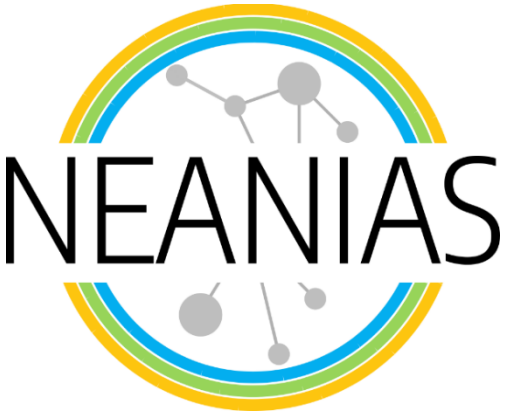

NEANIAS
Novel EOSC services for Emerging Atmosphere, Underwater and Space Challenges

Whitepaper. NEANIAS Best practices for interoperability and reuse

Simone Mantovani (MEEO), Laura Vettorello (MEEO) Kovács József (SZTAKI), Filomena Solitro (ALTEC)

NEANIAS is funded by European Union under Horizon 2020 research and innovation programme via grant agreement No. 863448

Disclaimer

NEANIAS is a Research and Innovation Action funded by European Union under Horizon 2020 research and innovation programme, via grant agreement No. 863448.

NEANIAS is project that comprehensively addresses the 'Prototyping New Innovative Services' challenge set out in the 'Roadmap for EOSC' foreseen actions. It drives the co-design, delivery, and integration into EOSC of innovative thematic services, derived from state-of-the-art research assets and practices in three major sectors: underwater research, atmospheric research and space research. In each sector it engages a diverse set of research and business groups, practices, and technologies and will not only address its community-specific needs but will also enable the transition of the respective community to the EOSC concept and Open Science principles. NEANIAS provides its communities with plentiful resource access, collaboration instruments, and interdisciplinary research mechanisms, which will amplify and broaden each community's research and knowledge generation activities. NEANIAS delivers a rich set of services, designed to be flexible and extensible, able to accommodate the needs of communities beyond their original definition and to adapt to neighboring cases, fostering reproducibility and re-usability. NEANIAS identifies promising, cutting-edge business cases across several user communities and lays out several concrete exploitation opportunities.

This document has been produced receiving funding from the European Commission. The content of this document is a product of the NEANIAS project Consortium and it does not necessarily reflect the opinion of the European Commission. The editor, author, contributors and reviewers of this document have taken any available measure in order for its content to be accurate and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document may be held responsible for any damage, financial or other loss or any other issue that may arise as a result of using the content of this document or any of the project outputs that this document may refer to.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 28 member states of the European Union. It is based on the European Communities and the member states' cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

Table of Contents

Abstract

The objective of this whitepaper is to report on the guidelines and best practices adopted to design, release and improve the level of FAIR (**F**indability, **A**ccessibility, **I**nteroperability, and **R**euse) of digital resources made available as part of the WP6 activities.

This document leverages on the valuable work and experience of technical partners as well as on the large availability of existing standards, specifications and implementations of FAIR services and validation tools.

1. Introduction

1.1. Context

The NEANIAS WP6 "Core Services Foundation and Implementation" establishes generic, cross-community services that amplify the potential of thematic services delivered by WP2-5, enabling Open Science, and facilitating the migration to the EOSC concepts, by streamlining access to cloud resources.

In 2016, the '**FAIR Guiding Principles for scientific data management and stewardship'** were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Being a concept transversal to all service categories, in the last years the FAIR principles have been mainly applied and exploited to the design and enable FAIR Data-as-a-Service in order to access the huge amount of data distributed across relevant research infrastructures. The data "FAIRification" together with the standardization bodies (e.g. OGC, ISO) and the rapid evolution of cloud technologies have contributed to the state-of-the-art of the data discovery and access services: reliable standard interfaces, including the progressive adoption of API provide a seamless access to Petabyte of data in the cloud.

The valuable expertise gained at European level on the DaaS layer is applicable to NEANIAS services: in particular, the challenging and ambitious plan to deliver within the same project

core services suitable for the implementation of under-water, atmospheric and space thematic services is the internal requirement and the main driver for the NEANIAS services FAIRification.

1.2. Contents and Rationale

This document introduces the FAIR principles and describes the best practices adopted within the NEANIAS project to deliver core and thematic services with an adequate level of FAIRness for integration in EOSC.

The best practices are strictly connected with the service architecture and specifications of the NEANIAS services [D6.1], [D6.4].

1.3. Structure of the document

In Section 2 the FAIR Interoperability Framework is presented to introduce the FAIR principles. Section 3 describes the EOSC onboarding and FAIR level assessment process for the NEANIAS core and thematic services. Best practices and guidelines adopted in NEANIAS are presented.

2. FAIR Interoperability Framework

This section introduces the FAIR principles and summarizes the main outcome from "Report from the EOSC Executive Board Working Groups (WG) FAIR and Architecture" and other relevant initiatives to achieve and assess the FAIRness level of digital resources.

2.1.1. FAIR principles

The GO FAIR community and [1] summarize the FAIR principles in a list of characteristics that data, services and infrastructure resources have to achieve for optimizing the FAIRness level:

- the components have to be **Findable**
  - F1. (Meta)data are assigned a globally unique and persistent identifier;
  - F2. Data are described with rich metadata (defined by R1 below);
  - F3. Metadata clearly and explicitly include the identifier of the data they describe;
  - F4. (Meta)data are registered or indexed in a searchable resource.
- the components have to be **Accessible,** user needs to know how data and services can be accessed, possibly including authentication and authorization.
  - A1. (Meta)data are retrievable by their identifier using a standardised communications protocol:
    - A1.1 The protocol is open, free, and universally implementable;
    - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary;
  - A2. Metadata are accessible, even when the data are no longer available.
- the components have to be **Interoperable,** data and services usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

  - I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation;
  - I2. (Meta)data use vocabularies that follow FAIR principles;
  - I3. (Meta)data include qualified references to other (meta)data.
- the components have to be **Reusable**, metadata and data should be well-described so that they can be replicated and/or combined in different settings.
  - R1. (Meta)data are richly described with a plurality of accurate and relevant attributes;
  - R1.1. (Meta)data are released with a clear and accessible data usage license;
  - R1.2. (Meta)data are associated with detailed provenance;
  - R1.3. (Meta)data meet domain-relevant community standards.

The elements of the FAIR principles are connected to each other and they summarize needed characteristics for data resources, services, vocabularies and infrastructures. They can be put into practice in any combination and incrementally in order to increase the FAIRness level.

There are at least four types of interoperability:

1. Technical interoperability is commonly defined as the "ability of different information technology systems and software applications to communicate and exchange data" and it can be extended by adding the "ability to accept data from each other and perform a given task appropriately and satisfactorily without the need for extra operator intervention". It is also related to software, services, workflows, protocols, hardware designs, etc and not only to the exchange of data.
2. Semantic interoperability can be defined as "the ability of computer systems to transmit data with unambiguous, shared meaning. Semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems" [2]. It requires a minimum set of metadata format that should be shared and understood between communities making part of the collaboration. Shared information should involve also describing artefacts to homogenize the interpretation and the exchange of resources.
3. Organisational interoperability, partners have to align standards, to improve the federation of resources, processes, business models and they have to align themselves on a common set of goals and benefits. The user communities have to be analysed to plan user-focused objectives.
4. Legal interoperability has to be improved for the reusability of data. It happens through an effort to reduce conflict, on license, between datasets that are provided by different sources. For further details see [3].

2.1.2. FAIR Data and Services

By definition to make the data and services FAIR means to apply the principles of findability, accessibility, interoperability, and reusability. FAIR data does not necessarily mean Open and Free, but to have an exhaustive and human-/machine-readable mechanism to describe and get access to the data resources. Likewise, the FAIR principles do not restrict recognition of legitimate reasons for shielding data.

The research communities and all stakeholders involved in the FAIR ecosystem have reached a clear and common understanding about the importance of providing FAIR data, the

extension of the concept also to services and software tools is still challenging and it requires an effort both to define guidelines and reference implementations to achieve the findability, accessibility, interoperability and reusability of digital resources. From the [4] it is possible to extract a brief list of actions for improving the FAIRness level into software:

- use a publicly accessible repository with version control, it improves the reproducibility of results generated by the software and facilitates software reusability. Versioning permits to track changes in software;
- add a license;
- register your code in a community registry;
- enable citation of the software;
- perform a Quality & Assessment process for your software.

Registries and repositories collect and manage access to information about resources and data and metadata. In this way, they offer services supporting the reuse requirements.

All the components of a project, or of a research ecosystem, should be inserted into a registry that has to support FAIR data sharing. A global registry for researchers is now becoming available via ORCID [5], a global registration and resolution system for persistent identifiers is available via Handles, and registries for metadata schemas and for concepts and vocabularies are also in a phase of definition. See the RDA-Force11 FAIRsharing [6] resource which links standards to databases, repositories and data policies and assigns DOIs to its records.

There are no standards yet for the assessment of registries. It would be useful to develop a set of standards to measure the FAIRness of registries, as well as other services.

Repositories take responsibility for long-term data stewardship by curating data and metadata.

2.2. Technical components and aspects

2.2.1. The technical ecosystem of FAIR Data

The concretisation of FAIR principles is achieving through the definition and implementation of a specific ecosystem. This is composed of several concepts and objects interacting with each other to reach the following FAIR recommendation (see [1]):

- **Rec.7**: Support semantic technologies. Interoperability needs semantic technologies to be developed, expanded and applied both within and across disciplines;
- **Rec.8**: Facilitate automated processing. The system should offer at their components some ways to communicate and to interact at multiple levels and across disciplines. Machine-to-machine communication has to be facilitated as much as possible.
- **Rec.9**: Develop assessment frameworks to certify FAIR services. Data services must be encouraged and supported to obtain certification, as frameworks to assess FAIR services emerge. An example is the CoreTrustSeal (CTS) for trusted digital repositories
- **Rec.22**: Use information held in Data Management Plans. DMPs describe information on the data and related outputs structured for future reusability. Investment should be made in DMP standards and tools to define common standards and support the information exchange across the FAIR data ecosystem.
- **Rec.23**: Develop FAIR components to meet research needs.
- **Rec.24**: Incentivise research infrastructures and other services to support FAIR data.

Investments, related to the listed recommendations, should take part in the sustainability strategy which defines the implementation of the project and, more in detail, of the service.

The main components of a FAIR ecosystem are policies, DMPs, identifiers, standards, repositories and the usage of a catalogue where they can be registered. We can briefly define them:

- policies, components definition and some functional aspects of their relationship;
- DMP, a generic index summarizing relevant information on the project, the defined Digital Object of the FAIR ecosystem and also their interaction. It should cover all outputs including the software and other research materials;
- identifiers (PID), they are assigned at the FAIR Digital Objects and at their components;
- standards, metadata definition (objective, value and formal rules) to formalize vocabularies, ontologies and information model;
- repositories, they are essential for the FAIR data ecosystem because they are needed to offer accessible and reusable data and metadata to interested users. Currently, many repositories store, manage and curate data and metadata and give access to it for users from specific disciplines. Services that allow researchers from many disciplines to deposit and publish data are emerging.

Components and their interaction should be validated and the data services should be certified in order to verify (and to improve) references to existing, or newly defined, standard. For a European founded project, the ecosystem should interact with the EOSC. The current technologies are going towards a highly distributed ecosystem where mechanisms of federation and collaboration have a strong role. The technical implementation has to include linking resources and process for agreement specification and standards. Distributed queries, a centralized catalogue and a Data Access service can be a good example for simulating a virtual integration of data and services. Some legal aspects have to be considered performing a kind of integration and the Federation can be a good practice to find a general agreement on functional and formal aspects. For example, on the integration of repositories and registries provided by different stakeholders or projects partners which are using different infrastructures, technologies and providers.

Figure 2 shows an example, extracted from [1], of interaction between components in the FAIR ecosystem.

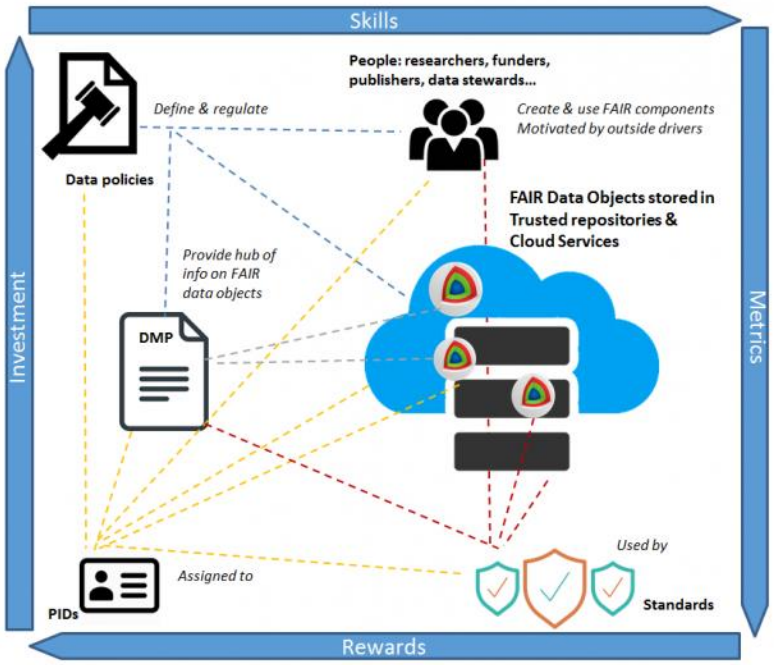

Figure 1: Interaction between ecosystem's components

Figure 2 explains that the FAIR data ecosystem achieves its objectives through a series of interacting components. It is clear the centrality of the FAIR Digital Object in the creation of global domain as a precondition for the Findability, Accessibility, Interoperability and Re-use of the system and process components.

2.2.2. FAIR Digital Objects

FAIR Digital Objects hold a central role in the definition and realisation of the FAIR ecosystem. They are assigned to protocols, research resources, software, services, infrastructures and data and their definition have to provide enough information to improve the possibility to find, use and cite the related entities.

In a FAIR ecosystem, the listed resources have also Persistent Identifiers (PIDs) and a set of metadata which was produced after a data curation life-cycle defining standards and vocabularies in compliance with the community approach. Software and algorithms should include the source, complete documentation, dependencies and licensing. The federation briefly described in the previous section would not be possible without specifications and standards realized for the definition of Digital Objects. It requires human resources with specific expertise as well as ad hoc defined data repositories and data services. This process should ensure the interoperability and the reuse of FAIR resources.

Before explaining more in details specifications and standards, it could be useful to summarize the idea of Digital Object going through the following list of concepts:

- Digital Objects defines the main component of a process they are necessary for a stable and reusable domain of data;

- Physically the data Digital Objects are described by a sequence of bit. This digitalization permits the storage in some repositories and the identification of a persistent identifier (PID) which has to be provided by an authorization system. Stable PIDs allow referencing to digital objects, for example in citations in publications.
- Metadata itself are Digital Objects;
- Digital Objects can include all kinds of digital information such as data, software, services, configurations, etc.

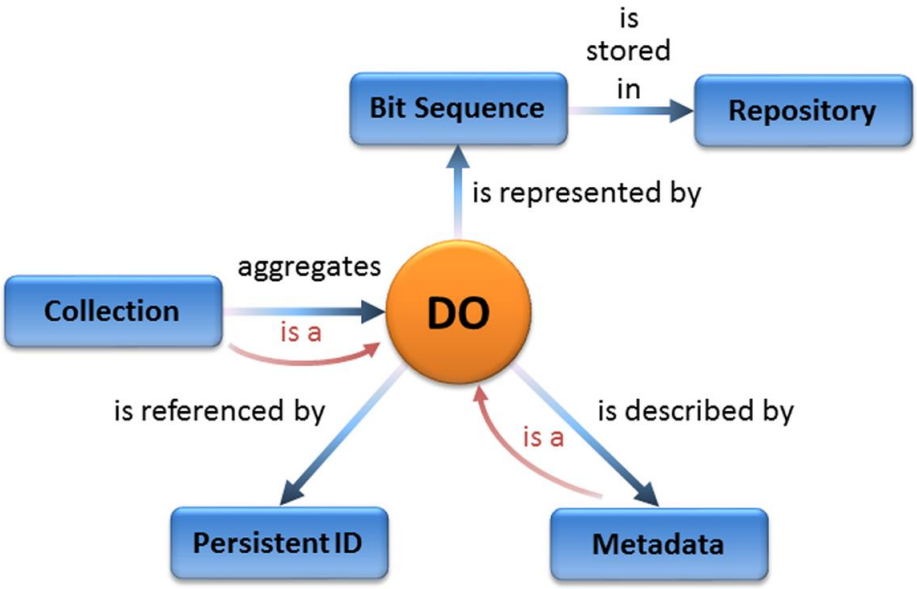

Figure 2: Digital Objects schema [2]

2.2.3. Data standards, metadata standards, vocabularies

Metadata is a central concept for FAIR data and service. It is needed to improve the findability and the re-use. It can be in a structured or simple format and it is recommended to choose a standard for all fields and component that could be existent or of a new definition. The project's ecosystem should be taken into consideration during the FAIRness decisional process because different standards, repositories and data centres could be considered more appropriate for the data and the services than others. Standards are easily findable with online resources similar to FAIRsharing [6] and different communities are working on convergence over a specific standard. An example is the Go-FAIR working group.

Examples of existing metadata and data standards for digital objects (e.g. [8], [9], [10], [11]) are mainly based on Dublin Core [11] and Datacite metadata schema [12] as for the OpenAIRE definition. In the NEANIAS project, we are collaborating with some experts related to this organization, the activities that we performed have been described in chapter 3.

Documentation has a key role in the FAIRness of metadata and data standards. It explains the vocabulary used to describe Digital Objects and it has to be findable and accessible by all the users.

3. FAIR Services and Data in NEANIAS

The NEANIAS core services' portfolio is detailed in deliverable D6.2 Core services Release #1, and related reports 6.3 Core Services Software Release Report. The logic behind the macro-categories (i.e. Sectors C1, C2, C3, C4) and fine granularity of the core services (e.g. C4.4 - Spatial Data Stores) reflects the twofold need of releasing FAIR core services per se and to facilitate the building of FAIR thematic services as concatenation of FAIR micro-services.

Among the NEANIAS services' portfolio, the good candidates for EOSC onboarding are selected according to the maturity level of services and successful assessment with respect to the EOSC onboarding process.

3.1. FAIR requirements for EOSC onboarding

The onboarding process into the European Open Science Cloud consists of several steps - from Provider registration to the Service management – that have been defined to establish the EOSC Catalogue and Marketplace as the reference one-stop-shop to facilitate interoperability and enhance the discoverability of service offer.

The requirements for service integration deal with the adoption of the metadata schemas used to register and describe both Service Provider and the service itself: mandatory and optional requirements are self-described in the metadata attributes per Provider and per Resource. Both API and user interfaces are available to support this process, as well as the validation tools to check the validity of the entries.

Moreover, a set of recommendations and reference implementations of micro-services are offered by EOSC to facilitate the integration of federation services such as

- the Authentication and Authorisation Infrastructure (AAI) that enables authenticated access to services and research data in EOSC
- the monitoring services that provide to service providers functionalities ready-to-use for service health status and service availability monitoring
- accounting services that can be re-used to store, collect and aggregate user accounting records from various services, such as Cloud, HPC and storage usage

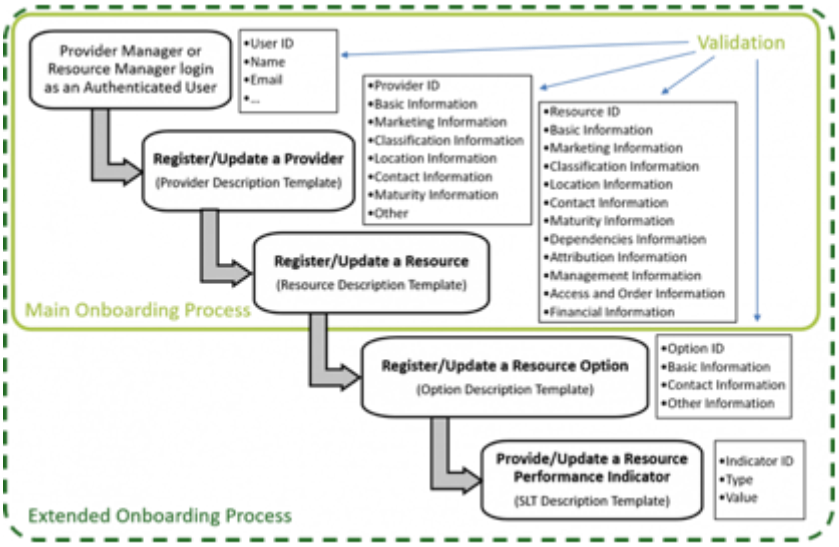

Figure 3: EOSC onboarding schema.

The detailed onboarding process, including mandatory and optional requirements, relevant documentation and tools are available in the EOSC portal[1].

3.2. FAIR level assessment

The FAIR level assessment requires the cross-check of an extensive list of requirements and recommendations. The figure below shows the twenty-seven recommendations from the European Commission Expert Group on FAIR data, including the above / below line that identify priority and supporting recommendations: it is event that the full assessment implies goes beyond the design and implementation of digital resources (i.e. data, services, infrastructures) with implications even in the sustainability and financial domain of the digital resources.

[1] https://eosc-portal.eu/providers-documentation/eosc-provider-portal-basic-guide

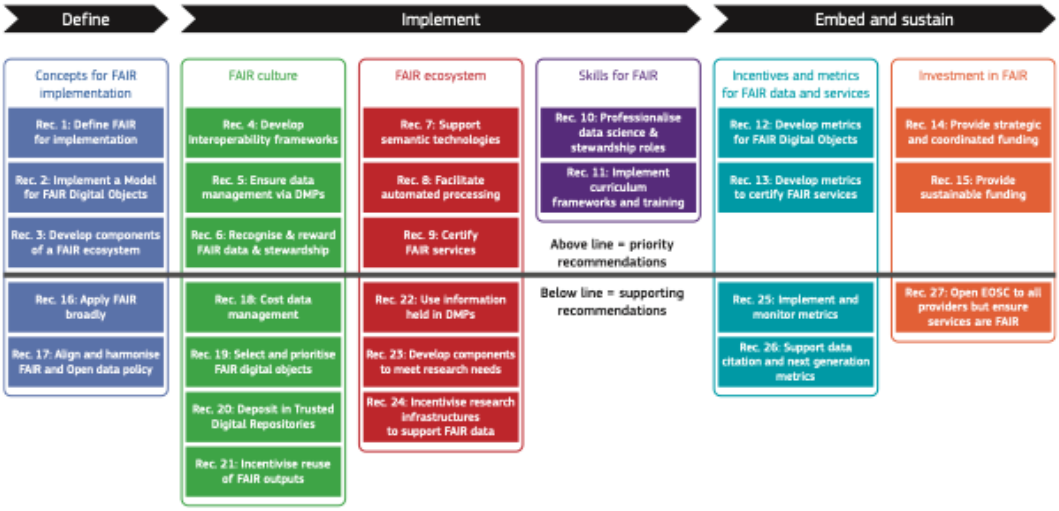

Figure 4: FAIR recommendations from European Commission Expert Group on FAIR data.

Within NEANIAS project activities, FAIR level assessment is narrowed down to the definition and implementation of FAIR culture and ecosystem, with a strong emphasis not only on the consolidated assessment for FAIR data but also on FAIR services.

To boost the FAIRification of NEANIAS services and boost the assessment process, a series of ad-hoc webinars with OpenAIRE experts are scheduled to promote the FAIR culture and initiate both service providers and researchers to the FAIR ecosystem. Relevant use cases are presented and bilateral meetings with NEANIAS service owners are scheduled to assess the FAIRness level and to define an action plan to address open issues and improve the NEANIAS services.

4. Best practices for interoperability and reuse

The best practices and guidelines adopted in NEANIAS for FAIR service development and improvement can be summarized (not exhaustive list):

- To increase the awareness and promote the use of existing FAIR services and the development of new FAIR services
- To encourage the setup of team with cross-domain skills to cover scientific, management, ICT, legal and all needed expertise to achieve the desired level of maturity for the services including Technology Readiness Level, FAIRness Level, reliability, cloud-ready level, ...
- To encourage the connection with FAIR experts
- To encourage the accurate design and implementation of FAIR- / EOSC-ready services, by addressing and considering
    - High modularity of micro-services to support the re-use of the component
    - Review of existing services in the EOSC marketplace to identify reusable service components e.g. integration of EOSC AAI, logging and accounting, data sharing, …
    - Use of existing and/or definition of metadata schemas

    - to support resources discovery (minimum)
    - to enable and improve service interoperability (rich)
    - To support access to digital resources, including the management of restrictions, access controls, license and citations
- To encourage the adoption of data curation and long term data preservation, including Permanent Identifier (PID) and Digital Object Identifier (DOI)
- To encourage the adoption of standards such as but not limited to
    - Data formats (nc, Cloud Optimized Geotiff, jp2, …)
    - OGC services (OpenSearch, Spatio Temporal Asset Catalogue, Web Map Service, Web Coverage Service, Web Processing Services, …)
    - Metadata (INSPIRE, ISO, …)
- To encourage the adoption of REST API to support machine-to-machine interfaces
- To encourage the adoption of and contribution to open-source software to evolve and enhance existing services
- To encourage the adoption of validation tools (applicable to both new and existing services) to test
    - Service interoperability
    - Data / metadata validation (DMP tools)
- To encourage the adoption of Continuous Integration / Continuous Development approach both for the improvement of service reliability and FAIRness level

References

[1] The FAIR Guiding Principles for scientific data management and stewardship, https://www.nature.com/articles/sdata201618.pdf
[2] FAIRsFAIR deliverable D2.1 Report on FAIR requirements for persistence and interoperability 2019. https://zenodo.org/record/3557381
[3] Doldirina, Catherine & Eisenstadt, Anita & Onsrud, Harlan & Uhlir, Paul. (2018). Legal Approaches for Open Access to Research Data. DOI: 10.31228/osf.io/n7gfa.
[4] RDA COVID-19 Working Group. RDA COVID-19 Recommendations and Guidelines on Data Sharing. DOI: 10.15497/RDA00052
[5] https://orcid.org/
[6] https://fairsharing.org/
[7] Digital Curation Centre's metadata directory. Available on https://www.dcc.ac.uk/guidance/standards/metadata [19.12.2020]
[8] RDA Metadata Standard Catalogue. Available on https://rdamsc.bath.ac.uk/ [19.12.2020]
[9] Library of Congress Standards. Available on https://www.loc.gov/librarians/standards [19.12.2020]
[10] Jenn Rilley and Devin Becker. Seeing Standards: A Visualization of the Metadata Universe. Available on http://jennriley.com/metadatamap/ [19.12.2020]
[11] DCMI Metadata Terms. Available on https://dublincore.org/specifications/dublin-core/dcmi-terms/ [19.12.2020]

[12] DataCite Metadata Schema 4.3. Available on https://schema.datacite.org/meta/kernel-4.3/ [19.12.2020]

List of acronyms

Acronym	Description
AAI	Authentication and Authorization Infrastructure
EOSC	European Open Science Cloud
DMP	Data Management Plan
DOI	Digital Object Identifier
FAIR	Findability, Accessibility, Interoperability, and Reuse
HPC	High Performance Computer
ISO	International Organization for Standardization
NEANIAS	Novel EOSC services for Emerging Atmosphere, Underwater and Space Challenges
OGC	Open Geospatial Consortium
PID	Permanent IDentifier
WG	Working Group