EOSC-Life

EOSC-Life: Building a digital space for the life sciences

D6.1 – FAIR Requirements Document

WP6 – FAIRification and Provenance Services
Lead Beneficiary: BBMRI and ERINHA/INSERM
WP leader: Isabelle Perseil (ERINHA/INSERM) and Petr Holub (BBMRI)
Contributing partner(s): INSERM, UOXF, LUMC, BBMRI-ERIC, University of Würzburg

Authors of this deliverable: **Isabelle Perseil, Susanna-Assunta Sansone, Peter McQuilton, Michel Dumontier, Marco Roos, Luiz Bonino, Rudolf Wittner, Petr Holub, Jörg Geiger**

Contractual delivery date: **29 Feb 2020**
Actual delivery date: **2 March 2020**
H2020-INFRAEOSC-2018-2

Grant agreement no. 824087
Horizon 2020
Type of action: RIA

Table of Contents

Executive Summary 3
Project Objectives 3
Detailed Report on the Deliverable 3
Background 3
Description of ongoing work 4
FAIR, what are the benefits and what is the state of play 4
Metrics for FAIR data 5
FAIRsharing, a community-driven service 5
FAIRassist, a community-driven service 6
Towards FAIRassist methodology 7
Provenance - basis for reproducibility and meaningful data reuse 7
FAIR Requirements 8
Requirements for EOSC-Life datasets 8
Requirements on provenance 10
Requirements on Common Provenance Model 10
Physical material and its processing 10
Data and its processing 11
Privacy requirements 12
References 12
Abbreviations 14
Delivery and Schedule 14
Adjustments 14

Executive Summary

This deliverable is the blueprint of the requirements needed to make European life science research data and infrastructures Findable, Accessible, Interoperable and Reusable according to the FAIR Principles. This first deliverable (i) focuses on database resources such as data repositories and knowledge bases, as well as (meta)data standards and data policies. In particular, how the FAIRsharing service helps and (ii) lays the groundwork for making datasets FAIR. The FAIRification efforts will be based on the diversity of the proposed datasets of the project, in order to set up a FAIRification methodology suitable for any type of data in life sciences.

This work will be tackled by subsequent deliverables via guidance from the FAIRassist tool.

Project Objectives

With this deliverable, the project has contributed to the following objectives:

a. Identification of the various datasets needs
b. Guidelines and services that help to enhance the FAIRness of some infrastructure
c. Foster the adoption of (meta)data standards as pillars of data FAIRness.
d. Design of a FAIRification methodology

Detailed Report on the Deliverable

Background

Community-developed standards, such as those for the identification, citation and reporting of data, underpin FAIR data and reproducible research, aid scholarly publishing, and drive both the discovery and the evolution of scientific practice. The number of these standardization efforts, driven by large organizations or at the grassroots level, has been on the rise since the early 2000s. Thousands of community-developed standards are available (across all disciplines), many of which have been created and/or implemented by several thousand data repositories. Nevertheless, their uptake by the research community has been slow and uneven mainly because investigators lack incentives to follow and adopt standards. Uptake is further compromised if standards are not promptly implemented by databases, repositories and other research tools, or endorsed by infrastructures. This is why, on this project, particular emphasis is placed on ongoing support and training (with planned project activities involving data experts).

As with any other digital object, standards, databases, repositories and knowledge bases are dynamic in nature, with a ‘life cycle’ that encompasses formulation, development and maintenance; their status in this cycle may vary depending on the level of activity of the developing group or community.

Within EOSC-Life, and more globally across the other disciplines and RIs, there is an urgent need for a service, like FAIRsharing, which enhances the information available on the evolving constellation of heterogeneous standards, databases, repositories and knowledgebases, that guides users in the selection of these resources, educates policy makers, such as publishers and funders, to recommend the relevant resources to their authors and awardees, and works with developers and maintainers of these resources to foster collaboration and promote harmonization. Such a common service is vital to reduce the knowledge gap among those involved in producing, managing, serving, curating, preserving, publishing or regulating data within the Life Science ESFRIs and beyond.

A further challenge for an ecosystem of resources based on FAIR principles is to achieve sufficient consistency for efficient machine-to-machine processing in the EOSC. Tools such as FAIRsharing.org already help users to identify emerging implementations, and FAIRassist.org provides communities with options to measure compliance with the (15 detailed) FAIR principles. In addition to these, we also need to ask what generic requirements, if any, can we define for EOSC-Life datasets to make a FAIR ecosystem efficient. In this general context, we started by investigating the requirements for EOSC-Life data repositories.

Description of ongoing work

FAIR, what are the benefits and what is the state of play

The FAIR data principles outline a set of key characteristics that make data easier to find and reuse in new applications. The benefits of making data FAIR have been articulated in several reports[1][2][3], and are broad in their impact. Anticipated benefits include:

- Improved effectiveness (precision and recall) in searching for relevant data
- Enhanced productivity by using higher quality data
- Augmented accuracy and reproducibility of findings by comparing to existing data
- Improved management and stewardship of digital resources
- Enhancement of the scientific and information technology infrastructure for discovery science.
- A savings of over 10 billion euros per year to the European economy

Numerous national and international initiatives to address the challenges posed in implementing the FAIR principles have emerged in recent years. Examples of these include EOSC-Life (this initiative), GO-FAIR, FAIRsFAIR, FAIRPlus[4], ELIXIR, and the European Commission.

- GO-FAIR is a stakeholder-driven and self-governed initiative to implement the FAIR data principles. It federates national initiatives, organisations, and individuals through bottom-up implementation groups focused on infrastructure, training, and cultural considerations.

[1] European Commission, Directorate-General for Research and Innovation, and PwC EU Services, Cost-Benefit Analysis for FAIR Research Data
[2] Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship"
[3] "CrowdFlower 2016 Data Science Report"
[4] FAIR CookBook (https://fairplus.github.io/the-fair-cookbook/intro.html#The-FAIR-Coobook-overviewl)

- FAIRsFAIR - Fostering Fair Data Practices in Europe - aims to develop guidelines for the implementation of the FAIR data principles throughout the research data life cycle, with an emphasis on research data repositories.
- ELIXIR is an intergovernmental organisation that coordinates, integrates, and sustains bioinformatics resources across its member states. ELIXIR is committed to coordinating these resources towards being FAIR research data infrastructures.
- The European Commission - together with stakeholders and assisted by the FAIR Data Expert Group - is working towards an action plan for FAIR data as part of the European Open Science Cloud initiative.

Metrics for FAIR data

The FAIR principles are a set of guidelines to increase the visibility and utility of data. They were carefully devised to be independent of any choice of implementation, so as to be resilient in the face of ever-changing technology. Nevertheless, the FAIR principles must be manifested in real digital resources, and the most widely used platform for making data available is through Web technology. As such, there have been several efforts (see RDA WG survey [5]) to evaluate the FAIRness of digital resources, spanning from questionnaires, to checklists, to automated assessments. Notably, the FAIR Metrics group[6], which have representation in WP6, have devised a set of computable metrics to assess FAIRness, and this has been implemented as part of an automated FAIR Evaluator tool that is available through FAIRsharing.org. As part of an ELIXIR Implementation Study led by Maastricht members in WP6, both the manual and automated assessments were utilised and evaluated with ELIXIR Core Data Resources[7] [8](). WP6 members are contributing to the development of a common set of assessment criteria and self-assessment model via the RDA FAIR Data Maturity Model Working Group[9].

FAIRsharing, a community-driven service

On April 2019, an article authored by the FAIRsharing team (led by the Oxford members in WP6) and 68 international authors (representing the FAIRsharing core adopters, advisory board members, and key collaborators) was published in Nature Biotechnology[10] illustrating how this community-driven service is as a core element of the FAIR-enabling ecosystem of resources. The article also highlights the role each stakeholder group must play to maximize the visibility and adoption of standards, databases, repositories and knowledgebases. This work is being used as the groundwork for EOSC-Life WP6.

FAIRsharing is an informative and educational resource that describes and interlinks community-driven standards, databases, repositories, knowledge bases and data policies. As of February 2020, FAIRsharing records detailed information on 1,353 standards, 1,331 databases and 130 data policies (of which 85 are from journals and publishers and 23 from funders), covering natural sciences (for example, biomedical, chemistry, astronomy, agriculture, earth sciences and life sciences), as well as engineering, humanities and the social sciences.

[5] https://docs.google.com/spreadsheets/d/14ojMSXVOITg3RoJn-PuDaPj8zuIGQz2Li-kl97HOBH4/edit#gid=0
[6] http://fairmetrics.org
[7] de Miranda Azevedo and Dumontier, “Considerations for the Conduction and Interpretation of FAIRness Evaluations.”
[8] https://doi.org/10.1162/dint_a_00051
[9] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[10] Sansone et al., “FAIRsharing as a Community Approach to Standards, Repositories and Policies”

FAIRsharing collects the necessary information to ensure that standards, databases, repositories and data policies align with the FAIR Principles: Findable (for example, by providing persistent and unique identifiers, and functionalities to register, claim, maintain, interlink, search and discover them), Accessible (for example, identifying their level of openness and/or license type), Interoperable as much as possible (for example, highlighting which repositories implement the same standards to structure and exchange data) and Reusable (for example, knowing the coverage of a standard and its level of endorsement by a number of repositories should encourage its use or extension in neighbouring domains, rather than reinvention).

Started in 2011 as BioSharing, which originally focused on the natural sciences only, it was renamed to FAIRsharing in 2017 to reflect its broader remit.

This resource has its roots in the MIBBI portal, which itself was launched in 2008[11]. Since 2018, FAIRsharing is one of the ELIXIR Recommended Interoperability Resources (RIR) and is a flagship output and endorsed recommendation of the RDA[12]. FAIRsharing is also recommended by the EOSC "Turning FAIR into Reality" report[13], as well as by ERC and Science Europe data management policies[14] [15]. Furthermore, FAIRsharing has already been adopted by a diverse set of stakeholders, representing academia, industry, funding agencies, standards organizations, infrastructure providers and scholarly publishers—both national and domain-specific as well as global and general organizations. Its community of core adopters, advisory board members, and/or key collaborators are listed at FAIRsharing[16].

FAIRassist, a community-driven service

Besides contributing in its primary role as a FAIR-enabling registry for standards, databases, repositories, knowledge bases and data policies, FAIRsharing also provides content to power other services, such as FAIR evaluator tools for datasets and other digital objects. Specifically, two prototypes have been published and connected to FAIRsharing: the FAIR Evaluator[17] [18] and the FAIRshake tool[19].

However, a commonly agreed FAIR evaluation processes, as well as the indicators or metrics needed to assess and evaluate the different digital objects are still a work in progress. Therefore, before a robust "FAIRassist" widget can be implemented (D6.4), we have launched it as a website[20] to list and describe the emerging resources for the assessment and/or evaluation of digital objects against the FAIR principles. This is not intended to be a comprehensive list of all groups, projects and organizations that tackle FAIRness or FAIRification, but as a preliminary scoping of material that will be consolidated in a manner digestible to a diverse set of stakeholders through the development of a digital assistant which will provide advice on how to make a data resource FAIR and to assess the level of FAIRness. The focus is on manual questionnaires, checklists and automated tests that help users understand how to achieve a state of "FAIRness", and how this can be measured and improved.

[11] Taylor et al., "Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations."
[12] RDA-Force11 FAIRsharing WG, "The FAIRsharing Registry and Recommendations."
[13] European Commission and Directorate-General for Research and Innovation, Turning FAIR Data into Reality.
[14] ERC Scientific Council, "Open Research Data and Data Management Plans."
[15] Science Europe, "Science Europe Guidance Document: Presenting a Framework for Discipline-Specific Research Data Management."
[16] "FAIRsharing | Communities."
[17] https://fairsharing.github.io/FAIR-Evaluator-FrontEnd
[18] Wilkinson et al., "Evaluating FAIR Maturity through a Scalable, Automated, Community-Governed Framework."
[19] Clarke et al., "FAIRshake," November 2019; Clarke et al., "FAIRshake," June 3, 2019.
[20] https://fairassist.org

Towards FAIRassist methodology

Work is underway, as collaboration between WP6 and WP2, to ensure EOSC-Life standards, databases, repositories, knowledge bases and data policies are described and interlinked in FAIRsharing. A dedicated Collection, which will enhance the discoverability of these resources will be released at month 36, as D6.3. Work will also continue to design and develop the FAIRassist widget, which will be released at month 48, as D6.4.

Provenance - basis for reproducibility and meaningful data reuse

Life science research is first and foremost based on the experimental exploration of biological samples and the analysis of the resulting data. Researchers using or reusing biological samples or data, replicating experiments with biological samples, or data or reviewing publications must be able to rely on their suitability for a particular purpose (e.g., analytical methods or data integration methods). Providing sample information (or metadata) on the collection, generation, processing, storage etc. to the researcher renders a well-founded decision on the suitability of the samples/data possible. Translation of laboratory findings to practical application in biomedicine requires exhaustive and consistent documentation of the developmental procedures and materials used. With the increasing complexity of research projects leading to diverse and sophisticated techniques, the reproducibility of observations, results and data analysis has become an exigent need. In recent years, several publications dealing with research quality have explored the reasons for a lack of reproducibility[21], and the resulting scientific and economical consequences[22]. Frequently, reproducibility issues in life sciences result from poor quality documentation of biological samples, related procedures, derived data and applied algorithms.

The W3C definition[23] states that provenance encompasses documentation about the *entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.* In the context of the life sciences, provenance can be perceived as comprehensive documentation that describes the whole scientific process, from the collection, generation, processing and analysis of biological material to respective data derivation, integration and analysis. Such information serves as a quality indicator and provides metadata on the reliability, trustworthiness and fitting-for-purpose of a described object.

Given this, provenance information must be regarded crucial to the implementation of FAIR principles. Provenance information details the procedures, techniques, equipment, materials and actors involved in collecting, accessioning, processing, handling, storing, shipping and disposing of biological samples. Results of measurements along with information on precision and reliability, the methods and devices used can be provided through provenance information enabling the researcher to assess fitness for use and saving the researcher from unnecessary experiments. By documenting the methods, algorithms and parameters of data analyses data can be re-evaluated, compared and combined for further analyses and meta-analyses.

Provenance information constitutes an unbroken chain of machine actionable information related to the history of an object described. The provenance model for the biotechnology domain, currently under development, is one of the upcoming deliverables of WP6 and will also

[21] Freedman and Inglese, "The Increasing Urgency for Standards in Basic Biologic Research"; Begley and Ioannidis, "Reproducibility in Science."
[22] Freedman, Cockburn, and Simcoe, "The Economics of Reproducibility in Preclinical Research."
[23] https://www.w3.org/TR/prov-overview/

be part of a standard which has been proposed to the ISO/TC 276. The ISO standard is intended to be applied by all kinds of stakeholders in biomedical research (pharmaceutical and biotech companies, research funders, hospitals, biobanks, academia, government, etc.) and will consist of four major components, 3 horizontal and a group of vertical standards: a) the generalized horizontal provenance information model to serve as a basis for implementing vertical standards for specific domains (see below), capable of interconnecting various parts of provenance information coming from various sources and organizations b) provenance information management requirements c) extensions dealing with authenticity and non-repudiation of provenance d) vertical models build on the horizontal standards corresponding to domain specific requirements, such as provenance documenting handling of biological samples, observations and measurements (sequencing, spectrometry, microscopy, assays etc.), data storage and processing (computational workflows) or data analyses (software, algorithms, visualization etc.). The ultimate goal of the standardization initiative is the interoperability of all elements of provenance information generated by the contributors.

Integrating a provenance standard into the scientific environment is of fundamental importance to achieve the aims mentioned: reproducibility of research, adequate use of samples, re-use of data, comparability of samples and data and by this improving the quality of research and leveraging translation of scientific findings to practical application.

The provenance model is currently being developed and standardized in the ISO TC/276 (Biotechnology) WG5 (Upstream data integration). The Part 1 has already been adopted as Preliminary Work Item PWI 23494-1 (“Provenance information model for biological specimen and data — Part 1: Design Concepts and General Requirements”) and it has now being updated to New Work Item in 36 months development track. Part 2 (Common Provenance Model) is being registered as PWI 23494-2 (“Provenance information model for biological specimen and data – Part 2: Common Provenance Model”).

FAIR Requirements

Requirements for EOSC-Life datasets

The vision of machine-actionable interoperability and reuse for datasets brought by the FAIR principles raised requirements for their properties and features such as globally unique and persistent identifiers for both data and their metadata, explicit reference to the data in the metadata, qualified references, rich provenance and adoption of relevant standards. In order to support the intended level of interoperability, especially across different domains, agreements on commonly used approaches are necessary. In this first version of this deliverable, we will focus on a set of requirements for FAIR datasets in terms of requirements for the repositories that host them. The reason for this approach is that by adjusting repositories to comply with the FAIR principles, we can make the data hosted in these repositories present some FAIR features without direct end-user intervention. In this way we facilitate the lives of the end-users by requiring less from them and increase the convergence on common approaches to data because the repositories implement these approaches on behalf of their users.

Normally data repositories behave more as document (file) management systems with some metadata about the files. In this scenario, it is relevant to focus on the metadata aspect of the FAIR principles, improving its support by data repositories.

The following list presents **requirements for FAIR repositories** that support the increase of FAIRness of the metadata and data hosted on them. In parenthesis we indicate which FAIR principle this requirement is related.

- Provide globally unique and persistent identifiers for both metadata and data (F1).
- Define a metadata schema that includes relevant content to support findability and reuse. This metadata schema should be extensible so different domains and applications can add more metadata items that are relevant for them (F2);
- Include in the metadata record the identifier of the data it describes using a well-defined predicate/property (F3);
- Index the datasets' metadata record allowing clients (human and computational) to search for datasets based on the metadata records' items (F4);
- Provide an accessibility method to both metadata and data based on an open, free and universally implementable protocol (A1.1);
- If necessary, support authentication and authorisation procedures for metadata and data accessibility (A1.2);
- Provide an explicit policy for the metadata persistence so that they can still be available when the data are no longer available (A2);
- Present metadata using a common language for knowledge representation, e.g., RDF/JSON-LD (I1);

The other **FAIR principles**, namely, F2 (the rich part of "rich metadata"), I2, I3, R1.1, R1.2 and R1.3 are an intrinsically responsibility of whoever is responsible for creation and curation of the metadata and data, not the repositories. However, even for these principles data repositories can implement features to help its users in following the FAIR principles as listed below:

- F2 - the richness part of "data are described with rich metadata" relates to decisions on how rich should be the metadata with respect to findability and reusability. In this case, data repositories could present to their users a list of additional metadata schemas used and/or defined through community agreed standards developed for that domain, that provide more information about the datasets such as provenance, relation with articles or other documentation for the data, technical aspects, etc.
- I2 - data repositories could guarantee that the vocabularies used in the metadata schema follow the FAIR principles.
- I3 - data repositories could guarantee that their metadata schemas include qualified references to other metadata and data.
- R1.1 - data repositories could provide and require a list of suitable data usage licenses as part of the dataset metadata.
- R1.2 - as in the suggestion for principle F2, data repositories could offer a number of community agreed provenance metadata schemas to be chosen by their users.
- R1.3 - data repositories could implement a connection to standard registries such as FAIRsharing to allow their users to indicate which domain-relevant community standards are used in their data and metadata.

In the next version of this deliverable we will add requirements and recommendations for EOSC-Life datasets focusing on data to be used by data creators and curators.

Requirements on provenance

Requirements on Common Provenance Model

DR.1 The provenance information model should capture, in a computable (machine-readable and processable) and reproducible way, all the events connected to the physical operations performed on the biological material and all the details of the data generation and data processing workflows, in order to allow the tracking and the backward reconstruction of the history related to sample processing and/or data generation and/or data processing.

DR.2 Provenance model should have "institution" entity, in order to capture institutional responsibility and also to support resolving distributed provenance. The model should, in a computable manner, define responsibility of institution and responsibility of individual persons (and possible delegation of responsibility from institution to a particular person).

DR.3 The provenance information model shall specify clear serialization guidelines as well as implementation guidelines to achieve interoperability of applications producing/consuming the provenance information.

DR.4 The provenance information model shall specify how to find and how to access the provenance information.

DR.5 The provenance information model should define the type (restrictions) on the provenance and its digital representations. Since provenance information can be represented as a graph, example of such a restriction is, e.g., a provenance graph shall be a directed acyclic graph.

DR.6 Provenance model shall support distributed provenance information, where different parts of it are stored and made accessible at a particular institution.

DR.7 Distributed provenance information must support both complete chain model (direct resolution of what are all responsible entities) as well as predecessor model (only previous step is resolved). The predecessor model is meant for scenarios where the chain model by its nature can make certain sensitive information available (e.g., showing the complete chain how highly pathogenic material is transferred across institutions routinely).

DR.8 Distributed provenance information must support opaque sub-chains of the complete provenance chain that are resolvable only at a particular responsible entity.

DR.9 Provenance model must support non-repudiation for all steps of processing.

DR.9.1 non-repudiation of non-sensitive information shall be made directly available as open part of the provenance information;

DR.9.2 non-repudiation of sensitive information must allow storing the sensitive information package only at the responsible entity, while still ensuring the non-repudiation property (using opaque parts of the provenance chain).

DR.10 Provenance information management shall technically support traversal of provenance information from parent to children. This shall be implemented as an optional feature that is only practically enabled for certain scenarios.

Physical material and its processing

DR.11 The provenance information model shall include generic and extensible support for describing (a) acquisition; (b) processing; (c) storage; (d) transport of biological material.

DR.12 The provenance information model shall support the following standards and community standards:

DR.12.1 Support for processes defined in Working Groups (WGs) 2 (possibly also 3 & 4 if relevant) of ISO/TC 276 and ISO/TC 212 WG 4.

DR.12.2 Support for existing methods describing pre-analytical sample processing: (a) CEN/TS 16826-1:2015, CEN/TS 16826-2:2015; (b) CEN/TS 16827-1:2015, CEN/TS 16827-2:2015, CEN/TS 16827-3:2015; (c) CEN/TS 16835-1:2015, CEN/TS 16835-2:2015, CEN/TS 16835-3:2015; (d) CEN/TS 16945:2016; (e) CEN/TS 16945:2016; (f) Standard PREanalytical Code (SPREC)[24]; (g) BRISQ[25].

DR.12.3 Support standards coming from current H2020 SPIDIA4P project (which is input into CEN standardization).

DR.13 The provenance information model shall allow the tracking of all the operations involving the biological material being processed, even if they are not directly part of the experimental protocol (e.g., retrieval from a certain lab, transportation, storage, etc., might not be part of experimental protocol).

DR.14 Material processing (e.g., cell culture staining, mechanical stimulation, etc.): the provenance information model shall be able to identify (a) the entities (entity ID, primary sample, eventual originating/deriving samples); (b) the processing method; (c) a link to the reference processing protocol, the performer, the processing parameters, the device(s), the software version, timestamp; (d) physical conditions; (e) post-analytical conditions; (f) execution logs; (g) result confidence; (h) reference to the donor consent (only if processing human material and if consent is needed in particular legal settings - further denoted as "if applicable")

DR.15 Material retrieval (e.g., thawing): the provenance information model shall be able to provide information about (a) the pre-analytical conditions; (b) donor identification; (c) entities identification; (d) location of the material to be retrieved; (e) the physical conditions at retrieval; (f) the physical conditions during transportation; (g) the performer, shipment details (sender, receiver, carrier), the timestamp; (h) reference to the donor consent (if applicable).

DR.16 Material storage: the provenance information model shall be able to locate the biological material, to provide (a) details about the physical storage conditions; (b) to detail the physical conditions at the arrival; (c) to identify the entities; (d) to record the physical conditions during transportation; (e) to identify the performer; (f) to provide the shipment details (sender, receiver, carrier), the timestamp; (g) reference to the donor consent (if applicable).

DR.17 The provenance information model shall contain link to the physical label identifying the biological material.

DR.18 Level of detail recorded in the provenance information (e.g., precision of timestamps or precision and frequency of temperature measurements) will depend on intended use of the biological material.

Data and its processing

DR.19 The provenance information model shall include generic and extensible support for data processing.

[24] Betsou et al., "Standard PREanalytical Code Version 3.0."
[25] Moore et al., "Biospecimen Reporting for Improved Study Quality (BRISQ)."

DR.20 The provenance information model shall support the following standards and community standards in the field of data processing:

DR.20.1 Support data generation and processing defined within WGs 3 & 4 of ISO/TC 276 and ISO/TC 212 WG 4.

DR.20.2 Support for workflow provenance, using commonly accepted workflow description language(s) such as Common Workflow Language (CWL)[26].

Note: Compatibility with ISO 8000-120:2016[27] needs to be clarified.

DR.21 Data retrieval (e.g., directly obtained data, processed data, etc.): the provenance information model shall be able to provide details about (a) data authorship; (b) acquisition information; (c) information necessary to verify non-corrupted status of the data; (d) retrieval information in a computable manner; (e) reference to the donor consent (if applicable).

DR.22 Data generation (e.g., image acquisition, cell culture measurements, etc.): the provenance information model shall be able to computationally describe (a) the acquisition protocol; (b) the performer; (c) the execution parameters; (d) the device(s); (e) the software version; (f) the timestamp of the execution; (g) acquisition logs; (h) reference to the donor consent (if applicable).

DR.23 Data processing (e.g., quantitative RT-PCR, statistical analysis): the provenance information model should be able to computationally describe (a) the analysis protocol; (b) the performer; (c) the analysis parameters; (d) the device(s); (e) the software version; (f) the timestamp; (g) execution log;(h) result confidence; (i) reference to the donor consent (if applicable).

Privacy requirements

This section specifies privacy related requirements on provenance information management.

DR.24 The privacy requirements only sets minimum requirements in order to help implementing privacy protection compliance.

DR.25 If applicable, the provenance information model shall provide means to detach personally identifiable information of the research participant contributing biological material and/or data, so that only a trusted party or authorized institution may re-identify the person (for purposes such as incidental findings).

References

Begley, C. Glenn, and John P.A. Ioannidis. “Reproducibility in Science: Improving the Standard for Basic and Preclinical Research.” *Circulation Research* 116, no. 1 (January 2, 2015): 116–26. https://doi.org/10.1161/CIRCRESAHA.114.303819.

Betsou, Fay, Roberto Bilbao, Jamie Case, Rodrigo Chuaqui, Judith Ann Clements, Yvonne De Souza, Annemieke De Wilde, et al. “Standard PREanalytical Code Version 3.0.”

[26] http://www.commonwl.org/

[27] “ISO 8000-120:2016 – Data Quality – Part 120: Master Data: Exchange of Characteristic Data: Provenance.” https://www.zotero.org/google-docs/?J6CPsC

Biopreservation and Biobanking 16, no. 1 (February 2018): 9–12. https://doi.org/10.1089/bio.2017.0109.

Clarke, Daniel J.B., Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, et al. "FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources." *Cell Systems* 9, no. 5 (November 2019): 417–21. https://doi.org/10.1016/j.cels.2019.09.011.

"CrowdFlower 2016 Data Science Report." CrowdFlower, n.d. https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.

ERC Scientific Council. "Open Research Data and Data Management Plans." ERC Scientific Council, July 3, 2019. https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf.

European Commission, and Directorate-General for Research and Innovation. *Turning FAIR Data into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data*, 2018. http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0618206ENN.

European Commission, Directorate-General for Research and Innovation, and PwC EU Services. *Cost-Benefit Analysis for FAIR Research Data: Cost of Not Having FAIR Research Data*, 2018. http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0219023ENN.

Freedman, L. P., and J. Inglese. "The Increasing Urgency for Standards in Basic Biologic Research." *Cancer Research* 74, no. 15 (August 1, 2014): 4024–29. https://doi.org/10.1158/0008-5472.CAN-14-0925.

Freedman, Leonard P., Iain M. Cockburn, and Timothy S. Simcoe. "The Economics of Reproducibility in Preclinical Research." *PLOS Biology* 13, no. 6 (June 9, 2015): e1002165. https://doi.org/10.1371/journal.pbio.1002165.

"ISO 8000-120:2016 – Data Quality – Part 120: Master Data: Exchange of Characteristic Data: Provenance." ISO/IEC, 2016.

Miranda Azevedo, Ricardo de, and Michel Dumontier. "Considerations for the Conduction and Interpretation of FAIRness Evaluations." *Data Intelligence* 2, no. 1–2 (January 2020): 285–92. https://doi.org/10.1162/dint_a_00051.

Moore, Helen M., Andrea B. Kelly, Scott D. Jewell, Lisa M. McShane, Douglas P. Clark, Renata Greenspan, Daniel F. Hayes, et al. "Biospecimen Reporting for Improved Study Quality (BRISQ)." *Journal of Proteome Research* 10, no. 8 (August 5, 2011): 3429–38. https://doi.org/10.1021/pr200021n.

RDA-Force11 FAIRsharing WG. "The FAIRsharing Registry and Recommendations: Interlinking Standards, Databases and Data Policies." RDA, October 11, 2018. https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases-wg/outcomes/fairsharing.

Science Europe. "Science Europe Guidance Document: Presenting a Framework for Discipline-Specific Research Data Management." Science Europe, January 2018. http://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf.

Taylor, Chris F, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A Ball, et al. "Promoting Coherent Minimum Reporting Guidelines

for Biological and Biomedical Investigations: The MIBBI Project." *Nature Biotechnology* 26, no. 8 (August 2008): 889–96. https://doi.org/10.1038/nbt.1411.

The FAIRsharing Community, Susanna-Assunta Sansone, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson L. Lister, and Milo Thurston. "FAIRsharing as a Community Approach to Standards, Repositories and Policies." *Nature Biotechnology* 37, no. 4 (April 2019): 358–67. https://doi.org/10.1038/s41587-019-0080-8.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, no. 1 (December 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Wilkinson, Mark D., Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, et al. "Evaluating FAIR Maturity through a Scalable, Automated, Community-Governed Framework." *Scientific Data* 6, no. 1 (December 2019): 174. https://doi.org/10.1038/s41597-019-0184-5.

Abbreviations

MIBBI	Minimum Information for Biological and Biomedical Investigations
ELIXIR RIR	Recommended Interoperability Resource
RDA	Research Data Alliance
DOI	Digital Object Identifier

Delivery and Schedule

The delivery is delayed:

No

Adjustments

Adjustments made:

None