D11.1 BIODIVERSITY AND ECOSYSTEM SUBDOMAIN IMPLEMENTATION SHORT TERM PLAN

Work Package	WP11
Lead partner	CMCC
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	31/12/2019
Submission date	28/02/2020

Deliverable abstract

This deliverable is the first analysis of the FAIRness of the Biodiversity and Ecosystem RIs that serves as guideline for the activities in the project. At the end of ENVRI-FAIR a new analysis will be done in order to evaluate the improvements and results and define the development strategy after the project.

The document is also a short guideline for all the RIs in which the main characteristics of the different RIs are summarized and where strengths, weaknesses, difficulties, synergies, overlaps and available resources are reported. We prepared the document with the aim to facilitate cross-RIs collaborations and exchanges in order to invest resources to develop as much as possible tools and procedures that can be used by multiple RIs.

In addition, two key sections have been developed in order to facilitate the analysis of the Biodiversity and Ecosystems RIs and to propose practical examples of collaboration and synergies:

- a description of the core competences of the different RIs that is useful to clearly identify the main expertise and competencies that other RIs can use and consult,
- the definition of common parameters, quantities or properties that the different RIs collect/measure/manage and that can be used as example of a practical collaboration.

ENVRI-FAIR (www.envri-fair.eu) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824068

DELIVERY SLIP

	Name	Partner Organization	Date
Main Author	Dario Papale	CMCC	
Contributing Authors	All WP11 participants		
Reviewer(s)	Keith Jeffery Zhiming Zhao		
Approver	Andreas Petzold	FZJ	28/02/2020

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	10/01/2020	First version for review	D. Papale
V 1.0	11/02/2020	Consolidated version	D. Papale

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the Project Manager at manager@envri-fair.eu and the main author at darpap@unitus.it.

GLOSSARY

A relevant project glossary is included in Appendix A. The latest version of the master list of the glossary is available at http://doi.org/10.5281/zenodo.3465753.

PROJECT SUMMARY

ENVRI-FAIR is the connection of the ESFRI Cluster of Environmental Research Infrastructures (ENVRI) to the European Open Science Cloud (EOSC). Participating research infrastructures (RI) of the environmental domain cover the subdomains Atmosphere, Marine, Solid Earth and Biodiversity / Ecosystems and thus the Earth system in its full complexity.

The overarching goal is that at the end of the proposed project, all participating RIs have built a set of FAIR data services which enhances the efficiency and productivity of researchers, supports innovation, enables data- and knowledge-based decisions and connects the ENVRI Cluster to the EOSC.

This goal is reached by: (1) well defined community policies and standards on all steps of the data life cycle, aligned with the wider European policies, as well as with international developments; (2) each participating RI will have sustainable, transparent and auditable data services, for each step of data life cycle, compliant to the FAIR principles. (3) the focus of the proposed work is put on the implementation of prototypes for testing pre-production services at each RI; the catalogue of prepared services is defined for each RI independently, depending on the maturity of the involved RIs; (4) the complete set of thematic data services and tools provided by the ENVRI cluster is exposed under the EOSC catalogue of services.

TABLE OF CONTENTS

BIODIVERSITY AND ECOSYSTEM SUBDOMAIN IMPLEMENTATION SHORT TERM PLAN

Short description of RIs involved and their focuses

The Biodiversity and Ecosystem (B&E) sub-domain is composed of seven RIs: AnaEE, DANUBIUS-RI, DiSSCo, eLTER, ICOS ERIC, LifeWatch ERIC and SIOS. They are characterized by clear niches, as core research focus or approach, with relatively small overlaps among RIs and high potential complementarity. This high diversity of scientific focuses and approaches and an even much larger diversity of the variables of interest of the RIs are surely a point of strength of the B&E subdomain, while introducing complexity when addressing specific variables of interest for the development of demonstration cases of data FAIRness. The different level of maturity of the RIs, ranging from operational ones to just entered in the roadmap, has contributed to increase the level of complexity, which has required effort and time in order to select common grounds of activities and produce demonstration cases of data FAIRness through integrative approaches of the seven RIs to address specific research questions.
A short description of the current state of the RIs involved in the Ecosystem and Biodiversity WP is provided, highlighting the focuses and organization of the data cycle. This information will be the basis for the activities in Task 11.3 to improve interoperability across RIs.

AnaEE

The Analysis and Experimentation on Ecosystems RI (AnaEE) provides to the scientific community a panel of distributed experimental, analytical and modeling platforms, focusing primarily on the impact of global change on functioning and productivity of continental ecosystems. The experimental platforms (from enclosed to open-air sites), distributed across different European climatic zones, soils and ecosystem types, including agroecosystems, allow testing climatic and management scenarios. Data and models derived from the AnaEE platforms allow the development of agricultural and ecological innovations ensuring food security, ecosystem service provision and contribution to bioeconomy.
In addition to the experimentation, analytical and digital services provided by the AnaEE platforms, transverse services will be offered by the AnaEE RI through three thematic centres, coordinated by a Central HUB: Technology Centre (TC), Data & Modeling Centre (DMC), Interface and Synthesis Centre (ISC). The objectives of these services are:

- developing the visibility and international use of platforms (Central Hub),
- improving the technological and experimental capacities of platforms, by harmonising and upgrading instruments and procedures, setting-up advisory groups of internationally recognized specialists, organizing workshops, developing innovative technology, training on new technology and procedures (TC),
- improving accessibility and interoperability of data, by coordinating the data management of AnaEE platforms, providing platforms with standardized procedures/tools for data processing and quality control, generating standard metadata indexed in an interoperable catalogue, developing semantics and offering a unique discovery portal for data exposure (DMC),
- stimulating modeling activities by establishing links between data and models and facilitating the access to a range of ecosystem models and model factories (DMC),
- establishing links with scientists and stakeholders, as well as in lobbying for new research programs, and in transfer activities and delivering scientific syntheses (ISC).

Data cycle description

The AnaEE data management strategy relies on a distributed architecture where the AnaEE platforms, through their Information Systems (IS), will contribute to feed a unique AnaEE portal for discovering data and model resources, herein the Portal, with standardized metadata. Each distributed IS will have to guarantee the security of its own system and the persistence of the managed data as well as data curation through common processes agreed with the DMC and the partners.
In the distributed infrastructure, data will go through two steps: data ingestion and curation at the platform level, data set publication.

Data ingestion and curation. The platforms acquire various datasets: continuous routine measurements of environmental conditions in the experimental platforms, project specific measurements and simulation data generated by models. These data are managed by ISs, which can be either developed at the platform or at a higher level (e.g. national). These ISs are responsible for data curation and facilitate exchange between platform manager and users. Data are documented with metadata, including elements needed to capture provenance and attribution information. The ISO19115/19139 INSPIRE compliant metadata standard is the minimum requirement for discovery characterization of AnaEE resources (data and services). The production of exploitation metadata following the EML (Ecological Metadata Language) standard is also recommended. Metadata production is done using a semantic annotation based on an ontology and a controlled vocabularies. The semantic approach is similar to both database and modelling resources to facilitate data/model interoperability as well as with other related infrastructures by the sharing and alignment of semantic standards as that of Ecosystem and Biodiversity sub-cluster of ENVRI-RIs.

Publication: Data sets will be extracted from the ISs using workflows able to generate data files on defined parameters (sites and experimental factors, variables, years, etc.) and standardized metadata (ISO and GeoDCAT) for their discovery and exploitation. Data sets will be identified by persistent identifiers (URIs, DOI, PID) and stored in open repositories hosted either by the data producer institutions or the DMC when institutions prefer to delegate this task to the DMC.

DMC will have the responsibility in managing semantic and quality standards. Metadata produced by AnaEE platforms will be harvested and managed by the DMC, in close cooperation with the AnaEE platforms to avoid duplication of effort. DMC is now developing the discovery portal providing its users with a search engine that will improve data findability and workflows to extract data files from the different ISs.

DANUBIUS-RI

DANUBIUS-RI – the International Centre for Advanced Studies on River-Sea Systems - will be a pan-European RI dedicated to interdisciplinary studies of the whole system from the source of the river to the coastal waters, with a focus on transitional zones such as estuaries, deltas and lagoons. It will support research addressing the conflicts between society's demands, environmental change and environmental protection in river–sea systems worldwide. Surface waters have a key role in global biogeochemical cycles, food and energy production, and societal wellbeing. They face pressures from natural and anthropogenic driven environmental perturbations at local and global scales.
European research is world-leading but fragmented, largely discipline-specific and often geographically isolated. The lack of interdisciplinary RIs has fuelled this fragmentation. DANUBIUS-RI will fill the gap, drawing on existing research excellence across Europe and enhancing the impact of European research. It will provide access to a range of European river-sea systems, facilities and expertise; a 'one-stop shop' for knowledge exchange in managing river-sea systems; access to harmonised data; and a platform for interdisciplinary research, inspiration, education and training.
A Science and Innovation Agenda, which is in an advanced stage of development, identifies four key research priorities to guide the development of the RI to implementation and

operation: climate change; water sufficiency; sediments and their management; and ecosystem health. They will be open to revision as the RI matures, and the aspiration is that DANUBIUS-RI will provide the facilities to advance interdisciplinary research on RS systems through collaboration between individual academic domains, ranging from the natural sciences to the social and economic sciences.
The architecture will include a Hub and a Data Centre (Romania), four subject-specific Nodes (led from the UK, Germany, Italy and Netherlands), a Technology Transfer Office (Ireland), an e-Learning Office (Spain) and Supersites at locations of high scientific importance and opportunity across Europe. The DANUBIUS Commons will drive the implementation of standardised procedures and quality control.
We are developing links with other European RIs for mutual benefit as well as to avoid duplication or overlap, including use of individual Supersites for observations and measurements by other environmental RIs.

Data cycle description

Data providers are all DANUBIUS-RI components: Observation Node, the Hub, Supersites, Analysis Node, Modelling Node, Impact Node and Technology Transfer Office. At the request of an authorized user, the resources for the data ingestion process are made available by the Data Centre.
These resources should be a specific application (to be developed) in the DANUBIUS cloud dedicated to data check and validation for long time preservation. This application should also assign all necessary metadata using metadata registry services to comply FAIR principles.
To the extent that the measuring devices (sensors) can be directly connected to the data collection application, user intervention will be minimal. Data files resulting from measurements or results of various processing, simulations or modelling processes are transferred to the Data Centre's computing system when submitted to the ingestion process, thus the possibility of data loss being minimized, data being safe, the degree of availability being the same as for all the information in the Data Centre. Once the data ingest application started, it will identify the type/context of data provided, ensure that the associated metadata are verified/ assigned in accordance with the standard requirements, thereby providing a primary data check. It is desirable that if the data is not compliant (e.g. lack of metadata), the application allows the user to manually add or select necessary metadata provided in accordance with the information from the metadata registry.
This application must contain procedures for all types of data or information provided by users. It is likely that this application, which will be adapted and developed over time, will require the use of artificial intelligence to ensure optimization of the process.
Thus, a series of procedures can be automated allowing for processing a large amount of data, while ensuring the correct assignment of metadata for each data set. In this procedure, for example, it is sufficient for the application to add information of the same type previously added to similar data, possibly from the same data provider, by using the results of a learning process. After this process, the data will be stored on the Data Centre systems in a temporary area. Once data gets in this area it will trigger a signal for data management team from the Data Centre in order to validate and include them in the data repository. For compliance it is necessary to check the data content at this level.

DiSSCo

The Distributed System of Scientific Collections is the new world-class Research Infrastructure (RI) for natural science collections, representing the largest ever formal agreement between natural history museums, botanical gardens, and collection-holding universities in the world.
Its mission is enabling Data-intensive frontier research through unified access to European natural science collections information and expertise.
The DiSSCo RI works for the digital unification of all European natural science assets under common curation and access policies and practices that aim to make the data easily Findable, more Accessible, Interoperable and Reusable (FAIR).
The value and impacts of the natural science collections assets are beyond question, but still, there is a gap between collection-based research and applied research in different domains

(human health, environmental challenges, food security, regional security, etc.). DiSSCo will enable the transformation of a fragmented landscape of the crucial natural science collections into an integrated knowledge base that provides interconnected hard evidence on the natural world, thereby connecting fundamental and applied research in response to Grand Challenges.
The scientific and technical approach to the DiSSCo mission can be very briefly summarised; DiSSCo sets the physical objects, the specimens, at the epicentre of the development of a robust, quality ensured and fit-for-purpose knowledge base for bio- and geo-diversity. This way, DiSSCo aims at putting Natural Science Collections-derived information at the very core of data-intensive bio- and geodiversity sciences.
By building the required economies of scale (i.e. the pool of resources to improve overall effectiveness), DiSSCo will significantly improve the role of Natural Science Collections in frontier scientific research. Furthermore, organisations will benefit from being able to better understand, describe and monitor the impact of their collections data in different scientific disciplines. Finally, organisations will be able to develop their specialisation and prioritisation strategies, within the wider DiSSCo community and in alignment with national primacies (e.g. Smart Specialisation Strategies), as well as developing and harmonising common research and innovation agendas.

Data cycle description

The DiSSCo data lifecycle is structured in five phases: data acquisition, data curation, data publishing, data processing and data use. All activities, applications, services and software tools of the DiSSCo infrastructure will be designed and implemented around these five major data lifecycle activities. For example, activities associated with digitisation of collections are a) pre-digitisation curation; b) imaging station(s) setup; c) imaging; [specimen] conservation; e) image processing; f) image archiving; g) optical character recognition; h) manual data entry and correction; i) data transcription; and, j) data publishing.
The data lifecycle begins with the acquisition of data through the principal activity of digitisation of physical specimens by a network of dedicated digitisation lines/factories. Digital specimen data is registered and curated within a data store(s) belonging to the DiSSCo infrastructure, either located in one of the DiSSCo Facilities or located as part of the DiSSCo Core. Curated data is published to DiSSCo users and parties external to the infrastructure, as well as directly to other services within DiSSCo. This sequence results in a natural partitioning into data acquisition, data curation and data publishing. In addition, DiSSCo provides services for further processing of data that can produce new data to be stored within the infrastructure. Finally, the broader research community uses DiSSCo data and can design digital experiments and analyses acting on the published data that produce results (new data), which in turn can be passed back into DiSSCo for curation, publishing and processing; thus restarting the lifecycle.

eLTER

The integrated European Long-Term Ecosystem, critical zone and socio-ecological systems Research Infrastructure (eLTER RI) is a new distributed Research Infrastructure which aims at integrating disciplinary natural sciences and holistic socio-ecological research approaches, to better understand ecosystem functioning across Europe´s environmental and socio-economic gradients. The focus is on making intensive use of the selected sites and platforms in terrestrial, freshwater and coastal ecosystems, for studying ecosystem responses, at multiple scales, to globally identified challenges related to ecosystem integrity, impact of climate change and endangered ecosystem services at a pan-European scale.
The objectives of the eLTER RI are to:

1. to advance access to a wide range of tailored, state-of-the-art sites and to their current and legacy data in a coordinated manner, in order to promote their capacities for service provisioning and accessibility to new users;

2. to foster and facilitate the development and uptake of harmonised protocols and methodologies for Findable, Accessible, Interoperable & Re-usable (FAIR) data services on Essential Ecosystem Variables (EEVs), and to add value to them through complementarity and synergies between other environmental research infrastructures and networks, e.g. by piloting co-location and co-design;
3. to enhance community building on ecosystem research and advance capacity building, support for innovations and improved technologies, by training site managers and researchers and engaging diverse stakeholders including industry, students, local residents and citizen scientists;
4. to provide national and European decision makers, natural resource managers and land use planners with quality controlled information tools for detecting the long-term efficiency of policies related to EU Directives and SDGs, e.g. related to biodiversity, agriculture, sustainable forest management, water and environmental quality, and to optimise the societal and policy relevance of eLTER RI;
5. to develop and disseminate novel methods for integrating ecosystem, critical zone, and socio-ecological research at Pan-European, *in situ* research sites, and to integrate *in situ* measurements and remote sensing observations.

Data cycle description

The implementation of the eLTER data cycles is in development and with the implementation of the eLTER RI more tools to support these data cycles will developed. Responsible parties for the data cycle are the data provider on the one side and eLTER as data infrastructure on the other. Currently quality controlled checked data are ingested into the eLTER Information System and provided through standardised service interfaces. The main steps in the eLTER data cycle are:

- Acquisition (data provider) - Data acquisition is conducted at the specific LTER sites and LTSER platforms either done continuously (e.g. continuous sampling using sensor and automated data acquisition systems), regular (e.g. sample based data acquisition of soil water chemistry) or periodically (e.g. recording of species frequency in vegetation plots by experts). Depending on the type of data collected different collection methods are applied. Based on the data type also the structure and format of the data might vary. eLTER aims to implement widely used community standards to ensure interoperability. The responsibility for data acquisition is with the data collector or data provider.
- Curation (data provider / eLTER RI) - Data curation procedures are split into different subprocesses, which either run on the data premises of the data provider or by the central eLTER data infrastructure. Data curation includes the following activities:
    - documentation of the research sites and issuing a deims.id for each research site as unique persistent for the generic observation location using DEIMS-SDR
    - conducting quality assurance procedures on the level of the acquired data and generate a quality controlled dataset to be shared with the community
    - documentation of the dataset using either local metadata infrastructure or DEIMS-SDR as centrally provided metadata infrastructure. Common metadata schemes need to the applied.
    - conducting basic semantic quality control using EnvThes to annotate metadata and variable naming in the datasets
    - definition of the data policy and data license if not defined centrally
- Processing (data provider / eLTER RI) - Generation of aggregated datasets following a common specification (e.g. daily means, daily sums). These additional datasets need to be documented on the common metadata infrastructure DEIMS-SDR.
- Publishing (eLTER RI) - Publishing of the datasets using B2SHARE (file based datasets, issuing PID/DOI) or the eLTER CDN (time series data). Here only quality controlled data are provided to the end user.
- Re-use (data user) - Any data user can discover (eLTER DIP) and access data on certain aggregation levels to conduct data analysis.

ICOS ERIC – Ecosystem component

The ICOS Ecosystem component is a distributed infrastructure with currently 80 stations measuring mainly the greenhouse gases exchange between ecosystems and atmosphere using the eddy covariance technique. Together with the ecosystem fluxes also micrometeorological parameters are monitored in the air and in the soil and ancillary data such green area index, biomass, soil C and N content, leaves nutrients also collected.
The main focus of ICOS Ecosystem is to monitor the role and effect on ecosystems in relation to climate variability and climate change, with an across-domains prospective since ICOS includes also an Atmospheric and a Marine component.
Ecosystem stations in ICOS are evaluated during a labelling process in order to ensure and verify the needed quality and robustness of the setup. This is done in collaboration with the Ecosystem Thematic Centre (ETC) that is the central facility coordinating the ecosystem component in ICOS. The ETC is located in Italy, Belgium and France and is also responsible for the processing of the data and definition of standards. All is done in collaboration with the thematic centres of the other ICOS components (ATC and OTC) and in coordination with the ICOS Head Office and ICOS Carbon Portal.

Data cycle description

The ICOS ecosystem stations collect data from multiple sensors (between 50 and 100) at frequencies in the range between 1 and 10 Hz. All these data are submitted in Near Real Time in half-hourly packages directly from the loggers to the ICOS Carbon Portal using cUrl commands and collected into daily data packages (the day after acquisition). The daily packages are registered at Carbon Portal with a PID and copied through the B2SAFE service as trusted repository. The metadata about the sensors, including their installation, calibration, management, disturbance and link to the dataset created and logger and file where they are stored are submitted to the Ecosystem Thematic Centre using a specific system named BADM (Biological, Ancillary, Disturbance and Metadata). The BADM is an international standard used in FLUXNET and developed in collaboration with AmeriFlux, the large network in the Americas that collect measurements comparable with ICOS Ecosystem. The main characteristic of the BADM system is the possibility to collect, store and distribute in a machine-readable format all the specific information needed to process the raw data and prepare the needed metadata.
The raw data and related metadata are then processed centrally by the ETC that access the raw data directly through the PID through the Carbon Portal. The products (NRT fluxes and yearly datasets) are processed in different clusters and then distributed by the ICOS Carbon Portal that translate also the relevant metadata from the BADM structure to the relevant global metadata standards. The data are directly accessible under an open CC4BY license.
Also the ancillary data are collected directly using the BADM system and where available the raw data. Raw data submitted to the ETC or Carbon Portal are the Digital Hemispherical Pictures and ceptometer measurements to calculate the green area index, the phenocam images to track phenology. The soil and vegetation samples are analysed in the ETC laboratories in France and stored for 20 years.

LifeWatch ERIC

LifeWatch ERIC is the eScience European Infrastructure for Biodiversity and Ecosystem Research, representing one of Europe's main line of response to biodiversity loss. It applies advancing ICT technologies, web networks and interconnects scientific communities and research centres internationally through its web-based research infrastructure.
LifeWatch ERIC: i. Provides access to data collected at a global level, ICT services, tools, storage capacity and computational power; ii. Connects physical observatories, research centres & scientific communities into a single web space accessible to all; iii. Offers researchers & stakeholders the resources to enact their own innovative scientific approach; iv. Empowers citizens to engage with science and contribute to their own well-being.
LifeWatch virtual research environments (VREs) capabilities facilitate the sharing and aggregation of data on biodiversity and ecosystems, their integration and analysis in advanced models, and provide the computational power to test scenarios of biodiversity organisation and conservation in future ecosystems under multiple drivers of change.

LifeWatch ERIC is building facilities to address these major challenges in all different ecological domains (i.e., terrestrial, freshwater and marine) and ecosystem types of the Biosphere according to strategic construction plans and co-construction actions directly developed with its stakeholders in order to fit their needs.

LifeWatch ERIC is a distributed and federated Infrastructure, with Common Facilities and Federated Thematic Transnational/National/Regional Centres. Common Facilities are: i. Statutory Seat and ICT-core, ii. Service Centre, iii. Virtual Lab and Innovation Centre.

Data cycle description

LifeWatch ERIC was originally not considered to be developed as a data repository but, rather, as an e-infrastructure to support the work of the scientists in the data discovery and data analysis.
In fact, and according to The Global Biodiversity Informatics Outlook (GBIO) Framework (Hobern et al. 2012) which identifies 20 components as essential elements of biodiversity informatics and organized as four layers: Culture, Data, Evidence and Understanding, LifeWatch ERIC is mostly allocated in the understanding with the main aim of building modeled representations of biodiversity patterns and properties, utilising any possible evidence, based on the following components: Multiscale species modelling; Trends and predictions; Modelling biological systems; Visualization and dissemination; and Prioritizing new data capture.
Therefore, it offers lots of such services to its users: annotating, managing, curating, storing, versioning, mapping and publishing of data. The (meta)data catalogues of LifeWatch ERIC collect and harvest information coming from all the LifeWatch National nodes, from the external RIs and from different Global Species Databases like GBIF, OBIS, Catalogues of Life, etc.
The metadata adopted in the different catalogues has been selected and agreed after an in-depth analysis process. For the dataset LifeWatch ERIC has adopted a specific EML 2.2.0 profile, for the Services an ISO19139/119 profile and a LW profile for the VREs has been created. All the metadata refer to specific Controlled Vocabularies developed or selected to assure the harmonization of the information.
The collected information has to pass through an appropriate curation process in order to avoid publishing (meta)data with low quality. LifeWatch ERIC in this sense uses different services for the syntactic, taxonomic and semantic checks of the (meta)data. The same services are supplied and made available in the Virtual Research Environments to maintain the data quality of LifeWatch ERIC users.
Only after the checks are passed, the (meta)data are published and are made accessible through the catalogues. The user can browse the information using different kinds of searches (free, with maps, semantical, etc..) and the same functionalities are available also with REST interfaces. Different output formats are available (xml, Json, etc.).
LifeWatch ERIC supplies different types of Virtual Research Environments in order to support the needs for analysis. Each VRE offers an Orchestrator that can combine different data types and computational services in a specific workflow. It is also providing distributed resources federating tools based on Blockchain technologies for data citation, provenance, accountability and ecosystem services valorization. that is the case of the recently launched so-called LifeBlock platform.

SIOS - Ecosystem component

SIOS is a regional observing system for long-term measurements in and around Svalbard addressing Earth System Science questions. SIOS integrates the existing distributed observational infrastructure and generates added value for all partners beyond what their individual capacities can provide. The observing system and research facilities offered by SIOS build on the extensive observation capacity and diverse world-class research infrastructure provided by many institutions already established in Svalbard.
The SIOS Data Management System (SDMS) relies on the principles of distributed data management. Datasets that are relevant for SIOS, as well as their associated discovery and

use metadata, are managed by several physically distributed Data Centres. Each Data Centre has its own set of Data Management Facilities for ingestion of new data (and associated metadata), maintenance of the data sets (including metadata) and for data curation.
The SDMS service will enable data submission, discovery, access, and use and preservation of SIOS relevant data sets and metadata across these physically distributed data centres.
There is no specific ecosystem component. However, biodiversity data are collected from different platforms for integration with other data in earth system studies.

Data cycle description

All data generated using the SIOS infrastructure shall be curated by a SIOS contributing data centre. Contributing data centres shall ensure proper life cycle management of datasets made available for SIOS. The actual cycle will depend on the type of data, mainly if the data are experimental or included in a long-term time series. In the latter case the data will go from acquisition through a quality control and validation procedure and reformatting before being registered in the data catalogue, issued with a DOI and published. The published metadata will then be harvested from an OAI-PMH metadata provider by the SIOS Data Management System (SDMS). The users can search for, evaluate and access the data through the SDMS search interface.

Current situation with respect to FAIRness and priorities

The RIs involved in the Ecosystem and Biodiversity subdomain have different levels of maturity in terms of the FAIRness of their data, metadata and processes. A first evaluation of the general FAIRness has been conducted, in collaboration with WP-7, in order to identify the strength and the weakness for each RI and set priorities in the developments that should be addressed in Task 13.3.

AnaEE

Until now, AnaEE is not compliant with FAIR principles.
Activities to improve Data FAIRness are ongoing, leveraging the following assets:

- Semantic tools set up by AnaEE within ENVRIplus;
- Cloud computing and storage infrastructure;
- Software development and system engineering advanced capabilities.

The main issue is related to the need for an updated and comprehensive state of the art of the AnaEE platforms. In this perspective, a preliminary survey on platforms data management has been prepared, but it is still not possible to administrate the online questionnaire, because of the lack of an official list of platforms correspondents. A second issue will be mobilization of data producers for metadata production and data set publication. This requires to raise awareness of data producers on FAIR principles and provide them user-friendly annotation tools and data publication procedures.

Strength in FAIRness

1) Competences over semantic vocabularies:
   - a shared AnaEE thesaurus has already been defined, which aims to provide a controlled vocabulary for the semantic description of the study of continental ecosystems and their biodiversity. It has been developed within the framework of the AnaEE-France infrastructure. Metadata processing pipelines are available as well, supporting automated semantic processing. These tools, set up within ENVRIplus, allow annotation of BDD and RDF data production and exploitation of this data for the generation of standardized metadata and data files (F,I).
   - an OBOE-based ontology used for semantic annotation of AnaEE data and modelling resources (F,I,R)
2) Availability of a powerful cloud infrastructure that allows data and services to be online 24/7 with geographical redundancy and scalable performances. Such an infrastructure will greatly support availability of data and services. (A)
3) Competences over software engineering and data publication. The DMC can offer its technological competences to the whole network to better engineer data and services into more reusable and interoperable components. In particular, the DMC has expertise on building Web APIs, multiplatform applications, and cloud-based services and micro-services. Furthermore, members of AnaEE have been involved in the development of other platforms, such as BioMA, and have extensive experience on building Interoperable and Reusable assets. (A,I)

Weakness in FAIRness

1) The data catalogue is still to be completed: resources right now are distributed among the various research platforms that form AnaEE. An extensive surveying activity to identify candidate resources for publishing is currently ongoing. Such an activity is, however, taking a long time and the DMC cannot enforce deadlines over the platforms. This organizational bottleneck is slowing down the data FAIRness plan and hindering the decision-making process. AnaEE needs to establish internal communication best practices to address this issue (F)
2) Need for the identification of the key metadata elements for provenance management and for shaping of the workflow to be used for their production. (R)

3) The development of the AnaEE Portal is ongoing: although the DMC is actively working on the AnaEE Portal, its release will take time; a preliminary release is expected in 2020, and until then data sets won't be accessible from other sources than the Information Systems currently maintained by individual platforms. (F,A)

Priority plan for the FAIRness improvement (next 2 years)

1) Full release of the AnaEE Portal: the DMC is currently developing the Portal using a stack of open source technologies. This activity will lead to a release in 2020 and a first version of the Portal will be deployed in the cloud. The Portal will provide users with a search engine to find data sets, multimodal access to data and metadata, and a set of indices to evaluate how compliant to FAIR principles the adhering research platforms are. (A,F)
2) Full survey of in-house resources: despite having identified several candidate resources for FAIR publication, the internal survey is far from being completed due to the aforementioned difficulties of organization and coordination. Over the next months we expect all research platforms to comply with the DMC requests. (F)
3) Implementation of the pipelines developed by AnaEE- France, starting from a first use case, selected from the datasets of the AnaEE platforms. Implementation has already been carried for some AnaEE-France and Italy databases and has been successfully tested on the Italian soil database. The data produced by the AnaEE platforms are managed in local information systems without prior (or with low) international standardization among platforms. The annotation of these data, using a common OBOE-based ontology will allow the production of homogenized semantically interoperable information (RDF triples) to be re-used by the AnaEE portal feeding and for the metadata and data set production. The process relies on the graph modeling of the observation/measurement (the observed entity (i.e.: soil sample), the characteristics (i.e.: pH), the value, unit, protocol) and of the experimentation context (location of the site, date of observation, treatments….). The implementation of the semantic pipelines will mostly consist (in addition to the declaration of the database access) of developing the semantic graph that models the observation and matching the graph with the database. The process is largely facilitated by the re-use of generic graph patterns. (F,I,R)

DANUBIUS-RI

DANUBIUS-RI is still at the beginning of the implementation phase – this means that there is still a lot of flexibility to implement FAIR principles. As a new infrastructure, it is easy to build the FAIR principle directly into the DANUBIUS Commons and ensure the enforcement of these principles, and there is the possibility possible to benefit by the experience from other RIs and ERICs with respect to FAIR principles. Examples of good practice as well as Standards that are being adopted by many of the other ERICs will be integrated and uptaken in the new infrastructure. In this way erroneous choices can be avoided by lesson learned. The following strengths and weakness are identified for the planned DANUBIUS-RI. Since DANUBIUS-RI is not yet operational we will describe below the strengths and weaknesses according to the projected deployment of the project.

Strength in FAIRness

1) Using the best practices of the partners in ENVRI-FAIR as well as other projects of similar magnitude will lead to the best possible structure of the data access/storage platform and data policy in general thus leading to compliance to FAIRness.
2) All services will be accessible through the DANUBIUS-RI portal, except for data ingestion from DANUBIUS-RI components. Data check/validation, LTP and FAIR considered on the data flow (F,A);
3) Data formats considered in DANUBIUS-RI have to be non-proprietary, unencrypted, uncompressed, open, documented using community standard, use common encodings (e.g. ASCII, Unicode) (A,I,R);

4) Data structure considered: raster data (as matrix), vector data provide for representation within a geographical information system (GIS) environment, gridded or array-oriented data using NetCDF libraries, hierarchical data (HDF formats), times series data (A,I,R)
5) DANUBIUS-RI standard considered: DATA repository will have to comply with the DANUBIUS-RI Commons and the INSPIRE guidelines, as well as other involved standards from ISO, Spatial Data Transfer Standard (SDTS), metadata standards (I, R).

Weakness in FAIRness

1) DANUBIUS-RI still does not have a complete overview of the working data flow. At this stage, there is still no data available, and most of the protocols need to be defined. It is therefore difficult to see where important problems in FAIR adoption could surface.
2) DANUBIUS Commons are still not finally defined and needed refinement beyond the preparatory phase, well into the implementation phase. The existing data already accumulated at DANUBIUS-RI consortium partners might not fulfil the FAIR principles.
3) The Modelling Node has to deal with data products connected with river sea systems domain. This can imply the use of different data formats from what, most likely, can be found in the other RIs structures (e.g. results of numerical models on unstructured grids). These mean that special treatment, new solutions and standards should be found and adopted. The specification of how data products have to be in line with the FAIR principles should be addressed.

Priority plan for the FAIRness improvement (next 2 years)

The DANUBIUS-RI in being in implementation phase, FAIRness improvement process will overlap with the development and implementation of the data services. In this respect, the proposed plan for FAIRness implementation is:

1) The first step would be to disseminate to all the members of the DANUBIUS-RI community the FAIR principles and make them aware of the advantages of implementing/complying to these principles on all data processing stages. As a result of this step, a team of trainers should be built to train the trainers in all DANUBIUS-RI components, making possible that the knowledge about the tools to implement FAIR principles shall reach all the researchers and users.
2) The next step is to define and agree, within the DANUBIUS-RI community and in compliance with ENVRI-FAIR outcomes, the minimum conditions of the Data Management Plan required to research projects carried out using the DANUBIUS-RI services.
3) From a technical point of view, DANUBIUS-RI will assemble a support team of experienced data curators from the staff of the Nodes and Supersites of the DANUBIUS-RI consortium. This team should work with the developers to write the requirements for the necessary software tools to manage the DANUBIUS-RI data types and metadata, in order to comply with the FAIR policies. These requirements should be used to develop/choose the tools for data ingestion, storage and sharing services.
4) This process will be followed by the development and installation of the software infrastructure in the testing phase; the support team (curators and developers) will stress the infrastructure in terms of compliance with the DANUBIUS-RI Data Policy and FAIR principles. The result will be a list of strong and weak points. This information will be communicated to all the DANUBIUS-RI components, get feedback, repeat the process to achieve the best solution for all Nodes, Supersites and users.
5) DANUBIUS-RI is willing to sustain the initiation of a common metadata repository for the biodiversity and ecosystem domains (and not only), which could lead to a new level of the common view for data harmonisation

DiSSCo

Within the ICEDIG project (a DiSSCo-linked project) a provisional DiSSCo Data Management Plan has been developed which is currently being reviewed by an external advisory board. Later it will also be sent for review by other stakeholders such as community expert groups, relevant project working groups, and pertinent international infrastructure technical teams. All comments and feedback will be incorporated into a single document to guide the development of DiSSCo infrastructure and services. The Data Management Plan is a living document reflecting the active data management planning and stewardship philosophy adopted by DiSSCo, with focus on achieving maximum openness and reusability of data, longevity of data and data preservation, and reproducible science. All sections contribute towards achieving and sustaining FAIRness, meaning that compliance with the procedures and rules set out in the plan must be maintained by DiSSCo and its contributing members throughout the lifetime of the infrastructure and throughout the lifecycle of identifiable data assets managed by the infrastructure.
Currently we do not have any data production system in place. The strengths and weaknesses will be then based on the future outcome of the current design and planning discussions in place directed by the Data Management Plan.

Strength in FAIRness

1) A distinct globally unique, persistent and resolvable identifier (this is currently in the planning phase) for all DiSSCo data. (F,A)
2) Open Source development process that provides access to the code and the software repository associated with the data. (A,R)
3) Digital Object Architecture (DOA) and Evolutionary Architecture based design model provides long term sustainability and modularity (F,I,R).
4) DiSSCo open access policy - "as open as possible, as closed as legally necessary" - reflects the dual aims of: i) maximising openness of data in the public interest (considering the central role of knowledge in solving societal challenges and innovation in generating economic growth); and ii) ensuring the application of legally required controls on access to sensitive information. (A,R)
5) Data for individual objects are retrievable by the object's handle using the DOIP 2.0. Data is also retrievable through the public REST (HTTP) API.
   DiSSCo uses JSON Schema as both internal and external representation of data and offers export to several popular formats such as Darwin Core Archive, ABCD XML documents and RDF. (A,I)
   Each object contains a minimum of mandatory terms consistent with its formal object type definition, with the possibility to include optional additional terms and enrichments as necessary. (F,A,I)

Weakness in FAIRness

1) While dealing with different types of physical specimens we need to find a balance between findability and usability. Potentially we will be dealing with 1.5 billion objects that will need to be digitized and identified. We will be adopting the PID kernel recommendation provided by RDA. Smaller kernel information is ideal for larger datasets but it might make it not useful in some cases.

Priority plan for the FAIRness improvement (next 2 years)

1) A Persistent Identifier framework and plan to ensure global unique, persistent linking, findability, and resolvability. DiSSCo and several other global partners are working towards a 'Natural Sciences Identifier' (NSId) scheme based on the Handle system in a joint international governance arrangement under the Alliance for Biodiversity Knowledge. (F,A)

2) A technical specification of an 'open Digital Specimen' (openDS) is needed. This specification will provide a definitive definition of what a Digital Specimen is, its logical structure and content, and the operations permitted on that. It integrates the framework of FAIR principles for digital objects (FAIR-DO) to ensure FAIRness and introduces native semantics into digital specimen and collection data. (F,A,I,R)

eLTER

The eLTER Information System (IS) aims to implement an open and FAIR access to data resulting from the LTER Europe network as well as the newly funded eLTER RI. The main focus in the first phase of the implementation of the eLTER IS was put on the aspect of discoverability in order to enhance the visibility and mobilise data from the network. The FAIR principles are the underlying guidelines for the design and implementation of the data infrastructure. A first version of the data management plan as well as strategy for the long term governance of the eLTER IS was developed within the eLTER (H2020) project.
Building on the concepts and tools developed in earlier projects (e.g. EnvEurope) within the eLTER[1] (H2020) project the main components for the data infrastructure were designed. This will be further developed in the eLTER RI implementation phase starting with 2020. eLTER implements the concept of federated data nodes providing metadata and data based on defined standards.
Currently basic data reporting workflows are established following a commonly defined data format. With the implementation of the eLTER the definition of common data products and the related workflows will be implemented.
eLTER fosters an open policy for data sharing and re-use also for the source data aiming to collect information on data usage (e.g. based on DOI or data citation). This means that the barriers for data publication and data re-use should be as little as possible and with the minimum time delay. The data license and policy is defined by the data providers but common recommendations developed in the eLTER (H2020) project are in place. In future also the publication of data using data journals will be discussed and evaluated.

Strength in FAIRness

1) Findability - research sites as well as the data(sets) generated at these sites are provided using standardised descriptions. Using eLTER DIP as generic catalogue all different data types can be discovered. In order to support discovery metadata following ISO 19119/19139 are generated by DEIMS-SDR and shared through standard OGC CSW and OAI-PMH interfaces. In addition, research site information is provided following the INSPIRE EF specification, thus providing the metadata in standardised and interoperable format. For metadata a unique UUID is generated resolving to DEIMS-SDR or the respective metadata provider. (F)
2) Accessibility - eLTER aims to implement open and service based access to data following common standards. This applies especially to the eLTER CDN where a core functionality is based on OGC SOS. For file based data open repositories such as B2SHARE are used to reduce the barriers for data access to end users and to foster the implementation and use of the services provided by the EOSC[2]. For research sites the DEIMS.ID issued and resolved by DEIMS-SDR is applied. For datasets either handles (e.g. B2SHARE Jülich) or DOIs (EUDAT B2SHARE) are generated to ensure unique identification. (A)
3) Interoperability - in order to provide a harmonised documentation and description of data and research sites a common vocabulary is used. EnvThes is based on SKOS and openly available. EnvThes aims to provide a source for the annotation of the metadata with keywords but also to provide a source for the naming of variables (parameters) used in the data sets. In addition, a common reporting format based on the recommendations of the UNECE ICP Integrated Monitoring programme was

[1] see https://www.lter-europe.net/elter
[2] see European Open Science Cloud

defined. These specifications were extended to the needs defined in the eLTER process. The specification follows the basic design of the Observation & Measurements (O&M) model. (I)

Weakness in FAIRness

1) Persistent identification - currently different PID systems are used to identify the research data objects in the eLTER context. Within the DEIMS-SDR metadata framework reference to this PIDs (e.g. PID handle, DOI, DEIMS.ID) is managed, but a full integration and automatic provision of PIDs is still missing. This includes also the identification of dynamic data provided via OGC SOS services where the current file based approach to assign PIDs is not working properly. (F,A,R)
2) Semantic integration - eLTER uses EnvThes to annotate metadata with keywords as well as starting to use the core vocabulary to define variable names in the datasets. This system of semantic annotation using different standard vocabularies (e.g. units, chemicals) needs to be extended and further integrated. (F,I)
3) Provenance documentation - basic documentation of quality assurance and data acquisition is currently only provided in a textual manner on the level of metadata. Especially when providing integrated data products (e.g. cleaned in-situ dataset on soil moisture measurements across a bio-geographic gradient) the provision of information on the data sources included and the methods applied (e.g. gap filling, data cleaning) is crucial. This information needs to be collected and documented using concepts of provenance tracking. The integration of provenance information when chaining data from different sources is an additional topic. (R)

Priority plan for the FAIRness improvement (next 2 years)

1) Common documentation of observation and experimentation facilities across different subdomain RIs. This includes the further development of the DEIMS.ID as common identifier for research sites and the establishment of a federated catalogue of research sites across the different RIs in order to identify co-location and spatial overlaps. (F,I)
2) Persistent identifiers - full integration of the system of persistent identifiers for the publishing and tracking the re-use of data in the framework of the eLTER RI. (F,A)
3) Extend the use and integration of common semantics (e.g. EnvThes) in the description and documentation of research sites, metadata and data(sets). This enhances the interoperability of the data. (I)
4) Provenance - establishment of standardised procedures (for core variables selected by the eLTER RI) for quality assessment and outlier detection. This needs to be automatically documented with the data. (R)

ICOS ERIC – Ecosystem component

Strength in FAIRness

1) All data objects from raw data up to elaborated products based on the observational data are assigned global, unique, resolvable and persistent identifiers using the Handle system which are resolved by dynamic landing pages that provide the relevant metadata, including download links to the data objects formatted in the community standard format (F,A,I)
2) All metadata is modelled in an ontology through OWL, using RDF and is published using an open SparQL endpoint, all metadata elements are exposed as linked open data where all metadata elements also have dynamic landing pages. the metadata system is fully versioned, supports data versioning, collections and dynamic citation generation (F,A)
3) All landing pages provide metadata in machine readable form using content negotiation (I)

Weakness in FAIRness

1) BADM metadata is a community standard that needs to be modelled in the OWL/RDF format and translated into the ISO19115 standard for export of metadata to external catalogues (F, I, R).
2) Fluxnet data is in a community defined columnar text format that needs to be translated into a more common usable format like the netcdf standard for better interoperability and use by the climate community (I)
3) Documentation of the metadata and data is not directly linked to the data objects and the workflow metadata (F, I, R).

Priority plan for the FAIRness improvement (next 2 years)

1) Elevate the BADM standard into a semantic web-based ontology and translate this into a ISO19115 based metadata export profile (I)
2) Provide the FLUXNET data in the cf1.7 compliant netcdf format (I)
3) integrate the data processing workflow into the metadata model including the description of processing, code, configuration and version used (F,R)
4) Develop new interfaces for the data collection in order to ensure real time metadata and data in a fully traceable and standard format (I,R).

LifeWatch ERIC

The LifeWatch ERIC e-Infrastructure has started the deployment phase and the goal to become FAIR-compliant both for data and web services is cross-cutting to the entire RI.
For this reason, LifeWatch ERIC has been investing more in the metadata, controlled vocabularies and ontologies definition and development. The actual catalogues are based on a well-defined set of (meta)data, controlled vocabularies and ontologies are used to harmonize the information, support the data and services discovery and interoperability. Still, however, this domain is fast evolving and it requires lots of effort to keep up with the tasks.

Strength in FAIRness

1) LifeWatch ERIC has a federated system of Data Centres, that allows data and services to be online 24/7 with geographical redundancy and scalable performance. (F,A)
2) Metadata standards and Controlled Vocabularies adopted and developed from the LifeWatch Network make the solid base in terms of Findability, Interoperability and Reusability. (F,I,R)
3) All the catalogues landing pages provide metadata in a machine readable format using content negotiation. (F,A,I,R)

Weakness in FAIRness

1) The system uses an internal unique identifier, a DOI management and generation still missing, but the procedures to install it are at the final stage. (R)
2) A DMP and Data Policies have already been developed at national level. However, they are now being discussed and implemented at LifeWatch ERIC level. (A,R)
3) Provenance has managed only at metadata level. (R)

Priority plan for the FAIRness improvement (next 2 years)

1) The first priority for the LifeWatch ERIC e-infrastructure is to integrate the management and production of the DOI for all the research objects in the architecture. (A,R)
2) To simplify the publication, the process of deploying an efficient Exposer/Connector tool is yet to be implemented. This tool will allow the different research groups to describe the metadata using automatic functionalities. (F,A)
3) DMPs and Data Policies have to be implemented and managed at LifeWatch ERIC level. (A)

SIOS - Ecosystem component

In its current state SIOS is performing reasonably well with respect to findability and accessibility, while there is a much lower level of maturity when it comes to interoperability and reusability.

Strength in FAIRness

1) The SIOS web portal has a search interface where datasets can be searched by means of several filters, such as full-text search, time intervals, geographical locations, institutions etc. (F)
2) Universally unique, persistent and resolvable identifiers, i.e. UUIDs and DOIs. (F)
3) A formally agreed data policy requiring full, open and free access to data. (A,R)
4) Well established standards are used for accessing metadata (OAI-PMH) and data (HTTP, OPeNDAP, WMS). (F,A,I,R)

Weakness in FAIRness

1) Not all data centres are using the standard protocols, and not all are providing download links in a way that permits indexing. (A)
2) Interoperability is severely limited by non-implementation of standard use metadata, vocabularies and file formats. (I)
3) Data provenance is rarely evident from the metadata. (R)
4) The use of data licences is inconsistent, and the data licence is not always referenced in a clear manner. (R)

Priority plan for the FAIRness improvement (next 2 years)

1) Urge the use of standard data protocols across the RI (A)
2) Work towards metadata standard compliance (F,A,I,R)
3) Inform metadata authors and encourage best practices in dataset documentation. (F,A,I,R)

Interoperability and accessibility improvement

Interoperability and accessibility of RIs data can be improved also by working on the data cycle starting from the data collection, for example, methods harmonization, additional common parameters, specific common technical metadata. The same is valid for tools for data production, handling, curation, tracking etc. that if harmonized and shared/co-developed, in addition to lead to saving of resources and higher quality will also improve the interoperability.

Core competencies identification

Each RI in the Ecosystem and Biodiversity subdomain has a set of core competencies that are at the basis of their creation and development and, in case of RIs collecting data, a set of core variables where the maximum expertise is developed. These high level competences could become available to the other RIs enhancing the cross-RI collaboration and contributing to the co-location (physical or of services). The seven Ecosystem and Biodiversity identified their most important competencies that are made available to the others and are listed in the following table.

RI	Core competence
AnaEE	Climate and management manipulation on terrestrial ecosystems and impact on ecosystem functions and their interactions (biochemical cycles, net productivity, trophic chain, population dynamic, biodiversity). Analytic and modeling tools to understand biodiversity/ ecosystem functions. Semantic vocabularies (thesaurus and ontology) and modelling for ecosystems and biodiversity characterization. Semantic annotation of databases for the generation of metadata or data sets and their publication.
DANUBIUS-RI	Monitoring water quality parameters (optically active constituents within the water column and temperature) through in-situ sensors and Sentinel 2 and 3 across the river-sea continuum in near real time and through image archive. Providing models' calibration/validation procedures and site-specific EO algorithms integrating on RSS remote sensing, in situ and modelled data.
DiSSCo	DiSSCo sets the physical objects - specimens - at the centre of the development of a robust, quality ensured and fit-for-purpose knowledge base for bio- and geo-diversity. The infrastructure works towards the digital unification of all European natural science assets under common curation, access policies, and best practices that aim to make the data FAIR.
eLTER	Monitoring and assessment of ecosystem changes in relation to relevant drivers and pressures applying a whole system approach including the socio-economic domain. Providing access to research sites and in-situ monitoring data to analyse ecosystem processes, their status and change focusing on the whole ecosystem including human interactions. The entire network provides local in-situ data across bio-geographic and socio-economic gradients of Europe for scientific analysis as well as input to data analysis workflows (e.g. validation and calibration of EO data products).
ICOS ERIC – Ecosystem component	The Eddy covariance technique is used to collect the core measurement in ICOS (GHGs fluxes between surface and atmosphere). All the aspects related to these measurements (from setup to final processing) at the core experience of ETC. Characterization of the eddy covariance target area in terms of physical properties such GAI, biomass, soil C content and vegetation nutrients (including the sampling design)

RI	Core competence
LifeWatch ERIC	The core niche of LifeWatch ERIC is the provision of e-Science facilities (FAIR-compliant data and web services), organized in virtual research environments (VREs), to researchers and other target users. This feature makes the infrastructure unique in the ESFRI landscape. Interoperability of data and all categories of web services is the focus of LifeWatch ERIC, whose domain specific solutions (VREs) can be also used by other research infrastructures dealing with the biodiversity and ecosystem research. It is also providing distributed resources federating tools based on Blockchain technologies for data citation, provenance, and accountability that is the case of the recently launched so-called LifeBlock platform.
SIOS – Ecosystem component	(Currently not anything we can think of)

Common parameters for case studies

The RIs involved in the subdomain ecosystems and biodiversity have, for their nature, a number of common parameters and variables that are collected, managed and processed by all of them since basic in the ecosystems characterization. The increase of the FAIRness can be reached also by improving the interoperability of the measurements and data use from the origin (collection). We decided to use a bottom-up approach, complementary to the analysis of gaps in the FAIRness done in the other WPs, in order to test in practice the issues present in a common and FAIR access to a variable monitored by most of the RIs. In a first phase each RI identified the five most important parameters that could be considered as a test case for an activity where practical steps to ensure across-RI accessibility, interoperability and use of the data are developed.

RI	Proposed common parameters
AnaEE	1) Soil water content and associated drivers (climate forcing/Soil properties/vegetation development, ecosystem type, species, ecosystem management). 2) Evapotranspiration and associated drivers (climate forcing/Soil properties/vegetation development, ecosystem type, species ecosystem management) 3) Biomass and associated drivers (climate forcing/Soil properties/vegetation development, ecosystem type, species ecosystem management) 4) Plant phenology and associated drivers (climate forcing, species) 5) Biodiversity notably phytoplankton measurement which was identified as a cross RI study-case in ENVRIplus
DANUBIUS-RI	Observable and traceable parameters from sentinel platforms (reproducible also through modelling) that include: 1) Chlorophyll a (Chl a) is a widely used parameter for water quality and ecosystem condition from local to global scales. This is established through a complex series of algorithms applied on a perpixel basis according to a preclassification of optical water type. Data are made available through webGIS portal. 2) Total suspended matter (TSM) provide measures of sediment loading useful in characterizing sediment load related to erosion, transport and deposition mechanism. Data can be made available through webGIS portal. Examples can be found through Copernicus Land Monitoring Service.

RI	Proposed common parameters
	3) Sea/Lake surface temperature (SSWT/LSWT) is limited by resolution of the Sentinel 3 radiometer 1000 m pixel resolution and limited to large lakes and coastal waters. As key variable to describe climate change influence on surface waters, Landsat 7&8 also provides a lower temporal resolution by higher spatial resolution (60m) opportunity for monitoring water
DiSSCo	1) Physical Specimen Identifier 2) Scientific name 3) Institution Code 4) Personal Profile (expertise, recorded by) 5) Image link (digitized object) 6) Extinction Status
eLTER	eLTER with the coming eLTER PLUS project aims to further integrate and mobilise data on ecosystem processes. Starting with DEIMS-SDR basic information on the availability of the data can be provided focusing on spatial, temporal and thematic extent of these measurements. Specifically eLTER PLUS will work on: 1) Biodiversity loss and related drivers at site scale - addresses the limitations imposed by insufficient temporal and spatial scales of biodiversity observations. Systematic integration of measurements collected at Europe's most advanced long-term ecological research sites is possible within eLTER PLUS. 2) Biogeochemical controls of ecosystem functions - focuses on investigating the interactions of carbon and nitrogen biogeochemical cycles across the European ecosystem types and biogeographic regions linking all spheres of ecosystems (atmo-, bio-, pedo-, hydro-sphere). Climate change (e.g. changes in rainfall and temperature patterns), agricultural management (e.g. fertilization, tillage, irrigation), timber harvest and the concentrations of airborne acidifying and eutrophying substances are important drivers of biogeochemical processes. 3) Soil water content and associated drivers - focusing on climate change and the expected to increase extreme events such as droughts and heat waves that will impact water availability and increase water demand for irrigation. 4) Socio-ecological systems - focuses on directly addressing societal questions and environmental policies at local, regional and continental scales.
ICOS ERIC – Ecosystem component	1) Soil water content is an important measurement that is also linked to the RS community. In ICOS it is measured by 2 to vertical profiles of sensors. 2) Radiation components, such short and long wave and PAR, incoming and reflected, are crucial to calculate the radiative balance and also linked to RS and modelling community. ICOS measures the different components at the ICOS stations. 3) Green Area Index is an important parameter to characterize the vegetation and it can be measured in different ways. In ICOS the GAI is measured in all the sites using Digital Hemispherical Pictures and/or ceptometers (plus destructive sampling). It is a variable complex to characterize and standardize.
LifeWatch ERIC	Because of its specific niche and core interests, LifeWatch ERIC does not have specific 'parameters' of interest, as it needs to be capable of processing virtually all parameters relevant to biodiversity and ecosystem research. However, LifeWatch ERIC has a major interest in genes, species, habitats and ecosystems characteristics, such as:

RI	Proposed common parameters
	1) Genes and species diversity and traits; 2) Species interactions and niches abiotic and biotic dimensions 3) Habitat and ecosystem functions, processes and services. All measurable parameters connected with the above characteristics can be used as test cases to check possible improvements in interoperability.
SIOS – Ecosystem component	1) Physical Specimen Identifier 2) Chlorophyll A concentration

From the analysis of the proposed parameters it has been decided to work on the following case studies. The exercise will be used to check in practice which are the main FAIR aspects that prevent data sharing among the RIs helping to define the priorities for the next phase of the project.

1) Soil Water Content: it is a key quantity measured in most of the RIs that have in-situ components but also handled by the e-infrastructures. It can be measured by different sensors, at different depths, with different time resolutions, calibrated by different methods and with reported in different units, different metadata and different variable names. During the case study the different RIs handling this variable will build a demonstrator where the data are findable across RIs and their use facilitated through a standardization of the information and data characteristics as well as the access of ancillary data, as soil characteristics, climate, vegetation cover, required to contextualize soil moisture data. This case study will probably involve the following RIs: AnaEE, eLTER, ICOS, SIOS and LifeWatch as integrating platform.
2) Species scientific names and identification: it is a basic information that all the infrastructures in the Biodiversity and Ecosystem domain have to handle. Although well-defined historically the scientific name and procedure for identification, like also the parameters connected to the species reporting have a degree of flexibility and uncertainty that can be solved in this case study in order to ensure cross-RI access to species information. This case study will probably involve all the RIs participating and a collaboration with the Catalogue of Life project (https://www.catalogueoflife.org/col/) where the DiSSCo RI is already active.
3) Site documentation: research sites and facilities are the core entity for observation or experimentation. RIs and research networks have documentations of their site network (e.g. DEIMS-SDR, AnaEE, SIOS, ICOS) but integrated, even if sites are co-located, is often missing. The same site can be listed with different name. The integration of existing information would enable the identification and management of co-located site as well as the variables observed/measured and further work on a consistent documentation and exchange of site information across RIs.

In addition to these three general case studies, the table above highlights a number of cases where 2 or 3 RIs can work on specific parameters that have in common and that can be used as case to increase the FAIRness. Tis will be evaluated during the next meetings and exchanges also in the context of Task 11.3.

Sharable tools

In addition to specific competences on measurements and processes characterizing the RI activities, RIs develop also general tools that cover specific processes and activities and that can be shared and made available to the others. This is also contributing to the FAIRness of the products and to the improvement in the RIs collaborations. In the analysis of the current situation, these tools have been proposed and identified as potentially interesting for other RIs of the subdomain:

AnaEE

1) Thesaurus (Anaeethes) and (Oboe-based) ontology for ecosystems and biodiversity characterisation.
2) Pipelines: a) semantic annotation of relational databases and b) exploitation of semantic data for the production and publication of standardized metadata (ISO19115/19139, GeoDCAT) and data sets (NetCDF). These 2 pipelines contribute to the ENVRI service portfolio (https://www.slideshare.net/EUDAT/data-for-science-service-portfolio)
3) AnaEE Portal Statistics (in development): being a data aggregator, the AnaEE Portal is providing the institutional users with a dashboard of metrics that include insights on the licensing adopted by the various platforms, the formats adopted, the data findability within the Portal (obtained via Usage Mining), and the provided metadata. These indicators are being evaluated for each research platform involved in the Portal and are providing a baseline for FAIRness evaluation.

DANUBIUS-RI

DANUBIUS-RI being in preparatory phase, the tools to share the data are not defined yet, leaving us the opportunity to adopt the best solution based on the experience of other RIs. However, studies indicate NetCDF is used for archiving data, standard tool available on the internet, could be used to describe, classify, and make data findable. The usual standards like CF Convention or CF Metadata will be applied.

DiSSCo

European Collection Objects Index will provide an API to access digitized natural science collections. Through DOIP (Digital Object Interface Protocol) this will provide access to natural science collections and related information

eLTER

The eLTER Information System enables provision and access to long term ecosystem monitoring data in order to support data analysis and forecasting. Within the eLTER project developed the basic architecture was designed and tools to support data providers as well as data users were developed. eLTER aimed to foster the availability to FAIR and open data which are key elements for Open Science. All components of the eLTER Information System can be accessed through the eLTER data landing page at https://data.lter-europe.net/.

1) Site and dataset registry: DEIMS-SDR (Dynamic Ecological Information System - Site and Dataset Registry; https://deims.org) to manage and discover metadata on long-term ecosystem research sites, along with the data gathered at those sites. This also includes people and networks associated with them. DEIMS-SDR contains information on a wide range of sites, providing a wealth of information, including each site's location, ecosystems, facilities, parameters measured and research themes. DEIMS-SDR applies OGC and ISO metadata and data exchange standards.
2) Common controlled vocabulary:, EnvThes (http://vocabs.lter-europe.net/edg/tbl/EnvThes.editor#http%3A%2F%2Fvocabs.lter-europe.net%2FEnvThes%2F) to provide a unified semantic backbone for the documentation and discovery of long term ecosystem monitoring data. Inter-alia, EnvThes is used by DEIMS-SDR to select common keywords. EnvThes is a SKOS vocabulary taking to account a range of conceptual models, like OGC O&M, OBOE

and SERONTO, with the focus to implement main design principles laid down in the Complex Properties Model (Leadbetter & Vodden 2016).

3) eLTER Central Data Node: (eLTER CDN, https://cdn.lter-europe.net/) to upload and publish time series data as OGC SOS service. It aims to support partners that do not have the resources and/or skills to host standardised data infrastructure, cloud-based data storage and services.
4) eLTER Data Integration Portal: (eLTER DIP, http://dip.lter-europe.net)) is the central catalogue to discover, access and visualise data sources provided through the data nodes registered in the eLTER DIP. Data provided as OGC SOS service can dynamically visualise and download using the 52°North Helgoland client.

In addition common specifications for metadata[3] and datasets[4] shared in the eLTER context has been developed.

ICOS ERIC – Ecosystem component

1) A full workflow for Eddy Covariance data processing from raw to final product has been implemented. This tool is mainly based on the ICOS setup but it is under development to process also other setups (already operational for some sensors and setup and shared on GitHub). The post-processing is already a standard with the AmeriFlux network (ONEFlux tool). Eddy covariance sites of other RIs could be processed with this tool and ensure full reproducibility, documentation and high level quality.
2) A tool for the Digital Hemispherical Pictures processing to perform a quality check and the calculate GAI and its variability is available for all the RIs that uses this method for GAI estimation. A similar tool is also available for ceptometer measurements.
3) Jupyter Lab VREs to support scientific discovery workflows allow for more transparent and reproducible production of elaborated products based on (ICOS) observations and ancillary data. The developed Jupyter notebooks themselves can be published and cited and can serve as documentation of the products, the data requirements, the processing code and the configuration data. Through the metadata and data access API the Jupyter code has direct access to all discovery information and the actual data and can directly publish the resulting data through the ICOS repository.

LifeWatch ERIC

1) (Meta)data, e-Services, VREs Catalogues: these tools are the main access to the LifeWatch Knowledge. The Catalogues software architecture is based on the GEONETWORK tool, and allow the access via Web interface or via machine-to-machine interfaces (API are available and documented). These catalogues are in continue evolution. It is available at http://www.lifewatch.eu
2) ECOPORTAL, is the LifeWatch ERIC catalogue of semantic resources. Ecoportal collects all the worldwide semantic effort done from the research groups in the ecological domain. It is a unique way to discover, align and use the existing semantic resources. There is the full implementation available at http://ecoportal.lifewatchitaly.eu
3) VREs, LifeWatch ERIC offers a series of Virtual Research Environment in order to support the work of the scientist in different domains and topics (Alien Species, Phytoplankton, Marine, RVlabs, ect..)

[3] see https://deims.org/models
[4] see https://www.lter-europe.net/lter-europe/data

SIOS – Ecosystem component

1) A validation tool for NetCDF/CF 1-6 and ACDD, see https://www.sios-svalbard.org/dataset_validation/form (login required). The tool is operational.
2) An Excel template generator for registration of data has been implemented and spreadsheet can be created and downloaded from the SIOS portal. This tool was intended to be used by researchers that do not have other tools for creating datasets (like Fish2Data, instrument formats etc.). The downloaded file contains variable names according to Darwin core and Units follow CF standard.
3) Conversion ASCII files to NetCDF – Rosetta http://tomcat.nersc.no/rosetta/

Data and technical metadata formats used in distribution

The last aspect analysed in order to prepare the short term plan of activities has been the format of data and metadata used for distribution to the users, in particular all the technical metadata providing detailed complementary information needed for a complete data understanding and use (e.g. location of the measurement, sensor model and history, calibrations, disturbances and maintenances etc.). The RIs that already defined the format(s) of data and metadata provided a short description and a link to an example with the aim to evaluate and ensure the cross-RI data use. From the analysis, actions to solve possible bottlenecks will be implemented.

AnaEE

1) Data are stored in local databases in formats which can vary from one platform to another (PostgreSQL format in France https://si-acbb.inra.fr/). NetCDF is one of the foreseen formats for published data sets.
2) Metadata annotation graphs are partly covered: object being measured, measured quantity, experimental setup, people. Provenance information (sensors, protocols, treatment) is still missing. The ISO19115/19139 INSPIRE compliant metadata standard is the minimum requirement for discovery characterization of AnaEE resources (data and services). The production of exploitation metadata following the EML (Ecological Metadata Language) standard is also recommended.

DANUBIUS-RI

The core data and metadata formats used to distribute the DANUBIUS-RI data are:

- Geospatial data (vector and raster data),
- tabular GIS
- attribute data
- Quantitative tabular data with minimal metadata
- Quantitative tabular data with extensive metadata
- Qualitative data
- Quantitative array-oriented data, self-describing

DANUBIUS-RI will aim to follow the standardizations offered by the Open Grid Forum (https://www.ogf.org) and the Organization for the Advancement of Structured Information Standards (OASIS -https://www.oasis-open.org/org). These formats offer worldwide standards for security, Internet of Things, cloud computing, energy, content technologies, emergency management and the European Legislation, Commission Regulation 1205/2008 of 3rd December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata.

DANUBIUS-RI will use and conform to published standards whenever possible. Several standards provide a guideline for DANUBIUS-RI and whenever suitable will be applied. We list below several of the current metadata standards that are considered relevant to DANUBIUS-RI:

- EML (Ecological Metadata Language) - the meta-data specification for the documentation and exchange of information about ecological data.
- GML (Geography Markup Language) - the data exchange format of OGC and INSPIRE
- CSML (Climate Science Modeling Language)
- ESML (Earth Science Markup Language)
- NcML (NetCDF Markup Language)
- MML (Marine Markup Language)
- SensorML (Sensor Model Language)
- CSDGM(Content Standard for Digital Geospatial Metadata)
- WaterML2 is a data exchange standard in Hydrology which can basically be used to exchange many kinds of hydro-meteorological observations and measurements.
- OGC framework Observation & Measurement
- OGS framework SWE (Sensor Web Enablement)
- Other OGC, such as TimeseriesML

DiSSCo

DiSSCo aims to mobilise, unify and deliver natural science (bio- and geo-diversity) information at the scale collected from the DiSSCo Facilities. For each level of digitisation a metadata specification will be applied. These two standards are currently under development: Minimum Information standard for Digital Specimens (MIDS) and Minimum Information standard for Digital Collections (MICS) standards. DiSSCo is also working towards openDS -- a new standard for digital specimens (representing a digitised physical specimen, contains information about a single specimen with links to related supplementary information). These standards will align with current standards such as ABCD 3.0 and EFG extension for geo-sciences and extending OBO Foundr's BCO.

eLTER

eLTER aims to build on a federated system of data provisioning nodes and thus to implement a variety of standards for the metadata and data provision.

1) Metadata for research sites, data products (activities) and datasets are provided using the OGC CSW 2.0 as well as the OAI-PMH protocol. Metadata for datasets are shared using ISO19115/19139 as well as EML (DataOne). Information on research sites is also provided using the INSPIRE EF specification to create a compliant GML file. In addition, research site centroid coordinates, bounding box and detailed boundaries are provided using OGC WMS and OGC WFS. Currently, a Rest-API is being developed for DEIMS-SDR that also allows to get the information entities (e.g. site records) as JSON blobs.
2) Identification of research sites using the DEIMS.ID[5]
3) Data(sets) are either stored as data files following the common data specification of eLTER[6] using a common open repository (e.g. B2SHARE) or as OGC SOS compliant data stream using the central data node facilities. For both cases, an O&M compliant data model is applied which in future will be extended to fully support semantic interoperability standards.

[5] see https://deims.org/docs/deimsid.html
[6] see https://www.lter-europe.net/lter-europe/data/eLTER_T3.4_VA_TemplateSpecification_V0.7_Short.pdf

ICOS ERIC – Ecosystem component

1) All metadata is modelled in RDF and is available through linked open data and the SparQL open endpoint. Metadata exchange through OAI-PMH is possible but deprecated as previous century solution. Data is available as plain ASCII files in CSV format with column headers, or as cf-compliant netcdf format. A simple but highly performing API provides direct (read-only) access to the data in binary format.
2) All (meta)data, including provenance, instrument and site information is provided using PIDs and where applicable for publishing also as DOIs. The system fully supports versioning of the metadata and data and supports also the minting of collections
3) All metadata on PID and DOI landing pages is provided through content negotiation as human readable HTML, XML, JSON-LD, Turtle

LifeWatch ERIC

The LifeWatch ERIC Catalogues are based on the GeoNetwork solution. GeoNetwork has different export options to export metadata. They operate on selected sets of metadata from the search results.
Xml, JSON, csv, pdf are examples of possible format outputs, in addiction is supported the metadata exchange format (MEF in short) that is a specially designed file format for the purpose of metadata exchange between different platforms.
A Semantic Search is at the deployment phase, that exposes the metadata in RDF format, making them available also through a SparQL open endpoint.

SIOS – Ecosystem component

1) The exchange protocol for discovery metadata within SIOS is OAI-PMH serving either GCMD DIF or ISO19115 (both are consumed). For physical environmental data SDMS is recommending to use NetCDF following the Climate and Forecast Convention and serving this over OPeNDAP in order to achieve streaming of data instead of file transfer. Biodiversity data can be accessed as csv or JSON. The Darwin Core metadata standard is recommended.
2) SIOS itself does not provide harvestable metadata. Metadata are provided by the contributing data centres. Minimum requirements are listed above, but some data centres will also provide metadata as JSON, DIF XML, ISO 19139 XML, Atom entry XML, DCAT (JSON-LD) and Datacite XML

Next steps and short term plan

The exercise through the case studies will be used to better identify the main issues and aspects to improve in terms of FAIRness. The results, together with the start of the Task 11.3 activities, will help to define the short term plan and to identify if there are common gaps that would allow the preparation of a Common Plan of development. This would be clearly a priority in the actions in order to organize a validation exercise toward the end of the project selecting more complex case studies and analysing the improvements obtained.
All the assets that will be developed in the process of improving the FAIRness, also through the case study, will be made available in the EOSC catalogue of services.

Conclusions

This is the first deliverable of the WP and for this reason it describes the current situation of the RIs involved, the starting point of the ENVRI-FAIR activities that will have the aim to improve the situation that will be evaluated with a similar report at the end of the project.
This first analysis of the status of the Biodiversity and Ecosystems RIs in terms of FAIRness and services highlighted a heterogeneous situation, with a large range of development status of the different RIs involved due in particular to different starting years. The strategies followed in terms of data description and access are different and the activities in the project will help to increase the level of inter-RIs compatibility and FAIRness.
It has been decided to create three case studies selecting two variables/properties (soil water content and species scientific names and identification) and the description of sites that are common across the RIs and where the improvement of FAIRness and cross-RI access will be tested.
The list of expertise developed in the framework of each RI and the description of the tools that can be shared with other RIs will be the basis of the discussions and collaborations among RIs in order to maximise synergies.

Appendix A: Glossary

List of acronyms

ABCD	Access to Biological Collection Data
ACBB	Agroecosystems, Biogeochemical Cycles and Biodiversity
AnaeeThes	AnaEE Thesaurus
API	Application Programming Interface
ATC	Atmospheric Thematic Centre of ICOS
B2SAFE	EUDAT Data Storage and Distribution System
B2SHARE	EUDAT Data Sharing Platform
BCO	Biological Collections Ontology
BDD	Behavior-driven development
BioMA	http://www.biomamodelling.org
DCAT (JSON-LD)	Data Catalog Vocabulary (JSON for Linking Data)
DEIMS-SDR	Dynamic Ecological Information System Site and Dataset Registry
DMC	Data and Modeling Centre of the AnaEE Rresearch Infrastructure
DOA	Digital Object Architecture
DOI	Digital Object Identifier
DOIP	Digital Object Interface Protocol
EFG	Access to Biological Collection Databases Extended for Geosciences
eLTER CDN	eLTER Central Data Node
eLTER DIP	eLTER Data Integration Portal
EML	Ecological Metadata Language
EnvThes	Environmental Thesaurus
EOSC	European Open Science Cloud
FLUXNET	Global network of eddy covariance sites
GCMD DIF	Global Change Master Directory Directory Interchange Format
GeoDCAT	Geospatial extension of DCAT (Data Catalog Vocabulary)
IS	Information System
ISC	Interface and Synthesis CeEntre of the AnaEE Rresearch Infrastructure
JSON	JavaScript Object Notation
LTER	Long Term Ecosystem Research
LTSER	Long Term Socio-Ecological Research
MICS	Minimum Information standard for Digital Collections
MIDS	Minimum Information standard for Digital Specimens
NetCDF	Network Common Data Form
NRT	Near Real Time
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBOE	Extensible Observation Ontology
openDS	Open Digital Specimen standard
OTC	Oceanic Thematic Centre of ICOS
PID	Persistent Identifiers

OBO	Open Biological and Biomedical Ontology
OGC SOS	Open Geospatial Consortium - Sensor Observation Service
OGC CSW	Open Geospatial Consortium - Catalogue Service Web
RDF	Resource Description Framework
RI	Research infrastructure
TC	Technology Centre of the AnaEE Research Infrastructure
SDMS	SIOS Data Management System
UNECE ICP	United Nations Economic Commission for Europe - International Cooperative Programme
URI	Uniform Resource Identifier