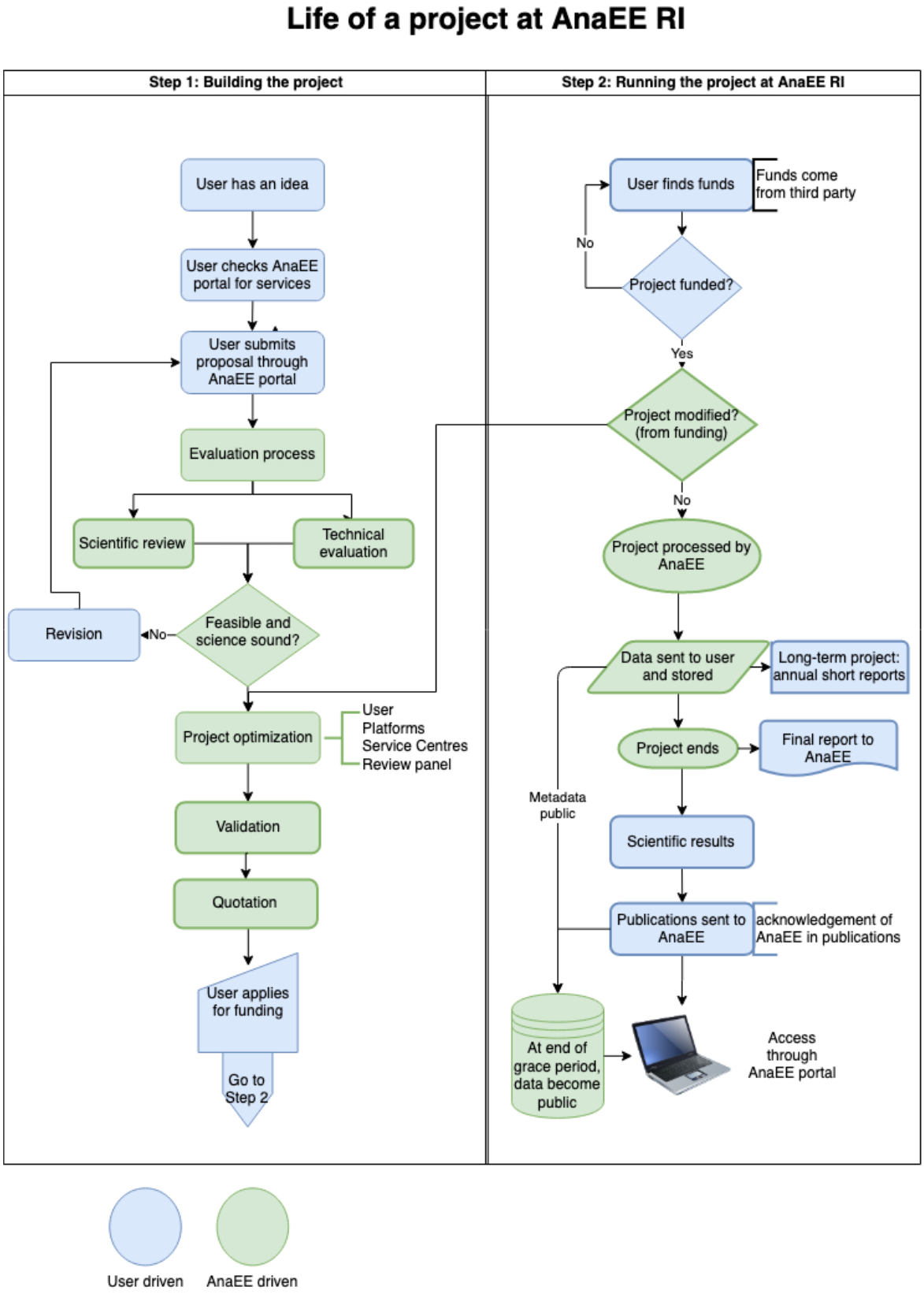

Figure 3: Description of the procedure for user project access to AnaEE platforms and services.

In Step 1, the user takes advantage of searching the web portal platforms and service catalogue for the most relevant platforms and services for the given project idea. He/she can also contact AnaEE directly

for advice. The user then submits a pre-proposal online through the web portal. The Central Hub facilitates that a scientific review is performed by the Project Review Committee. Upon a positive review, the Central Hub requests a technical feasibility check at the relevant platforms and Centres as well as their pricing for the suggested service taking into account the constraints related to data sharing. During this process, both the review panel and the platform(s) and Service Centre(s) representatives are urged to provide ideas for potential scientific and technical improvements (project optimization) of the project proposal. The Central Hub provides this feedback together with the review and the quotation for the suggested services to the user, who can now use this information and budget in the project proposal submitted to a funding body.

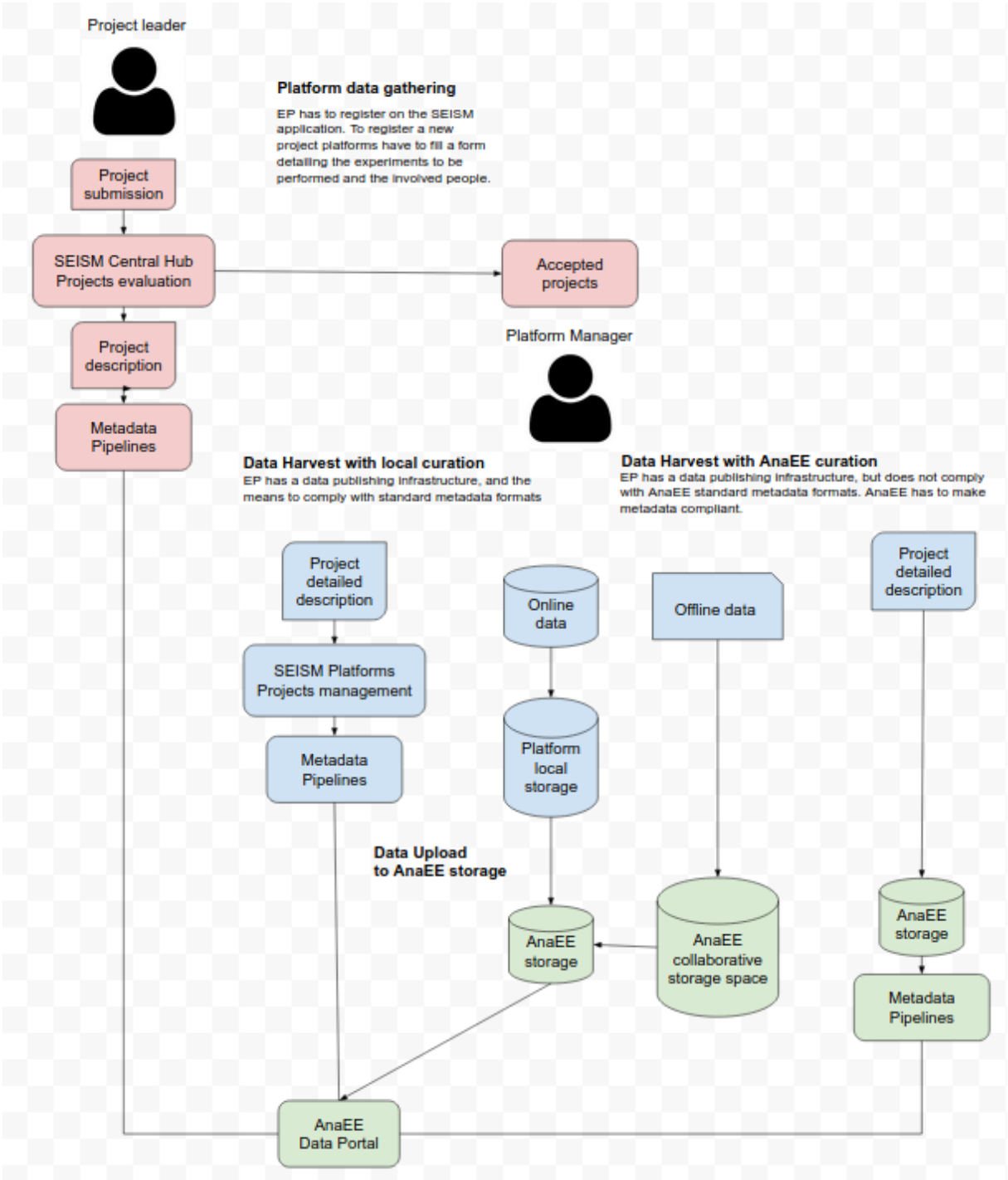

Figure 4: Description of the procedure for taking charge of the projects and the means implemented by the different actors in data management.

In Step 2, projects that succeed with obtaining funding can go into AnaEE processing and scheduling on the relevant platforms. If the project proposal was changed compared to the optimized and validated project proposal (in step 1) – this part of step 1 must be repeated. The final project proposal must include a Data Management Plan, dedicated to the project, compliant with general principles and best practices described in this document. Once the project has moved into AnaEE processing, the AnaEE DMC stores project metadata in a dedicated and restricted collaborative storage space, which will be accessible through its Web services by the research team and the platform team. The AnaEE processing consists in an iterative revision of the dataset that may follow the collection of the dataset; the goal of such revision is to ensure the meeting of a series of data quality criteria expressed in this document. Once a dataset is approved, it can be published and distributed on the AnaEE Web services. Optionally the owner of a dataset can ask for a so-called grace period, i.e. a delay between approval and publication during which the dataset will be available only to a restricted group of users. Access to publications is then facilitated through the web portal.
The Central Hub will collect annual reports of longer projects and final project reports when projects end, as well as the scientific results and papers resulting from the user projects.

1.5. Project Data Management Plan proposal

The diversity of data types produced within the AnaEE infrastructure requires to organize data management at each step of the data production process. The actors in this process are different at each step and the challenge is to clarify what responsibilities they have and what tools or services they will use to coordinate the management of AnaEE's data. While the bulk of the work is expected from research platforms and research teams involved in the project, AnaEE's centres will offer support on all data management tasks.
We will distinguish 2 types of Data Management Plans:
Background DMP: The "Background Data" includes all the data produced by the platform on a continuous basis and characterizes the environmental conditions of the ecosystem under study over the long term. The variables measured are the same for all projects hosted by the platform but these projects will be able to use these datasets over the project period or take into account longer periods if necessary (data prior to the project which constitutes a history of the life of the ecosystem under study).
Foreground DMP: The "Foreground Data" part corresponds to the variables measured over the period of a one-time project that is shorter in duration. They may correspond to measurements or sampling, analyses of these samples or observations of biodiversity.
The "Background Data" corresponds to the environmental conditions of the ecosystem and the "Foreground Data" to the effects of treatments applied to the ecosystem of our platforms during a time-limited project.
Two types of data management plan templates will be provided to platforms and project leaders. Platform managers will mainly contribute to the drafting of the "Background Data" and project leaders to that of the "Foreground Data".

Datasets are the central objects of AnaEE Data Management Policy: they are self-contained sets of information that include data and all the metadata required to re-use that data. Examples of dataset are: a database of historical and georeferenced observations, an archive of field observation, an archive of laboratory analysis results, a set of field or laboratory images, a set of system logs. A project dataset is composed by the "Foreground Data" and the "Background Data" over the project period.

The Data Management Plan of a candidate AnaEE project is a self-regulation document compiled by the project's regulating body describing:

- **Data collection workflow**: how data is acquired, processed, and packaged into a deliverable.
- **Data Conservation Plan**: how data is planned to be stored during the AnaEE evaluation and after its publication.

- **Data licensing policy**: how data should be accessed and used after publication, different dataset may have different licensing policies, although a coherent general project policy is expected.
- **Candidate datasets**: the data deliverables expected from the aforementioned data workflow.

1.6. Data Management Process implementation

Implementation of the DMP is necessary at both the central and local level. The final responsibility for implementing the DMP lies with AnaEE Central Hub. The Technological Centre and the Data and Modeling Centre will ensure the DMP is being implemented at the local level.
AnaEE Central Hub and Data Modeling Centre will support them by organizing the necessary training and the establishment of a data management working group. This working group of data managers and scientists will have a representative for each type of platform and take responsibility for the actualization of the DMP when needed.

With regard to data and metadata, the implementation process of AnaEE's scientific projects offers different tools and services to:

- Collect project description metadata at the project design and production stages
- Extract the "Background" and "Foreground" data
- Collect metadata of the project's completion
- Store the data produced
- Provide a portal for discovering and accessing datasets

This process must be flexible and compatible with the different practices of each platform. This is why we take into account the cases where the platforms will not use the metadata collection tool proposed by AnaEE or their current data sharing practices in thematic data warehouses.
Of course, the AnaEE infrastructure will invite, through training and support, the platforms to standardize their practices.

We identify the following main tasks in data management implementation:

- **Data description and collection**: the central task in data management, data collection includes also reuse of previously collected data. This first phase relies on research platforms and research teams involved in the project, and is supervised by the Central Hub and the Technology Centre.
- **Quality of the data management process**: the gathered data has to be curated with appropriate metadata, documentation, and other materials to improve its fruitability. The Central Hub and the Data Modeling Centre support research platforms in this task.
- **Storage and Backup Management**: storing data and assuring its availability over time is primarily a research platform responsibility to ensure backup. The AnaEE Technology Centre and Data and Modeling Centre support research platforms in arranging an adequate storage solution.
- **Meeting legal and ethical requirements**: the current legal frameworks around data management impose high standards that will be enforced by the Central Hub and implemented by research platforms
- **Data sharing management**: finally, the data needs to be findable and accessible both within AnaEE partners and among external organizations. The Data and Modeling Centre will offer services to store persistently, publish data and manage the access to said data. Research platforms are expected to use such services and to align their existing data sharing services with the AnaEE data catalog.

Activities may include, but are not limited to:

- Format conversion to comply with AnaEE format recommendations;
- metadata curation to comply with metadata standards supported by AnaEE;
- documentation curation for a better human understanding of the dataset;
- metadata curation to make it machine-readable;

- storage system revision to comply with AnaEE data management policies;
- API development or revision to comply with AnaEE guidelines.

1.7. Quality assurance process

We will distinguish 2 quality approaches concerning the management of AnaEE data. One of these approaches is a "product" quality approach concerning the quality of the datasets that are delivered to the research teams using our platforms. The 2nd approach is a "process" quality approach concerning the quality of the data management "process".
Indeed, the quality of the datasets is linked to the data management process but especially to the good experimentation practices applied on the platforms to observe, sample, correctly measure the ecosystem parameters and increase the reliability and robustness of the data sets. This role of training in good practices and their improvement is the responsibility of the Technology Centre.
As infrastructure data managers, we will focus on the quality of the process that enables the provision of datasets according to FAIR principles. It is therefore necessary to identify for each stage of the data life cycle, the human and material resources to be implemented, the criteria to be met to guarantee the FAIR principles and the indicators to verify that these criteria are met in order to improve them on a continuous basis.

Figure 5: Data life cycle.

In Table 2 are described the actions to be implemented by the different actors throughout the data life cycle to produce data sets compatible with the requirements of the FAIR principles:
To implement this quality approach, AnaEE has identified tools and services that enable the various actors in the data life cycle to meet the performance objectives of the criteria we have set. We will strive to set these improvement targets during periodic data management plan working groups involving representatives of the different types of platforms (Table 3).

Table 2: Actions to be implemented by the different actors throughout the data life cycle to produce data sets compatible with the requirements of the FAIR principles

		Data life cycle		**Planning research**	**Collecting data**	**Processing and analyzing data**	**Publishing and sharing data**	**Preserving data**	**Re-using data**
		Responsibility		CH, Platforms and Research team	TC, DMC, Platforms and Research team	DMC and Research Team	ISC, DMC, Research team	DMC and Research Team	DMC
Requirements to be FAIR		Resources or services							
Data description and collection or reuse of existing data		How?		SEISM, INRAE pipeline	SEISM, INRAE pipeline				
		Data types and formats		SEISM, INRAE pipeline	SEISM, INRAE pipeline				
Documentation and data quality		What metadata?		SEISM, INRAE pipeline	SEISM, INRAE pipeline				
		What are the measures to control data quality?		Objectives, criteria and indicators	Curation scripts production	Curation scripts catalog	Curation scripts catalog		Usage tracking

Storage and backup		How for data and metadata ?			Collaborative data storage space			DMC cloud infrastructure	
		Implementation of security and data protection?			DMC support to implement the adequate solution			CC IN2P3 mirroring	
Legal and ethical requirements, codes of conduct		RGPD?		CH, legal assistance			DMC Data Portal		Usage tracking
		Intellectual property?		CH legal assistance			DMC Data Portal		Usage tracking
		Ethics and codes of conduct?		CH legal assistance			DMC Data Portal		Usage tracking
Data sharing and long-term preservation		How and when do we share? Restrictions and embargoes?		CH legal assistance			DMC Data Portal; API portal	DMC cloud infrastructure and CC IN2P3 mirroring	DMC Data Portal; API portal
		How and where will the data to be retained over the long term					Data Portal	DMC cloud infrastructure	

		be selected?							
		Methods and software tools needed to access and use the data?				Data Portal; API portal	Data Portal; API portal	Data Portal; API portal	Data Portal; API portal
		How do you assign a PID?				Data Portal	Data Portal		

Table 3: Criteria, means and indicators of the AnaEE data

Data life cycle		Planning research	Collecting data	Processing and analysing data	Publishing and sharing data	Preserving data	Re-using data
Objectives		Increase the number of	Increase the reliability of	Increase datasets manipulatio	Increase number of datasets	Ensure 100% of data preserved	Provide reusable datasets across multiple scientific domains

		projects supported	data acquisition Increase a rich metadata collection	ns to process and analyse its	published and shared		
Criteria							Download rate over 5 years and 10 years. Claimant's scientific field.
Indicators							Number of downloads over 5 years/10 years. Percentage of origin of applicants.

1.8. Related documents, policies and procedures

Several documents contain statements that define the framework of this data management plan and therefore need to be referred to:

- AnaEE statutes
- AnaEE Technical and Scientific Description
- AnaEE Data Policy
- “Background” DMP template
- “Foreground” DMP template

2. DATA COLLECTION

In this section we present data collection guidelines and best practices adopted by AnaEE and its partners. This section is ever-evolving and research platforms should refer to the Technological Centre and the Data Modeling Centre for the latest updates and state of the art best practices.
AnaEE is running a survey over its research platforms to assess data collection de facto standards and best practices

2.1. Background data versus foreground data

According to the specificity of AnaEE's platforms, we can distinguish 3 types of data:

- observation data produced continuously on the platform over the long term,
- data produced during a project over the short to long term term,
- data from sample analysis.

As we described it earlier, long-term data can be considered as “background” data and are generated by the platform even if no particular project is in progress; data produced by observations or experimentation specifically installed for the needs of a project during the project and data from sample analysis are the “foreground” data that are the hosted project-specific data.
Since long-term data are data produced "continuously" over long periods of time, a data management plan dedicated to the measurement of these environmental variables is written and updated by the platforms that have this type of data.
When these platforms host short-term projects, they produce a dedicated data management plan for the project, referring to the long-term data management plan if it is part of the project deliverables.

2.2. Nature and types of data

The study of terrestrial and aquatic ecosystems requires the observation, instrumentation or sampling of the various compartments that make them up. In order to better understand the interactions between the different environments (atmosphere, soil, water, etc.) as well as the interactions between the different biodiversity compartments that occupy these environments, AnaEE's platforms use a wide variety of sensors, analytical instruments and sampling or biodiversity observation methodologies. For this reason, we will focus on accurately describing the types of long-term observational data from our platforms as well as the current analysis and measurement capabilities commonly used on our platforms (Background data, Table 2).
As "foreground" data is constantly evolving due to the diversity of the projects and the use of sometimes very innovative instrumentations and methodologies, these will be described in the data management plan dedicated to each project

Table 4: Table of Background Data Types for AnaEE Platforms

	Thematic data type category	Research domains	Data formats	Data files types
Physical data	Soil water content	Soil Science	CSV	
	Soil temperature	Soil Science	CSV	
	Water temperature			
	Meteorological data	Atmospheric Science	i.e. netCDF	
Chemical data				
Biological data				

2.3. Standards and methodologies in data collection

The perpetual evolution of instrumentation and the needs of ever more innovative research projects push to develop new experimental systems and new methodologies using the latest technologies. This is why we strive to standardize as much as possible the formats of our data and metadata by using the most widely used formats and best suited to the project and its research theme. In addition, recommended formats are compliant to the FAIR principles (see table 4).
These "foreground" data are described in the management plans dedicated to each project. The Technological Centre's mission is to ensure the standardization of the instrumentation and methodologies used to harmonize "upstream" the data formats and metadata produced. In addition, in relation to the Data and Modeling Centre, data and metadata formats will, if necessary, be harmonized "downstream" by providing a collaborative data curation catalog of conversion scripts to the most interoperable common formats.
Interoperability of formats remains the essential condition for any type of machine to be able to reuse them and the various AnaEE data management stakeholders will seek to work towards this goal in a continuous improvement process.

2.4. Metadata standards

Data and metadata standards and formats are a key aspect for technological and semantic data operability in order to make data discoverable for promoting international and interdisciplinary access and use of research data.
The standards and formats used during data collection by the platforms and research teams are very varied. This is why AnaEE proposes and develops software tools (SEISM application) to simplify data and metadata collection (TC), semantically annotate metadata and make them compatible with the semantic web (INRAE pipeline) deployed locally or by the DMC. These tools facilitate data sharing and long term access by structuring datasets and their metadata to make them fully discoverable, interoperable and machine-readable.
Thanks to the semantic annotation tools developed by AnaEE (INRAE pipeline), all metadata standards will be transformed into RDF graphs compatible with the semantic web and any type of machine (Table 2).
Metadata standards are integrated into the project tracking tool (SEISM). Project leaders and platform managers will select the most suitable standards available to describe all the elements of the project (Table 5) in a metadata standards catalog provided by AnaEE.
AnaEE expects the project DMP to identify the datasets produced by the project, herein candidate datasets, and to include the following basic information for each candidate dataset that will undergo AnaEE evaluation:

- Abstract: a short textual description of the candidate dataset.
- Owner: the person or organization who legally owns the data therein provided.
- Contact person: one or more individuals responsible for the communication with AnaEE for that candidate dataset.
- Means of accessing the data: if the dataset is served with an API, an endpoint will be needed, otherwise if the dataset consists of downloadable files, one or more download links, if the data is hosted on a cloud provider, adequate access keys should be provided; these links and credentials should remain active for the whole revision of the dataset.
- Format and structure: all the information required to open, navigate, and access the dataset’s structure.

These metadata are the minimum metadata to be associated with a dataset. It is therefore necessary to expand this base with much richer metadata to describe the full production context of a dataset.

2.5. Documentation and metadata collection

AnaEE acknowledges that for a dataset to be reusable by a large community of users, it is essential to describe all the conditions under which the data were generated. The user therefore needs to have a clear and as precise as possible description of the context of the measurement, described with controlled vocabularies and associated documentation. In this section we provide an outline of documentation and metadata curation best practices AnaEE expects from its research platform.

The AnaEE infrastructure develops tools (SEISM and INRAE pipeline) to structure the metadata collection from the conception of the scientific project and during the implementation of the experimental protocol (TC and Central Hub). The collection of a minimum set of core metadata is a mandatory requirement for all datasets, for their discovery, while rich metadata is also recommended to ensure a full interoperability and reusability of data, including most of the provenance elements. In any case, the metadata should explicitly include the persistent identifier of the data it describes.

A project monitoring and resource management tool (SEISM) facilitates project leaders and platform managers in collecting metadata by describing the experimental protocol, the equipment used and the associated documentation, the measurement chain and its lifecycle (maintenance, calibrations, ...) that allows to measure experiment variables, the human resources involved...

In a second step, this metadata is processed in a semantic annotation pipeline to provide metadata in the form of RDF graphs compatible with the semantic web, to ensure data findability and interoperability (INRAE pipeline).

The DMC will provide national platforms with guidelines and operational tools (thesaurus, ontologies, etc.) to implement metadata and data standardization.

These tools provide discovery metadata as well as technical metadata to the Data Portal managed by the DMC in the form of a catalogue. The semantic metadata is pushed to AnaEE data storage to benefit from the same persistent unique identifier as the dataset. Catalog provided by the DMC, through a search engine, allow data users to quickly find links to data from the discovery metadata.

In order to facilitate metadata structuration and improve interoperability, AnaEE recommend metadata standards

Table 5: Recommended metadata standards for data.

Metadata standard	Description
Darwin-Core Archive (DwC-A)	**Biodiversity informatics data standard (https://github.com/gbif/ipt/wiki/DwCAHowToGuide)**
NetCDF CF	**Climate and Forecast Metadata (http://cfconventions.org/)**
ISO 19115/19139	ISO/TS 19139:2007 defines Geographic MetaData XML (gmd) encoding, an XML Schema implementation derived from ISO 19115 (https://www.iso.org/standard/32557.html)
EML	Ecological Metadata Language (https://knb.ecoinformatics.org/tools/eml)
CSW	OGC Catalogue Services Specification - **2.0.2** OGC 07-006r1
DCAT	Data Catalog Vocabulary (https://www.w3.org/TR/vocab-dcat-2/)
SoilML	ISO 28258:2013(en) Soil quality — Digital exchange of soil-related data

2.6. Recommended persistent formats for sharing, reuse and preservation of data

Given the diversity of scientific themes related to ecosystem studies, AnaEE's platforms produce different types of data, including raw data, geospatial data, images, soundtracks, videos and documents.

It is therefore necessary to list the formats (Table 6) used to move towards standardization in common and standard formats that can be used by the greatest number of people. If these formats are different, it is strongly encouraged to provide software tools to convert these files.

While allowing for uploading and treatment of any form of raw data coming from publishers, DMC fosters and has preference for structured formats in implementing its internal storage mechanisms, as they allow for better automatic treatment regarding collection, update retrieval and data quality verification.

In particular, Relational Databases allow for extraction of a portion of the dataset, not forcing potential users to download the whole of it. Also Cloud storage(e.g. Azure Blob Storage or Amazon S3) or NoSQL databases.

Moreover, the adoption of APIs to publish data would allow Research Infrastructure to have a layer of abstraction on top, permitting internal refactoring or information reorganization (e.g. migration from a RDBMS to a NoSQL architecture), not impacting the publication on the DMC Data Portal as long as the APIs contract is respected.

Table of recommended and accepted formats by data type:

Table 6: Recommended and accepted data format per data type.

Type of data	Recommended format	Other acceptable formats for data preservation
Quantitative tabular data with extensive metadata	XML; JSON	
Quantitative tabular data with minimal metadata		
Geospatial data	GML; netCDF; GeoJSON; **KMZ; GPX;**	SHP; GeoTIFF; OWS; WFS;
Qualitative data	XML; **SQL;**	**CSV; XLS; TM8; m38; n38; DAT; LOG; MDB**
Digital image data		Jpeg; PNG; BMP; TIFF;
Digital audio data		mp3
Digital video data		mp4
Documentation and scripts		

2.7. Data quality assurance

AnaEE's platforms provide long-term data (background) and project data (foreground). The data acquisition processes are very clearly defined for background data but are very specific for foreground data. The data quality process is monitored by the DMC according to the criteria and indicators defined in section 1.7 to ensure that all means have been implemented to produce data according to FAIR principles.
We will focus here on the quality of the data produced in terms of continuity and reliability of the measurement chain used, taking into account sensor drifts and missing data.
For this reason, a data quality index can be provided by the platforms for both types of data, provided that the quality criteria are clearly defined and that the research and platform teams provide the raw data as well as the various scripts for processing the data and calculating the quality index. Users will be able to apply the processing scripts according to the criteria that meet their expectations or they will be able to propose other criteria and other scripts to evaluate the quality of the data.
In addition, the events in the laboratory notebook related to data acquisition will have to be specified so that users will be able to identify the data disturbed by particular events (machine breakdowns, human interventions ...).
The AnaEE evaluation of dataset quality will be performed by using the endpoints provided, that are expected to host up-to-date data.

3. STORAGE AND BACKUP

In this section we will describe currently adopted practices for data storage management and requirements that research platforms are expected to meet in their data management activities. As for the previous technical sections, also this section is in continuous evolution and may be revised to fit the de facto standards in storage and backup solutions.
AnaEE acknowledges that cloud storage services can nowadays be considered a commodity and believes that this trend is not going to change in the near future due to existing market trends and initiatives such as the EOSC.
AnaEE encourages research platforms in preferring such technologies over on-premise data centers for a number of reasons that include, but are not limited to: scalability, security, replicability, resilience, and operational cost. AnaEE's platforms are responsible for ensuring the secure long-term storage of the data they collect. They can liaise with the DMC to help them implement appropriate solutions.. This data is then pushed at an appropriate frequency to the AnaEE data storage.
For each project, the research team and the platform managers open a closed-access collaborative data storage dedicated to the project. This allows for versioning of the dataset, secure backup and easy open access as soon as the embargo period is over. This collaborative storage is provided by the DMC that will only have to open the access to datasets collected during the project.

3.1. Persistent solutions

AnaEE DMC provide a long-term storage to centralize all the data production from AnanEE Platforms. This storage is a cloud storage (e.g. Azure Blob Storage or Amazon S3). To secure this backup, a mirror backup will be deployed at CC IN2P3 which guarantees a long term perennial backup.AnaEE divides persistence solutions into three categories:

- **File storages:** unstructured storage solutions that are used to store serialized files and are typically accessed with protocols such as ftp, sftp, or smb. May or may not have files arranged in a directory hierarchy.
- **AnaEE Data storage :**

- **Document databases**: loosely structured databases that may host a large number of semi structured documents in formats like JSON, XML, or CSV and usually provide an API that allows querying with SQL-like languages, optionally they can be accessed with an ad-hoc developed API that provides a more user-friendly facade. This solution is fit for IoT and application data.

3.2. Assurance of adequate storage capacity

The capacity needed to store data and metadata is assessed at the project design stage, taking into account the data formats generated and the frequency of acquisition sufficient to demonstrate the effects predicted by the scientific hypotheses.
It is up to the research platforms to coordinate the deployment of adequate storage solutions and to maintain them over time, sustaining all the related costs.
The AnaEE DMC hosts datasets on its cloud storage to grant its availability and preservation over time. More specifically, the DMC will store data on a cloud solution that grants replication, high fault tolerance, scalability, and allows to periodically create snapshots to prevent data loss in the unlikely event of catastrophic system failures.

3.3. Responsibilities for back-up and recovery

AnaEE assumes that each adhering platform has storage space sized according to the specific nature of their activity. As such, platforms are responsible also for replication and backup of their datasets. Platforms are expected to either demand these responsibilities to a third-party provider or provide secure storage space locally or adequate disaster recovery procedures that will allow for the reconstruction of data sets and their metadata in the unlikely event that data storage fails.
AnaEE's DMC ensures that the latest data sets are backed up on its storage space but cannot guarantee the recovery of data that has not been transferred to the storage space if the frequency of data transfers is not regular.

3.4. Risks and mitigations regarding data security, assurance to secured access

Access and security of the data sets produced by the AnaEE platforms is guaranteed by the DMC data storage. Of course, the security of these accesses is a strong criterion and is provided by a cloud storage (e.g. Azure Blob Storage or Amazon S3).
Data stored locally at the platform level must be secured without direct access to the external network.
It remains for the moment difficult to control and monitor the use of data sets that are shared freely. However, some projects for the development of tools to monitor the use of datasets initiated within the ENVRI-FAIR project, whom AnaEE is participating to, could be launched in the near future in order to monitor compliance with citation conditions or to evaluate the impact of a dataset on the scientific community by quantifying its reuse.

4. SELECTION AND PRESERVATION

4.1. Data to be retained or destroyed for contractual, legal, or regulatory purposes

The datasets generated by the AnaEE infrastructure will initially be provided in their entirety, i.e. the raw data, associated metadata, and the scripts for processing the raw data. As long as it is possible to maintain this permanent storage, it will be maintained for as long as possible. However, if this storage reaches a high cost for a low rate of reuse, an evaluation of the minimum dataset required for its exploitation will be set up in order to define the most optimal compromise between storage cost and interest of the dataset. The AnaEE Interface and Synthesis Centre may undertake

actions to valorize these datasets in order to verify that this dataset can be reused by other research fields or other public and private actors.
Some clashing needs could arise from a legal standpoint: for example, complying to GDPR may force DMC to grant to publishers the "right of oblivion", but deleting published data could compromise the persistence implied by the use of DOIs. These conflicts must be regulated by the legal part of the Consortium.

4.2. Foreseeable research uses for the data

It is difficult to assess the potential of a dataset at the time of publication. The AnaEE infrastructure will conduct campaigns to assess the impact of the datasets it produces on the scientific community. Indicators on the diversity of users within the scientific community, but also towards the public decision making spheres and private companies will allow to improve this impact and to carry out dissemination actions by the ISC to the actors who would not have understood the interest of the data produced by the AnaEE RI.

4.3. Long-term preservation plan

Preservation of the AnaEE infrastructure datasets is ensured by the DMC data storage and it mirror storage in the CC IN2P3.

5. DATA ACCESS AND SHARING

In this section we describe the AnaEE policy on accessing and sharing approved datasets and resources in general. The content of this section applies to datasets and resources that have been approved for publication by AnaEE.

5.1. Data publication workflow and policies

Once a dataset has met its quality requirements and has been approved for publication, AnaEE will assign it systematically a Persistent Identifier in the form of a Digital Object Identifier (DOI) as stated in the AnaEE Scientifical Technical Document. The proposing project will also be asked to provide a bibliographic reference to a companion paper or a technical report to perfect the publication information.
When both the DOI and the reference will be ready, the dataset will be published.
The dataset's owner may optionally negotiate with AnaEE a so-called grace period, i.e. a delay between approval and actual publication during which only the dataset's metadata will be visible to the community. The grace period request must be presented before publication to the revisers and the AnaEE central Hub and motivated. Acceptable motivations may include, but are not limited to:

- delays in the publication process of a companion paper or other dataset-related material;
- external organizational constraints imposed by third parties;
- pre-existing non-disclosure agreements with third parties.

Grace periods are meant to be exceptions and can last up to a year. After the end of the grace period the dataset will be published automatically in its integrity.
Datasets are published with the license specified in their corresponding project DMP, hence the responsibility for choosing an appropriate license relies on the proposing party described in the next 5.3 section.

Then, in order to allow findability and accessibility of the data production of the AnaEE infrastructure, the DMC provides a Data Portal to access the datasets produced during research projects.

This portal allows the scientific community and the general public (see Figure 2) to easily retrieve the datasets and associated publications thanks to discovery metadata. In addition, the data users will be able to find the associated cleaning, processing and analysis scripts as well as the models developed by the community. These scripts and models will be published in the AnaEE Developer Portal, which is also managed by the DMC.

5.2. Data publication tools

AnaEE encourages its research platforms to get their own data publication systems where possible. Recommended systems are CKAN, GeoNode, GeoNetwork, and in general all OSGEO supported products. Resources to be published on AnaEE services will then be linked and accessed through the original publisher's platform, preventing duplication issues and simplifying the enforcement of licenses and user access policies. Research platforms that don't have their own data publication system may upload their datasets into the services provided by the AnaEE DMC.
AnaEE will provide primarily two online data publication tools: the Data Portal and the API portal. Both systems are Web based, have federated user access with the AnaEE Identity Provider, and are maintained by the AnaEE DMC on its cloud space.
The data portal provides a Catalog of all datasets published by the RI, in the form of a Web Application. Such an application will allow the user to browse the catalog, search the catalog leveraging its metadata, access data and metadata, and preview the data. The same tool will also allow AnaEE platforms to manage dataset revisions and to keep track of the AnaEE evaluation process.
The Data Portal is built on top of the CKAN data management system and serializes its data in a data warehouse and a file storage hosted in the DMC cloud space. Various formats can be foreseen for uploading data (refer to relevant section in this document), and the underlying tools can provide some degree of automatic discovery of structure in said data to allow for map plots, chart plots, tabular views and querying. Metadata curation and upload is however the responsibility of the publishing platforms.
The API portal is built on top of MIcrosoft's API Management technology and provides a catalog of available Restful APIs developed by AnaEE and its partners, examples of such APIs may include, but are not limited to programmatic access to data warehouses, simulation models, and Artificial Intelligence models. The API portal allows its user to register to APIs and to request authorization keys to query the said APIs. Publishing a new API on the API portal requires a research platform to host all its components and to provide the DMC with an OpenApi 3.0 specification along with adequate credentials to access the API.
THe API portal is online at the following address: https://anaee-api-portal.developer.azure-api.net
Both the Data Portal and the API portal allow for advanced user authorization management: users can be organized in groups and have different individual or group privileges on the resources therein hosted. Resource access and visibility on these portals can be granted or restricted, depending on the use case, either by the resource owner or by the system administrator.

5.3. Data Licensing Policies

AnaEE encourages the adoption of liberal licenses that allow users to access the data and use it to develop new data products and hopefully contribute to the advancement of human knowledge. AnaEE encourages also third parties who may build new products on top of the published resources to share their results like AnaEE did with its data, hence the preferred license for AnaEE datasets is Creative Commons Attribution-Share Alike (CC BY-SA 3.0). AnaEE may reject dataset proposals that have license clauses that may hinder their fruition in the research community.
The AnaEE recognizes that for some third party stakeholders, the preferred license, as well as other liberal licenses, may be incompatible with their organizational and commercial constraints. In order not to impede this type of partnership, AnaEE therefore encourages its research platforms to also

adopt dual licensing or open licensing delay policies to meet the needs of the open science community and other organizations involved.

The project DMP must specify a publishing license for each candidate dataset. AnaEE expects the project AnaEE acknowledges four types of projects with different data policy requirements:

- **Academic observation**: initiatives mostly funded by AnaEE partners themselves that aim at creating long lasting resources whose main aim is expanding academic knowledge in general and that can serve multiple other research projects.
- **Academic projects**: deliverable-oriented projects supported by European, national, regional or local funds, these projects have a discrete lifespan and the deliverable licensing must meet the funding body's requirements.
- **Academic/Private consortium cooperation**: projects wherein private sector actors are involved either as work package units or funders and may impose some constraint on data publishing.
- **Private sector**: projects entirely funded by a private sector operator that may involve non disclosure agreements, hence limiting the possibility of publishing project data.

These four project stereotypes do not differ from each other in terms of AnaEE governance or general framework, but may produce artifacts with different licensing policies to meet their funder's requirements. The adoption of a non-liberal license, i.e. all rights reserved, for the produced datasets should therefore be supported by a brief report indicating the reasons why a liberal license is not viable.

AnaEE strongly supports Open Access publication in all of its forms, however AnaEE acknowledges that there exists a wide array of Open Data licenses, some of which are incompatible with each other.

AnaEE deems acceptable all the major open data licenses, namely: CC-0, CC-PDM, CC-BY-ND, CC-BY-NC-ND, CC-BY, CC-BY-SA, CC-BY-NC, CC-BY-NC-SA, OCD-PDDL, ODC-BY, ODC-ODbL, OGL 2.0, and OS OpenData.

It is important however to stress that these licenses are not equivalent, and picking an appropriate one is far from a trivial decision. These licenses have different features that are summarized in the following table:

	Permissions			**Requirements**					**Prohibitions**
License	Reproduction	Distribution	Derivative Works	Notice	Attribution	Share Alike	Copyleft	Lesser Copyleft	Non-Commercial
CC0	X	X	X						
CC-PDM	X	X	X						
CC-BY-ND	X	X		X	X				
CC-BY-NC-ND	X	X		X	X				X
CC-BY	X	X	X	X	X				
CC-BY-SA	X	X	X	X	X	X			
CC-BY-NC	X	X	X	X	X				X

CC-BY-NC-SA	X	X	X	X	X	X			X
ODC-PDDL	X	X	X						
ODC-BY	X	X	X	X	X				
ODC-ODbL	X	X	X	X	X	X			
OGL 2.0	X	X	X	X	X				
OS OpenData	X	X	X	X	X	?			

In general, AnaEE encourages the usage of the most liberal licenses, which are CC-0, CC-PDM, and ODC-PDDL. Adopting these liberal licenses allows AnaEE's end users to build and publish derivative products for research and business purposes as well, while less liberal licenses may prevent derivative products from being published with Open Data licenses or from being developed at all, hence making them not reusable. The following table shows the compatibility matrix between source data licenses and derivative product licenses.

	Permissible License for derivative												
Original License	CC0	CC-PDM	CC-BY-ND	CC-BY-NC-ND	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	ODC-PDDL	ODC-BY	ODC-ODbL	OGL 2.0	OS OpenData
CC0	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
CC-PDM	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
CC-BY-ND	-	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY-NC-ND	-	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY	N	N	Y	Y	Y	Y	Y	Y	N	Y?	Y	Y	Y
CC-BY-SA	N	N	N	N	N	Y	N	N	N	N	N	N	N
CC-BY-NC	N	N	N	Y	N	N	Y	Y	N	N	N	N	N
CC-BY-NC-SA	N	N	N	N	N	N	N	Y	N	N	N	N	N
ODC-PDDL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ODC-BY	N	N	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
ODC-ODbL	N	N	N	N	N	N	N	N	N	N	Y	N	N
OGL 2.0	N	N	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
OS OpenData	N	N	N?	N?	N?	Y	Y?	Y	N	Y?	Y	N	Y

In addition to these limitations, some licenses are not compatible between them, preventing the merging of data published under different licenses. For instance a data set published with CC-0 License cannot be integrated with data taken from a CC-BY-ND one, despite both of them being considered open data licenses. In other cases the merging of two data sets is not forbidden, but its reuse is limited by the less liberal license, forcing the author of the merged data to publish it under the less liberal of the two licenses. For the sake of clarity, we present in the following table the acceptable publishing licence for an hypothetical data product obtained by merging, with no further processing, two data sets.

	Second License												
First License	CC0	CC-PDM	CC-BY-ND	CC-BY-NC-ND	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	ODC-PDDL	ODC-BY	ODC-ODbL	OGL 2.0	OS OpenData
CC0	No restrictions	No restrictions	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	No restrictions	ODC-BY	ODC-ODbL	OGL 2.0	OS OpenData
CC-PDM	No restrictions	No restrictions	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	No restrictions	ODC-BY	ODC-ODbL	OGL 2.0	OS OpenData
CC-BY-ND	-	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY-NC-ND	-	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY	CC-BY	CC-BY	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	CC-BY	CC-BY	ODC-ODbL	CC-BY	OS OpenData
CC-BY-SA	CC-BY-SA	CC-BY-SA	-	-	CC-BY-SA	CC-BY-SA	-	-	CC-BY-SA	CC-BY-SA	ODC-ODbL	CC-BY-SA	OS OpenData
CC-BY-NC	CC-BY-NC	CC-BY-NC	-	-	CC-BY-NC	-	CC-BY-NC	CC-BY-NC-SA	CC-BY-NC	CC-BY-NC	-	CC-BY-NC	OS OpenData
CC-BY-NC-SA	CC-BY-NC-SA	CC-BY-NC-SA	-	-	CC-BY-NC-SA	-	CC-BY-NC-SA	CC-BY-NC-SA	CC-BY-NC-SA	CC-BY-NC-SA	-	CC-BY-NC-SA	OS OpenData
ODC-PDDL	No restrictions	No restrictions	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	No restrictions	ODC-BY	ODC-ODbL	OGL 2.0	OS OpenData
ODC-BY	ODC-BY	ODC-BY	-	-	ODC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	ODC-BY	ODC-BY	ODC-ODbL	ODC-ODbL	OS OpenData
ODC-ODbL	ODC-ODbL	ODC-ODbL	-	-	ODC-ODbL	ODC-ODbL	-	ODC-ODbL	ODC-ODbL	ODC-ODbL	ODC-ODbL	ODC-ODbL	OS OpenData
OGL 2.0	OGL 2.0	OGL 2.0	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	OGL 2.0	ODC-BY	ODC-ODbL	ODC-ODbL	OS OpenData
OS OpenData	OS Open Data	OS Open Data	-	-	OS OpenData	OS OpenData	?	?	OS OpenData	OS OpenData	?	OS OpenData	OS OpenData

The adoption of less liberal licenses such as CC-BY-ND, CC-BY-NC-ND, CC-BY-SA, CC-BY-NC, CC-BY-NC-SA, and ODC-ODbL limits the reusability of data and therefore prevents it to be considered truly FAIR. While AnaEE will accept these licenses, it will require proposing platforms to justify them with clear and well documented constraints. A third party can negotiate an embargo to keep a competitive and commercial advantage over products from the AnaEE Infrastructure but they will have to commit to making the data that generated these products public at the end of the negotiated embargo period. Different or bespoke licenses can be accepted if adequately supported by concrete motivations. Dual Licensing is acceptable and encouraged to maximize the impact of the published data.

5.4. User Access and authorization policies

AnaEE expects from adhering research platforms to be open towards user access federation and committed towards improving accessibility to its scientific data. Research platforms are also expected to manage their user accounts within an identity provider compliant with OpenID and LDAP standards.

AnaEE has its own Identity Provider compliant with LDAP and OpenID protocols. Such a system is used to manage user access to all services maintained by AnaEE at DMC level. Research platforms may request to federate their own Identity Providers with the AnaEE one to allow their users to authenticate on all AnaEE services with no need of registering additional accounts.

Research platforms can also use, under request to the DMC, to use the AnaEE Identity Provider as a login method for their services.

On top of user identification, AnaEE can also provide User Authorization to systems supporting the OAuth 2.0 protocol and accept authorization federation with OpenAuth 2.0 compliant authorities. This is meant to allow a finer and more robust control over user access to critical resources.