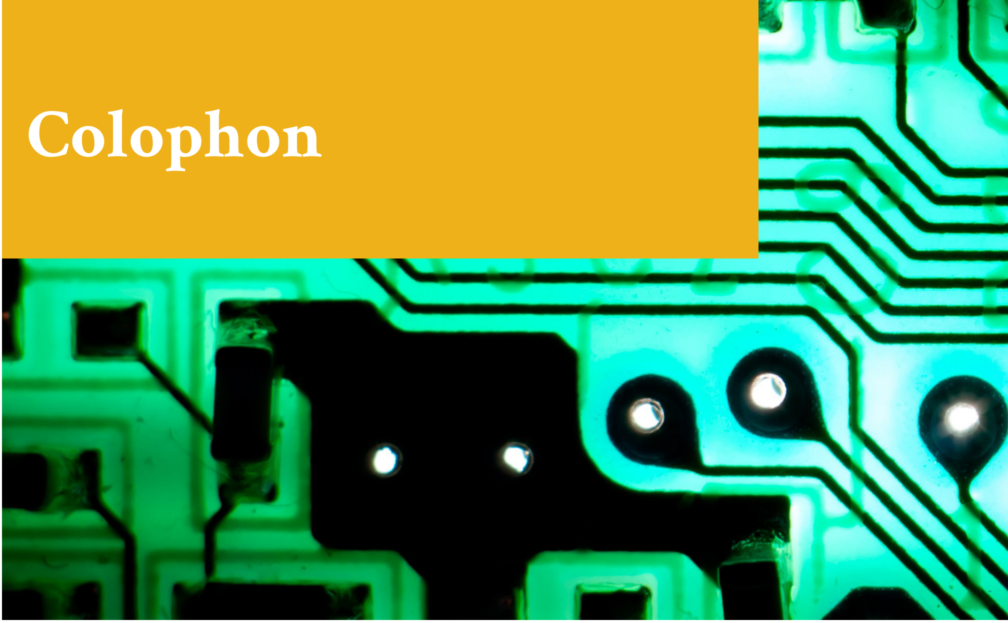
EUROPE’S RESEARCH LIBRARY NETWORK

FAIRness of Repositories & Their Data

A Report from LIBER’s Research Data Management Working Group

Colophon



ABOUT LIBER

LIBER *(Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries)* is the main network for research libraries in Europe. Founded in 1971, LIBER has grown steadily to include some 450 national, university and special libraries.

Together we work to represent the interests of European research libraries, their universities and their researchers.

HOW TO CITE THIS REPORT

Ivanović, D., Schmidt, B., Grim, R., Dunning, A. FAIRness of Repositories & Their Data: A Report from LIBER's Research Data Management Working Group (2019). Retrieved from https://doi.org/10.5281/zenodo.3251593

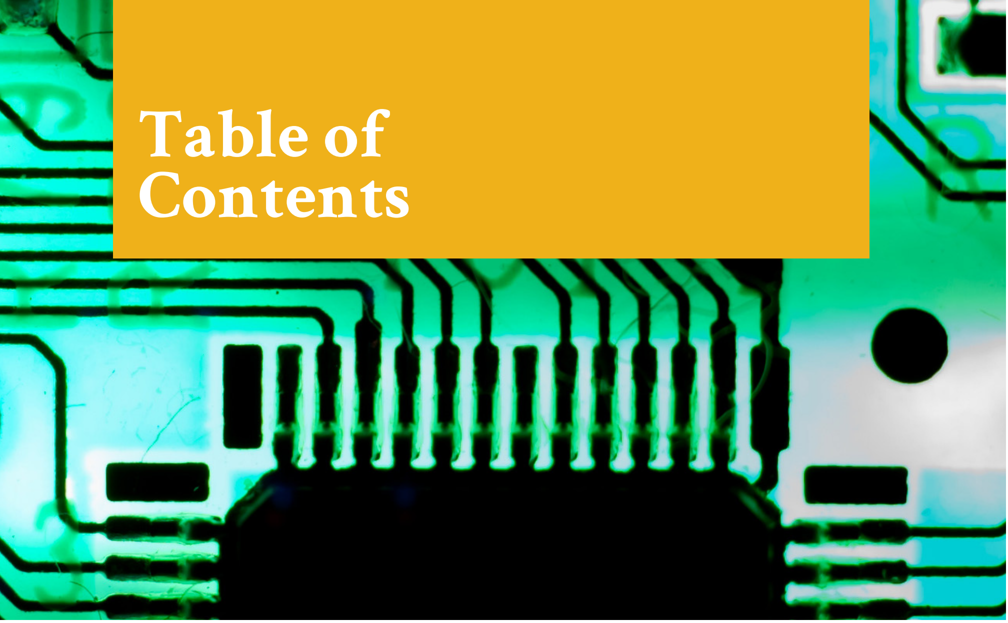


Table of Contents

Executive Summary

This document presents the results of two surveys: one aimed at managers and librarians of repositories, and one directed at technical staff working on repositories.

GOOD PRACTICES

The good practices found in the questionnaire responses are listed in the Discussion section on p. 23. Some of those good practices are:

- DOI, Handle, URN, URI or locally generated numbers should be used as permanent identifiers for metadata records and data.
- Well-known standardized global vocabularies such as ISO vocabularies for country and language codes, as well as COAR[1], OpenAIRE[2], and DataCite[3] vocabularies for publication/ resource types, access status and roles should be applied as much as possible.
- Repositories should preserve information about data provenance stored in metadata: creator, institutions - publishers, source, mail address, publication year, production year, geo-location, data collector, data manager, distributor, editor, funder, producer, rights holder, sponsor, and supervisor.

At the same time, the surveys highlighted some misunderstanding of the FAIR Principles, and misleading implementations.

RICH METADATA MODELS

The definition of what constitutes a rich metadata model is not well defined, and this leads to some misunderstanding of the F2 FAIR principle. In this survey, a majority of respondents said they used a rich data model but 12 of the analyzed repositories had 13 or fewer mandatory fields. Of these, eight had seven or fewer mandatory fields.

MACHINE READABILITY

Nearly 80% of respondents said their repositories completely comply with the I1 FAIR principle: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation). This means humans and computers should be able to exchange and interpret each other's data[4].

Data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings. In order to ensure this, it is critical to use (1) commonly used controlled vocabularies, ontologies, thesauri and (2) a well-defined framework to describe and structure (meta)data. However, 45% didn't answer or said they didn't know whether their repository could display metadata in some semantic web technology such as OWL, RDF notation. Five repositories offer metadata in a semantic web technology, while four plan to implement this feature. The remaining seven

respondents reported that there was no such possibility in their repository.

METADATA PROVENANCE

Although provenance of (meta)data (R1.2) should be described in a machine-readable format, some implementations of this FAIR principle include a free-text provenance description or an attached file which describes provenance.

MISSING INFRASTRUCTURE

Taking into account that the I2 FAIR principle is often missed and quite complicated for implementation (see Figure 2, p. 10 and Figure 5 p. 12), an infrastructure/platform/service which could help in this implementation should be a top priority of EU and other funding programs.

[1] *https://www.coar-repositories.org/activities/repository-interoperability/coar-vocabularies/deliverables*
[2] *https://guidelines.openaire.eu/en/latest/literature/field_publicationtype.html*
[3] *https://schema.datacite.org/meta/kernel-4.2/include/datacite-resourceType-v4.xsd*
[4] *https://www.go-fair.org/fair-principles/i1-metadata-use-formal-accessible-shared-broadly-applicable-language-knowledge-representation*

INTRODUCTION

Data repositories play a crucial role in the evolution of Open Science. The FAIR Data Principles establish how to make data Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). The FAIR principles are as follows:

TO BE FINDABLE

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

TO BE ACCESSIBLE

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

TO BE REUSABLE

- R1. meta(data) have 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 their provenance.
- R1.3. (meta)data meet domain-relevant community standards.

METHODOLOGY

Based on the FAIR Data Principles, two questionnaires were created. The first (hereafter #Q1 - see Appendix #1) targeted repository managers and/or librarians and consisted of 40 questions. The second (hereafter #Q2 - see Appendix #2) targeted technical staff responsible for repository development and maintenance and consisted of 25 questions.

Members of LIBER's Research Data Management (RDM) Working Group[5] circulated the questionnaires between December 2018 and February 2019. Responses were collected from managers and/or librarians of 29 repositories for the first (#Q1) questionnaire.

In addition, technical staff responsible for the development and maintenance of 14 repositories (Table 1) responded to the second (#Q2) questionnaire. In 11 cases, repositories filled out both #Q1 and #Q2.

In this report, the responses for both questionnaires have been merged and analyzed to gain a comprehensive picture about FAIRness at the level of repositories and their data.

ABOUT THE RDM WORKING GROUP

The RDM Working Group operates as part of LIBER's Strategic Direction on Research Infrastructure, which in turn is one of the pillars of LIBER's 2018-2022 Strategy.

The group collects good practices and lessons learned in the area of RDM in libraries, and collaborates with other initiatives to evaluate and support skills development.

It is chaired by Birgit Schmidt, Head of Knowledge Commons at Göttingen State and University Library, and Rob Grim, Economics (Data) Librarian at Erasmus University Rotterdam.

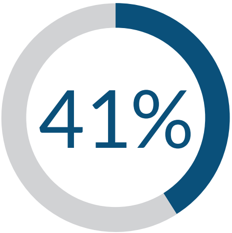
Work on this survey was led by working group member Dragan Ivanović, Associate Professor at the University of Novi Sad. Significant contributions were made by Alastair Dunning, Head of Research Data Services at TU Delft and Head of 4TU.Centre for Research Data, Birgit Schmidt and Rob Grim.

[5] *https://libereurope.eu/strategy/research-infrastructures/rdm*

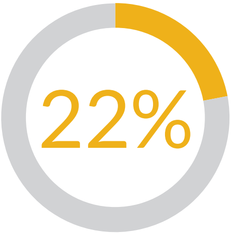
Survey Respondents

- 32 repositories took part in the survey;
- DSpace is the most popular repositories software platform;
- Research data, publications and software are most commonly stored by the repositories.

Among the collected responses, 26 repositories were classified as Institutional, 5 as belonging to Library/Museum/Archives and 5 as Publication repositories.



Of analyzed repositories are based on DSpace (13/32).



Are implemented as an in-house solution (7/32).

In addition to repositories using DSpace or in-house solutions, three are based on EPrints, two on the Pure platform, two on Dataverse and one each on Figshare, Diva consortium, Samvera, Invenio, and Digitool. Managers of four repositories stated they plan to migrate to another solution due to some shortcomings of the currently used solution.

Target communities for the analysed repositories include researchers, academic staff and students, although most repositories represented in the survey are open for all citizens. Research data, publications and software are most commonly stored by the repositories. In a few cases, multimedia files (audio, video, images), musical compositions, teaching materials, patents and data documentation can be also stored.

Data Security

- Very important topic for the surveyed repositories;
- Detection of corrupted files, data backup, recovery, and long-term preservation mechanisms;
- TLS encrypted transfer of sensitive data.

Data security is an important topic for the surveyed repositories. To improve data reliability, redundant media, and the distribution and replication of data between different servers are used. Other methods include:

- Hash values of submitted files verified on a regular basis;
- Functionalities to detect corrupt files;
- Three copies of every item (metadata + associated files) saved by the data center;
- Procedures in place to prevent anyone from accessing read and edit permissions over repository's content;
- Regular auditing activities (log analysis).

Storage facilities are provided by university data centers or data are replicated in several geographical locations. Data are encrypted and/or the servers are secured by ssh key access and firewalls. If sensitive data are stored in the repository, any transfer to a client's machine or to other nodes in the repository servers' network is always TLS encrypted. Also, there are intentions to train researchers to anonymize data or to do that automatically, in order to comply with the EU General Data Protection Regulation (GDPR).[6]

There are also data backup and recovery mechanisms in the analyzed repositories, both incremental and full daily backups. In some repositories, data and databases are backed up according to a 30-day rolling protocol. Backups are stored on a different server.

Respondents from three repositories (#Q1) have a CoreTrustSeal certificate[7] - a core level certification based on the DSA-WDS Core Trustworthy Data Repositories Requirements catalogue and procedures. Three more are in the process of applying for this certificate. This universal catalogue of requirements reflects the core characteristics of trustworthy data repositories. Also, 17 of the analyzed repositories (#Q1) perform long-term preservation and three more plan to do so.

[6] *https://eugdpr.org*

[7] *https://www.coretrustseal.org*

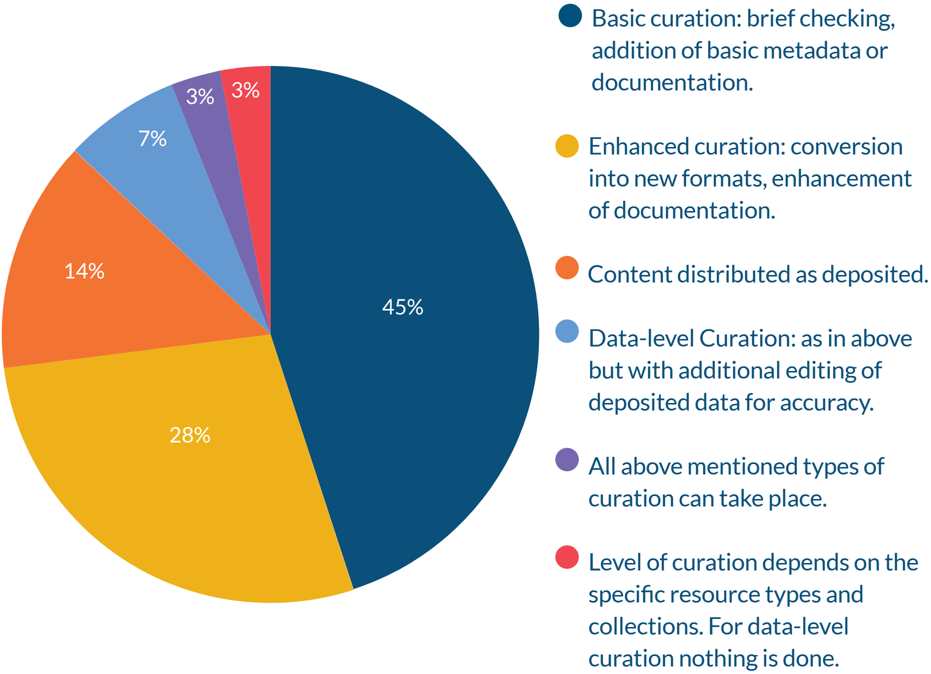
Data Curation & Quality Control

- Curation approaches include brief metadata checking, conversion to new formats, enhancement of documentation;
- Metadata templates, mandatory and optional metadata are defined by repositories.

When asked about relevant types of the level of performed curation (see Figure 1), nearly three-quarters of analyzed repositories said they performed one of two curation approaches:

1. Basic curation – brief checking, addition of basic metadata or documentation
2. Enhanced curation – conversion to new formats, enhancement of documentation

Figure 1: Repository Level of Curation Performed (29 responses)



For some of the analyzed repositories (#Q1), there are metadata templates for new records and librarians validate the information provided by researchers. For others, quality control is the responsibility of the depositor/author. The majority of analyzed repositories contain both mandatory and optional metadata fields. Some repositories also offer the possibility to link a data management plan (DMP) with a dataset.

FAIRness of Data

- F2, A1.2, I2, I3, R1, and R1.3 FAIR principles not completely implemented by all analyzed repositories
- F2, A1.2, I2, and R1.3 FAIR principles are quite complicated to implement
- Checklists to help ascertain FAIR compliance as well as standardization for the use of vocabulary services are missing.

Some of the respondents to the first questionnaire (#Q1) said their repository did not fully comply with the following FAIR data principles (Figure 2):

- F2. data are described with rich metadata.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- I2. (meta)data use vocabularies that follow the FAIR data principles.
- I3. (meta)data include qualified references to other (meta)data.
- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.3. (meta)data meet domain-relevant community standards.

Respondents did not consider these six listed FAIR principles as highly important for the repository (or domain) community (Figure 3) but they did declare an intention to work towards becoming compatible with the FAIR principles F2, I2 and R1.3 (Figure 4).

On the other side, the technical staff who responded to #Q2 indicated that they consider the F2, A1.2, I2, and R1.3 FAIR principles quite complicated to implement (Figure 5). Also, they suggested Data FAIRport, GO FAIR materials, CORE Trust Seal, iRODS, JSONSCHEMA, REACT, Hydra, Fedora, and DSpace as useful sources/platforms/frameworks for implementation of the FAIR principles.

Respondents said they did not feel particular tools were missing for the implementation of the FAIR data principles, even though some FAIR data support mechanisms had to be developed in-house. However, they did say that checklists to help ascertain FAIR compliance as well as standardization for the use of vocabulary services were missing. Additionally, they stated that trainings, guides and best-practice examples were needed more than tools

Figures 2a & 2b: Repository Data Complies With the FAIR Principles

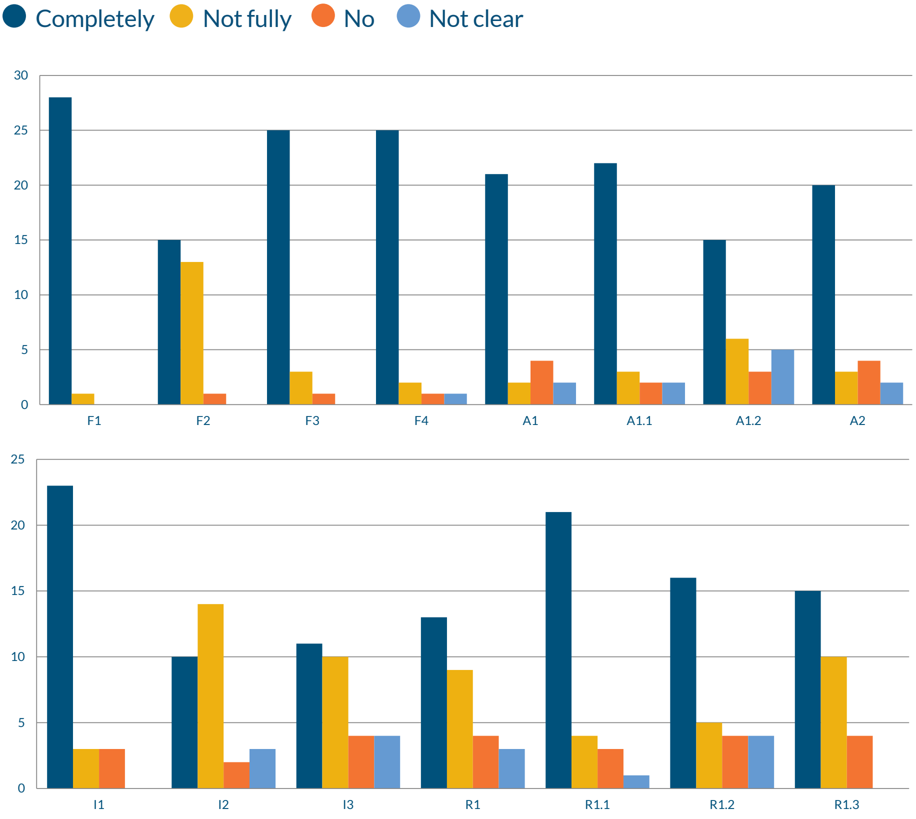


Figure 3: Importance of the FAIR Principles for the Community

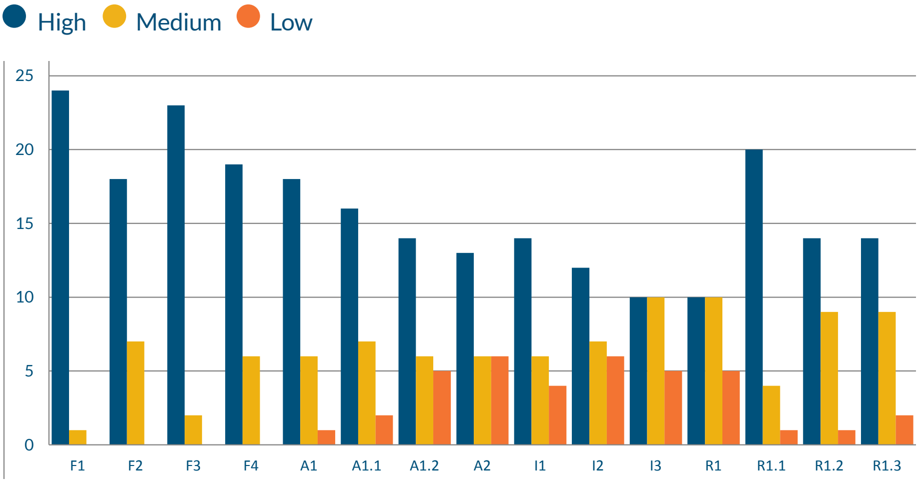


Figure 4: Intention to Implement the FAIR Principles in Future

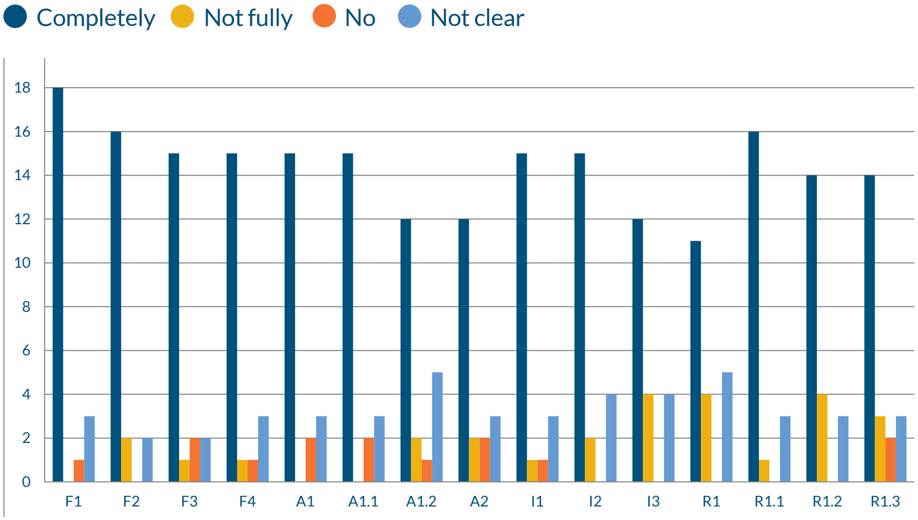
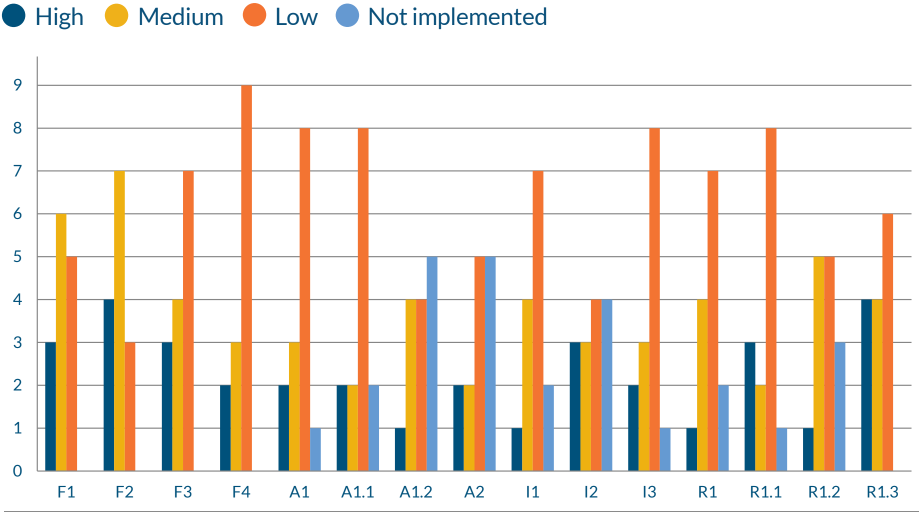


Figure 5: Complexity of Implementation of the FAIR Principles



In #Q1, respondents reacted to the question “Do you monitor the FAIRness of deposited data? If so – how?”. Eight of 29 analyzed repositories monitor data by librarians/staff. Among those, only one responded reported that they systematically monitor of FAIRness of data, while a few others said they planned to improve this aspect of their repository. Four repositories monitor/control FAIRness of deposited data using the editor and a defined metadata set and format. One respondent stated “We don’t know how to monitor it”. Eleven of 29 responses were “No”, while three were “Not yet”.

Findabilty (FAIR)

- DOI, Handle, URN, URI or locally a generated number used as permanent identifiers for metadata records and datasets
- All analyzed repositories have a search web page

All repositories, except one, use DOI, Handle, URN, URI or a locally generated number as permanent identifiers for metadata records. These permanent identifiers are not included in the metadata of five repositories, but inclusion of identifiers in the metadata set is planned for two of the five. The remaining analyzed repositories store permanent identifiers as a metadata.

The usage of DOI, Handle and URN is also dominant as permanent identifiers of datasets associated with metadata records. However, eight of the analyzed repositories don't use any permanent identifier for datasets. Moreover, five state that metadata records and data sets are treated as one entity and it is not possible to have several data sets associated with one record. One respondent recognized the lack of permanent identifiers as a drawback of the repository and stated it is possible to have several datasets associate with one record in his/her repository, although there is only one identifier associated with a metadata record. Almost all analyzed repositories which use permanent identifiers for datasets have already implemented a solution to store dataset identifiers in metadata records, except two which plan to do so.

From the view of technical staff (#Q2), all repositories have a search web page and one supports federated search via the SRU/W protocol. Metadata and data can be indexed and searched via Google and other web search engines for nine analyzed repositories.

Accessibility (FAIR)

Most analyzed repositories:

- Have policies to retain metadata and to remove data
- Support the OAI-PMH protocol for harvesting dataset metadata
- Have authentication in place for humans and machines accessing repository data

More than two thirds of analyzed repositories (#Q1) have a metadata retention policy (Figure 6), and a good majority has a removal of data policy in place (Figure 7).

Figure 6: Is There a Policy for Metadata Retention in Your Repository?

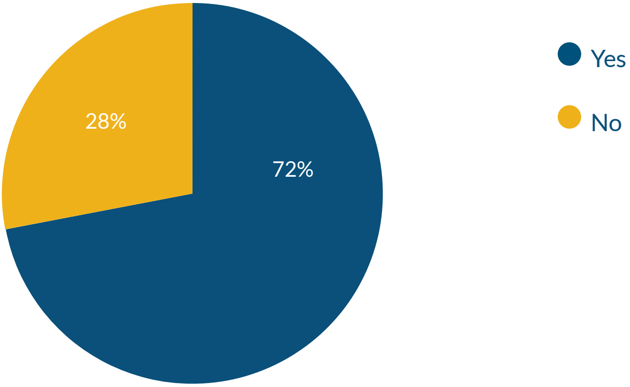
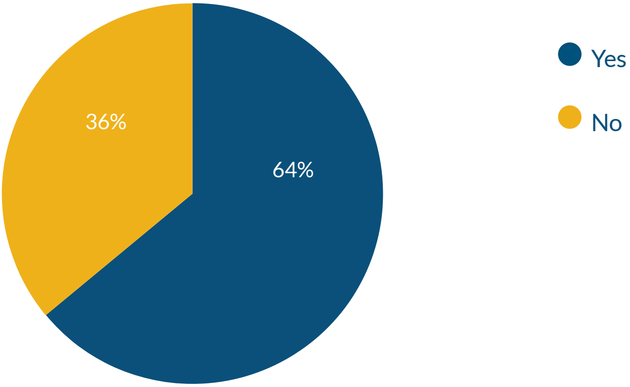


Figure 7: Do You Have Policies Regarding the Removal of Data?



#Q2 respondents provided technical information about access to data for humans and machines.

Half of the analyzed repositories (#Q2) don't have any restriction for accessing deposited data. If irregular data consumption is detected, two repositories have an IP-based restriction which can be used to block access from specific IP addresses/crawlers. Also, the Robots Exclusion Protocol (the robots.txt file) is used to instruct web robots/ crawle4rs. Moreover, time-limited embargoes can be applied for some data by depositors. In four cases, repositories use either a username/password (for people) or an authentication key (for machines) to verify the identity of the person/machine downloading the data.

All analyzed repositories except one support the OAI-PMH protocol for harvesting dataset metadata. Furthermore, two repositories have implemented the OAI-ORE protocol for exporting dataset and its metadata, while one repository uses the ResourceSync protocol for this purpose. One repository offers a JSON-based format with serialized Linked Data (JSON-LD) embedded in landing pages. Moreover, one repository applies the WebDAV and the iRODS protocol for the needs of access data by third parties, while two repositories implement and one is going to implement a REST API for this purpose. One repository offers XML-encoded byte stream for accessing data by third parties.

Depending on the nature of the data, some specific software could be needed in order to open and use data. These are usually standard tools: spreadsheet applications, etc. However, one of the analyzed repositories creates a web-renderable surrogate version for most deposited file formats (e.g. jpg or tiff) which allow them to be viewed / rendered in the browser, while for some obscure file formats the user must download and use other tools to open them.

Interoperability (FAIR)

The analyzed repositories:

- Most commonly support the Dublin Core (83%), DataCite (10%) and DDI (7%) metadata formats
- Use global vocabularies for country and language codes, publication types, access status and roles, and often non-FAIRness customized vocabularies for subjects, scientific fields, temporal qualifiers, publishers and funders
- Mostly support establishing non-qualified links between data stored in the system

Managers and librarians who responded to #Q1 provided information about the richness of metadata, used metadata formats and standards, FAIRness of vocabularies, and linking of data in their repositories. Most analyzed repositories use simple Dublin Core (24 of 29), followed by DataCite (3 of 29) and DDI (2 of 29). Although all Dublin Core fields are optional, repositories define their own list of mandatory fields through a user interface. Twelve of the analyzed repositories had 13 or fewer mandatory fields. Of these, eight had seven or fewer mandatory fields.

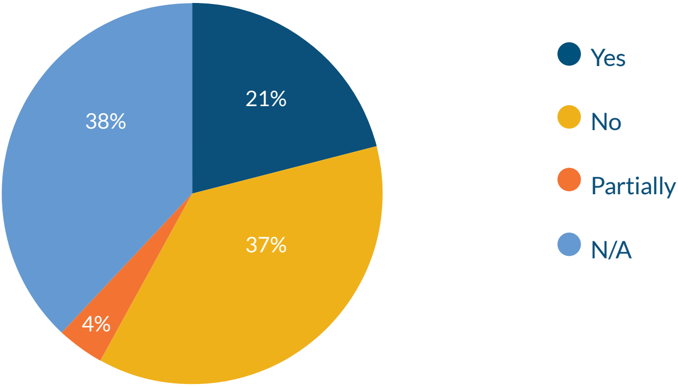
SEMANTIC WEB TECHNOLOGY

Thirteen out of 29 respondents didn't answer or said they didn't understand/know whether their repository could display metadata in some semantic web technology such as OWL, RDF notation. Five repositories offer metadata in a semantic web technology, while four plan to implement this feature. The remaining seven respondents reported that there was no such possibility in their repository.

GLOBAL VOCABULARIES

The analyzed repositories use ISO standardized global vocabularies for country and language codes, as well as COAR, OpenAIRE, and DataCite vocabularies for publication types, access status and roles. Moreover, some use customized vocabularies for subjects, scientific fields, temporal qualifiers, publishers and funders. Just 21% say their customized vocabularies are fully FAIR, as shown in Figure 8.

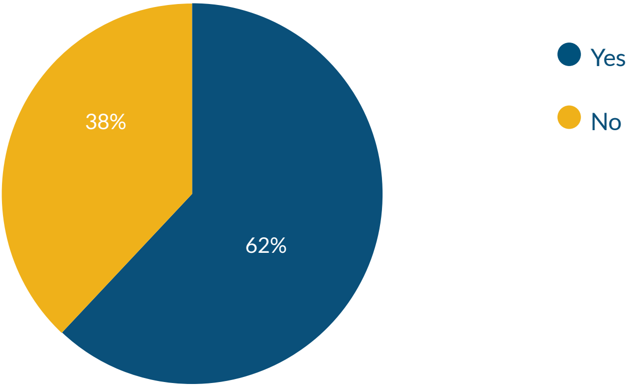
Figure 8: Can Someone Find, Access & Re-Use Your Customized Vocabularies?



Twenty-seven of 29 analyzed repositories allow establishing links between data stored in the system, for instance between datasets, or between datasets and software or publications. Repositories which have implemented the Dublin Core format typically use dc:relation for establishing such links. However, these relations are not qualified. Although all related records could have defined a type of the record, this is not enough to define the intention of the relation. For instance, a relation between publication and dataset could be: Publication cites dataset, Publication describes dataset, etc.

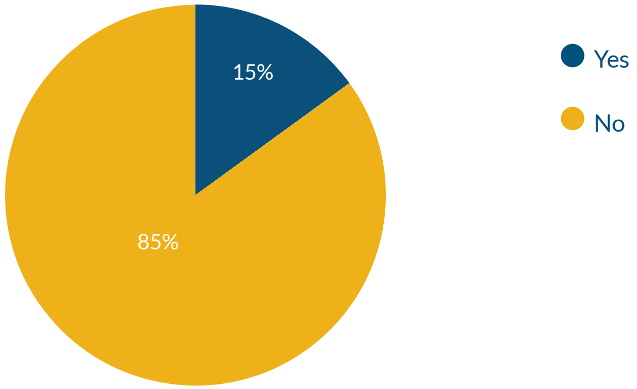
A small majority (62%) of analyzed repositories do offer a "How to cite" option (Figure 9).

Figure 9: Does Your Repository Offer a "How to Cite" Option for Data?



However, over 80% of all analyzed repositories don’t offer citation graphs or another type of analysis tool for data citations and relations (Figure 10).

Figure 10: Does Your Repository Display Citation Graphs or Some Other Analysis Tool for Data Citations and Relations?



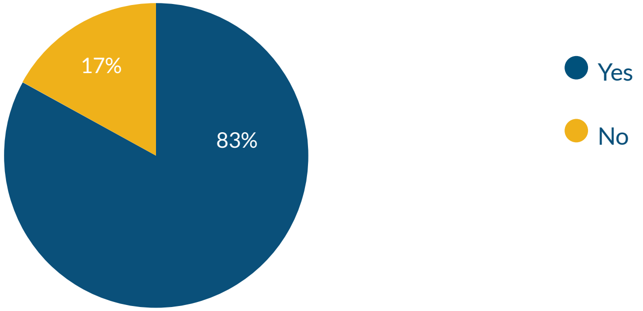
Reusability (FAIR)

- Creative Commons licences are the most popular usage license associated with the data stored in analyzed repositories
- Open, closed and embargoed data are published by the repositories which responded to the survey
- Various ethical or legal issues can have an impact on data sharing via repositories
- Almost 90% of the analyzed repositories preserve more or less information about data provenance

Data quality information which could be preserved in the analyzed repositories include data provenance, peer-reviews and the confidence level when Optical Character Recognition (OCR) is used. Some respondents stated that data quality control is assured by library liaison, in collaboration with data submitters.

The majority of analyzed repositories preserve or link to the usage license associated with the data (Figure 11). Mostly these are Creative Commons licences. Some repositories publish data with Selected Software Licenses, Open Government Licenses, CC Zero, MIT, GPL, Apache, GNU LGPL 2.1, GNU GPL 2.0, ODbL, DbCL, and Open Data Commons Attribution Licence. The information about licenses is stored in metadata. Repositories based on Dublin Core format use the dc.rights.license element for this purpose. One repository modeled usage restrictions in the objects themselves (part of data), if for instance only one part of an object is restricted (like a single photograph in a book). Repositories show the license logo/image/icon, or a hyperlink and/or short license information are displayed alongside the dataset file name or in a separate tab on the dataset-page with information about the Terms of Use.

Figure 11: Does Your System Preserve or Link to Usage Licenses Associated With Data?



OPEN, CLOSED, EMBARGOED DATA

Ten percent (3 of 29) only publish open, publicly distributable research data. The remaining repositories also publish embargoed and closed data. An embargo period can be invoked to comply with the RDM policy of the university, or at the depositor's request. Most repositories don't preserve reasons for embargo periods. Moreover, a strong majority has no limit on embargo periods. Six do define limits as follows: five years (one repository), three years (two repositories), two years (two repositories), and one year (one repository).

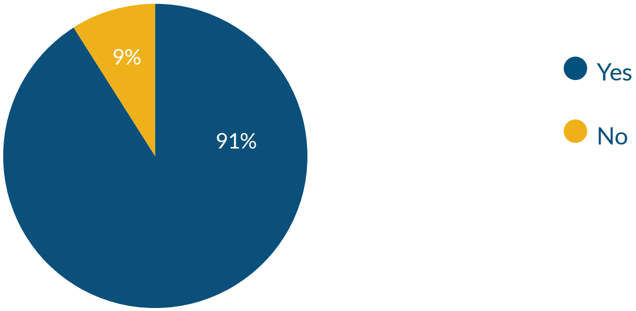
ETHICAL & LEGAL ISSUES

Most analyzed repositories deal with ethical or legal issues which impact on data sharing. Some of the stated approaches for dealing with those issues are as follows:

1. Data submitters who have created datasets by re-using pre-existing resources must indicate this in the submission form. A librarian confirms that value-added elaboration of pre-existing data resources has taken place and that there are no database copyright violations.
2. If the dataset contains privacy-sensitive data (personal data covered by GDPR) and is not anonymized, it can only be accepted with informed consent from the subjects. The data can then only be made available on Restricted Access and/or can be accessed by request to the depositor only. Submitters should have anonymised all observations including privacy-sensitive data if there is no informed consent from the subjects.
3. Depositors have to agree to the following deposit agreement: a) The depositor must own the data or have the right on behalf of the owner/s to deposit the data and make it publicly available under CC-BY licence (subject to any embargo period); b) The data must not break any law e.g. data protection; c)The data does not breach any commercial or legal agreement.
4. Legal/ethical statements can also be included in the metadata record (data protection, ethical approval, commercial constraints, sensitive information).

More than 90% of analyzed repositories directly or indirectly request informed consent for data sharing collected from the data creators (Figure 12).

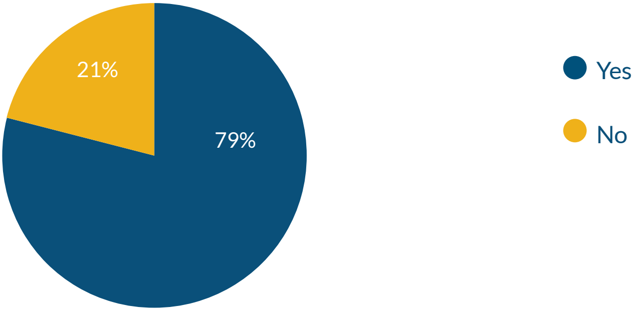
Figure 12: Is Informed Consent for Data Sharing Collected From Data Creators?



Almost 90% of the analyzed repositories (26 of 29) preserve some information about data provenance. Those information could be stored in metadata, as a free-text provenance description, or as a provenance description file assigned with the dataset. Metadata could include: creator, institutions - publishers, source, mail address, publication year, production year, geo-location, data collector, data manager, distributor, editor, funder, producer, rights holder, sponsor, and supervisor.

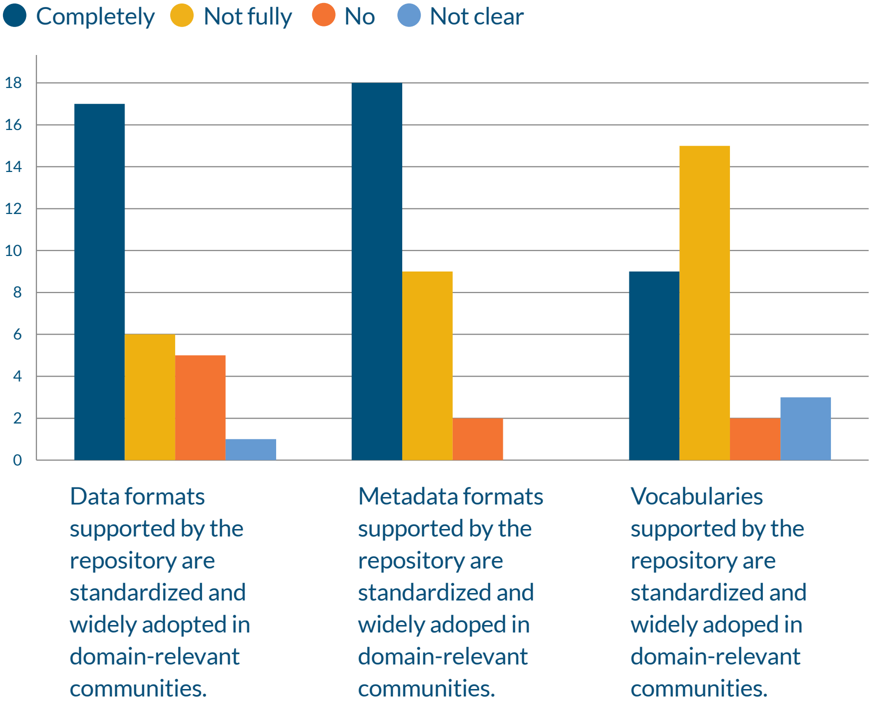
More than three quarters of the analyzed repositories (#Q2) already support, or will support, dataset versioning (Figure 13).

Figure 13: Can Datasets Be Versioned?



A majority of the #Q1 respondents say their repositories support standardized and widely adopted data and metadata formats in certain domain-relevant communities (Figure 14). However, they only partly agree that vocabularies used in their repositories are standardized and widely adopted in the certain community (Figure 14). The majority of #Q2 respondents agree that their repositories support standardized and widely adopted data access protocols in certain domain-relevant communities.

Figure 14: Standardized and Widely Adopted Data Formats, Metadata Formats and Vocabularies in Domain-Relevant Communities



DISCUSSION

- *Good practices for the implementation of FAIRness at the level of repositories*

DOI, Handle, URN, URI or a locally generated number can be used as permanent identifiers for metadata records and data. Those permanent identifiers should be included in the metadata. Several data objects could be associated with one metadata record. Each data object should have its own permanent identifier which should be also stored in metadata. Digital repositories should have a search web page, but could also support federated search via the SRU/W protocol. Also, repositories could enable crawling of open access metadata and data by Google and other web search engines.

Metadata retention policies, as well as removal of data policies should be defined. If irregular data consumption is detected, repositories should support IP based restriction which can be used to block access from specific IP addresses/ crawlers. Also, the Robots Exclusion Protocol (the robots.txt file) could be used to give instructions about their closed or embargoed data to web robots/crawlers. Username and password for persons or authentication key for machines could be used in order to ascertain the identity of the person/machine downloading the data. Repositories could support exporting of metadata/data via OAI-PMH, OAI-ORE and ResourceSync protocols. Moreover, metadata/data could be exposed to third parties through REST API, WebDAV, or iRODS. Repositories could create a web-renderable surrogate version (e.g. jpg or tiff) for most deposited file formats which allow them to be viewed / rendered in the browser.

Repositories should be based on, or at least should support exporting metadata to, some standardized metadata format such as Dublin Core, DataCite or DDI. The list of optional and mandatory metadata should be defined (prescribed or not by standardized metadata format).

Well-known standardized global vocabularies (such as ISO vocabularies for country and language codes, as well as COAR, OpenAIRE, and DataCite vocabularies for publication types, access status and roles) should be applied as much as possible. However, if some customized vocabularies are used, FAIRness of these vocabularies should be implement by the certain repository.

Establishing qualified links between data stored in the system should be supported. "How to cite" option should be supported.

Data quality information which could be preserved in the repositories are: data provenance, peer-reviews associate with datasets, and OCR confidence. Also, data quality control could be assured by library liaison with data submitters.

Repositories should preserve the usage license associated with the data. Repositories should support Creative Commons licences, although publishing data under some other well-known licences should be supported by repositories as well. The information about license should be stored in metadata.

Usage restrictions in the objects themselves (part of data) could be supported, if for instance only one part of an object is restricted (like a single photograph in a book). Repositories should show license logo/image/icon, hyperlink and/or short license information displayed alongside the dataset file name or in a separate tab on the dataset-page with informations about Terms of Use.

Repositories should support publishing open, closed and embargoed data. Although an embargo period can be invoked for the reason of compliance with universities RDM policies, funder and publisher policies, it could be just the depositor decision. Moreover, repositories could preserve a reason for embargo period and could limit the longest (maximal) period for an embargo.

Moreover, repositories have to deal with various ethical or legal issues that can have an impact on data sharing via repositories. Data submitters who created datasets by elaborating pre-existing resources should indicate that in a submission form. A librarian might help to confirm that value-added elaboration of pre-existing data resources has taken place, and that there are no database copyright violations. The data must not break any privacy, commercial or legal agreement.

If the dataset contains privacy-sensitive data (personal data covered by the GDPR) and it is not anonymized, it can only be accepted when there is informed consent from the subjects. The data can then only be made available on Restricted Access and/or can be accessed by request to the depositor only. Informed consent from the subjects (legal/ethical statements) should be included in the metadata record: data protection, ethical approval, commercial constraints, sensitive information. If there is no informed consent from the subjects regarding privacy data, submitters should have anonymised all observations including privacy-sensitive data. The depositor must own the data or have the right on behalf of the owner/s to deposit the data and make it publicly available under some licence (after any embargo period). Furthermore, repositories should request informed consent for data sharing collected from the depositor.

Repositories should preserve information about data provenance which could be stored in metadata. Metadata could include: creator, institutions - publishers, source, mail address, publication year, production year, geo-location, data collector, data manager, distributor, editor, funder, producer, rights holder, sponsor, and supervisor. Also, repositories could support dataset versioning.

- *Analysis of misunderstandings of the FAIR principles and misleading implementations*

Some managers/librarians stated that the F2 (F2. data are described with rich metadata.) principle is not fully implemented in their repositories. However, there are managers who stated it is completely fulfilled, although the metadata model is based on simple Dublin Core (15 elements). Although all Dublin Core fields are optional, repositories define its own list of mandatory fields through the user interface. However, some of those repositories have 7 or less mandatory fields. Richness of the metadata model is not well defined, thus the F2 FAIR principle could be misunderstood.

Overall 23 of 29 respondents (managers/librarians) stated their repositories completely comply with the I1 FAIR principle (I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation). This means humans and computers should be able to exchange and interpret each other's data. It means data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings. In order to ensure this, it is critical to use (1) commonly used controlled vocabularies, ontologies, thesauri and (2) a well-defined framework to describe and structure (meta)data. However, respondents stated that they are using local customized vocabularies which do not follow the FAIR principles. The RDF extensible knowledge representation model, OWL and JSON LD could be used to describe and structure datasets. Moreover, just a few of respondents stated that formats used in their repositories can be expressed in some semantic web technology (OWL, RDF notation).

Ideally, provenance of (meta)data (R1.2) should be described in a machine-readable format. However, there are implementations of this FAIR principle which include free-text provenance description or attached file which describes provenance.

Only a few of the responding technicians stated that their repositories exposed metadata through a REST API, although Best Practice #24[8] of W3C Data on the Web Best Practices suggests a REST API as good example of usage of Web Standards as the foundation of APIs.

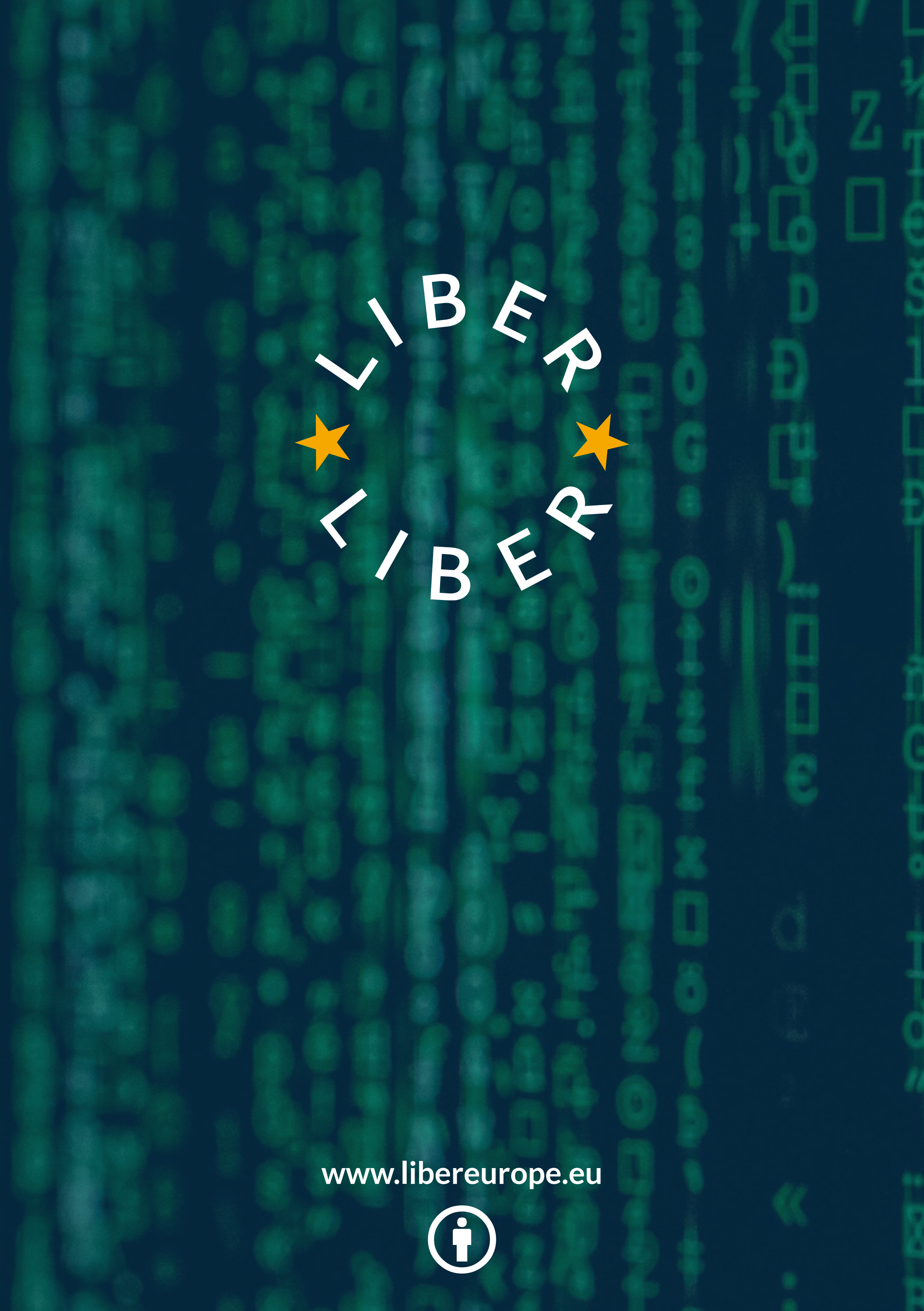
- *Analysis of misunderstandings of the FAIR principles and misleading implementations*

The FAIR principles usually not met from repositories managers/librarians point of view in their repositories are F2, A1.2, I2, I3, R1, and R1.3. Also, managers/librarians would like to implement in the future the following FAIR principles F2, I2 and I3. On the other side technicians find F2, A1.2, I2, and R1.3 FAIR principles quite complicate for implementation. Also, technicians find Data FAIRport, GO FAIR materials, CORE Trust Seal, iRODS, JSONSCHEMA, REACT, Hydra, Fedora, and DSpace useful sources/platforms/frameworks for implementation of FAIR principles. Checklist to help ascertain FAIR compliance as well as standardization for the use of vocabulary services are missing on the market from the technicians point of view. Taking into account that the I2 FAIR principle (I2. (meta)data use vocabularies that follow FAIR principles) is often missed and quite complicated for implementation, infrastructure/platform/service which could help in this implementation should be a top priority of EU and other funding programmes.

LITERATURE

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[8] *https://www.w3.org/TR/dwbp/#APIHttpVerbs*



LIBER

LIBER

www.libereurope.eu