EOSC-SYNERGY

EU DELIVERABLE: D3.3

Intermediate report on technical framework for FAIR principles implementation

Document Identifier:	EOSC-SYNERGY-D3.3
Date:	27/08/2020
Activity:	WP3
Lead Partner:	DANS
Document Status:	APPROVED
Dissemination Level:	PUBLIC
Document Link:	

https://drive.google.com/file/d/1wtNdZeb-hl3RI9s5yCRcZU99Gwgw4w4x/

Abstract:

This deliverable introduces the provisional recommendations for assessing data FAIRness, and 'FAIR enabling' data repository features, coming from FAIRsFAIR. It also provides a roadmap for implementation of FAIR requirements, and details about architecture, requirements, and other technical considerations related to Software- and Service-Quality Assurance in line with the focus of the EOSC-SYNERGY project.

EOSC-SYNERGY receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857647

I. Copyright Notice

II. Delivery Slip

	Name	Partner/Activity	Date
From	Wilko Steinhoff	DANS/WP3	27/08/2020
Reviewed by	**Moderator:** **Reviewers:** Valentin Kozlov, Mário David	KIT/WP2 LIP/WP3	27/08/2020
Approved by		PMB	28/08/2020

III. Document Log

Issue	Date	Comment	Author/Partner
1	07/05/2020	Document creation and structure	Gerard Coen/DANS
2	06/07/2020	Added use cases	W.Steinhoff/Tykhonov/DANS
3	24/07/2020	Added additional details about DIGITAL.CSIC	Isabel Bernal/CSIC
4	24/07/2020	Added additional details about Worsica	Alberto Azevedo/LNEC
5	28/07/2020	Added additional details about Worsica	Samuel Bernardo/LIP
6	29/07/2020	Added details about F-UJI tool	W.Steinhoff/DANS
7	30/07/2020	Added tables, and figures. Edited structure and added annexes	Gerard Coen/DANS
8	31/07/2020	Added additional details about DIGITAL.CSIC	Isabel Bernal/CSIC
9	05/08/2020	DIGITAL.CSIC Pilot, FAIR assessment tools, DIGITAL.CSIC FAIR evaluator	Fernando Aguilar/CSIC
10	18/08/2020	Revision, references and formatting	W.Steinhoff/DANS
11	19/08/2020	Added overview of FAIR Assessment tools	Tykhonov/DANS
12	28/08/2020	2nd revision	W.Steinhoff/Gerard Coen/DANS

IV. Authors

Authors (Organisation)	Gerard Coen (DANS), Wilko Steinhoff (DANS), Vyacheslav Tykhonov (DANS), Isabel Bernal (CSIC), Fernando Aguilar (CSIC), Alberto Azevedo (LNEC), Samuel Bernardo (LIP)

V. List of Acronyms

Acronym	Description
AAI	Authentication and Authorization Infrastructure
ABCD	Access to Biological Collection Data
API	Application Programming Interface
CI/CD	Continuous Integration / Continuous Delivery
CTS	Core Trust Seal
DDI	Data Documentation Initiative
DO	Digital Object
EOSC	European Open Science Cloud
FAIR	Findable Accessible Interoperable Reusable
FsF	FAIRsFAIR (INFRAEOSC-5c project)
JePL	Jenkins Pipeline Library
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KPI	Key Performance Indicator
OAIS	Open Archival Information System
OAS	OpenAPI Specification
PID	Persistent Identifier
PMB	Project Management Board
PO	Project Office
RDA	Research Data Alliance
RDF	Resource Description Framework
RDM	Research Data Management

REST	REpresentational State Transfer
RoP	Rule(s) of Participation
SAFE	Standard Archive Format for Europe
SQA	Software Quality Assurance
TDR	Trustworthy Digital Repository
TOC	Table of Contents
URL	Uniform Resource Locator
WP	Work Package
YAML	YAML Ain't Markup Language

Table of Contents

Executive Summary

Expanding the capacity of the EOSC poses the challenge of making sure that any scientific data and thematic services made available can be considered to implement the FAIR principles. These principles advocate for making research data, software and other digital outputs of research as Findable, Accessible, Interoperable and Reusable as possible.

Safeguarding these FAIR principles during implementation, validation and monitoring requires FAIRness assessment and validation of the system where the data will be stored (repositories), the metadata describing the data, and the datasets or digital objects. The co-dependencies between (meta)data, digital objects, and their repository environment are explained. A distinction is also made between different types of research data repositories [Section 2.1].

The requirements of a FAIR data repository, taken from FAIRsFAIR, are described. Also a set of data assessment metrics for programmatic assessment has been defined [Section 2.2 and 2.3] originating from FAIRsFAIR and the recommendations from RDA. To be able to ensure FAIRness automatically is essential to be capable of processing the big volume of data and digital objects in general produced in Europe continuously, as well as enabling objective mechanisms without human intervention. Two candidate automated tools are discussed [Section 2.4].

A roadmap is given on how to go from theory to practice for FAIR implementation within EOSC-SYNERGY task 3.3 [Section 3]. It aims to provide guidance to make data comply with FAIR. The implementation plan is divided into 3 phases: Evaluate, Establish and Endorse, in which a total of 12 steps have been identified. These will be carried out on the basis of three identified use cases, namely: DIGITAL.CSIC, WORSICA and SOCIB.

1. Introduction

1.1 Project context

The European Open Science Cloud (EOSC) programme [R13] brings together institutional, national, and European initiatives to jointly develop a 'science commons' where data are Findable, Accessible, Interoperable, Reusable (FAIR) [R3] and where research-enabling (and other) services are made available throughout the European Union. The EOSC will enhance the possibilities for researchers to find, share and reuse publications, data, and software leading to new insights and innovations, higher research productivity, and improved reproducibility in science.

The EOSC-SYNERGY project [R14] extends the EOSC programme in nine participating countries by harmonizing policies and federating relevant national research e-Infrastructures, scientific data and thematic services, bridging the gap between national initiatives and the EOSC. The project introduces new capabilities by opening national thematic services to European access, thus expanding the EOSC offer in the Environment, Climate Change, Earth Observation and Life Sciences. This is supported by an expansion of the capacity through the federation of compute, storage and data resources aligned with the EOSC and FAIR policies and practices, meaning more compute and storage for researchers, and more datasets and tools to expand avenues of research.

1.2 Scope of this task

Expanding the capacity of the EOSC poses the challenge of making sure that any scientific data and thematic services made available implement the FAIR principles. This requires assessment and validation of the following elements for FAIR principles: the system where the data will be stored (repositories), the metadata describing the data, and the datasets or digital objects. Open data and data management policies that call for the long-term storage and accessibility of data are becoming more and more commonplace in the research community. With it the need for Trustworthy Digital Repositories (TDRs) to store and disseminate data is growing. TDRs capable of curating FAIR data for researchers are a critical requirement for a functioning EOSC [R18].

The FAIRsFAIR project [R15] is engaged in the development of policies and practices that will turn the EOSC into a functioning infrastructure. However the focus in FAIRsFAIR is strongly oriented to the 'traditional' concept of a research data repository. While the development of various FAIR assessment frameworks for repositories, data, and other digital objects has enjoyed substantial activity over the last few years, software and data services have largely been neglected. Also regarding scientific communities it is necessary to establish the boundaries of FAIR - what can be considered as 'FAIR enough'

for different research disciplines and domains. This requires evaluation of the recommendations coming from FAIRsFAIR; as well as consideration of FAIR-related recommendations being generated by other initiatives in the EOSC and FAIR ecosystems; and, close cooperation with repositories and service providers to help them navigate these rapidly moving developments.

Due to the policy oriented focus of FAIRsFAIR, detailed guidance on compliance at the technical level is not offered by the project. As partners in the EOSC programme, EOSC-SYNERGY builds upon the work of FAIRsFAIR by evaluating the applicability of proposed recommendations for the thematic services and repositories involved in this project at a technical level. In doing so, the EOSC-SYNERGY project aims to: expand the EOSC capacity by adding FAIR-compliant data and data services to the EOSC Portal; support thematic services and research data repositories to adhere to the FAIR principles and EOSC Rules of Participation (RoPs) [R21]; and establish a technical framework at the level of implementation, validation, and monitoring which facilitates FAIR data integration to the EOSC. Therefore we aim to provide technical capacities to services (especially data services) to support FAIR principles in different ways.

The objective of this intermediate report is to introduce the provisional recommendations for assessing data FAIRness, and FAIR enabling data repository features, coming from FAIRsFAIR and any other relevant initiatives (RDA, GO FAIR, etc.). It also provides a roadmap for implementation of FAIR requirements, a timeline for work on the task, and details about architecture, requirements, and other technical considerations related to AAI, Software- and Service-Quality Assurance in line with the focus of the EOSC-SYNERGY project.

2. Implementing the FAIR Principles

The FAIR Guiding Principles for Scientific Data Management and Stewardship [R11] emerged in 2016 from a cross-organisational, interdisciplinary effort to refine the availability and usability of digital research data and other digital outputs of research, regardless of their public availability [R12]. At their highest level, the FAIR principles advocate for making research data, software and other digital outputs of research as Findable, Accessible, Interoperable and Reusable as possible. The FAIR principles are created for the consideration of all those who search for, create, manage, share, and use research data. Given the extent to which data handling is (or is becoming) automated, the FAIR principles are also written with machine-actionability in mind (e.g., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention [R16]), 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 [R17]. Annex I lists an overview of the FAIR principles.

The FAIRsFAIR project [R15] is engaged in the development of global standards for FAIR certification of repositories and the data within them, contributing to those policies and practices that will turn the EOSC into a functioning infrastructure. It aims to contribute to the FAIR-oriented dimensions of the Rules of Participation (RoP) [R15] and regulatory compliance for participation in the EOSC. The EOSC governance will use these FAIR aligned RoPs to establish whether components of the infrastructure (such as thematic services, research data repositories, data) implement the FAIR principles.

The recommendations and tools coming from the FAIRsFAIR project are intended to work on various levels. FAIRsFAIR is involved in producing the requirements which repositories should meet with a particular focus on the technical standards for repository interoperability [R37]. Within the project, efforts are being made to develop a FAIR assessment framework for data services. The "*FAIR Certification (of Repositories) - WP4*" [R38], aims to contribute guidance on the implementation of FAIR in relation to certification, and the related metrics, assessment and tools needed to support validation. This work focuses primarily on: FAIR Objects, FAIRness evaluations of individual datasets; FAIR-enabling repositories, supporting the co-development and implementation of certification schemes for data repositories, building on existing framework; and FAIR object and data repository complementarity.

2.1 FAIR Objects and FAIR enabling environments

In order to guarantee the 'FAIRness' of a dataset or digital object it is necessary to evaluate aspects related to the object itself, and the repository where it is stored. The co-dependencies between (meta)data, digital objects, and their repository environment are

introduced below. Across different scientific disciplines and domains both researchers and their repositories work with different assumptions about what constitutes a 'digital object'. The understanding and role of 'metadata' among these communities and repositories also varies. Below is a conceptual model for understanding what can be considered as a FAIR dataset or object which has been taken from FAIRsFAIR D4.2 [R20] and is a modified version of a similar conceptual model originally presented in the '*Turning FAIR Data into Reality*' report [R29].

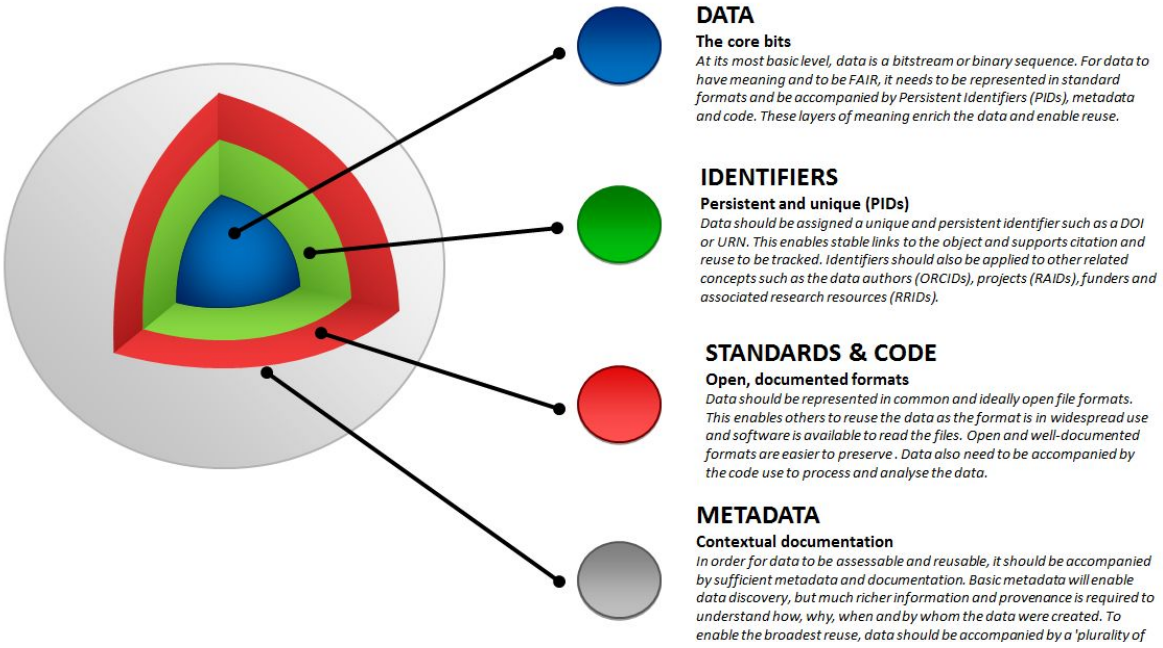

Diagram 1: A model for FAIR Digital Objects (coming from FAIRsFAIR D4.2 [R20])

There are many different definitions of research data available. One definition of research data is: 'The recorded factual material commonly accepted in the scientific community as necessary to validate research findings' [R43]. Research data covers a broad range of types of information, and digital data can be structured and stored in a variety of file formats. Metadata is often defined as "data on data" and describes specific relationships or details between things on these data, it is stored in the data repository database and can be managed and modified by both researchers and data librarians. The task of data librarians is to improve original metadata created by researchers and add more information to increase the chance of their datasets to be found by submitting a relevant request. Metadata can be harvested by using standard protocols like Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [R39] and allows the transfer of any metadata format that has a valid, addressable schema/namespace.

The line dividing data (as the original target for collection/creation) from its supporting metadata is often quite blurry. How to handle data and its supporting metadata does not have a well established clear and consistent practice e.g. some standards support data and

associated metadata contained within a single file. In this case we observe two major groups, one holding metadata about the entire dataset; and the other holding the actual data records (e.g. DDI [R30], ABCD [R31]). A repository is a searchable and queryable interfacing entity that is able to store, manage, maintain and curate Digital Objects [R19]. Repositories create their own 'business information' which include policies, procedures, and other documentation. Through their internal practices and operations repositories generate their own metadata, known as 'process metadata'. Process metadata could range from 'policy review/approval' to 'format risk updated' for example. Technical metadata such as 'validation of a checksum' or 'file format migration completed' might be stored and managed with the object metadata. Both process and technical metadata generated by repositories are important as they can be seen to either enable FAIRness directly, or provide supporting evidence of FAIRness.

Repositories are part of a wider data service ecosystem. Within the context of the EOSC there are different types of repositories, working with different volumes and types of data, managing different ingest and storage requirements, and serving different communities. The FAIR requirements integrated into the EOSC RoPs should therefore remain general enough to be applicable to a broad range of repositories.

In order to produce a framework and outputs which can be later adopted and implemented by actors outside of the EOSC-SYNERGY project, we have classified the repositories involved in the project into two groups (Table 1). This simple classification is intended to serve as a repository classification in the scope of EOSC-SYNERGY, as an aid for clustering use cases within the project, and as a tool for repositories interested in using the framework produced within the project to understand which examples are most relevant for themselves.

General Repository	Thematic Repository
This covers: Institutional research data repositories (e.g. UT DataDOI, Edinburgh DataShare, Harvard Dataverse); National research data repositories (e.g. EASY, Digital.CSIC, ANDS); Multidisciplinary research data repositories (e.g. Figshare, ZENODO, Dryad).	This covers: Disciplinary research data repositories: community oriented (DRYAD ,PANGAEA, BioSharing, TROLLing, DataONE); Thematic research data repositories (e.g. Scientific Drilling Database); and repositories supporting data processing services oriented to thematic areas (e.g. WORSICA, SOCIB).

Table 1: A basic repository classification for the EOSC-SYNERGY project

2.2 Requirements for enabling FAIR in repositories

The requirements of a FAIR data repository as described below (Table 2) are taken from FAIRsFAIR '*D2.3 Set of FAIR data repositories features*' [R8], an output of FAIRsFAIR. The requirements listed below have a particular focus on the technical standards for repository interoperability. Interoperability between repositories and with other components of the EOSC is essential. The importance of repository interoperability will also be specifically addressed as part of the roadmap for implementation [section 3.2].

The requirements cover either an organisational or a technical level or both. Each requirement has been linked to one of the four overarching FAIR principles as indicated in parentheses after each recommendation. In contrast to automated checking of FAIRness of Data Objects (see below section 2.3), a tool to programmatically assess the FAIRness of Data Repositories is not yet available.

FAIRsFAIR requirements for FAIR repositories

Organisational requirements

List of the requirements which can not be implemented on the technical level of data repositories but are targeted at service level agreements, further agreements between users and repositories or communities and data providers.

- The repository itself should have a PID (FA)
- The repository needs to be listed in registries of repositories (F)
- Explicit data deletion policy - explicit roles and responsibilities (I)
- Different access policies for different versions of the data (A)
- Technical support for predefined file formats (I)
- Reuse of community standards and ontologies from public registries (FI)
- Use of PIDs as the manifestation of a data policy (I)
- Only mint one PID per digital object, collection or what one wants to identify (IR)
- Explicit data policies (like versioning and dynamic data) and PID policies in human and machine interoperable way (FAIR)
- Documentation of interfaces and APIs (FAIR)

Technical requirements

Below is a list of technical features which are proposed to improve the FAIRness of data repositories. Currently, those features do not suggest any specific implementation or technology. However, if implemented, they are expected to improve the interoperability between data repositories.

- Metadata for digital objects:
  - The repository should provide metadata in different formats, which can be harvested by different search engines (I)
  - Metadata should be provided as RDF, including JSON-LD. Based on these machines can provide human-friendly presentations/visualisations by resolving the URIs and retrieving the human-readable labels (I)
  - Providing metadata at the level of files, variables, attributes, individual cells, granularity to be decided by the repository (I)
  - Gather provenance metadata on digital objects and files upon upload(IR)
  - Provide masks and ways to quickly upload metadata (I)
  - Demand fine-grained metadata from data providers (FI)
  - Implement community standards (FI)
  - Automatic ontology suggestions and lookup (FI)
  - Landing pages should be machine-interpretable or implement content negotiation, have metadata in different formats (FI)
  - HTTP header should contain technical metadata about the DO (FI)
- Machine-readable and interpretable metadata about repository itself (I)
- Expose (Meta) Data Model (in machine-readable form) (I)
- PID policies
  - PID for each digital object or file (I)
  - Use global persistent identifiers (I)
  - The target of PID should be inferable by machines from PID metadata itself, employ PID information types or Linked Data type (I)
- Data object and file requirements
  - Connect compute infrastructures and data repositories (to avoid commuting data) (I)
  - Subsetting of data (I)
  - Technical support for predefined file formats (including complex data formats like netCDF), with a preference for open file formats (FI)
- Machine-readable license (R)
- The repository should provide a search interface or be linked to aggregating services that enable findability (F)

Not directly linked to FAIR

During the development of these requirements by the FAIRsFAIR team, requirements were identified which indirectly affect FAIR. Although those features do not directly contribute to enabling FAIR in repositories, they are mentioned below for completeness.

- Depending on the nature of the repository and the types of data that it houses, the repository should:
  - Support dynamic data sets (e.g. time series data)
  - Sent notifications to the creator if similar data appears elsewhere
  - Publication tracker for associated datasets
  - Have clear Service Level Agreements
  - Allow citation of reuse of partial data or single elements of datasets
  - Have downloadable citations (e.g. RIS, BibTeX) that point to the data
  - Variety of access restrictions
  - Tombstone[1] procedure
- The repository search interface should have high usability.
- Repository staff should:
  - Provide training on APIs
  - "Spend time being a researcher to better understand the challenges they have making data available in a way that supports findability."

Table 2: Requirements for FAIR repositories (coming from FAIRsFAIR D2.3)

2.3 FAIRness of Digital Objects & Assessment

The establishment of indicators and metrics for assessing compliance of datasets and digital objects with the FAIR principles, as well as the development of the necessary tools to test and validate the metrics are among the many rapidly evolving areas of FAIR. The following set of metrics for programmatic assessment have been defined by the FAIRsFAIR project. Table 3 lists 13 data assessment metrics [R2] proposed by FAIRsFAIR. At present, the metrics address the FAIR principles, with the exception of A1.1, A1.2 (standardised communication protocol), and I2 (FAIR vocabularies). Each metric has an identifier that can be used as a reference in other projects, such as FAIR assessment tools.

[1] A "tombstone page" is a special type of landing page describing an item which has been removed or made inaccessible from a collection [R33].

A detailed overview of the process of the FAIRsFAIR project, and cooperation with the RDA FAIR Data Maturity Model WG [R22], in order to adopt and adapt the metrics below can be seen in the paper '*From Conceptualization to Implementation: FAIR Assessment of Research Data Objects* [R40]; more information on the further development of the metrics is available in the corresponding publication '*FAIRsFAIR Data Object Assessment Metrics*' [R2].

Identifier	Name
FsF-F1-01D	**Data** is assigned a globally unique identifier.
FsF-F1-02D	**Data** is assigned a persistent identifier.
FsF-F2-01M	**Metadata** includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
FsF-F3-01M	**Metadata** includes the identifier of the data it describes.
FsF-F4-01M	**Metadata** is offered in such a way that it can be retrieved by machines.
FsF-A1-01M	**Metadata** contains access level and access conditions of the data.
FsF-A2-01M	**Metadata** remains available, even if the data is no longer available.
FsF-I1-01M	**Metadata** is represented using a formal knowledge representation language.
FsF-I1-02M	**Metadata** uses semantic resources.
FsF-I3-01M	**Metadata** includes links between the data and its related entities.
FsF-R1-01MD	**Metadata** specifies the content of the data.
FsF-R1.1-01M	**Metadata** includes license information under which data can be reused.
FsF-R1.2-01M	**Metadata** includes provenance information about data creation or generation.
FsF-R1.3-01M	**Metadata** follows a standard recommended by the target research community of the data.
FsF-R1.3-02D	**Data** is available in a file format recommended by the target research community.

Table 3: FAIRsFAIR Data Objects Assessment Metrics and their identifiers [R2].

Briefly, the Research Data Alliance (RDA) Working Group, called "FAIR Data Maturity Model WG" and established in January 2019, aimed to develop a common set of core indicators to check the FAIRness of digital objects focused in both data and metadata. As a result of the work performed, a set of guidelines and a checklist related to the implementation of the indicators have been produced which formed an important core of the above metrics.

The guideline document "*FAIR Data Maturity Model: Specification and Guidelines*" [R26], specifies prioritized indicators for assessing adherence to the FAIR principles. These indicators are designed for re-use in evaluation approaches and are accompanied by guidelines for their use. The guidelines are intended to assist evaluators to implement the indicators in the evaluation approach or tool they manage. Therefore these guidelines should also be taken into account.

The recommendations aim to be generic and open-enough to be adaptable to the communities' needs, since flexibility is one of the core of the FAIR principles. Experts from diverse disciplines have participated in the definition of those indicators, and they also include a priority evaluation to define the level of importance, being tagged as: *'Useful'*, *'Important'* or *'Essential'*.

2.4 FAIR assessment tools state-of-the-art

Due to the importance of the FAIR principles adoption for projects, organizations and institutions, diverse tools and services are being developed and available online. Although the number of solutions provided is high, one of the goals of EOSC-SYNERGY is providing a framework to evaluate FAIRness automatically. This is essential to be capable of processing the big volume of data and digital objects in general produced in Europe continuously, as well as enabling objective mechanisms without human intervention.

Some organizations and associations like RDA provide manual evaluators that allow data managers to know how well their digital objects are FAIR, based on questionnaires or checklists. This has some drawbacks since manually answers are limited by the knowledge of the people answering, it can include some misinformation or over-estimate the status of the digital objects. The granularity of this kind of questionnaire is diverse, and some of these include more generic questions than others, which ask at different levels of details. Table 4 provides a list of state-of-the-art FAIR assessment tools of which many are still under development. Note that the majority of the tools listed are self-assessment tools. Only the last three tools are automated assessment tools. The topic of FAIRness metrics is still under active discussion in the FAIR community.

No	Tool	Organization	Status	Link
1	ANDS-NECTAR-RDS-FAIR data assessment tool	ARDC	Tool	[Link]
2	DANS-FAIRdat	DANS-KNAW	Tool, pilot	[Link]

3	DANS-Fair enough?	DANS-KNAW	Tool, pilot	[Link]
4	The CSIRO 5-star Data Rating tool	CSIRO	Tool, V1.0	[Link]
5	FAIR Metrics Questionnaire	The FAIR Metrics Group	Document	[Link]
6	Stewardship Maturity Mix	NOAA's CICS-NC, NOAA's NCDC	Document	[Link]
7	FAIR Evaluator	GO FAIR, LUMC CBGP, IDS, OeRC, IQSS	Document	[Link]
8	Data Stewardship Wizard	ELIXIR NL/CZ	Tool	[Link]
9	Checklist for Evaluation of Dataset Fitness for Use	Assessment of Data Fitness for Use WG (WDS/RDA)	Document	[Link]
10	RDA-SHARC Evaluation	SHARC IG (RDA)	Poster	[Link]
11	WMO-Wide Stewardship Maturity Matrix for Climate Data	The SMM-CD WG	Document	[Link]
12	Data Use and Services Maturity Matrix	The MM-Serv WG	Document	[Link]
13	FAIRAWARE Assessment Tool	DANS-KNAW	Tool, pilot	[Link]
14	F-UJI (FAIRsFAIR Research Data Object Assessment Service)	FAIRsFAIR project	Automated tool	[Link]
15	FAIR evaluator	DIGITAL.CSIC	Automated tool	n.a.
16	FAIR Evaluation Services	FAIRmetrics and FAIRsharing groups	Automated tool	[Link]

Table 4. Overview of FAIR Assessment tools and related links

Due to the diversity in terms of formats, standards, data and metadata types, it is relatively easy to have general indicators of the digital objects FAIRness, but harder to get more details as some indicators are treated at a disciplinary level. There are expected to be differences between what FAIR means for one community versus another, this leads to the notion of 'FAIR Enough'. Research communities, as well as service providers, are expected to play a role in defining and agreeing which indicators of FAIRness might be minimized, adapted, or ignored.

The goal of EOSC-SYNERGY is to move the state-of-art beyond manual assessment tools. This will be achieved through a framework that will run assessment tests automatically to get some measurements in terms of FAIR. To this end the project is identifying tools and capabilities to be integrated in the framework. The more advanced service available is the "FAIR Evaluation Services", an online tool developed by a consortium associated with *FAIRsharing.org*. It has been designed with a modular approach, giving the chance to external users to define their own FAIR Maturity Indicator tests based on APIs. It also allows to create complete collections of tests (e.g. disciplinary) and evaluate resources selecting the proper indicators for each community. To do so, the digital object can be provided using its globally unique identifier (DOI, Handle, etc.). This open and modular approach facilitates the participation of the communities not only adding their tests but also publishing the results of the evaluations done transparently. The code is open source and it is available on GitHub [R35].

As part of the work to develop global standards for FAIR certification of repositories and the data within them, FAIRsFAIR is working on the production of an automated FAIR data assessment toolset and badging scheme. (See STEP 9 of the roadmap [section 3.2] for more detail on the relationship between badging and certification in both FAIRsFAIR and EOSC-SYNERGY).

The '*FAIRsFAIR Research Data Object Assessment Service*' (F-UJI) web service has been made available to programmatically assess FAIRness of research data objects based on the Data Objects Assessment Metrics (Table 3). An assessment of the FAIRness of data or digital objects is performed based on aggregated metadata. This includes metadata embedded in the data (landing) page, metadata retrieved from a PID provider (e.g. Datacite content negotiation), and other services (e.g. re3data [R28]). The program is written in the programming language Python and uses the OpenAPI Specification (OAS) and is published under the MIT License. OAS defines a standard, language-agnostic interface to RESTful APIs. The RESTful API enables integration with other applications. This means the API can be queried during validation and monitoring of digital objects which live in a research data repository. This will help to ensure that research data objects remain FAIR over time. F-UJI can be run locally or in a Docker container. The GitHub repository for the tool is available at: https://github.com/FAIRsFAIR/fuji [R41]. Currently no Dockerfile is available from the GitHub repository.

It should be noted that while an expected outcome of the work in FAIRsFAIR is badging, the current tool does not provide any validation or certification, instead it provides a 'checklist' of the assessment metrics. The outcome of the tool is a JSON file that describes the status for each FAIRsFAIR metric identifier (Table 3) and a test-status: "`pass`" or "`fail`". An example of a data object assessment conducted using F-UJI is shown in Annex II.

A few issues should be considered when adopting this tool as a service to perform automated FAIR assessments:

- FAIR assessment must go beyond the data object itself. FAIR enabling services and repositories are vital to ensure that research data objects remain FAIR over time. Therefore, machine-readable services (e.g., registries) and documents (e.g., policies) are required to enable automated tests.
- In addition to repository and services requirements, automated testing depends on clear machine assessable criteria. Some aspects (rich, plurality, accurate, relevant) specified in FAIR principles will still require human mediation and interpretation.
- The tests can only focus on generally applicable data/metadata characteristics until domain/community-driven criteria have been agreed upon (e.g., appropriate schemas and required elements for usage/access control, etc.). For example, for some of the metrics (e.g. on I and R principles), the automated tests it performs, only inspect the 'surface' of criteria to be evaluated. Therefore, tests are designed in consideration of generic cross-domain metadata standards (e.g. Dublin Core, DCAT, DataCite, schema.org, etc).

Another tool is being developed by DIGITAL.CSIC, EOSC Synergy partner. DIGITAL.CSIC as a repository aims to provide a service to enable self-assessment for checking how well a digital resource provided by researchers fulfil the FAIR principles. The FAIR evaluator for DIGITAL.CSIC is a tool (prototype) being developed in Python that interacts with the repository to extract information to decide if a digital object published there satisfies the metrics defined by the list of FAIR indicator provided by the "FAIR Data Maturity Model WG" at the Research Data Alliance (RDA). The document "*FAIR Data Maturity Model: Specification and Guidelines*" [R26] specifies a prioritized list of indicators for the FAIR assessment as discussed earlier.

In different groups, following the four FAIR principles (Findable, Accessible, Interoperable, Reusable), this tool implements technically the indicators to work with DIGITAL.CSIC and provides points to know how your digital object fulfills that principle, obtaining an overall rating. The tool is intended not only to show how FAIR your data is, but also to guide you how it can be improved, providing a set of recommendations, and guidelines to address the different characteristics defined in the indicators. Although this tool was originally designed to interact with DIGITAL.CSIC only, it can be generalized to work with any other repository, including generic tests, and re-define methods at technical level if needed.

3. Roadmap

To go from theory to practice for FAIR implementation, EOSC-SYNERGY task T3.3 aims to provide guidance to make data comply with FAIR. The work will build upon the recommendations for FAIR practices and policies mainly delivered by the FAIRsFAIR project while also considering the output from other projects and initiatives which touch on this topic. The implementation of the FAIR data principles implies the timely analysis of the outcomes on FAIR practices and policies, and the design of a technical framework to support those best practices operationally at the level of: implementation, validation and monitoring, as shown graphically in diagram 3:

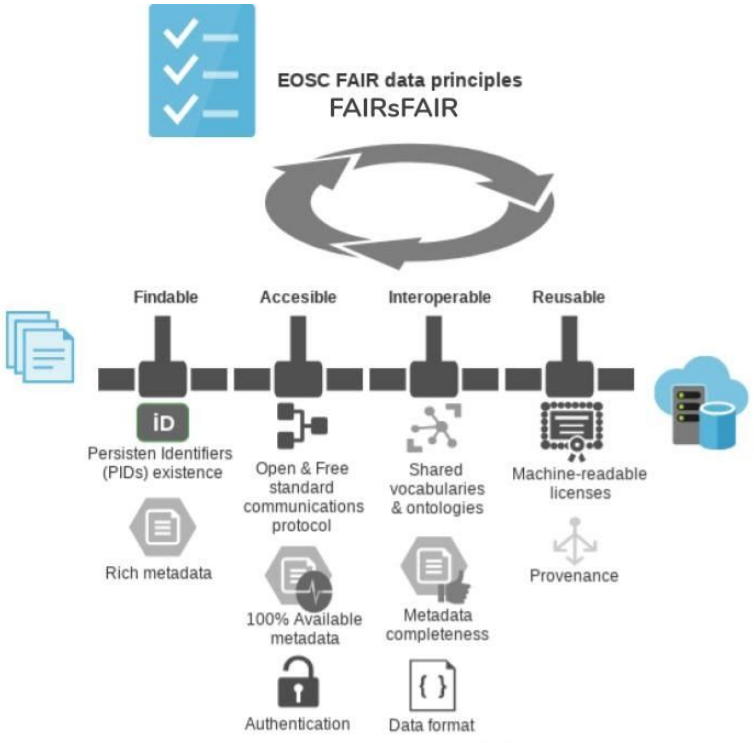

Diagram 3: Supporting and fostering data FAIR adoption in cooperation with FAIRsFAIR

3.1 The EOSC-SYNERGY quality framework

EOSC-SYNERGY is developing a quality based approach to foster the adoption of EOSC software and services, which will improve, promote and reward quality. This approach is described in deliverable D3.1 [R23] and relies on automated quality assessment of relevant quality criteria.

Quality assessment is an important trait. It allows to build higher trust that items submitted to assessment meet their requirements, and contributes to the maintainability,

stability and sustainability. Finally, it contributes to facilitate the collaboration between developers and promotes good practices.

Task T3.1 is developing two sets of quality criteria described in the documents "*Software Quality Assurance baseline document*" [R24] and "*Service Quality Assurance baseline document*" [R25].

The EOSC-SYNERGY quality assurance framework currently being developed by task T3.2 (SQAaaS) will leverage the quality criteria developed by task T3.1 to enable the automated quality assessment of software and services. The SQAaaS framework architecture is introduced in deliverable D3.1. The framework leverages CI/CD pipelines to automate quality verification and delivery. The core of the framework is fully generic and can be applied to a wide range of Information Technology situations where automation is required.

Task 3.3 will extend the current sets of quality criteria beyond software and services by identifying relevant quality criteria for automated FAIR assessment. Furthermore task T3.3 aims to identify tools suitable for automated FAIR assessment. Whenever possible the relevant capabilities provided by those tools will be integrated using the pipelines being developed by task T3.2, thus extending the EOSC-SYNERGY quality assessment functionality beyond software and services to include FAIRness. Task T3.3 will also exploit the use of the T3.1 and T3.2 developments related to software and services in the context of data repository services.

3.2 Implementation plan

The work in task T3.3. is divided into 3 phases:

EVALUATE (M1-9)

This phase deals with the setup of the task, clarification of the scope, as well as the completion of necessary landscaping work, establishing synergies and partnerships with internal and external project stakeholders.

ESTABLISH (M10-24)

This phase deals with ensuring the readiness with project partners to have a common baselines for the FAIRification work. A common baseline will allow for an equitative benchmark for assessment, the evaluation of recommendations coming from FAIRsFAIR, and also the tracking of progress made to implement FAIR over the lifetime of the project among the partners involved.

ENDORSE (M25-30)

This phase deals with the aspects of validation and monitoring of the FAIR implementation and also the communication and exploitation of the FAIR-related outputs of the project, primarily the framework of guidance, tools and services for FAIR implementation.

Phase One: EVALUATE (M1-9)

STEP 1) Use case identification: Based on '*D4.1 – Best Practices Elicitation including Data Management Plans*' we have identified three use cases (e.g. Digital.CSIC, WORSICA and SOCIB) to be involved in the work on implementing the recommendations coming from FAIRsFAIR. A detailed overview of the use cases currently involved in the work is available in Annex III, which provides ample background and technical information about the repositories. The pool of use cases involved in the project have been divided into two clusters. The first cluster of use cases includes both 'fully fledged' research data repositories (e.g. Digital.CSIC) and also thematic services who are actively involved in storing or sharing research data and/or already committed to making their data FAIR (e.g. WORSICA). A second set of use cases has also been identified to be evaluated later in the project lifecycle (see STEP 8). The second cluster of use cases includes those thematic services principally involved in: generating dynamic data (e.g. using instruments); generating or processing data in the petabyte range with high sustained ingest rates; and/or providing services on top of data stored elsewhere.

STEP 2) Repository classification: As addressed in section 2, in order to produce a framework and outputs which can be later adopted and implemented by actors outside of the project and to ease work internally we have split repositories using a simple classification.

During this process we also considered criteria related to: the purpose of the repository; the community(s) served; the (perceived) maturity of the repository; the role of data sharing, storage, and transfer played by the repository; and aspects related to the four V's of Big Data (Volume, Variety, Velocity and Veracity) [R32].

STEP 3) Scope clarification and roadmap development: This step involved the identification of the internal and external project dependencies in order to generate synergies and establish a feasible roadmap for the work. Some of the dependencies considered when building the roadmap can be seen in Annex IV.

Phase One expected outputs:

- Use cases defined (M6)
- Repository classification established (M7; see Table 1)
- Synergies and dependencies identified (M8; see Annex IV)
- Implementation roadmap developed (M9)

Phase Two: ESTABLISH (M10-24)

STEP 4) Establish technical readiness: An important aspect of the EOSC-SYNERGY project is the harmonization of EOSC-relevant thematic initiatives and use cases in relevant scientific areas (Earth Observation, Biomedicine, Astrophysics and Environment) providing open research data and services. Before assessing the FAIRness of repositories and services a baseline technical readiness level needed to be established. This involved making sure that the first set of use cases are in compliance with the Software- and Service-QA (Quality Assurance) baselines [R36] proposed by T3.1. This work has involved facilitating use cases to align their systems in a number of ways including where applicable: upgrading to the most recent version of Dataverse [R46] or DSpace [R47], integrating with federated AAI ('*EGI AAI Check-in service*' [R42]), and simplifying automated repository deployment through the use of Linux containers for validation and adoption.

STEP 5) Establish the baseline level of repository FAIRness: As described in section 2, in order to guarantee the FAIRness of a dataset or 'digital object' it is necessary to evaluate aspects related to the object itself, and the repository where it is stored. This step involves the evaluation of the 'first cluster' of use cases to establish the presence of 'FAIR enabling' features as per section 2.2. The outcome of this will be used as a first indication whether the recommendations provided from FAIRsFAIR are general enough to be applicable to all repositories in the project, and to validate the automation strategy of EOSC-SYNERGY. This will enable a better understanding of the relevant FAIR properties.

STEP 6) Establish the baseline level of 'object' FAIRness: Using the most suitable FAIR assessment tools previously identified [section 2.4], the FAIRness of research objects, datasets and collections stored by repositories will be assessed. This will produce feedback on the suitability of the tools for integration with the quality pipelines framework being developed by T3.2, and/or the implementation of further changes to FAIRify the

systems and processes of the use cases. The adaptation and integration work to be performed together with T3.2 forms the second core aspect of the technical framework to be addressed in STEP 7.

STEP 7) Development of a draft technical framework: Having worked in close collaboration with the first set of use cases to evaluate the previously identified FAIR assessment tools and the FAIR requirements, a first implementation of the technical framework will be developed. This framework will integrate the suitable FAIR assessment tools with the EOSC-SYNERGY quality pipelines being developed by T3.2 (further detail is provided in STEP 11). By M19 is also expected to have integrated the 1st reference implementation of the data repositories features coming from FAIRsFAIR. The reference implementation is primarily focused on improving interoperability.

STEP 8) Expanding the collaboration to the second set of use cases: The work will be expanded to the second set of use cases. The intention of leaving this set of use cases until later in the project lifecycle is two-fold:
1) Experience and feedback from the first three use cases will be available making the second set of use cases easier to address;
2) Working with these use cases at a later stage enables leveraging the fine tuned recommendations, tools and other useful resources such as further FAIRsFAIR outputs.

Phase Two expected outputs:

- This document D3.3. '*Intermediate report on technical framework for FAIR principles implementation*' (M12)
- 1st evaluation of FAIR enabling repository features completed (M14)
- 1st assessment of FAIR data or digital object features completed (M14)
- Draft technical framework for FAIR implementation (M24).
- Badge issuing (M3.4) should be implemented (M24). Aligned to work on badge issuing from FAIRsFAIR project also.

Phase Three: ENDORSE (M25-30)

STEP 9) Validation, Badging and Certification: Work to validate compliance with criteria and requirements is a core aspect of both EOSC-SYNERGY and FAIRsFAIR. The automated pipelines developed by T3.2 will ensure that repositories and services connecting to the EOSC can be badged as being "EOSC-SYNERGY ready" in terms of Software- and Service-QA. A state of the art whitepaper regarding digital badge issuing technologies has already been published by T3.2 and work on this topic is already quite advanced: https://digital.csic.es/handle/10261/206348. The equivalent badging mechanism as an outcome of the automated FAIR assessment introduced in section 2 is also planned but not expected to be delivered by FAIRsFAIR WP4 until August 2021 (*D4.5 Report on FAIR data assessment toolset and badging scheme*). It is the ambition that T3.3. will facilitate the integration of automated FAIRness badging in the pipelines developed in T3.2, and/or as part of the portfolio of tools and services to be made available via the technical framework.

Current JePL library [R44] developed in T3.2 allows the integration of tools from several build systems and programming languages. The current implementation defines SQA criterias that are linked with the correspondent set of tools [R45]. Those are launched within a predefined docker container provided by the user or selected over a set of suggested images.

In T3.2 continuous integration (CI) context, the provided framework expects a solution that can be deployed using docker images. Also SQAaaS web portal interface expects a REST API following OpenAPI standard to release a better binding between services in the end. The F-UJI web service fits perfectly on this solution.

It is not considered possible to fully automatically certify repositories as TDRs given the level of human-machine interaction needed in research data management, and the role of data stewards in facilitating data FAIRness. Nevertheless we aim to ensure that the technical aspects which repositories must comply with to achieve CoreTrustSeal certification will be met and can be automatically tested as part of the EOSC-SYNERGY (& FAIRsFAIR) badging process [R20].

STEP 10) EOSC Portal integration: In cooperation with T3.2 generic pipelines and workflows will be further developed to aid the integration of further repositories and digital objects into EOSC common catalogs, aiming at integration of this framework in the EOSC Portal. This step is aligned with the goal of making further FAIR- compliant data available for research.

STEP 11) Development of a final technical framework: The final release of the technical framework for FAIR is expected to cover the following aspects:

- Guidance for the integration of machine-actionable features to automate the FAIR assessment and validation process using standard protocols.
- Automated tools for the assessment of FAIRness at the level of both repositories and digital objects which can be integrated into internal repository workflows, *or* as part of the layer facilitating interoperability between repositories integrated to the EOSC, *or* the necessary pipelines for connecting to the EOSC Portal etc.
- Clear examples of technical solutions for the implementation of FAIR for repositories. Covering the most common repository platforms Dataverse and DSpace, also with examples for custom developments. T3.3. aims to be specific rather than technology agnostic - providing concrete implementation examples while trying to cover as broad a range of possibilities as possible.
- Information on: the architecture and dependencies for using the EOSC-SYNERGY pipeline orchestrator; making repositories and datasets FAIR through facilitating inclusion in EOSC common catalogs; the process for integrating with the EOSC Portal.

- Monitoring will be modelled on the CI/CD approach whenever a dataset or digital object is modified, it will be re-checked in order to certify its current FAIRness level. Other monitoring issues like data citation and reusability indexes will be taken into account.

Implementation of requirements which do not relate directly to FAIRness (e.g. AAI) will be dealt with by WP2.

STEP 12) Maintenance (sustainability) and alignment with the EOSC Rules of Participation: The last activity as part of this task will be to ensure the continued functioning and maintenance of any tools, services or infrastructure included in the framework. It will also seek to make sure that all of the use cases involved in the FAIR implementation task meet the FAIR-related EOSC RoPs.

Phase Three expected outputs:

- Relevant use cases achieve FAIR compliance (M26)
- Final technical framework for FAIR implementation (M26)
- D3.5 Final report on technical framework for EOSC FAIR data principles implementation (M27)
- At least 8 FAIR-compliant data collections available in the EOSC (M30)
- FAIR badging achieved for the use cases involved in the task (M30)
- FAIR compliant services, repositories and tools will be available through the EOSC Portal (M30).

4. Current status of the work & next steps

As per the roadmap presented in section 3, we are currently working with three use cases to evaluate the FAIRness requirements coming from the FAIRsFAIR project, both for the FAIRness assessment digital objects, and investigating the presence of 'FAIR enabling' repository features. These use cases cover both national repositories for long-term preservation, and services which have a research data-repository implementation installed (covering both 'General Repository' and 'Thematic Repository' examples from Table 1).

The work is currently in Phase Two and spread across Steps 4, 5, and 6. As part of the drive to improve FAIRness, and also associated with the complexity of adapting services which are in production, the work in Step 4 to ensure a common technical baseline has been challenging given that each of the use cases work with a different repository platform. As described in Annex III, the use cases cover repositories using DSpace, Dataverse, and custom-development repository software.

The CSIC institutional repository already complies with a wide range of FAIR requirements, in particular as far as appropriate and granular usage of standard metadata, PIDs assignation and deployment of open, free and interoperable protocols are concerned. In addition, the repository covers most of organizational considerations and other issues not directly linked to FAIR principles (shown in section 2.2. above).

DIGITAL.CSIC participation in EOSC-SYNERGY builds on the following motivations:

- Complete alignment with FAIR principles;
- Enhance its role as data provider in EOSC;
- Get exposed to latest trends and emerging technologies as regards automation processes in line with FAIR principles and develop new services for CSIC community and external users;
- Test selected APIs and other technologies that will enable automated processed on top of its collections, in particular those with datasets and research software and
- Better prepare the renewal of Data Seal of Approval, now merged with other repository certifications in CoreTrust Seal

DIGITAL.CSIC already offers a significant volume of research outputs (around 210,000 items of which 62% are open access) which gathers large usage statistics (altogether around 100 million downloads of single files since 2009, with current average monthly downloads in the region of 1,400,000). Interest and usage of CSIC collections are out of doubt given this data, however, other than OAI-PMH and SWORD technologies, it currently lacks APIs and other tools that may enable users to process automated activities on top of its collections. Therefore, one principal motivation is to start enriching the repository infrastructure with selected APIs in order to enhance the set of services offered to the institutional community and to facilitate reusability of its contents. Further, a complete

alignment with FAIR principles will be achieved once the repository will be able to integrate such APIs and at the same time this upgrade will maximize opportunities for DIGITAL.CSIC to play a more active role in the EOSC ecosystem. Last but not the least, these improvements will be instrumental to a successful outcome in the planned application for CoreTrust Seal.

Aiming to enhance the DIGITAL.CSIC functionalities to enable machine-actionable features related to FAIR principles, a prototype version of the repository based on DSpace-CRIS 7 has been deployed. This has two main goals: on one hand, to have a more modern and robust system capable of adopting CI/CD methodologies, as well as to exploit the SQaaS proposed by WP3. On the other hand, to enable APIs and other flexible features to automate processes like FAIRness checks, metadata extraction, data retrieval, etc.

The new versions of DSpace-CRIS have a modular approach based on dockers, so they can be easily deployed and increase the number of component instances depending on the demand. This will enable scalable features for the new version of DIGITAL.CSIC. Furthermore, it can be integrated with the SQaaS, testing any new development automatically. Having API methods to manage the repository almost completely, a number of processes can be automated. Combining metadata extraction and digital object retrieval, different FAIRness tests can be implemented and automated. Selecting representative datasets published at DIGITAL.CSIC, the roadmap proposed in section 3 will be addressed in the production version of the repository, but with special attention to the new version to be deployed, which will facilitate the automatic interaction and increase the flexibility of actions.

The repository of WORSICA is currently not in production since a number of larger changes are being implemented to the systems as part of ambitious planning to broaden and upscale their data intensive activities. Since the recommendations for enabling FAIR in repositories is a manual exercise this can be undertaken by the WORSICA team directly and is now underway. An automated assessment of the Object FAIRness of datasets in the WORSICA repository is currently not possible but the development team are working on establishing both unique links and PIDs for datasets which will in turn improve FAIRness and permit the use of the F-UJI tool among others.

The repository of SOCIB has just recently finished a self-assessment to apply for CoreTrustSeal Certification. It will become available for participation in EOSC-SYNERGY starting from M13, however the information sharing to establish SOCIB among the first set of use cases is already taking place.

References

R1	Scott Bradner; Key words for use in RFCs to Indicate Requirement Levels, (1997): https://www.ietf.org/rfc/rfc2119.txt
R2	Devaraju, Anusuriya, Huber, Robert, Mokrane, Mustapha, Cepinskas, Linas, Davidson, Joy, Herterich, Patricia, … White, Angus. (2020, July 10). FAIRsFAIR Data Object Assessment Metrics (Version 0.3). Zenodo. http://doi.org/10.5281/zenodo.3934401
R3	FAIR Principles: https://www.go-fair.org/fair-principles/
R4	Worsica: https://github.com/WorSiCa, http://worsica.lnec.pt/
R5	Experimental Digital.CSIC: http://193.146.75.184:3000
R6	DSpace-CRIS: https://wiki.lyrasis.org/display/DSPACECRIS
R7	Wittenburg et al., 2018: Digital Objects as Drivers towards Convergence in Data Infrastructures; http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11
R8	D2.3 Set of FAIR data repositories features; https://doi.org/10.5281/zenodo.3631527
R9	EGI AAI Check-in service
R10	Crosas, M. Ph.D. @ European DataverseWorkshop 2020, Tromso, Norway: Fair Principles and Beyond: Implementation in Dataverse; https://scholar.harvard.edu/files/mercecrosas/files/fair-dataverse-tromso.pdf
R11	Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18
R12	Turning FAIR Into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data. https://op.europa.eu/s/n1qc FAIRsFAIR (2020). Retrieved from https://www.fairsfair.eu/
R13	European Open Science Cloud (EOSC) programme; https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
R14	EOSC-SYNERGY project website: https://www.eosc-synergy.eu
R15	FAIRsFAIR project website: https://www.fairsfair.eu/the-project
R16	GO FAIR Glossary: https://www.go-fair.org/resources/glossary/
R17	FAIRsFAIR D3.4 Recommendations on practice to support FAIR data principles: https://doi.org/10.5281/zenodo.3924132
R18	FAIRsFAIR D5.3 Report on the First Synchronisation Force Workshop: https://doi.org/10.5281/zenodo.3629159
R19	RDA Data Foundations Terminology: https://smw-rda.esc.rzg.mpg.de/dft-2.0.html#Repository
R20	FAIRsFAIR D4.2 Repository Certification Mechanism: a Recommendation on the Extended Requirements and Procedures: https://doi.org/10.5281/zenodo.3835698
R21	European Open Science Cloud Rules of Participation Version 0.2 (29 January 2020): https://repository.eoscsecretariat.eu/index.php/s/QWd7tZ7xSWJsesn#pdfviewer
R22	FAIR Data Maturity Model WG: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

R23	EOSC-SYNERGY D3.1 Software Maturity baseline (29 June 2020): Internally available.
R24	Pablo Orviz, Alvaro Lopez, Doina Cristina Duma, Mario David, Jorge Gomes, Giacinto Donvito, “A set of Common Software Quality Assurance Baseline Criteria for Research Projects”, 2017, http://dx.doi.org/10.20350/digitalCSIC/12543 (Github repository: https://github.com/indigo-dc/sqa-baseline)
R25	Orviz Fernández, Pablo ; Mario David; Jorge Gomes; Joao Pina; Samuel Bernardo; Campos Plasencia, Isabel ; Germán Moltó; Miguel Caballer, “EOSC-Synergy: A set of Common Service Quality Assurance Baseline Criteria for Research Projects”, June 2020: DOI http://dx.doi.org/10.20350/digitalCSIC/12533 (Github repository: https://github.com/EOSC-synergy/service-qa-baseline)
R26	RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. Research Data Alliance. DOI: 10.15497/RDA00050
R28	Re3data webportal: https://www.re3data.org
R29	Directorate-General for Research and Innovation (European Commission), 2018: Turning FAIR into reality; https://doi.org/10.2777/1524
R30	Data Documentation Initiative: https://ddialliance.org/Specification/
R31	Access to Biological Collection Data: https://abcd.tdwg.org
R32	The Four V's of Big Data: https://www.ibmbigdatahub.com/infographic/four-vs-big-data
R33	Best Practices for Tombstone Pages: https://support.datacite.org/docs/tombstone-pages
R35	FAIR Maturity Evaluation service: https://github.com/FAIRsharing/FAIR-Evaluator-FrontEnd
R36	A set of Common Software Quality Assurance Baseline Criteria for Research Projects: https://github.com/indigo-dc/sqa-baseline
R37	FAIR Practices: Semantics, Interoperability, and Services - WP2: https://www.fairsfair.eu/fair-practices-semantics-interoperability-and-services
R38	FAIR Certification (of Repositories) - WP4: https://www.fairsfair.eu/fair-certification
R39	The Open Archives Initiative Protocol for Metadata Harvesting: https://www.openarchives.org/pmh
R40	Anusuriya Devaraju, Mustapha Mokrane, Linas Cepinskas, Robert Huber, Patricia Herterich, Jerry de Vries, Vesa Akerman, Hervé L’Hours, Joy Davidson, Michael Diepenbroek. “From Conceptualization to Implementation: FAIR Assessment of Research Data Objects” (2020, pending)
R41	FAIRsFAIR Research Data Object Assessment Service: https://github.com/FAIRsFAIR/fuji
R42	EGI AAI Check-in Service: https://wiki.egi.eu/wiki/AAI
R43	University of Edinburgh: https://www.ed.ac.uk/information-services/research-support/research-data-service
R44	https://github.com/indigo-dc/jenkins-pipeline-library/tree/release/2.1.0
R45	https://indigo-dc.github.io/jenkins-pipeline-library/release_2.1.0/user/step_by_step/sqa_criteria.html
R46	Dataverse Project: https://dataverse.org
R47	DSpace: https://duraspace.org/dspace

Annex I - List of FAIR Principles

Overview of the FAIR Principles [R3]:

Findable	
F1	(Meta) data are assigned globally unique and persistent identifiers
F2	Data are described with rich metadata
F3	Metadata clearly and explicitly include the identifier of the data they describe
F4	(Meta)data are registered or indexed in a searchable resource
Accessible	
A1	(Meta)data are retrievable by their identifier using a standardised communication protocol
A1.1	The protocol is open, free and universally implementable
A1.2	The protocol allows for an authentication and authorisation where necessary
A2	Metadata should be accessible even when the data is no longer available
Interoperable	
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2	(Meta)data use vocabularies that follow the FAIR principles
I3	(Meta)data include qualified references to other (meta)data
Reusable	
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

Annex II - F-UJI assessment example

Example request command to start an assessment on localhost for persistent identifier: “https://hdl.handle.net/10261/134880”:

```
$ curl -X POST "http://localhost:1071/fuji/api/v1/evaluate" -H  "accept: application/json" -H  "Authorization: Basic dXNlcm5hbWU6cGFzc3dvcmQ=" -H "Content-Type: application/json" -d "{\"object_identifier\":\"https://hdl.handle.net/10261/134880\",\"test_debug\":true}"
```

F-UJI API Response with references to the FAIRsFAIR metric identifiers:


```
[
 {
   "id": 1,
   "metric_identifier": "FsF-F1-01D",
   "metric_name": "Data is assigned a globally unique identifier.",
   "output": {
     "guid": "https://hdl.handle.net/10261/134880",
     "guid_scheme": "handle"
   },
   "score": {
     "earned": 1,
     "total": 1
   },
   "test_debug": [
     "INFO: Unique identifier schemes found ['handle', 'url']",
     "INFO: Finalized unique identifier scheme - handle"
   ],
   "test_status": "pass"
 },
 {
   "id": 2,
   "metric_identifier": "FsF-F1-02D",
   "metric_name": "Data is assigned a persistent identifier.",
   "output": {
     "pid": "https://hdl.handle.net/10261/134880",
     "pid_scheme": "handle",
     "resolvable_status": true,
     "resolved_url": "https://digital.csic.es/handle/10261/134880"
   },
   "score": {
     "earned": 1,
     "total": 1
   },
   "test_debug": [
```

```
      "INFO: Persistence identifier scheme - handle",
      "INFO: Retrieving page http://hdl.handle.net/10261/134880",
      "INFO: Content negotiation accept=text/html, application/xhtml+xml,
status=200",
      "INFO: Found HTML page!",
      "INFO: Object identifier active (status code = 200)"
    ],
    "test_status": "pass"
  },
...
```


Figure AII-1: Excerpt of the JSON response from F-UJI.

Annex III - Repositories & Thematic Services (Detailed information on use cases)

Digital.CSIC (Experimental)

DIGITAL.CSIC (https://digital.csic.es/) is CSIC institutional repository and falls under the umbrella of the Unit of Scientific Information Resources for Research (URICI), the Central institutional department that coordinates CSIC Libraries and Archives Network (http://bibliotecas.csic.es/). DIGITAL.CSIC is managed by a Central Office consisting of librarians and engineers within URICI.

DIGITAL.CSIC was launched in January 2008 with the aim to organize, preserve and provide open access to CSIC research outputs. However the main focus is put on recent CSIC outcomes (last two decades), DIGITAL.CSIC also identifies, collects, describes and gives open access to past works so as to document and maintain the scientific memory of the institution. Older contents in DIGITAL.CSIC date as far back as 1912.

The repository kicked off with an early version of DSpace software (1.x) and over the last decade a number of version upgrades were accomplished. In 2015 DIGITAL.CSIC moved to DSpace-CRIS v4.3, which is the one that the repository has at present. For many years DIGITAL.CSIC used ORACLE and in 2019 a migration to Postgres was completed following an institutional policy shift.

DIGITAL.CSIC has been organized taking into account the multidisciplinary nature of the Research Council. Thus, the repository shows a hierarchical structure in 8 research areas (namely, Agricultural Sciences, Biology and Biomedicine, Chemistry Sciences, Food Sciences, Materials Sciences, Natural Resources, Physics, and Social Sciences and Humanities) and each extinct and existing CSIC institutes falls under one of them. At a lower level, each institute grows its own collections within DIGITAL.CSIC, most of which are output types driven - journal articles, conference contributions, books and chapters, datasets, theses, software, patents and so on- however a growing number of thematic collections can be found, too. A detailed account of all DIGITAL.CSIC policies and services is available at https://digital.csic.es/dc/politicas-servicios.jsp.

With close to 210,000 research outputs available, DIGITAL.CSIC is the largest institutional repository in Spain. Around 62% of its contents are made available open access, and the remainder of items are either metadata-only items and embargoed items. With nearly 12,000 datasets available, this output type ranks third in the classification of most widely represented types of research outputs, following journal articles (127,000 items) and conference contributions (21,500). Most contents are added to DIGITAL.CSIC through the so-called Mediated Archiving Service offered by the repository's Central office and CSIC Network of Libraries to institutional researchers. This service accounts to more than 90% of

monthly submissions and the remaining % distributes between researchers self-archiving and automatic submissions from publishers.

It is important to note that since April 2019 CSIC has a Green Open Access Mandate (https://digital.csic.es/handle/10261/179077) which requires all institutional researchers to deposit and make their research outputs open access (in concrete, peer reviewed publications and associated datasets) in DIGITAL.CSIC. Compliance with FAIR Principles is explicitly stated in the mandate. A number of institutional services are offered to CSIC scientific community to meet the requirements of the mandate and compliance is linked to annual institutional assessment exercises.

By default, DIGITAL.CSIC uses Qualified Dublin Core to describe its items. Additionally, it provides functionalities at item level to export the metadata in other formats (BibTex, csv and DataCite). The repository has also integrated a number of controlled vocabularies by default, including COAR Vocabulary for research output types, FundRef for funding agencies, ISO 639 for languages, LCSH and VIAF for publishers, and openAire guidelines to state access rights and projectIDs. Further, some collections have been described by making use of subject or discipline specific controlled vocabularies such as EarthChem Vocabularies http://www.earthchem.org/resources/vocabularies for geological datasets, Library of Congress Subject Headings for several Humanities Collections and GEONAMES for datasets with geolocalization. On another front, a number of metadata elements values include persistent identifiers or URLs, for instance, DOIs of publications and associated publications and datasets, and funding agencies, machine readable licences, ORCIDs of works authors are also packaged in an extended metadata element, too. Examples of items: https://digital.csic.es/handle/10261/180630, https://digital.csic.es/handle/10261/194041

The repository provides Handle IDs to all items that are deposited on its infrastructure. In addition, the repository mints DOIs to a selected range of output types, including datasets, research software and preprints. DOIs asignation started in 2016 and URICI, the department where DIGITAL.CSIC belongs to, is responsible for covering the costs associated with DataCite institutional membership. Through this membership other CSIC related initiatives are minting their own DOIs, including SOCIB, the Unit of Marine Technology (UTM) and the researcher community from CSIC Royal Botanical Garden that participates in GBIF. This DOI institutional service and obligations of beneficaries are explained at http://digital.csic.es/dc/politicas/politica-asignacion-dois.jsp.

DIGITAL.CSIC published its research data management policy in 2013 and has been revised and enhanced several times over the last years, http://digital.csic.es/dc/politicas/politicaDatos.jsp. DIGITAL.CSIC started to accept datasets in 2010 and the development of this policy came to meet the most pressing needs from institutional researchers, namely, compliance with data sharing policies by journals and the willingness to share data from long term projects. Since 2019 through the institutional

mandate researchers are also asked to deposit and make their datasets available following FAIR Principles. Most datasets available fall under the broad category of “long tail of research” and DIGITAL.CSIC is not meant to be a big data infrastructure. However, the repository hosts more and more collections of datasets with a significant volume (a few dozens GB per file). As per discipline, there is a particularly strong representation of datasets resulting from projects on Natural Resources, Agricultural Sciences, Humanities and Biology.

DIGITAL.CSIC provides a set of services related to research data management to institutional users. Services available include hosting and curation of datasets, DOI minting, training to researchers, laboratory support staff and librarians (http://digital.csic.es/dc/formacion-dc.jsp), support on copyright and other legal issues (http://digital.csic.es/dc/copyright/), support in the preparation of data management plans (DMPs), http://digital.csic.es/dc/politicas/preparacion-planes-gestion-datos.jsp and awareness raising and compliance with FAIR Data Principles http://digital.csic.es/dc/politicas/adhesion-principios-fair.jsp.

Last but not least, DIGITAL.CSIC got awarded with Data Seal of Approval (DSA) late 2015 and we aim to apply for CoreTrustSeal in 2021 as soon as the new institutional long term preservation system unfolds. DIGITAL.CSIC application falls under the preservation strategy implemented by the CSIC IT Central Services.

Concerning involvement in EOSC, in 2018 DIGITAL.CSIC started a collaboration with EPOS European Plate Observing System project (https://epos-no.uib.no/epos-tna/facilities) so as to include in its infrastructure its collection of datasets by a group of researchers from CSIC Institute of Earth Sciences Jaume Almera. This collaboration has already resulted in the aggregation of such datasets in EPOS Multi-scale laboratories data catalog, https://acc.epos-msl.uu.nl.

In EOSC-SYNERGY DIGITAL.CSIC contributes with 4 different collections of datasets and new value added services, in line with Open Science and FAIR Data Principles will be implemented:

- One collection with data about global climate change and drought trends
- One collection with Archeological/Anthropological datasets
- One collection on Entomology data
- One collection on Earth Sciences, in particular seismic activity

In the selection of new value added services to make available on top of DIGITAL.CSIC collections of datasets a special focus has been put on enabling new functionalities around metadata, formats, and interoperability.

Over the last months we have completed the first new integration, that of Scholix. This emerging standard is the consensus achieved by a number of organisations — journal publishers, data centres, global service providers — to create an open global information

ecosystem to collect and exchange links between research data and literature. DIGITAL.CSIC has implemented Scholix so as to expose relations between its datasets and scientific literature in services like DataCite Event Data and ScholeXplorer. This implementation primarily rests on the existence of good metadata in the records of datasets available in DIGITAL.CSIC as they are vital for the standard to establish relationships between datasets and other associated datasets and publications. The PIDs of such associated works are to be found in one relation metadata element in the records of the datasets available in the repository. Scholix can also work in case associated outputs lack DOIs as long as they have other identifier such as PMCID, arXiv ID..An example of this implementation is https://scholexplorer.openaire.eu/#/detail/60%7Ce41c4e330068a8b49ad8b06e85b45ce6.

This emerging standard shows a lot of potential for users to easily discover and access not only relevant datasets but also get a much more accurate contextual information via provided links to associated works. Nowadays, Scholix is consumed differently in the two existing services (DataCite Event Data and Scholexplorer) but a wide range of possibilities may be promising, for instance, to better track datasets created by a given institution, regardless of the repository that hosts them, and to assess the citation-based impact of datasets. As said, the existence of complete metadata in the datasets records is vital as a first step in order to exploit the benefits of Scholix.

Other developments are currently underway, in particular:

1. The experimental DIGITAL.CSIC [R5] use case is a docker-compose version of DSPACE-CRIS [R6] v.7 with the following modules: solr, postgreqsl database, dspace core under tomcat8, angular UI.
2. The configuration of OAI-PMH is being extended in the current version of DIGITAL.CSIC to include DataCite schema.
3. Open ID Connect is not natively supported by Dspace-CRIS and no developments have been found. However, there are some plugins with other ID mechanisms that can be observed.

SOCIB

SOCIB (Sistema d'Observació i predicció Costaner de les Illes Balears) is a national marine research infrastructure included in the Spanish Large Scale Infrastructures (ICTS) Map.

It is made up of a network of facilities and equipment devoted to marine observation, acquisition of data, processing, analysis, operational numeric modelling, and dissemination of multidisciplinary marine information in a systematic and consistent manner. It openly provides oceanographic data in real time and prediction services in support of operational oceanography, answering the needs of a wide range of society's scientific, technological, and strategic priorities in the context of climate and global change.

The mission of SOCIB is to develop a coastal ocean observing and forecasting system which provides free, open, quality controlled and timely streams of data to:

- Support research and technology development on key internationally established topics
- Support operational oceanography and associated marine technology development in the Balearic Islands and in Spain
- Support the strategic needs of society in the context of global change

The general goal of the SOCIB Data Centre is to provide users with a system to locate and download the data of interest and to visualize and manage the information. Following SOCIB principles, data need to be (1) discoverable and accessible, (2) freely available, and (3) interoperable and standardized.

WORSICA

WORSICA (Water mOnitoRing SentInel Cloud plAtform), a service that integrates remote sensing and in-situ data for the determination of water presence in coastal and inland areas, applicable to a range of purposes from the determination of flooded areas to the detection of large water leaks in major water distribution networks. WORSICA is a one-stop-shop service to provide access to customized remote sensing services based on Copernicus data, currently applied to the detection of the coastal water land interface and the inland water detection (for large water infrastructure leak detection). The WORSICA service was developed by Laboratório Nacional de Engenharia Civil (LNEC), in Lisbon, Portugal, in the scope of the INCD project, just for national usage. During the EOSC-SYNERGY, the WORSICA service will be improved, with EOSC IT services available in the marketplace (https://marketplace.eosc-portal.eu - AAI, Dataverse, etc.), and integrated in the Ibergrid computing infrastructure to better serve the European scientific community. The service will continue to be maintained by LNEC (http://worsica.lnec.pt), and the new version will be accessed at the end of the EOSC-SYNERGY project in http://worsica.ncg.ingrid.pt.

WORSICA can be used for the detection of coastlines, coastal inundation areas and the limits of inland water bodies. It can also be applied to a range of other purposes, from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water distribution networks. This freely available service enables the user communities to generate maps of water presence and water delimitation lines in coastal and inland regions. In particular, the service helps to promote 1) the preservation of lives during an emergency, supporting emergency rescue operations of people in dangerously inundated areas, and 2) the efficient management of water resources targeting water saving in drought-prone areas.

The major processing tasks of WORSICA service are to handle satellite and drone imagery, and also in situ data for the evaluation of the water features in the images.

Therefore the main type of data used by the service are geo-referenced images and shapefiles, ascii and netcdf files with auxiliary data from field surveys. The data can be retrieved from public repositories (https://scihub.copernicus.eu), or private/proprietary data, such as drone surveys. The service downloads the needed data (or the data is uploaded by the users), afterwards the images are processed and some intermediate and final products are produced.

The WORSICA service tries to promote the FAIR principles for data management. Nevertheless, the dissemination of the data is dependent on the user authorization. If the user authorizes that his private data (e.g. drone survey), should be publicly available to everyone, all the data and metadata used and produced by the WORSICA workflow will be sent to Dataverse. On the other hand, if the user does not authorize the dissemination, only the metadata of the data used (and produced) will be sent to Dataverse.

First, The Worsica Thematic Service [R4]. It has been selected because of the Dataverse repository installation into their services. Dataverse version is 4.20

Data will be available using a Dataverse endpoint that complies with FAIR principles [R10].

Using Dataverse is possible to follow the OAIS (Open Archival Information System) reference model keeping data objects with its associated representation information. Data formats will follow standards defined by ESA such as SAFE (Standard Archive Format for Europe).

https://earth.esa.int/SAFE/

https://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/preservation-strategies/preservation-strategies.pdf

Annex IV - Overview of synergies & dependencies

EOSC Synergy related deliverables:
D2.1 - Roadmap for integration of national capacities into the EOSC (M5)
D4.1 - Best Practices Elicitation including Data Management Plans (M6)
D3.1 – Software Maturity baseline (M10)
D4.2 - First prototype of EOSC thematic services implementation (M12)
D3.2 – First prototype of Service Integration platform (M15)
D4.3 - First release of EOSC thematic services (M24)
D3.4 – Final release of Service Integration platform (M29)
D4.4 - EOSC Thematic services validation report (M30)

FAIRsFAIR related deliverables:
D2.3 - Set of FAIR data repositories features (Published Jan 2020, available here: https://doi.org/10.5281/zenodo.3631528)
Assessment report on 'FAIRness of services' (Published Feb 2020, available here: https://doi.org/10.5281/zenodo.3688761)
D4.2 - Repository certification mechanism: a recommendation on the extended requirements and procedures (Published May 2020, available here: https://doi.org/10.5281/zenodo.3835698)
FAIRsFAIR Data Object Assessment Metrics (Published July 2020, available here: https://doi.org/10.5281/zenodo.3934401)
D2.6 - 1st reference implementation of the data repositories features (Feb 2021)
D2.7 - Framework for assessing FAIR services (August 2021)
D4.5 - Report on FAIR data assessment toolset and badging scheme (August 2021)
D2.9 - 2nd reference implementation of the data repositories features and client application (Feb 2022)

Other:
EOSC Rules of Participation WG & Outputs
- European Open Science Cloud Rules of Participation Version 0.2 (29 January 2020) Draft EOSC RoP (v0.2)

EOSC Portal