Grant Agreement No.: 732638
Call: H2020-ICT-2016-2017
Topic: ICT-13-2016
Type of action: RIA

D2.01: Initial Guidelines on Data Management

Work package	WP 2
Task	Task 2.7
Due date	30/06/2017
Submission date	07/03/2018
Deliverable lead	Steve Taylor (IT Innovation)
Version	1.0
Authors	Steve Taylor (IT Innovation) Adrian Quesada Rodriguez (Mandat International) Lucio Scudiero (Mandat International)
Reviewers	Peter Van Daele (imec)

Abstract	
Keywords	

DISCLAIMER

COPYRIGHT NOTICE

ACKNOWLEDGMENT

Co-funded by the European Union

Co-funded by the Swiss Confederation

This deliverable has been written in the context of a Horizon 2020 European research project, which is co-funded by the European Commission and the Swiss State Secretariat for Education, Research and Innovation. The opinions expressed and arguments employed do not engage the supporting parties.

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:		**R**
Dissemination Level		
PU	Public, fully open, e.g. web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to FED4FIRE+ project and Commission Services	

** R: Document, report (excluding the periodic and final reports)*

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

Co-funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

This deliverable addresses questions concerning research data storage and data protection. It contains a first draft set of principles and suggested practices by which Fed4FIRE+ aims to uphold the principles of the H2020 Open Data Pilot. In addition, this deliverable contains guidelines, requirements and a plan for Fed4FIRE+ to support General Data Protection Regulations (GDPR). This deliverable also serves as a briefing paper on the state of the art concerning open research data standards, practices and storage.

EC H2020 regulations state that projects should produce an initial Data Management Plan (DMP) at month 6. It is actually not possible for Fed4FIRE+ to provide an initial data management plan concerning the data that will be generated within the project, because of the unique nature of Fed4FIRE+: it is a platform for experimentation by third party experimenters, and it is the experimenters, not the project, who generate the research data. Each experiment that generates open research data will need its own DMP, and Fed4FIRE+'s role is to support experimenters in creating a DMP for their research data. This is the focus of the data management section of this deliverable, and a template DMP with guidance is provided to assist experimenters.

A critical concern for data management is support for General Data Protection Regulations (GDPR), so that citizens' personal data is respected and protected. To address this in Fed4FIRE+, this deliverable provides a set of guidelines for protection of personal data, requirements for Fed4FIRE+ and a plan to ensure that the design of the Fed4FIRE+ federation and architecture is compatible with support for the GDPR.

The key recommendations for Open Research Data Support are as follows, and take the form where the major points below are the key principles, and the minor points for each principle indicate the proposal to address it.

- Encourage experimenters to be as open as possible but do not prevent them from closing data.
  - To encourage opening data, experimenters' costs for opening data are covered by the project (up to limits TBD). Costs are payable at the end of the experiment, once the experimenter has opened their data.
  - Experimenters can elect to close data at any point before publication, citing valid reasons. If they do this, they will not be eligible for cost claims associated with opening experiment data.
- Experimenters must understand the implications of opening data at the outset of their experiment
  - Experimenters who declare an intention to open data should create an initial DMP. This is a simple version of the DMP and not intended to be time consuming.
  - At the end of the experiment, the experimenter should fill in a final DMP. This is a more detailed version of the initial DMP that describes the data from the experiment and how it is stored.
  - Payment for opening data will contingent on both DMPs being filled in, as well as the data being placed in an open repository.
- Fed4FIRE+ should provide an approved set of data repositories that have been evaluated for suitability, and specify the requirements for a Fed4FIRE+ approved repository.
  - The key requirements for the repositories broadly correspond to support for the FAIR principles:

Co-funded by the Horizon 2020 Framework Programme of the European Union

  - Digital Object Identifiers (DOIs) to uniquely identify data
  - The repository needs to provide evidence that it is likely to be there for the long term.
  - Data integrity needs to be protected. The repository needs to provide evidence as to how it will protect the data it stores from compromise or loss.
  - The repository needs to export the metadata of data it stores, so it can be indexed by the popular open data search engines
  - The repository needs to be flexible in the license terms it uses for data it hosts, so that experimenters (who actually own the data) can choose a license.
- It is recommended that Zenodo be approved by Fed4FIRE+, and can serve as the default data repository.

The recommendations regarding support for GDPR are as follows. These requirements concern the authentication information of users. Fed4FIRE+ will not, in principle, use personal data in any of the foreseen experimentations, and personal data will only be obtained and processed as necessary to enable the authentication of users and the provision of the required identity management and single sign-on services.

- **Project data management:** The system must automatically record all internally generated data, storing these data into the Fed4FIRE+ platform, while minimizing the collection of personal data.
- **Data back-ups:** Back-up operations will be carried out periodically, so as to ensure the continuity of the system and prevent the loss of data.
- **Authentication of identities:** The whole system will collect different types of data and it will be designed to ensure the privacy and trust of the users. In order to do this, each identity accessing the system will be authenticated and appropriately authorised to be able to use it. Where necessary (e.g. when the system is used to process health data), strong authentication (e.g. two-factor authentication, double opt-in, biometric recognition, etc.) methods must be supported.
- **De-activation of authentication credentials:** Personal authentication credentials shall be de-activated if they have not been used for at least six months (except in case of technical authorization).
- **Purpose limitation:** As set out by article 5 of the GDPR, Fed4FIRE+ will process personal data only for security purposes, unless the data controller configures the system to pursue other legitimate, specific and explicit purposes, determined at the time of collection of the data.
- **Data accuracy and updating:** Personal data which are inaccurate or incomplete, having regard to the purposes for which they were collected or processed, will be erased or rectified as set out by article 5 of the GDPR.
- **Security of processing:** Fed4FIRE+ will protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access through the implementation of technical and organisational measures as required by article 32 of the GDPR.
- **Data breach information:** The Fed4FIRE+ system must immediately inform its users of any breach to personal data leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed as required by articles 33 and 34 of the GDPR.

Co-funded by the Horizon 2020 Framework Programme of the European Union

- **Encryption by default:** As provided by Article 32 of the GDPR, encryption will be applied to all stages of handling data, including in communication, storage of data at rest, storage of keys, identification, access, as well as for secure boot process.
- **Right of access:** The Fed4FIRE+ system shall support the data controllers in providing to every data subject, without excessive delay or expense, confirmation as to whether or not data relating to him/her are being processed and information as to: the purposes of the processing; the categories of data concerned; the recipients to whom the data are disclosed; the envisaged period of storage for the data; and the existence of automated decision-making processes within the system. The legal source of this requirement is article 15 of the GDPR.
- **Right of erasure:** The Fed4FIRE+ platform must ensure that the right of erasure exercised by data subjects towards the data controller is enforced, when the conditions set out by article 17 of the GDPR are met.
- **Data portability:** As detailed by article 20 of the GDPR, the Fed4FIRE+ platform must be able to support the data controller in responding to requests for data portability lodged by the data subjects. This entails that the data subject shall receive the data in a structured, commonly used and machine-readable format.
- **Regular monitoring of security:** The Fed4FIRE+ platform will regularly monitor the system's status in terms of security for personal data as required by article 32 of the GDPR.

Co-funded by the Horizon 2020 Framework Programme of the European Union

TABLE OF CONTENTS

DISCLAIMER 2
COPYRIGHT NOTICE 2
ACKNOWLEDGMENT 2
EXECUTIVE SUMMARY 4
TABLE OF CONTENTS 7
1 INTRODUCTION 9
2 INITIAL DATA MANAGEMENT PLAN 10
2.1 RATIONALE 10
2.2 GOALS 12
2.2.1 Key Principles 13
2.2.2 Specific Detailed Goals 16
2.3 STRATEGY 17
2.3.1 Incentives for Open Data and Policy for Closed Data 17
2.3.2 Identification of Research Data (for Findability) 17
2.3.3 Open Licensing (for Reusability) 18
2.3.4 Search Support (for Findability) 20
2.3.5 Integrity Protection (for Reusability) 21
2.3.6 Metadata (for Findability, Interoperability & Reusability) 21
2.3.7 Long-Term Storage in Repositories (for Accessibility & Reusability) 27
2.4 DMP TEMPLATE 34
2.5 OPEN DATA PROCESS 42
3 RISK & COST MANAGEMENT FRAMEWORK 44
3.1 RISK MITIGATION 46
3.1.1 RE1: Compromise of Data & RE2: Loss of Data 46
3.1.2 RE3: Data is not findable 46
3.1.3 RE4: Data is not accessible 46
3.1.4 RT1: Data is lost or compromised whilst in the testbed domain 46
3.1.5 RF1: Data compromise brings the federation into disrepute 47
3.1.6 RR1: Unauthorised access to / alteration of / deletion of data stored in repository 47
3.1.7 RU1: Data is not authentic or has been altered 47
3.2 COSTS 47
4 RELEVANT DATA PROTECTION REGULATIONS FOR FED4FIRE+ 48
4.1 KEY GDPR PRINCIPLES 48
4.2 KEY EPRIVACY DIRECTIVE PRINCIPLES 50
4.3 INITIAL EXAMINATION OF PRIVACY REQUIREMENTS 50
4.4 RELEVANT DATA PROTECTION REQUIREMENTS 52
4.5 RELEVANT PRIVACY RISKS AND RISK MITIGATION MEASURES 53
4.6 GENERAL OUTLINE OF THE STRATEGY 54
5 CONCLUSIONS: KEY RECOMMENDATIONS 55
5.1 OPEN RESEARCH DATA 55
5.2 GENERAL DATA PROTECTION REGULATION 56
6 REFERENCES 58
7 BIBLIOGRAPHY 61
8 APPENDIX 1 – ZENODO PRINCIPLES 65

Co-funded by the Horizon 2020 Framework Programme of the European Union

8.1 BEST EFFORT PRINCIPLES 65
8.2 FAIR PRINCIPLES 65
8.2.1 To be Findable: 65
8.2.2 To be Accessible: 65
8.2.3 To be Interoperable: 66
8.2.4 To be Reusable: 66

Co-funded by the Horizon 2020 Framework Programme of the European Union

1 INTRODUCTION

This deliverable addresses questions concerning research data storage and data protection. It contains a first draft set of principles and suggested practices by which Fed4FIRE+ aims to uphold the principles of the H2020 Open Data Pilot. In addition, this deliverable contains guidelines, requirements and a plan for Fed4FIRE+ to support General Data Protection Regulations (GDPR). This deliverable also serves as a briefing paper on the state of the art concerning open research data standards, practices and storage.

It is often an open question as to what constitutes “research data”. A set of categories of data has been determined by the US National Science Board: observational, computational, experimental, and records [National Science Board 2005], which may be used as starting points for classifications of research data. For Fed4FIRE+ purposes, our major concern is the opening and preservation of the data generated within the experiments run on the Fed4FIRE+ platform, and our chosen term for this is “Fed4FIRE+ open experiment data”.

EC H2020 regulations state that projects should produce an initial Data Management Plan (DMP) at month 6. It is actually not possible for Fed4FIRE+ to provide an initial data management plan concerning the data that will be generated within the project, because of the unique nature of Fed4FIRE+: it is a platform for experimentation by third party experimenters, and it is the experimenters, not the project, who generate the research data. Each experiment that generates open research data will need its own DMP, and Fed4FIRE+’s role is to support experimenters in creating a DMP for their research data. This is the focus of the data management section of this deliverable, and a template DMP with guidance is provided to assist experimenters.

A critical concern for data management is support for General Data Protection Regulations (GDPR), so that citizens’ personal data is respected and protected. To address this in Fed4FIRE+, this deliverable provides a set of guidelines for protection of personal data, requirements for Fed4FIRE+ and a plan to ensure that the design of the Fed4FIRE+ federation and architecture is compatible with support for the GDPR.

The deliverable is structured as follows.

- It provides a discussion of the rationale for, strategy to support, and a template for, Data Management Plans (DMPs). It determines a set of requirements for data storage repositories and evaluates candidate repositories for suitability.
- It provides an initial analysis of the risks and costs associated with long term data management, with mitigations to address the risks.
- It provides guidelines and a plan to assist Fed4FIRE+ and its experimenters comply with General Data Protection Regulations (GDPR).
- It concludes by making recommendations for the data management strategy and GDPR for Fed4FIRE+.

Co-funded by the Horizon 2020 Framework Programme of the European Union

2 INITIAL DATA MANAGEMENT PLAN

2.1 RATIONALE

Retention and open access to research data (Open Research Data – ORD) is increasingly becoming an essential element for successful and impactful research activities. ORD supports the concept of open science, which aims to reinforce the basic scientific principles of openness, transparency and reproducibility (paraphrased from [Nosek 2015]), and by making the data scientific publications are based on open by default, these principles are upheld.

Both the EC and the OECD have investigated mechanisms to reinforce the principles of open science. The EC has consulted widely to determine its position, a major result being the Results of the Public Consultation on Science 2.0 [EC Science 2.0 2014], which investigated how science needed to change given new technologies and changes in attitudes, especially how the new technologies can enable and reinforce science's core principles of openness, transparency and reproducibility. The name "Science 2.0" has now been replaced by "open science", better illustrating the core principle of the initiative. The OECD has determined policy to clarify and support open science [OECD 2015a]. The slide set "Open Science: The Policy Challenges" [OECD 2015b] asserts that science should be a global public good, accessible by and for the benefit of all humankind. The OECD defines open science by what elements it should include:

- Open access to scientific publications
- Open and "intelligent" access to research data (and materials)
- Open access to digital applications and source code
- Open access for scientists, the public and commercial companies

[OECD 2015b]

Responding to the Public Consultation on Science 2.0 [EC Science 2.0 2014], the Royal Society concurs with the principles of open science, backed up with open data:

- "Science has benefited from open practices throughout history."
- "Open science offers public and civic, economic and international benefits. Making data open can improve public engagement, enabling the public to engage more easily in the process and results of science."
- "Openness should be the default for research unless precluded by fundamental ethical and legal requirements (such as privacy, security, safety and confidentiality)."
- "Data-intensive science is also becoming a driver for economic growth and development."

[Royal Society Science 2.0 2014]

Also in response to the EC's Public Consultation on Science 2.0, HEFCE et al determined key principles regarding open science data. These are discussed in detail below in Section 2.2, but the first principle provides further evidence as to the benefits of open research data in support of open science:

- "Principle 1: Open access to research data is an enabler of high quality research, a facilitator of innovation and safeguards good research practice."

[HEFCE et al 2016]

Co-funded by the Horizon 2020 Framework Programme of the European Union

To support the principles of open science, funding bodies want to make a public good from not only the science funded by public money, but also the research data underpinning it. In 2011, the leader of a special issue of the journal "Science" covering "Dealing with Data" concluded:

> *"We must all accept that science is data and that data are science, and thus provide for, and justify the need for the support of, much improved data curation." [Hanson 2011]*

Since this assertion, there has been considerable effort made to enable scientific data curation and openness. The EC has been instrumental in this effort, by supporting progress towards openness that has continued for the past ten years. [Dechamp 2016] illustrates the progression from 2008, beginning by advocating open access to publications created in research projects to the current time (2017) where open access to data by default is advocated. At the current time, initiatives such as the H2020 Open Research Data Pilot [H2020 ORD 2016a] and similar national policies are driving institutions to develop research data management policies and infrastructures for storing data to make it publicly available after the original research. Advocating retention of research data with open access (unless there is a good reason not to) and the need for data management plans with justified costs are all because the funding bodies see value in research data being made accessible and reused. The key principle of the EC's approach to ORD is that research data be "as open as possible, as closed as necessary", indicating that the ambition is that the data be as open, discoverable and interoperable as possible, but if there are good reasons for not opening research data, it may be kept closed.

Borgman [Borgman 2012] identifies four motivating scenarios for data sharing that additionally support the case for open science:

(1) to reproduce or to verify research,

(2) to make results of publicly funded research available to the public,

(3) to enable others to ask new questions of extant data, and

(4) to advance the state of research and innovation.

In addition, the so-called "fourth paradigm" – data intensive scientific discovery - has been proposed by Hey [Hey 2009], as an evolution of the existing approaches to science.

FIRE Experimental Facilities generate an ever increasing amount of research data that provides the foundation for new knowledge and insight into the behaviour of Future Internet (FI) systems. It is essential that with the increasing quantities and complexities of data that FIRE offers, effective data management strategies for users are adopted and are supported by tools, services and infrastructures that can implement the plans' objectives and practices. Organisations using FIRE must be encouraged to maximise the value of the research they do using experimental facilities. Research data that is curated, managed and available achieves this end by creating a valuable asset that can provide an underpinning for vibrant and healthy research communities using experimental infrastructures.

Given all these motivations, there is seen to be considerable value in opening data to support science. The Fed4FIRE+ consortium has committed to participation in the Pilot on Open Research Data in Horizon 2020 to offer open access to its scientific results reported in publications, to the relevant scientific data and to data generated throughout the project lifetime in its numerous demonstrators. The Federation proposed by the consortium will generate a significant amount of research data through the experiments running on the testbed facilities that participate in the Federation, and the consortium considers that by giving open access to scientific results and publications backed up by associated open research data, important breakthroughs can be sped up, giving European industry and academia advantages in knowledge and competiveness.

Co-funded by the Horizon 2020 Framework Programme of the European Union

2.2 GOALS

The primary goal of Fed4FIRE+ data management is: to provide mechanisms to enable open research data from the outputs of experiments running within the Fed4FIRE+ testbed federation to be made publicly available in open access form.

In response to the EC's Public Consultation on Science 2.0 [EC Science 2.0 2014], HEFCE et al [HEFCE et al 2016] determined key principles regarding open science data. These serve as a set of high-level principles from which to determine the key goals, objectives and requirements concerning open data management for Fed4FIRE+. The principles are listed as follows (in italics), accompanied by notes (in plain text) regarding their impact on Fed4FIRE.

- Principle 1: Open access to research data is an enabler of high quality research, a facilitator of innovation and safeguards good research practice. This is the key motivating principle, and consistent with the EC's advocacy of ORD: research data should be considered as important a research output as the publications it supports.
- Principle 2: There are sound reasons why the openness of research data may need to be restricted but any restrictions must be justified and justifiable. Fed4FIRE+ can adopt the standard valid reasons why data is not opened, but should not be prevented from considering other reasons. The standard valid reasons for keeping data closed include:
  - Commercial sensitivity
  - Personal information
  - Sensitive information
  - Confidential information
  - Excessive cost to make it open
  - Dependence on other material with restrictive conditions
- Principle 3: Open access to research data carries a significant cost, which should be respected by all parties. Fed4FIRE+ has an objective to encourage experimenters to open data, and costs to open data are likely to be considered barriers to opening data by experimenters. Therefore Fed4FIRE+ experimenters should not be disadvantaged financially if they want to open their experiment data.
- Principle 4: The right of the creators of research data to reasonable first use is recognised. Fed4FIRE+ experimenters may be permitted to delay opening data if they wish to exploit its potential, but the overarching principle for Fed4FIRE+ is to encourage experimenters to open data, so there should remain incentives for experimenters to do so, even if data is embargoed for a time.
- Principle 5: Use of others' data should always conform to legal, ethical and regulatory frameworks including appropriate acknowledgement. Users of ORD need to respect the legal and ethical conditions of the open data they use. The open experiment data created within Fed4FIRE+ needs to be licensed by its creator (usually the experimenter), and the terms of the license determine the rights users have for the data. Fed4FIRE+ should provide support in choosing an appropriate license to experimenters (when it is needed).

Co-funded by the Horizon 2020 Framework Programme of the European Union

- Principle 6: Good data management is fundamental to all stages of the research process and should be established at the outset. Data management should not be considered as an afterthought to research, rather it should be planned beforehand. For Fed4FIRE+, a DMP should be written as part of an experiment proposal (unless the experimenter chooses to keep their data closed and provides adequate reasons). This ensures that the experimenter has considered the implications of opening their experiment data, especially in terms of protecting the data from loss, compromise or mis-identification during the course of the experiment. This principle needs to be set against the effort required to create a DMP, so the DMP at the start of the project should be lightweight.
- Principle 7: Data curation is vital to make data useful for others and for long-term preservation of data. Data curation is vital for long term preservation, and is performed by a data repository. Many repositories exist at the present time (see [Re3Data] for a directory of ORD repositories), and it is recommended that Fed4FIRE+ evaluate and select a set of existing repositories, rather than re-invent the wheel by creating its own. This evaluation is discussed in detail later.
- Principle 8: Data supporting publications should be available by the publication date and in a citable form. For Fed4FIRE+, some experiment data may not accompany a publication, so the first clause only need apply in the situation where data accompanies a publication. Data certainly needs to be in a citable form, and the citation must be valid for the long term. This does not mean simply providing a hyperlink to the data, as these are at risk of becoming invalid (e.g. dead links) over time as server addresses change.

Also responding to the Public Consultation on Science 2.0 [EC Science 2.0 2014], the Royal Society recommends best practice on the maintenance, discoverability, intelligibility and reusability of open science data.

- "Open science requires the effective communication of data: they must be accessible and readily located; they must be intelligible to those who wish to scrutinise them; they must be assessable to that judgments can be made about their reliability and the competence of those who created them, and they must be supported by explanatory metadata (data about data)."

[Royal Society Science 2.0 2014]

2.2.1 Key Principles

2.2.1.1 Fed4FIRE+ Experiment Data Should be Open by Default But Closed is Permitted - Given Valid Reasons

This is a common principle. HEFCE et al [HEFCE et al 2016] state that data should be opened is possible but there may be sound reasons why it should be restricted. These reasons need to be justified and justifiable. The Royal Society [Royal Society Science 2.0 2014] concur: openness should be the default for research, but opt outs are permissible for valid reasons. There is no reason for Fed4IRE+ to deviate from this principle, therefore Fed4FIRE+ experiment data should be open by default, but if there are good reasons, experimenters may close it.

According to Dechamp [Dechamp 2016], the top three reasons for opt-out for EC funded projects are:

- Intellectual property rights – e.g. the data may contain commercially sensitive information, and opening it may impair the protection of IPR.
- Privacy – e.g. the data contains personal or sensitive information.

Co-funded by the Horizon 2020 Framework Programme of the European Union

- Opening might jeopardise project's main objective – e.g. the project may be primarily concerned with developing commercial technology and opening the data prevents effective exploitation of the technology.

HEFCE et al [HEFCE et al 2016] describe similar reasons why data may be kept confidential:

- The data is commercially sensitive or contains confidential information
- The data contains personal or sensitive information
- There would be excessive costs to make the data open
- The data depends on other material with restrictive conditions

Some of these reasons will require additional explanation, e.g. to quantify the costs of opening data.

2.2.1.2 Incentives for Openness Should be Provided

Researchers should be encouraged to be open about their research. The Royal Society [Royal Society Science 2.0 2014] cite incentives for openness, e.g. funding and publishing should require open data, and academic recognition should be based on open data as well as publishing. In addition, HEFCE et al assert that research data should be considered as important a research output as the publications it supports. Therefore Fed4FIRE+ will provide incentives for experimenters to provide open data.

2.2.1.3 The FAIR Principles Must be Supported

The EC advocates the *FAIR* principles for research data management [Wilkinson 2016]. FAIR is an acronym for *Findable*, *Accessible*, *Interoperable* and *Reusable*. **Error! Reference source not found.** summarises the principles (from [Wilkinson 2016]). The Royal Society proposes a variation of FAIR with extensions:

> *"Intelligent openness is the optimal way to achieve open data. The first joint G8 Science Ministers and Presidents of national academies of sciences meeting in 2013*[1] *outlined that for the system of open data to work, data must be:*
>
> a. *Discoverable*
> b. *Accessible*
> c. *Intelligible*
> d. *Assessable*
> e. *Usable"*
>
> [Royal Society Science 2.0 2014]

For the purposes of Fed4FIRE+, given that it is an EC-funded federation, the FAIR principles will be used.

[1] G8 Science Ministers Statement: https://www.gov.uk/government/news/g8-science-ministers-statement

Co-funded by the Horizon 2020 Framework Programme of the European Union

The FAIR Guiding Principles

To be Findable:
- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

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) 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

2.2.1.4 Data Integrity Must be Protected

The integrity of Fed4FIRE+ open experiment data needs to be protected, meaning that once data is made available, it must not be altered or deleted. The data needs to be verifiably the same as that originally made available by its creator, otherwise it is worthless at best and damaging at worst, because any subsequent experiments will be made on different data than the original experiment, violating the “reusability” FAIR principle.

It has been known for the original owner of the data to want to alter it after opening it, e.g. if they discover a flaw in their scientific argument. If the publication is accepted even with a flaw in its argument or methodology for example, the underpinning data must be consistent with the publication so that the flaw may be verified and addressed by the community. The data and the publication it underpins need to be independently immutable, and also bound together immutably. Once accepted for publication, neither the publication nor the data can be changed, and their relationships must be unbreakable.

Co-funded by the Horizon 2020 Framework Programme of the European Union

2.2.2 Specific Detailed Goals

More detailed goals, which all contribute to the primary goal above, are listed as follows.

(1) Fed4FIRE+ experimenters must be encouraged to make their experimental data open, but must not be discouraged from experimenting using the platform when there is a genuine reason to keep the data confidential. This is in accordance with the EC's policy of "as open as possible, as closed as necessary", and Fed4FIRE+ aims to follow the principle of "open by default".

(2) Data created within Fed4FIRE+ experiments and opened must conform to the EC's FAIR principle (Findable, Accessible, Interoperable, Reusable).

Findable:

I. Each Fed4FIRE+ open experiment data set must be uniquely identifiable.

II. Fed4FIRE+ open experiment data must be discoverable using standard methods, e.g. search engines that operate on keyword searches.

III. Fed4FIRE+ open experiment data must be discoverable by both humans and machines, e.g. humans doing internet searches and machines crawling for relevant metadata.

Accessible:

IV. Fed4FIRE+ open experiment data must accessible to anyone who needs it.

V. Fed4FIRE+ open experiment data must be accessible for the long term.

Interoperable:

VI. Fed4FIRE+ open experiment data should not use closed or proprietary formats.

VII. Fed4FIRE+ open experiment data should be described using standardised vocabularies, and the vocabularies' definitions should be referenced in the metadata.

Reusable:

VIII. Fed4FIRE+ open experiment data must be licensed such that reuse is permitted.

IX. Fed4FIRE+ open experiment data must be citable.

X. Any data that Fed4FIRE+ open experiment data derives from must be cited by the Fed4FIRE+ open experiment data.

(3) Data stored as a result of Fed4FIRE+ experiments must have its integrity protected, e.g. it must not be possible for unauthorised parties to delete or tamper with it.

(4) Making data open and FAIR should be as easy and straightforward as possible for experimenters

XI. Workloads for opening data, adding any metadata necessary to support FAIR and placing in a repository should be as lightweight as possible.

XII. Fed4FIRE+ should assist experimenters as much as possible in making their data open and FAIR.

Co-funded by the Horizon 2020 Framework Programme of the European Union

2.3 STRATEGY

This section contains plans to address the objectives described in section 0.

2.3.1 Incentives for Open Data and Policy for Closed Data

It is proposed that the experimenters can select whether they want their data closed in their proposals in their proposal, and if so, they need to provide reasons for this. If experimenters do not explicitly request to close the data to be closed, it is assumed to be open, and experimenters need to be clearly informed of this.

It is also proposed that experimenters who opt to have their data open get paid more, to cover their costs for making the data open. This should also encourage experimenters to open data whenever possible. This is in line with HEFCE et al's [HEFCE et al 2016] principle 3: There should be funding to cover the costs of opening data.

The experimenters that choose to provide open data (i.e. the default case) will need to fill in a draft Data Management Plan) at the time of application.

We need to cover the case where experimenters can close data at any time during their experiment. If an experimenter wants to close the data when they have previously declared it to be open, they are permitted to do so, but need to provide adequate reasons. In addition, they will lose the extra money for their experiment. To simplify payment processing, it is proposed to make the payment for experimentation either at the end of the experiment, or to have the payment in two parts, one before the experiment, and the second when the experiment ends. Having a payment at the end of the experiment means that it can be calculated based on the experimenter releasing any open data they have promised.

Fed4FIRE+ needs to determine the amount that can be claimed by experimenters in support of open data. It is not clear whether this should be a fixed limit, a percentage of an experiment budget or another mechanism. What is important is that experimenters understand at the outset that they will get more money if they open their experiment data.

2.3.2 Identification of Research Data (for Findability)

Fed4FIRE+ open experiment data needs to be uniquely identified worldwide, so unique identifiers need to be attached to every open data set created within Fed4FIRE. The clear best choice, and therefore recommendation, for Fed4FIRE+ is the Digital Object Identifier (DOI) standard [DOI], [Paskin 2010], as it is by far the most widely adopted and already accepted as the de facto standard for identification and citation of academic material.

The DOI is an ISO standard [ISO 26324: 2012] that provides permanent and unambiguous identification for digital objects. The DOI identifier is bound to metadata describing the object and other properties like where the object is located. A key benefit of the DOI is that the object's location can change over time (e.g. move web site), and the object's DOI identifier can remain the same – only the metadata associated with the object needs to change. This makes DOI identifiers more stable than URLs to locate objects.

DOIs are assigned by Registration Agencies, who are managed by the International DOI Foundation (IDF). A DOI has a prefix and a suffix separated by a forward slash (/), e.g.: doi:10.1000/182. The prefix identifies the Registration Agency, and the suffix is managed by the individual Registration Agency to uniquely identify the object within it. A Registration Agency can update the metadata for the DOIs they manage at any time, e.g. when a URL changes. DOIs can be resolved using doi.org, e.g.:

https://doi.org/10.1000/182

Co-funded by the Horizon 2020 Framework Programme of the European Union

will read the metadata associated with the DOI and redirect to the resource's location stored in the metadata.

2.3.3 Open Licensing (for Reusability)

Licensing of Fed4FIRE+ open research data is necessary to set out the terms under which the data may be used. Experimenters should be free to choose the terms under which they license their data, but it is often the case that they will need support in choosing a license that best meets their requirements. In addition, it is an objective that licensing should emphasise open science and data and encourage reuse, so Fed4FIRE+ will provide guidance for the experimenters on selection of the right license terms based on their wishes, whilst illustrating the cases where licensing terms can restrict openness and reuse.

The Open Definition "provides a set of principles that define "openness" in relation to data and content" [Open Definition], and references a set of licenses compatible with these principles. The Open Definition:

> *… makes precise the meaning of "open" in the terms "open data" and "open content" and thereby ensures quality and encourages compatibility between different pools of open material.*
>
> *It can be summed up in the statement that: "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."*
>
> [Open Definition]

Ball [Ball 2014] provides a guide to the three major license conditions:

- Attribution – acknowledgement of the creator of the licensed work. This is by far the most common term, as it sets out who owns the rights to the licensed work.
- Copyleft – any derived works must use the same license terms as the licensed work. This is a term used by licenses that enforce a wish to propagate a certain condition, and is known as a "viral" license. For example, the GPL [GPL v3] advocates that software should be open, and includes strong copyleft, meaning that any work derived from a GPL-licensed library must also be open. Copyleft can be a restriction to commercial exploitation, so including copyleft terms in a license may restrict reuse.
- Non-commercial – any derived works must not be commercial. This is another viral license term. It does not restrict the license used for any derived works, but does prohibit any commercial use.

Licenses also often include statement that there is no warranty and a limitation of licensor liability. Many software licenses are based around one or more of these terms. Some licenses (e.g. [FreeBSD]) are no more than attribution in the form of a copyright statement plus a statement of no warranty and a limitation of liability.

Creative Commons [Creative Commons] provides a set of standardised licenses that are popular due to their clarity. Creative Commons is currently at version 4.0, and all subsequent discussion and recommendation will be based on version 4.0. All Creative Commons (CC) licenses have a statement of no warranty and limitation of liability, and each CC license is differentiated by inclusion or exclusion of the following terms.

- BY – Attribution. This is used in all CC licenses, and means that the original author of the licensed work is identified.

Co-funded by the Horizon 2020 Framework Programme of the European Union

- SA (Share Alike) – Copyleft – any derived works must use the same license as the licensed work. This type of license is often referred to as “viral” because once this license is used any derived work from then on must use the same license.
- ND (No Derivatives). No derivatives are permitted. This is highly restrictive, as only the licensed work is permitted. This term is not compatible with the Open Definition because it prohibits derivatives and thereby contradicts the Open Definition’s permission to modify the licensed work.
- NC (Non-Commerical). Commercial use is prohibited. This term is partially viral in that any derived works do not have to be licensed under the same terms, but they must be non-commerical. This term is not compatible with the Open Definition because it prohibits commercial use, and this contradicts the Open Definition’s permission to use the work for any purpose.

Some terms are mutually exclusive, e.g. ND is not compatible with SA (because SA means derived works and ND prohibits derived works), and all licenses use BY for attribution. The actual licenses are:

- CC-BY – Attribution alone. The licensee can do anything with the licensed work as long as they acknowledge the creator of the licensed work. This is the most permissive CC license, and therefore is likely to bring maximum usage. (The following licensed omit the attribution clause in their description because it is the same for each.)
- CC-BY-SA – Attribution and Share Alike. Derived works must use CC-BY-SA.
- CC-BY-ND – Attribution and No Derivatives. The licensed work must be distributed by any licensee unchanged and complete, but commercial use is permitted. This license is not compatible with the Open Definition.
- CC-BY-NC – Attribution, Non-Commercial. Derived works must be non-commercial, but may choose their license. This license is not compatible with the Open Definition.
- CC-BY-NC-SA – Attribution, Share Alike, Non-Commercial. Derived works must be non-commercial and use CC-BY-SA-NC. This license is not compatible with the Open Definition.
- CC-BY-NC-ND – Attribution, No Derivatives, Non-Commercial. The licensed work must be distributed unchanged and complete, and under non-commercial terms. This is the most restrictive of all six CC licenses, and is not compatible with the Open Definition.

If an experimenter wishes to select their own license, they are free to do so. If experimenters need assistance in deciding on a licence for their data, Fed4FIRE+ will advocate the CC licenses, because they are clear and easy to understand, as well as providing for the major stipulations for licensing of data. Fed4FIRE+ will assist experimenters in selecting the most appropriate CC license for their purposes (the CC provides a license chooser wizard on their website[2]). In the case that the experimenter objects to all CC licenses, other licenses should be investigated.

[2] https://creativecommons.org/choose/

Co-funded by the Horizon 2020 Framework Programme of the European Union

There is an additional set of terms that the CC provides, under which works can be distributed. This is the so-called "CC0 waiver[3]", and means that the owner of the work waives (gives up) all their rights and titles to a work distributed under CC0. Any work distributed under CC0 is explicitly public domain, meaning that anyone can do anything with the work, and there is no requirement for acknowledgement to the creator of the work. Distributing a work under CC0 is irreversible – anything distributed under CC0 cannot later be distributed under different terms, because once it is in the public domain, it cannot be removed. Because of the irreversibility, CC0 will not normally be recommended to experimenters, unless it is their express wish that their work be public domain and they require no attribution.

2.3.4 Search Support (for Findability)

Once stored in a repository, the data needs to be findable by e.g. searching. The primary mechanism for this from the data creator's perspective is to add metadata at deposition time so as to enable searching via e.g. keywords or tags.

Dedicated research data and article search engines harvest and aggregate metadata from repositories to provide a search service. The de facto standard for metadata harvesting is The Open Archives Initiative Protocol for Metadata Harvesting:

> *"The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata."*
>
> [OAI-PMH]

In order to search repositories, researchers can use dedicated search engines for open access repositories, and major search engines are described below. It is therefore a requirement that any repository chosen by Fed4FIRE+ makes available the metadata describing its contents to these open data search engines.

OpenDOAR

The Directory of Open Access Repositories portal provides search functionality to find repositories[4] based on criteria such as subject types for data, and also to search the repositories' contents themselves for data based on keywords[5]. In its own words, it is:

> *"…OpenDOAR provides a quality-assured listing of open access repositories around the world. OpenDOAR staff harvest and assign metadata to allow categorisation and analysis to assist the wider use and exploitation of repositories. Each of the repositories has been visited by OpenDOAR staff to ensure a high degree of quality and consistency in the information provided …"*
>
> [OpenDOAR]

[3] https://wiki.creativecommons.org/wiki/CC0_FAQ
[4] http://www.opendoar.org/find.php
[5] http://www.opendoar.org/search.php

Co-funded by the Horizon 2020 Framework Programme of the European Union

OpenAIRE

A portal providing access to EC-funded open research data (with a search function for the actual data based on keywords[6]) is the OpenAIRE [OpenAIRE]. This is an aggregator of research output (papers and data), and provides an access portal for many OpenAIRE-compliant data repositories (where the actual research data is stored). A user can simply go to the OpenAIRE search page, type a keyword query and survey the results categorised by:

- publications;
- research data;
- projects;
- people;
- organizations; or
- data providers.

DataCite

DataCite [DataCite] provides a search function at its webpage[7]. This enables users to use keywords to search for open research content. The results can be filtered by resource type (e.g. data, text, image, software), publication year or repository.

Re3Data

Re3Data [Re3Data] is a directory of data repositories. It is searchable, but can only find repositories, not data in the repositories. Because Re3Data can only find repositories, it is only necessary for a repository to export metadata to Re3Data, only register with it.

2.3.5 Integrity Protection (for Reusability)

Data must be protected from decay, intentional and unintentional alteration or deletion once it is in the repository. These requirements are the responsibility of the repository, so will be key criteria for selection of data repositories for Fed4FIRE+. The requirements for repositories are described in Section 2.3.7.1.

2.3.6 Metadata (for Findability, Interoperability & Reusability)

Metadata for Fed4FIRE+ open experiment data is needed for a number of reasons:

- to enable data to be easily found, e.g. via a keyword search;
- to enable interoperability, e.g. describing the format specification of the data bundle so that it can be understood;
- to enable reuse – in addition to the interoperability reasons above, e.g. license terms governing reuse need to be specified.

The metadata should be created by the experimenter and submitted along with the data for storage. Whilst it is technically possible that the repository owner can add metadata, the experimenter is much better placed to describe their own experiment data, so therefore it should be a requirement of funding for opening data that the experimenter submit adequate metadata when uploading to a repository.

[6] https://www.openaire.eu/search/find?keyword=

[7] https://search.datacite.org/

Co-funded by the Horizon 2020 Framework Programme of the European Union

2.3.6.1 Metadata Requirements

The repository providers typically determine which metadata should or must accompany data submissions, and follow already existing metadata schemas for description of research data. Fed4FIRE+'s selection of a provider will in part be determined by which metadata schemas they support, and any metadata schema chosen by a provider should have the following properties.

- It is mature and well-adopted. The schema must have a strong chance of survival, and be interoperable and compatible with the widest number of users and tools possible.
- It supports data identification using Digital Object Identifiers.
- It supports a basic set of information that is commonly used to find research data. An example of such basic information[8] is :
  - Title
  - Author/contributor name(s)
  - Author/contributor ORCID iD(s)
  - Abstract
  - Keywords
  - Licence (e.g. CC BY)
  - Identifier (ideally DOI)
  - Publication date
  - Version
  - Institution(s) (of the authors/contributors)
  - Funder(s) (ideally with grant references; can also be "none/not externally funded").
- It supports information on the format of the data.
- The basic information is optionally extensible to include e.g. domain-specific data.
- It should be lightweight enough so as not to discourage experimenters from opening data.

Possible candidates for metadata standards can be those cited by [Metadata Standards Directory WG] in the "General Research Data" Section, or those referenced by the UK Digital Curation Centre [DCC][9]. These are discussed and evaluated for suitability next.

2.3.6.2 Evaluation of Metadata Standards

CERIF

The Common European Research Information Format (CERIF) [CERIF] is advocated as the EU's preferred choice to describe research activity. It is complex, and attempts to describe all aspects of a research project [Russell 2012]. As such, it is rather heavyweight for the purposes of simply describing experiment data, and there is the fear that its complexity will act as a barrier to encouraging experimenters to open data.

[8] https://wwwf.imperial.ac.uk/blog/openaccess/2016/02/19/less-is-more-metdata-schema-discovery-research-datary-of-research-data/

[9] http://www.dcc.ac.uk/drupal/resources/metadata-standards

Co-funded by the Horizon 2020 Framework Programme of the European Union

Data Packages

Data Package [Data Packages] is:

> *"… a simple container format used to describe and package a collection of data. The format provides a simple contract for data interoperability that supports frictionless delivery, installation and management of data.*
>
> *Data Packages can be used to package any kind of data. At the same time, for specific common data types such as tabular data is has support for providing important additional descriptive metadata – for example, describing the columns and data types in a CSV."*
>
> [Data Packages]

As such, it can be suitable for creating packages of data, but not for describing the data to assist searching and interoperability.

DCAT

The Data Catalog Volcabulary (DCAT) is

> *"DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.*
>
> *By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation."*
>
> [DCAT]

Given this, DCAT is suitable as a vocabulary aimed at the repository owners for interoperability, rather than at experimenters.

Dublin Core

Dublin Core [Dublin Core] is a well-established standard for basic, domain-agnostic metadata describing resources of many different types. It was published as an ISO standard in 2009, and has been revised in 2017 [ISO 15836-1:2017]. Dublin Core provides a set of 15 basic elements suitable for describing a resource and its attributes [DCES 1.1], shown in the table below.

Label	Definition
Contributor	An entity responsible for making contributions to the resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Creator	An entity primarily responsible for making the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Description	An account of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Language	A language of the resource.
Publisher	An entity responsible for making the resource available.
Relation	A related resource.
Rights	Information about rights held in and over the resource.
Source	A related resource from which the described resource is derived.
Subject	The topic of the resource.
Title	A name given to the resource.
Type	The nature or genre of the resource.

Table 1: Dublin Core Metadata Element Set [DCES 1.1]

Each element is optional and can be repeated if needed, e.g. if there are multiple related resources. It can be seen from the table that the key elements needed for description of Fed4FIRE+ open experiment data are present, and given that Dublin Core is mature and well-adopted, so if a repository supports Dublin Core, this contributes to its case as a Fed4FIRE+ approved repository.

DataCite

The specific application of DOIs to research data is supported by DataCite [DataCite]. This is an international organisation whose members include the British Library[10], the TU Delft Library[11], the California Digital Library[12] and German National Library of Science and Technology[13], and whose purpose is to "*help the research community locate, identify, and cite research data with confidence*"[14]. DataCite has a rich community of members[15], some of which allocate DOIs for research data:

> *DataCite has 42 members and more than 800 data centers around the world*
> - *Allocating members: they allocate DOI names and use the Registration Agency of DataCite in their capacity as allocating*

[10] https://www.bl.uk/
[11] http://www.library.tudelft.nl/en/
[12] http://www.cdlib.org/
[13] https://www.tib.eu/en/
[14] https://www.datacite.org/mission.html
[15] https://www.datacite.org/members.html

Co-funded by the Horizon 2020 Framework Programme of the European Union

agents. They work actively with data centers and users for the purpose of issuing DOIs.

- *Non-allocating members: those who support DataCite's mission but do not wish to allocate DOI names. They work actively with DataCite and the wide research data community, but do not act as allocating agents.*

[DataCite About]

DataCite provides a profile for metadata to describe research data [DataCite Metadata Kernel_v4.0]. This provides basic fields that are harvested by search engines so that the research data can be easily found, and in a consistent way. The metadata profile is an XML schema (an example is provided at the DataCite schema website[16]), and the key fields are given below.

Mandatory	Recommended	Optional
Identifier	Subject	Language
Creator	Contributor	Alternate ID
Title	Date	Size
Publisher	Related identifier	Format
Publication year	Description	Version
Resource Type	GeoLocation	Rights

Table 2: DataCite Metadata Schema Fields (from [DataCite Tech])

As with Dublin Core, DataCite is clearly applicable to the needs of Fed4FIRE+, so if a repository supports DataCite, this contributes to its case as a Fed4FIRE+ approved repository. It is mature, is well-adopted and has strong support from major establishments worldwide. It supports data identification using DOIs and covers the major fields needed to identify and search for open research data. In addition, it is relatively simple to create an XML file containing DataCite metadata.

OAI-ORE

The Open Archives Initiative Object Reuse and Exchange [OAI-ORE] is a standard for describing aggregations of web resources, e.g. collections of photos, links etc. As such, it is not directly applicable for description of open research data, but may be applicable later on for describing aggregations of open research data.

Observations and Measurements

The Open Geospatial Consortium has provided a schema for Observations and Measurements [OGC] concerning measurements and location-based information. It is therefore not directly suitable for describing experiment metadata, but may be used to provide additional information, particularly as it additionally includes standards to describe sampling techniques.

[16] https://schema.datacite.org/meta/kernel-3/example/datacite-example-full-v3.1.xml

Co-funded by the Horizon 2020 Framework Programme of the European Union

PREMIS

The PREMIS (Preservation Metadata: Implementation Strategies) Data Dictionary [PREMIS, PREMIS Dict] has the purpose of describing the information needed for long-term preservation and usability of data. It was created by the PREMIS working group in the mid-2000s and is now managed by the US Library of Congress. It enables description of the original structure of a data item, including the environment in which it was created and originally functioned. The environmental information is useful because of the obsolescence of formats and operating environments, so the information should enable either migration to a new data format or emulation of the original environment. PREMIS also enables recording of provenance information for a data item – the history of the item in the form of events that affected it.

The top level schema of PREMIS is shown in Figure 1. The main entity is Object, which represents the items to be preserved. Objects can represent distinct real-world items that are preserved (the so-called Intellectual Entity), or digital Representations of them (e.g. in the form of files). Objects can aggregate other Objects – for example an Intellectual Entity can have multiple Representations, each having different files. An Environment is a special class of Object – this is the technical stack of hardware, software and other dependencies needed to interpret the Representations. The Environment itself consists of files and other digital items, so it is therefore an Object. Events record actions operating on Objects. Actions may optionally alter the Objects, create new relationships with other Objects or may just record access requests to the Object. Whether Objects are immutable is an important design consideration but is outside the scope of the specification. Often the question of data immutability is addressed by the domain-specific community in which the data was created or is used. Events record the history of actions on an Object, so therefore contribute to the provenance record of an Object. Agents are actors who execute actions on Objects (which generate Events to record them). Rights Statements describe the rights that determine whether Agents (either specific or types) can have access to Objects, and the terms under which access is permitted (e.g. license terms).

PREMIS is provided as an XML schema, and an example PREMIS metadata XML file is shown at the Library of Congress website . PREMIS has applications within Fed4FIRE+ for recording the origin of the data (e.g. its creation environment), plus representations needed to transform the data into a meaningful resource (e.g. formats and environment) and tracking the provenance of data created in Fed4FIRE+. However, it is heavyweight – the PREMIS dictionary [PREMIS Dict] is 283 pages. As a result, PREMIS is a good candidate for advanced experimenters who want to provide additional information, but should not be mandatory.

2.3.6.3 Conclusion

For the purposes of identification and finding data, Dublin Core and DataCite are clearly the most applicable standards, so a repository that supports either may be considered a candidate for Fed4FIRE+. DataCite also provides the advantage that its website provides a search engine for open research data.

Co-funded by the Horizon 2020 Framework Programme of the European Union

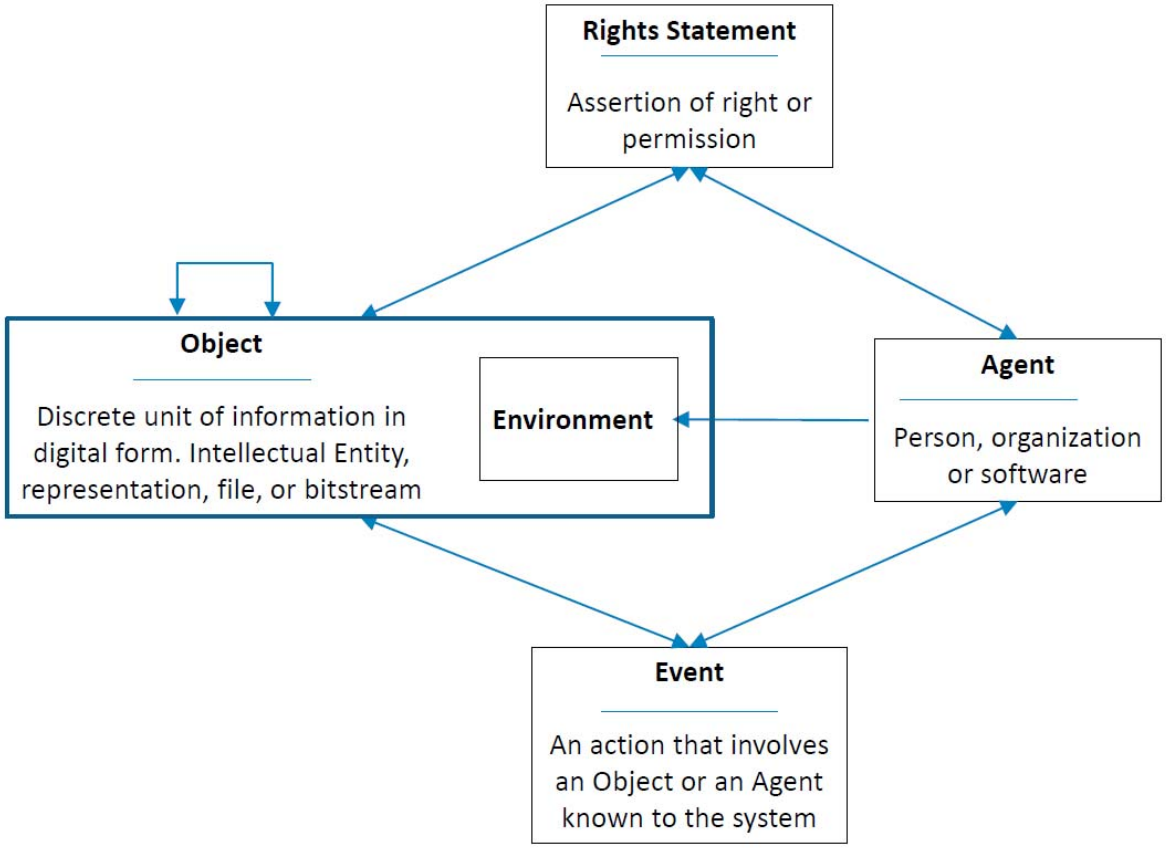

Figure 1: PREMIS Entities (from [PREMIS Figs])

2.3.7 Long-Term Storage in Repositories (for Accessibility & Reusability)

In order for research data to be persistent and open-access, it needs to be stored in a repository. There are already many existing repositories whose explicit purpose is to support open science by providing storage, open access and discovery to research publications and supporting data. A directory of such repositories can be found at Re3Data [Re3Data], and a search engine for EC-funded open research data is at the OpenAIRE Portal [OpenAIRE].

It is impossible to survey all repositories for suitability due to their number, so this section will provide requirements for repositories (with justification) to assist evaluation of a repository. Fed4FIRE+ does not mandate a single repository, but will provide recommendations of repositories based on evaluation against the requirements. If an experimenter wishes to upload Fed4FIRE+ open experiment data to a different repository than those recommended, they are free to do so, as long as the repository is compliant with the requirements.

2.3.7.1 Repository Requirements

This section describes the requirements for selecting a repository for Fed4FIRE+ open experiment data.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in these requirements are to be interpreted as described in [RFC 2119].

Co-funded by the Horizon 2020 Framework Programme of the European Union

(1) *Requirement 1 – DOIs as Unique Data Identifiers.* The repository must be an authorised issuer of DOIs, and must issue a DOI to identify each data deposition. DOIs are a well-established and proven mechanism to provide archival citations for digital resources, and DOIs are not vulnerable to link rot[17].

(2) *Requirement 2 – Evidence of Repository Longevity.* The repository must provide evidence of its sustainable future existence over the long term. This could be evidence of funding for the foreseeable future, a sustainable business operation model. Further evidence of the potential for the repository's longevity e.g. that the repository is well-adopted and well-used is desirable. This requirement is necessary because the data needs to be stored for the long term, so the repository it is stored in must be sustainable over the long term.

(3) *Requirement 3 – Commitment to Data Integrity Preservation.* The repository must publish policies as to the level and type of their efforts concerning protection from data decay, alteration or loss, including what will happen to data the event of the repository going out of business. This is so that users of the repository can judge the repository's methods and commitment to data integrity preservation. The policies should provide evidence that at least best efforts are made to protect the integrity of the data.

(4) *Requirement 4 – Descriptive Metadata Specification.* The repository must determine or choose a metadata specification that is adequate to enable the data to be found (e.g. provide keywords, title, description). The metadata specification must also unambiguously identify the creator of the data set (e.g. name, affiliated organisation) and the creation date. The metadata specification should provide descriptions of data types or formats, especially if they are proprietary. From the evaluation of metadata standards in Section 2.3.6.2, a repository's support for Dublin Core or DataCite is considered advantageous in its case as a Fed4FIRE+ approved repository.

(5) *Requirement 5 – Storage & Export of Descriptive Metadata.* The repository must store descriptive metadata immutably bound to the data it describes. The metadata must be made available to open research data search engines (e.g. those described in Section 2.3.4). The metadata exported to search engines should be in a standardised format (e.g. Dublin Core or DataCite) and exported automatically using standardised protocols (e.g. OAI-PMH). Storing and exposing descriptive metadata is necessary to enable the data it describes to be found easily.

(6) *Requirement 6 – Flexibility of Licensing.* The repository must not mandate a single license for all its content. This is because the data creators need some say in deciding the license terms of their data. It is acceptable to provide a set of licenses from which the data creator can choose. Any license stipulations by the repository should be compatible with the Open Definition. This is because the overall objective of storing data in a repository is to open it.

2.3.7.2 Evaluation of Candidate Repositories

This section contains an evaluation of two candidate repositories with a view to providing at least one repository as a starting point for Fed4FIRE+. More repositories may be evaluated and added as is deemed necessary.

Dryad

The Dryad Digital Repository is:

[17] https://en.wikipedia.org/wiki/Link_rot

Co-funded by the Horizon 2020 Framework Programme of the European Union

> *"… a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of datatypes."*
>
> [Dryad]

Requirements Assessment

Dryad's policies [Dryad Policies] provide information to asses Dryad against the repository requirements. All quotations in this assessment, unless otherwise attributed, are from [Dryad Policies].

Requirement 1 – DOIs as Unique Data Identifiers. The preamble to Dryad's policy concerning Dryad's objectives states:

- "Guiding these Terms of Service are Dryad's aims to:
  - […]
  - assign and provide Digital Object Identifiers (DOIs) to repository content;"

This is adequate evidence that DOIs are used within Dryad.

Requirement 2 – Evidence of Repository Longevity. Dryad's policies concerning sustainability state:

- "… Dryad's governance and business model are designed to provide for long-term organizational stability and viability, ensuring that revenues from Data Publishing Charges (DPCs) cover the Repository's core operating costs (including curation, storage, and maintenance). Thus, long-term funding for Dryad is not dependent on a small number of grants, or the continued largesse of a single host institution."

This is adequate evidence of the sustainability of Dryad.

Requirement 3 – Commitment to Data Integrity Preservation. Dryad's policies concerning its procedures for preservation state:

- "Dryad aims to preserve the originally submitted Content indefinitely. Steps to ensure long-term availability include:
  - Persistent identification in the form of DOIs.
  - Replication of data and metadata.
  - Periodic verification that stored content remains uncorrupted. Upon ingest, an MD5 checksum is calculated and recorded for each submitted data file. Integrity checks are run nightly.
  - If data objects are found to be corrupted, affected data will be restored from known-good copies as appropriate.
  - To prevent data loss, multiple copies of content are kept at different sites.
  - Prior to release, files are manually inspected by a Dryad curator. Any issues that affect preservation (i.e., file corruption, etc.) are addressed before release.
  - Ingest, curation, and publication actions are documented by provenance metadata.
  - Dryad content is replicated through participation in the DataONE network[18].

[18] https://www.dataone.org/

Co-funded by the Horizon 2020 Framework Programme of the European Union

- To the greatest extent possible, personnel create documentation to reflect their activities. We strive to ensure that more than one employee can perform each critical function of the repository."

- In addition, Dryad's policy on sustainability states: "… Dryad's participation in the DataONE network ensures that all its data will be available through other institutions if the Dryad organization ever dissolves."

This is adequate evidence that data stored within Dryad is protected for the long term or at risk from the Dryad going out of business.

Requirement 4 – Descriptive Metadata Specification. Dryad is not specific about the metadata necessary, because its assumption is that data is bound to a publication stored in a different publisher's repository, and the publication will reference the data. It states in its FAQ[19]:

- "Dryad welcomes data files associated with any published article in the sciences or medicine, as well as software scripts and other files important to the article."
- "Dryad works with journals to integrate article and data submission, streamlining the submission process."
- "[submitters need to] Provide titles, descriptions and keywords for your datafiles, to make the data more discoverable and to assist in understanding the relationship of the datafile to the publication."

Fed4FIRE+ needs to cater for data being findable independent of publications, so that Dryad relies on data bound to publications is problematic.

Requirement 5 – Storage & Export of Descriptive Metadata. It is clear that Dryad stores some metadata. Dryad provides a search within its website based on keywords from the publications, but the metadata does not appear to be exported to external open data search engines. This is most likely due to the operating model of Dryad where data is bound to a publication, which can be found externally. For Fed4FIRE+, this approach is problematic.

Requirement 6 – Flexibility of Licensing. Dryad uses the Creative Commons CC0 rights waiver for all data it hosts meaning that the data is explicitly public domain and the creator has waived all their rights to it. This policy is directly contradictory to the requirement that the repository must not mandate a single license.

Costs: Dryad charges a fixed fee of $US 120 for each submission.

Given the problems described above. Dryad cannot be considered as a repository for Fed4FIRE+

Zenodo

Zenodo [Zenodo], hosted by CERN, is recommended as the de facto OpenAIRE-compliant data repository, but other repositories are permissible provided they are compliant with the OpenAIRE standards:

> *"Researchers working for European funded projects can participate by depositing their research output in a repository of their choice, publish in a participating Open Access journal, or deposit directly in the OpenAIRE repository ZENODO – and indicating the project it belongs to in the metadata. Dedicated pages per project are visible on the OpenAIRE portal."*[20]

[19] https://datadryad.org/pages/faq

[20] https://www.openaire.eu/support/faq/openaire-faq

Co-funded by the Horizon 2020 Framework Programme of the European Union

Zenodo's implementation of the FAIR principles is described in [Zenodo Principles], and is included in Section 8. Zenodo provides Digital Object Identifiers (DOIs) [DOI] for all uploaded data bundles.

Requirements Assessment

This assessment uses [Zenodo Principles] and [Zenodo Policies] as its primary sources.

Requirement 1 – DOIs as Unique Data Identifiers. Zenodo mints and assigns DOIs to data submissions. From [Zenodo Principles] on FAIR Principles:

- "F1: (meta)data are assigned a globally unique and persistent identifier
  - A DOI is issued to every published record on Zenodo."

This is clearly adequate evidence that Zenodo issues DOIs for the data it stores.

Requirement 2 – Evidence of Repository Longevity. The Longevity section of [Zenodo Policies] state:

- "Retention period: Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least."

In the "Accessible" section, [Zenodo Principles] state:

- "A2: metadata are accessible, even when the data are no longer available
- Data and metadata will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least."

Metadata are stored in high-availability database servers at CERN, which are separate to the data itself.

Given that Zenodo is hosted by CERN, which is large, well-funded and has a research program planned for at least the next 20 years, Zenodo is deemed to have sufficient chances of longevity for the long term.

Requirement 3 – Commitment to Data Integrity Preservation. In the section concerning Longevity, [Zenodo Policies] states:

- "Versions: Data files are versioned. Records are not versioned. The uploaded data is archived as a Submission Information Package. Derivatives of data files are generated, but original content is never modified. Records can be retracted from public view; however, the data files and record are preserved.
- Replicas: All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. Data files are kept in multiple replicas in a distributed file system, which is backed up to tape on a nightly basis.
- Retention period: Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
- Functional preservation: Zenodo makes no promises of usability and understandability of deposited objects over time.
- File preservation: Data files and metadata are backed up nightly and replicated into multiple copies in the online system.

Co-funded by the Horizon 2020 Framework Programme of the European Union

- Fixity and authenticity: All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.
- Succession plans: In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories."

The [Zenodo FAQ] has a specific question on data preservation:

- "Is my data safe with you / What will happen to my uploads in the unlikely event that Zenodo has to close?
- Yes, your data is stored in CERN Data Center. Both data files and metadata are kept in multiple online and independent replicas. CERN has considerable knowledge and experience in building and operating large scale digital repositories and a commitment to maintain this data centre to collect and store 100s of PBs of LHC data as it grows over the next 20 years. In the highly unlikely event that Zenodo will have to close operations, we guarantee that we will migrate all content to other suitable repositories, and since all uploads have DOIs, all citations and links to Zenodo resources (such as your data) will not be affected."

Files are redundantly stored in multiple locations, are also backed up and have frequent integrity tests in the form of checksums. In addition, Zenodo has a succession plan in the event of its closure. These factors provide adequate evidence of Zenodo's commitment to the integrity of the data it hosts.

Requirement 4 – Descriptive Metadata Specification & Requirement 5 – Storage & Export of Descriptive Metadata. The Content section of [Zenodo Policies] states:

- "Metadata types and sources: All metadata is stored internally in JSON-format according to a defined JSON schema. Metadata is exported in several standard formats such as MARCXML, Dublin Core, and DataCite Metadata Schema (according to the OpenAIRE Guidelines)."

The Access and Reuse section of [Zenodo Policies] states:

- "Access to data objects: Files may be deposited under closed, open, or embargoed access. Files deposited under closed access are protected against unauthorized access at all levels. Access to metadata and data files is provided over standard protocols such as HTTP and OAI-PMH."

Zenodo exports metadata in Fed4FIRE+'s preferred standards, Dublin Core and DataCite. OAI-PMH is used to export metadata, which can be harvested by open research data search engines. Therefore both requirements 4 and 5 are deemed satisfied.

Requirement 6 – Flexibility of Licensing. The Content section of [Zenodo Policies] states:

- Licenses: Users must specify a license for all publicly available files. Licenses for closed access files may be specified in the description field.

In the Reusability section, [Zenodo Principles] state:

- "R1.1: (meta)data are released with a clear and accessible data usage license
  - License is one of the mandatory terms in Zenodo's metadata, and is referring to a Open Definition license.
  - Data downloaded by the users is subject to the license specified in the metadata by the uploader."

Co-funded by the Horizon 2020 Framework Programme of the European Union

The metadata that describes any data stored in Zenodo is CC0 license, forcing it to be public domain and non-copyright. The Access and Reuse section in [Zenodo Policies] states:

- "Metadata access and reuse: Metadata is licensed under CC0, except for email addresses. All metadata is exported via OAI-PMH and can be harvested."

Licensing only metadata as CC0 is acceptable because the metadata's purpose is to advertise and describe the type and purpose of the data so that it may be reused. The owner of the data is still free to choose a license for the actual data they upload.

That the user specifies a license for their data is mandatory. Zenodo recommends that the license is compliant with [Open Definition] (i.e. using one of the licenses in [Open Definition Licenses]). Therefore, Zenodo gives the user a choice of license, so requirements 6 is deemed satisfied.

Costs: Zenodo is free at the point of use for most submissions. [Zenodo Terms] state:

- "Content may be uploaded free of charge by those without ready access to an organized data centre."

Other Factors:

Zenodo also provides automated reporting to the EC for open data stored within it, so evidence of the commitment from Fed4FIRE+ and experimenters to provide open experiment data can be easily tested.[21]

Zenodo also supports collections, and it is possible that a collection could be created for the Fed4FIRE+ experimental community. This would be a focal point for the Fed4FIRE+ open experiment data, even though it should be possible to upload the data to any compatible repository.

2.3.7.3 Recommendation

Clearly Zenodo is the obvious choice for the first Fed4FIRE+ approved data repository. Others may follow as appropriate, provided they are compliant with the requirements in Section 2.3.7.1.

[21] http://help.zenodo.org/features/

Co-funded by the Horizon 2020 Framework Programme of the European Union

2.4 DMP TEMPLATE

This section contains an annotated copy the H2020 Data Management Plan provided in [H2020 ORD 2016b]. The H2020 DMP template has many complex questions, which may be daunting or off-putting to experimenters and could negatively influence their decision to open data. The guidance notes in [H2020 ORD 2016b] point out that the DMP is a living document, and the questions need not all be answered at the beginning of the data management process, nor in great detail. It is recommended (see e.g. [H2020 ORD 2016b]) that DMP be reviewed periodically, and for the case of a Fed4FIRE+ experiment, whose lifetime is in the order of months, it seems reasonable that a basic DMP be assessed when judging the experiment proposal, and a more detailed version be reviewed after the experiment, upon which reimbursement of the experimenter's data management costs will depend.

The following table contains the H2020 Data management questions, with an indication of which questions need to be answered[22] in the initial DMP, which must accompany the experiment proposal if the experimenter wishes to be eligible for reimbursement of data opening costs, and the final, more detailed DMP, which needs to be submitted in order to claim the data opening costs. The table contains guidance notes indicating the information needed in Fed4FIRE+.

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
0	**Experiment Information**			
	Name of Experiment	Y	Y	
	Names of Experimenters	Y	Y	
	Experimenters' Organisations	Y	Y	
	Fed4FIRE+ Call ID	Y	Y	
	Experiment Start Date	Y	Y	
	Experiment End Date	Y	Y	
	Fed4FIRE+ Testbeds	Y	Y	
	Fed4FIRE+ Sponsor	Y	Y	
1	**Data Summary**			

[22] Y = mandatory to answer question, O = optional to answer, N/A = not applicable in Fed4FIRE+

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	What is the purpose of the data collection/generation and its relation to the objectives of the project?	Y	Y	This should be the abstract of experiment from proposal including objectives of collecting the experiment data.
	What types and formats of data will the project generate/collect?	Y	Y	Initially this can be an estimate. In the final DMP this should be a statement of the formats, so it can go into the metadata.
	Will you re-use any existing data and how?	O	Y	If any external data is anticipated before the experiment starts, state it here. If any external data has been used during an experiment, it must be stated, along with any license terms or stipulations.
	What is the origin of the data?	Y	Y	This is the expected source of the data before the experiment runs, and the actual source of data once the experiment is complete.
	What is the expected size of the data?	O	Y	Initially this can be an estimate. In the final DMP this should be the actual size of the data.
	To whom might it be useful ('data utility')?	O	O	If there are any expected users of the data, state them.
2	**FAIR data**			
2.1	*Making data findable, including provisions for metadata*			
	Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?	Y	Y	Initially, this should be a statement committing that the experiment data will be discoverable. When the experiment is complete, the experiment data's Digital Object Identifier (DOI) and metadata should be cited. Fed4FIRE+ advocates DOIs for Fed4FIRE+ open experiment data, and payment of costs associated with opening data will be contingent upon the experimenter placing the research data in a recognised repository, and sharing the DOI issued by the repository with the Fed4FIRE+ Federator.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	What naming conventions do you follow?	O	Y	Initially this can be optional, although it is recommended to think of the naming conventions before the data is collected. After the experiment, this should cite the naming conventions used.
	Will search keywords be provided that optimize possibilities for re-use?	Y	Y	This should always be YES - there will be or are keywords for search terms. The keywords should be stated here.
	Do you provide clear version numbers?	O	O	Version numbers should be stated if applicable.
	What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.	Y	Y	Initially, this should be citations of the metadata schemas that are planned to be used, with indications of what will go into the fields (e.g. the title of the experiment etc). After the experiment, this should be a citation to the actual metadata used for the data.
2.2	*Making data openly accessible*			
	Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.	N/A	N/A	This question is not applicable in Fed4FIRE's context. The question of whether data is intended to be opened or not is a gateway question and only if the answer is "yes", will the DMP be filled in at all.
	How will the data be made accessible (e.g. by deposition in a repository)?	Y	Y	The default position of Fed4FIRE+ is that the answer to this is to be placed in an open repository.
	What methods or software tools are needed to access the data?	O	O	If there are any special tools or methods needed to access the data (e.g. commercial software tools that can open the data's format), state them here.
	Is documentation about the software needed to access the data included?	O	O	If software tools are needed, cite the documentation.
	Is it possible to include the relevant software (e.g. in open source code)?	O	O	If possible, include or cite the software tools (e.g. sourceforge location)

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.	Y	Y	Initially, state the anticipated repository location by name and URL. For the final DMP state the actual name and URL of the repository and cite the data's DOI. Fed4FIRE+ anticipates providing a recommended repository list, but users are free to choose their own repository as long as it complies with the requirements of Fed4FIRE+.
	Have you explored appropriate arrangements with the identified repository?	O	O	For Fed4FIRE+ recommended repositories, this should not be necessary for the experimenters because Fed4FIRE+ has already investigated the repository's storage arrangements.
	If there are restrictions on use, how will access be provided?	N/A	N/A	The default position of Fed4FIRE+ is that data should be open, not restricted, so this should not apply.
	Is there a need for a data access committee?	N/A	N/A	
	Are there well described conditions for access (i.e. a machine readable license)?	N/A	N/A	
	How will the identity of the person accessing the data be ascertained?	N/A	N/A	This is the responsibility of the repository.
2.3	*Making data interoperable*			
	Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?	Y	Y	The default position for Fed4FIRE+ is "yes - the data will be (or is) interoperable". This section should be a statement of commitment by the experimenter that the data will be (or is) interoperable.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?	Y	Y	Initially, this should be a statement of the formats intended for the data, together with citations of their definitions if applicable (e.g. RFCs etc). For metadata, the experimenter should cite the anticipated metadata schemas by URL. After the experiment is complete, it should be a statement of the actual formats used, as well as citations to metadata schemas.
	Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?	N/A	N/A	Addressed in the above question.
	In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?	O	O	Description of the mappings, if applicable.
2.4	*Increase data re-use (through clarifying licences)*			
	How will the data be licensed to permit the widest re-use possible?	Y	Y	Initially, this should be a statement of the intended license, which at least must permit open access. Once the experiment is complete, the data must be licensed under terms that permit open access, and the license must be named here.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.	O	O	The default position is that the data must be deposited in a Fed4FIRE+ compatible public repository within two months of the completion date of the experiment, to allow time to prepare and upload the data. If there are any special requirements for additional embargo time, they must be stated in either the initial or final DMP. Should any additional embargo time be needed by the experimenter, the Federator will judge the justification of the request, and if it is granted, will extend the publication deadline for that data. Only once the data is uploaded in the repository and it is uploaded before the deadline, will any payments of open data costs be made to the experimenter.
	Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.	N/A	N/A	The data should be reusable by third parties.
	How long is it intended that the data remains re-usable?	O	O	The default position is in perpetuity, but there may be practical limits on the time that the data is made available, and if they are known, they can be stated here.
	Are data quality assurance processes described?	O	O	If any QA procedures are observed, they should be stated - it is in the interest of the experimenter to describe these, as they will help the reusability of the data.
3	**Allocation of resources**			
	What are the costs for making data FAIR in your project?	Y	Y	The experimenter can claim additional costs for opening data over and above their experiment budget, up to a specified limit. In order to claim the costs, the experimenter must provide an indication in the initial DMP and the actual costs in the Final DMP.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).	N/A	N/A	Fed4FIRE+ will reimburse costs for opening data up to a fixed limit (to be decided by the Federation Board), once the experimenter has uploaded their open data into a repository and provided the DOI and metadata to the Federator.
	Who will be responsible for data management in your project?	Y	Y	The person responsible for the data management should be named in both the initial and final DMP.
	Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?	O	O	This is the responsibility of the repository. The repository should provide a long term data retention policy that describes how long data is kept for, as well as any notification procedures for disposal.
4	**Data security**			
	What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?	N/A	N/A	This is the responsibility of the repository. The experimenter may base their choice of repository on its reputation and any guarantees a repository provides regarding security and integrity.
	Is the data safely stored in certified repositories for long term preservation and curation?	N/A	N/A	This is the responsibility of the repository. The decision the experimenter needs to make is which repository, and this decision needs to be based on an assessment of the repository's chances of preserving the data long term.
5	**Ethical aspects**			
	Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).	Y	Y	Legal, ethical and data protection issues must to be described in the initial DMP that forms part of the experimenter's proposal before the experiment runs, together with procedures for correct compliance with the applicable laws including the implications of storing the data for the long term in an open repository.

Co-funded by the Horizon 2020 Framework Programme of the European Union

Sect-ion	DMP Category and Question	Initial DMP	Final DMP	Fed4FIRE+ Guidance Notes
	Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?	Y	Y	The experimenter must specify methods for acquiring informed consent in their initial DMP.
6	**Other issues**			
	Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?	O	O	If other DMP procedures are used, the experimenter should state them.

Co-funded by the Horizon 2020 Framework Programme of the European Union

2.5 OPEN DATA PROCESS

This section contains a first draft process for open experiment data, involving an experimenter and the Federator.

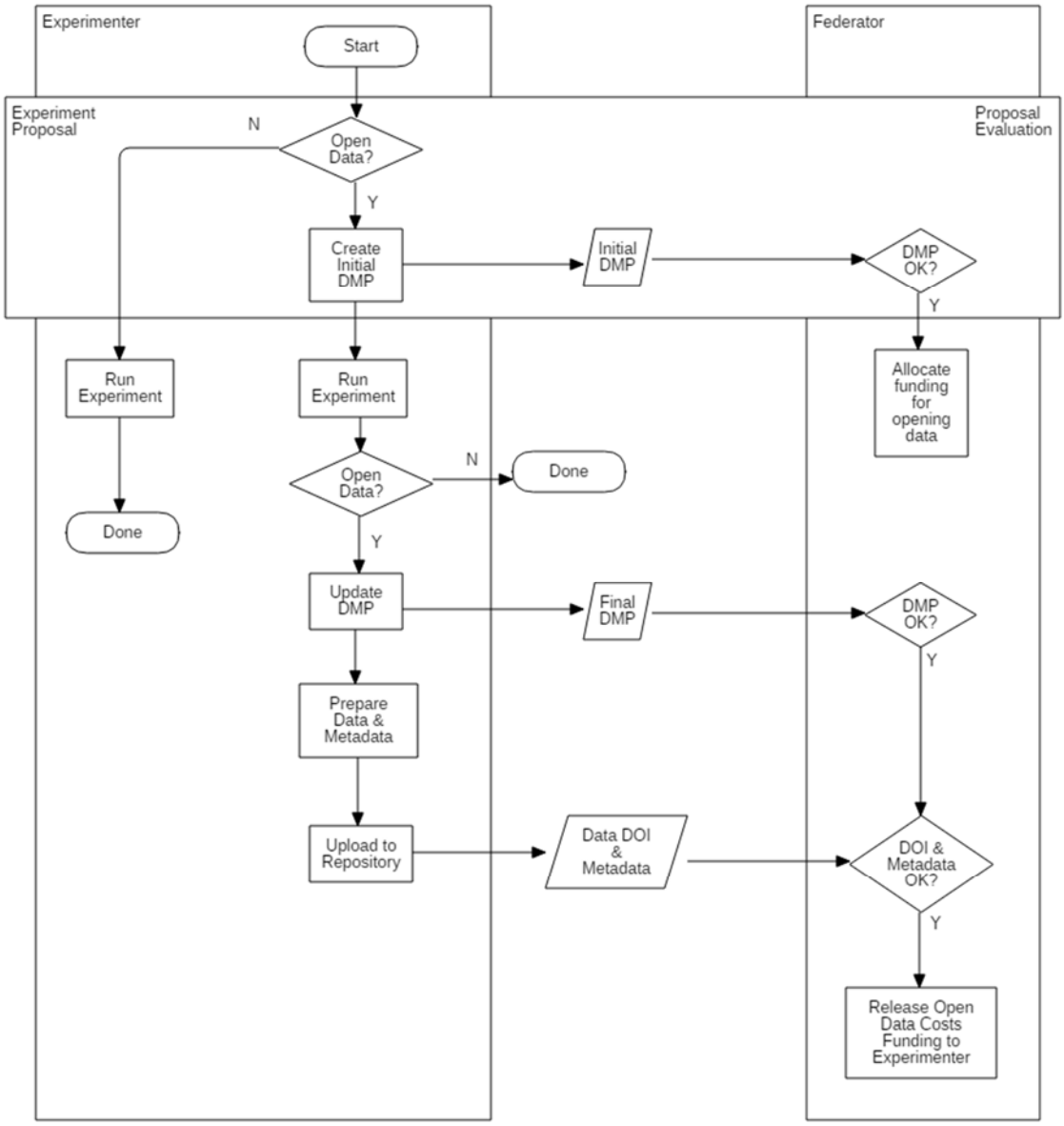

Figure 2: Open Data Process

During their preparation of a proposal in response to a Fed4FIRE+ open call, the experimenter makes a decision in principle whether they want to open data23. If they know from the outset that they cannot or will not open data, they can notify the Federator giving valid reasons. The experimenter then just runs their experiment.

[23] It should be noted that the process described in Figure 2 is a separate (but related) decision to the decision to grant funding to support the experiment itself. The process described here is solely concerned with opening data

Co-funded by the Horizon 2020 Framework Programme of the European Union

If the experimenter agrees in principle to open data (understanding that they are not obliged to and can opt out at a later stage) they fill in an initial DMP, containing the appropriate information as described in Section 2.4, and submit this to the Federator as part of the proposal evaluation. If the Federator approves the DMP (i.e. simple checks to determine whether all the key fields are filled in sensibly), the Federator allocates the funding for opening data. The Federator does not give the funding to the experimenter at this stage, just allocates it to the experiment.

The experimenter runs their experiment, and when the experiment is complete, the experimenter has another opportunity to open the result data. If they decide not, they notify the Federator giving valid reasons. In this case the experimenter does not get the funding for opening data so the Federator can release it back to fund other experiments.

If the experimenter decides to open their result data, they update their DMP to make a final version, which they submit to the Federator. The experimenter prepares the data package containing their result data and associated metadata, and uploads this to a repository. The experimenter submits the metadata and the DOI that is issued by the repository to the Federator, who check the DMP and the DOI & metadata. If all are in order, the Federator releases the funding to the experimenter. If there are any problems, the Federator will notify the experimenter, who can make the necessary adjustments.

and whether to supply funds to support this. The decision to support the experiment itself is outside the scope of this task and deliverable.

Co-funded by the Horizon 2020
Framework Programme of the European Union

3 RISK & COST MANAGEMENT FRAMEWORK

The open sharing of research data poses certain risks for different stakeholders in the federation, and the risks need to be identified and mitigated in order to provide confidence that open sharing of data is not damaging.

To identify the risks and appropriate mitigation approaches, we need to determine the answers to the following questions:

- Who are the key stakeholders in the federation?
- What are the key assets of each stakeholder?
- What are the realistic scenarios of data sharing?
- What are the interests of the stakeholders and are there any conflicts of interest between the stakeholders?
- What are the risks to each stakeholder (and their assets) in each scenario?
- How can the risks be mitigated?

This section contains an initial risk analysis for the situation of experimenters opening their data and placing it in a repository for long term storage and availability to other users. It builds on a stakeholder and risk analysis from the Fed4FIRE project [Taylor 2016], which is directly relevant to the federated testbed situation of Fed4FIRE+.

The key stakeholders in the federation are the three major federation participants (experimenters, testbeds & federator), plus any data repository owners and potential users of experiment data.

- Experimenters. These are the owners of the experiment data generated as a result of the experimenters running experiments on testbeds.
- Testbed owners. The testbeds upon which an experiment is run are the places where the data is initially generated.
- Federator. The federator is responsible for setting up the experiment, and has an overall responsibility for operating the federation. As such, the federator is not directly connected with the experiment data, but has an interest in maintaining the integrity and reputation of the federation, which includes the protection of experimenters' data.
- Repository owners. This is the owner of the repository where the data is stored. They have the responsibility of protecting the data and making it available in the long term.
- Users of data. This is any potential user of the data. They wish to find data that is relevant to their purposes and need to trust that the data is faithful to that originally created by the experimenter.

The table below shows an initial analysis of the stakeholders associated with Fed4FIRE+, their interests in open research data and the risks they face. The risks are identified with tags, so that they may be referenced later in the section on how they may be mitigated.

- RE – risks to experimenters
- RT – risks to testbeds
- RF – risks to the federator
- RR – risks to the repositories holding Fed4FIRE+ experiment data
- RU – risks to potential users of Fed4FIRE+ experiment data

Co-funded by the Horizon 2020 Framework Programme of the European Union

Stake-holder	Experimenters	Testbed Owners	Federator	Repository Owner	Data Users
Assets	• Data from experiments • Integrity as experimenters	• Testbed Reputation • Testbed Resources	• Federation Reputation	• Reputation of repository • Storage Infrastructure	
Interests	• Run experiments that provide valuable results to them • Exploitation of the experiment results	• Enhanced reputation as a result of providing a good service	• Preserve and enhance Federation's reputation • Keep the federation relevant • Recruit new experimenters and testbeds • Encourage experimenters open research data	• Promote open science by hosting open research data	• Find relevant & useful open research data
Risks	• RE1: Compromise of Data • RE2: Loss of Data • RE3: Data is not findable • RE4: Data is not accessible	• RT1: Data is lost or compromised whilst in the testbed domain	• RF1: Data compromise brings the federation into disrepute	• RR1: Unauthorised access to / alteration of / deletion of data stored in repository	• RU1: Data is not authentic or has been altered

Table 3: Fed4FIRE+ Stakeholders, Assets and Risks

Motivating scenarios can be taken from Borgman [Borgman 2012], as described in Section **Error! Reference source not found.** above. They are repeated here:

(1) to reproduce or to verify research,
(2) to make results of publicly funded research available to the public,
(3) to enable others to ask new questions of extant data, and
(4) to advance the state of research and innovation.

The practical scenario for opening data is to place it in a repository that makes it publicly available and findable for others to use.

Co-funded by the Horizon 2020 Framework Programme of the European Union

3.1 RISK MITIGATION

By far the majority of the risks are associated with data loss or compromise, either by accident, due to decay or intentional misbehaviour. The risk is addressed from the different perspectives of the stakeholders, and the major mitigating factor is the interest each stakeholder has in protecting their reputation through taking reasonable steps to protect the data from unauthorised access or decay.

3.1.1 RE1: Compromise of Data & RE2: Loss of Data

Protection from data compromise and loss is mostly handled by the repositories. The repositories have a vested interest in preserving their reputation as trusted stores of data who protect its integrity. It is up to Fed4FIRE+ help experimenters choose repositories who have a strong probability of protecting their data, and this is one of the guiding principles of the repository requirements in Section 2.3.7.1.

The requirements in Section 2.3.7.1 and the subsequent analysis of repositories in Section 2.3.7.2 illustrate the risks and how the repositories address them, for example replication, checksums and digital signatures.

3.1.2 RE3: Data is not findable

The experimenter must supply adequate metadata (e.g. title, creator, keywords etc) to describe their data, and the repository must make this metadata available in standardised formats to open research data search engines so that other users can find it.

The metadata must accurately reflect the data, otherwise it does not stand a chance of being found. This is the responsibility of the experimenter, although the repository may evaluate the metadata and suggest additional keywords to widen the chances of the data being found, for example.

The data repository must export the metadata of the data it holds to open research data search engines so that the data may be found by users using the search engines.

3.1.3 RE4: Data is not accessible

This risk can take a number of forms, from temporary inaccessibility due to e.g. server outages, or permanent inaccessibility due to the data provider going out of business or closing down the data repository.

The repository must provide policies that determine the accessibility of data, especially in the event that the repository is closed. Usually this will take the form of a migration plan to other repositories.

3.1.4 RT1: Data is lost or compromised whilst in the testbed domain

This risk is similar to RE1 and RE2, but from the perspective of the testbed. The testbed will suffer major reputation damage if it becomes known that experiment data was lost or compromised while in their charge, so the testbed has a clear interest in protecting the data as much as possible.

Co-funded by the Horizon 2020 Framework Programme of the European Union

3.1.5 RF1: Data compromise brings the federation into disrepute

This risk is from the perspective of the federator, the federation manager. The federator needs to make clear to all service providing stakeholders in the federation that protection of experiment data is important, and that protecting the data is in each of their interest (they should understand this already). There may be terms and conditions that govern the service providers (e.g. the testbed operators) that enforce the policy of protection of experimental data.

Some data repositories are outside the federation, and here the mitigation strategy is to ensure that the repositories chosen are robust and reliable, and that they have their own interest in protecting data, covered next.

3.1.6 RR1: Unauthorised access to / alteration of / deletion of data stored in repository

This risk is the impact on a repository of loss or compromise of data while it is in their charge, mainly reputation damage. Most reputable repositories have taken steps to address this, and the requirements in Section 2.3.7.1 determine the criteria for selecting external repositories for Fed4FIRE+.

3.1.7 RU1: Data is not authentic or has been altered

This risk is the impact on the potential user of data should that data not authentic. A user may download a dataset, believing it to be accurate and genuine, and base their interpretation or experiment on it. If that data is inaccurate, then the user's results will also be inaccurate.

The user needs to operate traditional scientific scepticism (which should be standard practice) to verify the data that they get from data repositories and understand that the data is supplied at their own risk.

The user should have reasonable expectations that the provider of any data will have taken reasonable steps to protect the data whilst in their charge, and this has been covered in previous risks.

3.2 COSTS

The costs of data storage will depend significantly on the choice of repository. From the analysis above, Zenodo provides storage (up to a limit) free of charge, and Dryad has a flat fee of $120. The costs therefore are mainly concerned with the preparation of the data for open access. This is mitigated from the experimenter's point of view by the key principle of encouraging experimenters to open data by subsidising their costs to prepare the data for open access, over and above the funding the experimenters get for running their experiment.

It is an item of further wok to determine if any limits can be placed the federation's support for these costs. It is expected that there will need to be limits, because the Federator needs to allocate budget to cover opening data costs for experimenters.

Co-funded by the Horizon 2020 Framework Programme of the European Union

4 RELEVANT DATA PROTECTION REGULATIONS FOR FED4FIRE+

Following Article 16 of the Treaty on the Functioning of the European Union, which is the legal basis for the adoption of data protection rules in the EU, the European Union legislator adopted the Regulation 679/2016 (hereinafter "GDPR") to protect the fundamental right to data protection and to guarantee the free flow of personal data between Member States. The logic of the adoption of a GDPR was to prevent disparities between Member States in terms of procedures and sanctions, harmonizing the data protection in the EU.

This Regulation is complemented by Directive 2002/58/EC on privacy and electronic communications (ePrivacy Directive), which concerns the processing of personal data and the protection of privacy in the electronic communications sector and states specific requirements concerning the protection of personal data and privacy of the users of electronic communication services. In the context of Fed4FIRE+, the following principles are of relevance:

4.1 KEY GDPR PRINCIPLES

(1) The GDPR applies to the **processing of personal data wholly or partly by automated means** and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.

(2) The GDPR have an **extra-territorial reach**, meaning that its rules apply not only to controllers or processors established in the European Union, but also to entities having their establishment in a third country, if they:

a) Offer goods or services, irrespective of whether a payment of the data subject is required, to data subjects in the Union (e.g. a US-based social network); or

b) Monitor the data subjects' behaviour, as far as their behaviour takes place within the Union (e.g. email tracking service providers)

(3) **Personal data cannot be processed without a legal ground**. This usually entails that the data subject has to give his/her **consent** to the processing of his or her personal data for one or more specific purposes; however, different legal grounds may apply, in different instances, which could exempt controllers or processors from collecting the data subject's consent. This holds true when personal data processing:

a) is necessary for the performance of a **contract** to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract (e.g. when transferring connected cars' data to an external provider of maintenance services, as agreed with the car's owner through a contract);

b) is necessary for **compliance with a legal obligation** to which the controller is subject (e.g. a Union, national or regional law setting out rules and obligations for cities within smart cities' programs);

c) is necessary in order to protect the vital interests of the data subject or of another natural person (e.g. when deploying IoT devices for emergency health care purposes);

d) processing is necessary for the performance of a **task carried out in the public interest** or **in the exercise of official authority** vested in the controller (e.g. when personal data processing is necessary to manage a tax system);

Co-funded by the Horizon 2020 Framework Programme of the European Union

e) processing is necessary for the purposes of the **legitimate interests** pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child (the discipline of the legitimate interest still vary across EU Member States and needs a case by case assessment).

(4) Consent should be free, unambiguous, informed, prior and demonstrable by the data controller, meaning that it must be documented somehow (also electronically, e.g. by means of a log).

(5) In any event, data subjects must be informed about the processing undergone by their personal data before the processing starts or, when data are not collected from the data subjects themselves, within a reasonable period, in any event no later than the first communication or the first disclosure to the public, when such activities are foreseen (e.g. in a smart city context, by complete information notices published on the cities' websites, by icons displayed on the users' devices, by signs on the street in correspondence of IoT sensors or cameras).

(6) Data protection principles (*i.e.* data minimization, purpose limitation, data accuracy, storage limitation etc.) must always be respected; a data controller may have a legal ground to process personal data (e.g. the data subject's consent), yet it may still run the processing in breach of one of the key data protection principles, which would make the personal data processing unlawful and, potentially, trigger a sanction by competent authorities. This is the essence of the principle of accountability.

(7) The principle of data protection-by-design is now set in law. It requires the controller to implement "technical and organizational measures appropriate to the processing activity being carried out and its objectives, such as data minimization and pseudonymisation, in such a way that the processing will meet the requirements of [the] Regulation and protect the rights of (…) data subjects";

(8) Same goes for the principle of data protection-by-default, that refers to the amount of data collected, retention period, extent of the processing, data accessibility etc. Essentially, "*the controller shall implement appropriate measures for ensuring that, by default, only (…) personal data (…) which are necessary for each specific purpose of the processing are processed*".

(9) Clear procedures must be in place to ensure data subjects' rights, namely:

a) Right of access;

b) Right to rectification;

c) Right to erasure;

d) Right to restriction;

e) Right to data portability

f) Right to object

(10) Procedures to handle and notify **Data Breaches** to Data Protection Authorities and Data Subjects concerned must be in place.

(11) Stakeholders must delete **raw data** as soon as they have extracted the data required for their data processing.

Co-funded by the Horizon 2020 Framework Programme of the European Union

4.2 KEY EPRIVACY DIRECTIVE PRINCIPLES

(1) Where the ePrivacy Directive provides for a specific rule applicable to natural and legal persons in relation to processing in connection with the provision of publicly available electronic communications services in public communication networks, it prevails over the general rule set out by the GDPR ("*Lex Specialis derogat generali*"- **Principle of Specialty**);

(2) Electronic Communication Services and Networks must be secured through appropriate technical and organizational measures **(Security**);

(3) The confidentiality of communications and the related traffic data by means of a public communications network and publicly available electronic communications services, must be ensured **(Confidentiality);**

(4) Access to, or storage of, information into the users' devices must be authorized by the users with a specific consent, unless it is "*strictly necessary in order to provide a service explicitly requested by the subscriber or user*" (also known as "cookie law", **Prior Consent**);

4.3 INITIAL EXAMINATION OF PRIVACY REQUIREMENTS

The design of the system architecture is a crucial phase to ensure the security and privacy of the information processed therein. In fact, according to the GDPR, "*the controller should adopt **internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default**. Such measures could consist, inter alia, of **minimising the processing of personal data**, **pseudonymising personal data as soon as possible**, **transparency with regard to the functions and processing of personal data**, **enabling the data subject to monitor the data processing, enabling the controller to create and improve security features**. When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations*".

Moreover, the system must be embedded with **appropriate technical and organizational measures to ensure a level of security appropriate to the risk**, including inter alia as appropriate:

- the pseudonymisation and encryption of personal data;
- the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;
- the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;
- a process for regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing;
- **policies and procedures to periodically test the security** resilience of a system (e.g., penetration tests, vulnerability assessments, etc.) and carry out the relevant remediation activities;
- a well-defined internal procedure to alert the system administrators when any **data breaches take place.**

Co-funded by the Horizon 2020 Framework Programme of the European Union

Fed4FIRE+ aims to ensure the systematic adaptation and improvement of its experimental framework to fully comply with the personal data protection principles. Following an examination of these principles, the following Privacy Requirements (“PR”) are drawn from the GDPR amongst those relevant for Fed4FIRE+. These requirements are adapted to the foreseen architecture, based on the assumption that Fed4FIRE+ would not be focused on performing personal data processing activities, and would only require such data to authenticate users and provide the related identity management and single sign-on services.

Co-funded by the Horizon 2020
Framework Programme of the European Union

4.4 RELEVANT DATA PROTECTION REQUIREMENTS

- **Project data management:** The system must automatically record all internally generated data, storing these data into the Fed4FIRE+ platform, while minimizing the collection of personal data.
- **Data back-ups:** Back-up operations will be carried out periodically, so as to ensure the continuity of the system and prevent the loss of data.
- **Authentication of identities:** The whole system will collect different types of data and it will be designed to ensure the privacy and trust of the users. In order to do this, each identity accessing the system will be authenticated and appropriately authorised to be able to use it. Where necessary (e.g. when the system is used to process health data), strong authentication (e.g. two-factor authentication, double opt-in, biometric recognition, etc.) methods must be supported.
- **De-activation of authentication credentials:** Personal authentication credentials shall be de-activated if they have not been used for at least six months (except in case of technical authorization).
- **Purpose limitation:** As set out by article 5 of the GDPR, Fed4FIRE+ will process personal data only for security purposes, unless the data controller configures the system to pursue other legitimate, specific and explicit purposes, determined at the time of collection of the data.
- **Data accuracy and updating:** Personal data which are inaccurate or incomplete, having regard to the purposes for which they were collected or processed, will be erased or rectified as set out by article 5 of the GDPR.
- **Security of processing:** Fed4FIRE+ will protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access through the implementation of technical and organisational measures as required by article 32 of the GDPR.
- **Data breach information:** The Fed4FIRE+ system must immediately inform its users of any breach to personal data leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed as required by articles 33 and 34 of the GDPR.
- **Encryption by default:** As provided by Article 32 of the GDPR, encryption will be applied to all stages of handling data, including in communication, storage of data at rest, storage of keys, identification, access, as well as for secure boot process.
- **Right of access:** The Fed4FIRE+ system shall support the data controllers in providing to every data subject, without excessive delay or expense, confirmation as to whether or not data relating to him/her are being processed and information as to: the purposes of the processing; the categories of data concerned; the recipients to whom the data are disclosed; the envisaged period of storage for the data; and the existence of automated decision-making processes within the system. The legal source of this requirement is article 15 of the GDPR.
- **Right of erasure:** The Fed4FIRE+ platform must ensure that the right of erasure exercised by data subjects towards the data controller is enforced, when the conditions set out by article 17 of the GDPR are met.
- **Data portability:** As detailed by article 20 of the GDPR, the Fed4FIRE+ platform must be able to support the data controller in responding to requests for data portability lodged by the data subjects. This entails that the data subject shall receive the data in a structured, commonly used and machine-readable format.
- **Regular monitoring of security:** The Fed4FIRE+ platform will regularly monitor the system's status in terms of security for personal data as required by article 32 of the GDPR.

Co-funded by the Horizon 2020 Framework Programme of the European Union

4.5 RELEVANT PRIVACY RISKS AND RISK MITIGATION MEASURES

Fed4FIRE+ will not, in principle, use personal data in any of the foreseen experimentations, and personal data will only be obtained and processed as necessary to enable the authentication of users and the provision of the required identity management and single sign-on services. Since a number of privacy risks can still be identified despite this reduced level of processing, several mitigation measures must be implemented to tackle each risk, namely:

Privacy Risk	Recommended Mitigation Measures	
Missing end-to-end encryption: The traffic is transmitted without being encrypted	Adoption of encryption by default and by design.	
Use of unsecure of obsolete cypher suites: The traffic is encrypted, but the encryption methods have known vulnerabilities	Permanent review of the encryption methods.	
Man-in-the-middle attacks: Circumvention of mutual authentication between client and server by an attacker	Implementation of strong authentication and tamper detection mechanisms.	
Missing transparency of service storage method: Data stored within a third-party-service may be leaked because the user has no control over storage security	Avoid use of third-party-services.	
Security vulnerabilities in service backend: The backend deployed by a service provider may be susceptible to security vulnerabilities	Scheduled and continuous testing and updating of all backend elements.	
Traffic analysis: Information is leaked and exploited through passive eavesdropping and analysis of encrypted transmission	Encryption of both metadata and packet data and routing under trusted third parties.	
DNS request leakage: Secure DNS is not used and DNS requests are visible to everyone	Adoption of Secure DNS by default and by design.	

Co-funded by the Horizon 2020 Framework Programme of the European Union

4.6 GENERAL OUTLINE OF THE STRATEGY

In the context of the efforts towards ensuring Fed4FIRE+ fully complies with the relevant data protection regulations described below, the following general strategy will be pursued through its upcoming phases. Firstly, MI will adopt a vigilant role throughout the initial phases to ensure that personal data collection is minimized. Once the architecture is stabilized, a Privacy Impact Assessment will be performed, and once completed each identified risk will be addressed with adequate mitigation measures; this approach will be further complemented with the regular monitoring of the infrastructure by the partners and Personal Data Officer of each testbed. Finally, third party end-users of the platform will be invited to assess the personal data protection and the platform services through the diverse stages of the process.

Co-funded by the Horizon 2020
Framework Programme of the European Union

5 CONCLUSIONS: KEY RECOMMENDATIONS

This section concludes the deliverable by summarising its key recommendations. These are divided into two major sections: open research data and GDPR.

5.1 OPEN RESEARCH DATA

This section contains a proposal that will be submitted to the Fed4FIRE+ Federation Board for discussion and approval. It follows the form where the major points below are the key principles, and the minor points for each principle indicate the proposal to address it.

- Encourage experimenters to be as open as possible but do not prevent them from closing data.
  - To encourage opening data, experimenters' costs for opening data are covered by the project (up to limits TBD). Costs are payable at the end of the experiment, once the experimenter has opened their data.
  - Experimenters can elect to close data at any point before publication, citing valid reasons. If they do this, they will not be eligible for cost claims associated with opening experiment data.
- Experimenters must understand the implications of opening data at the outset of their experiment
  - Experimenters who declare an intention to open data should create an initial DMP. This is a simple version of the DMP and not intended to be time consuming.
  - At the end of the experiment, the experimenter should fill in a final DMP. This is a more detailed version of the initial DMP that describes the data from the experiment and how it is stored.
  - Payment for opening data will contingent on both DMPs being filled in, as well as the data being placed in an open repository.
- Fed4FIRE+ should provide an approved set of data repositories that have been evaluated for suitability, and specify the requirements for a Fed4FIRE+ approved repository.
  - The key requirements for the repositories broadly correspond to support for the FAIR principles:
    - Digital Object Identifiers (DOIs) to uniquely identify data
    - The repository needs to provide evidence that it is likely to be there for the long term.
    - Data integrity needs to be protected. The repository needs to provide evidence as to how it will protect the data it stores from compromise or loss.
    - The repository needs to export the metadata of data it stores, so it can be indexed by the popular open data search engines
    - The repository needs to be flexible in the license terms it uses for data it hosts, so that experimenters (who actually own the data) can choose a license.
  - It is recommended that Zenodo be approved by Fed4FIRE+, and can serve as the default data repository.

Co-funded by the Horizon 2020 Framework Programme of the European Union

5.2 GENERAL DATA PROTECTION REGULATION

This section contains recommendations to the Fed4FIRE+ board concerning support for GDPR, in the form of the requirements in Section 4.4. These requirements concern the authentication information of users.Fed4FIRE+ will not, in principle, use personal data in any of the foreseen experimentations, and personal data will only be obtained and processed as necessary to enable the authentication of users and the provision of the required identity management and single sign-on services.

- **Project data management:** The system must automatically record all internally generated data, storing these data into the Fed4FIRE+ platform, while minimizing the collection of personal data.
- **Data back-ups:** Back-up operations will be carried out periodically, so as to ensure the continuity of the system and prevent the loss of data.
- **Authentication of identities:** The whole system will collect different types of data and it will be designed to ensure the privacy and trust of the users. In order to do this, each identity accessing the system will be authenticated and appropriately authorised to be able to use it. Where necessary (e.g. when the system is used to process health data), strong authentication (e.g. two-factor authentication, double opt-in, biometric recognition, etc.) methods must be supported.
- **De-activation of authentication credentials:** Personal authentication credentials shall be de-activated if they have not been used for at least six months (except in case of technical authorization).
- **Purpose limitation:** As set out by article 5 of the GDPR, Fed4FIRE+ will process personal data only for security purposes, unless the data controller configures the system to pursue other legitimate, specific and explicit purposes, determined at the time of collection of the data.
- **Data accuracy and updating:** Personal data which are inaccurate or incomplete, having regard to the purposes for which they were collected or processed, will be erased or rectified as set out by article 5 of the GDPR.
- **Security of processing:** Fed4FIRE+ will protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access through the implementation of technical and organisational measures as required by article 32 of the GDPR.
- **Data breach information:** The Fed4FIRE+ system must immediately inform its users of any breach to personal data leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed as required by articles 33 and 34 of the GDPR.
- **Encryption by default:** As provided by Article 32 of the GDPR, encryption will be applied to all stages of handling data, including in communication, storage of data at rest, storage of keys, identification, access, as well as for secure boot process.
- **Right of access:** The Fed4FIRE+ system shall support the data controllers in providing to every data subject, without excessive delay or expense, confirmation as to whether or not data relating to him/her are being processed and information as to: the purposes of the processing; the categories of data concerned; the recipients to whom the data are disclosed; the envisaged period of storage for the data; and the existence of automated decision-making processes within the system. The legal source of this requirement is article 15 of the GDPR.

Co-funded by the Horizon 2020 Framework Programme of the European Union

- **Right of erasure:** The Fed4FIRE+ platform must ensure that the right of erasure exercised by data subjects towards the data controller is enforced, when the conditions set out by article 17 of the GDPR are met.
- **Data portability:** As detailed by article 20 of the GDPR, the Fed4FIRE+ platform must be able to support the data controller in responding to requests for data portability lodged by the data subjects. This entails that the data subject shall receive the data in a structured, commonly used and machine-readable format.
- **Regular monitoring of security:** The Fed4FIRE+ platform will regularly monitor the system's status in terms of security for personal data as required by article 32 of the GDPR.

Co-funded by the Horizon 2020
Framework Programme of the European Union

6 REFERENCES

[Ball 2014] Ball, A. (2014) How to License Research Data. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf

[Borgman 2012] Borgman, C.L., 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), pp.1059-1078.

[CERIF] http://www.eurocris.org/cerif/main-features-cerif

[Creative Commons] Creative Commons: About The Licenses, https://creativecommons.org/licenses/, retrieved 2017-02-13.

[DataCite] DataCite https://www.datacite.org/

[DataCite About] About DataCite. https://www.datacite.org/assets/datacite.pdf. Retrieved 2017-05-23.

[DataCite Metadata Kernel_v4.0] DataCite Metadata Working Group. (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. DataCite e.V.http://doi.org/10.5438/0012.

[DataCite Tech] DataCite Technical Introduction. https://www.datacite.org/assets/tech_intro.pdf. Retrieved 2017-05-23.

[Data Packages] Data Packages, http://frictionlessdata.io/data-packages/

[DCAT] Data Catalog Volcabulary (DCAT) W3C Recommendation 16 January 2014, https://www.w3.org/TR/vocab-dcat/

[DCC] Digital Curation Centre (DCC), http://www.dcc.ac.uk/

[DCES 1.1] Dublin Core Metadata Element Set, Version 1.1 http://dublincore.org/documents/dces/

[Dechamp 2016] Jean-François Dechamp, Open by default: the challenges of research data in Europe, European Commission, Directorate-General for Research & Innovation, @OpenAccessEC, 3rd LEARN Workshop, 28 June 2016, Helsinki

[DOI] Digital Object Identifier, https://www.doi.org/index.html

[Donnelly 2011] Donnelly, Martin & Jones, Sarah. (2011) DCC Checklist for a Data, Management Plan v3.0. Retrieved 18 May 2017, from http://www.dcc.ac.uk/webfm_send/431

[Dryad] The Dryad Digital Repository. http://datadryad.org/. Retrieved 2017-06-08.

[Dryad Policies] Dryad Digital Repository – Policies. http://datadryad.org/pages/policies/. Retrieved 2017-06-09.

[Dublin Core] The Dublin Core Metadata Initiative (DCMI). http://dublincore.org/specifications/

[EC Science 2.0 2014] European Commission, Validating the results of the Public Consultation on Science 2.0: Science in Transition, 2014, https://ec.europa.eu/research/consultations/science-2.0/science_2_0_final_report.pdf

[FreeBSD] The FreeBSD Copyright. https://www.freebsd.org/copyright/freebsd-license.html

[GPL v3] GNU General Public License, version 3.0. https://www.gnu.org/licenses/gpl-3.0.en.html

[Hanson 2011] Hanson, Brooks, Andrew Sugden, and Bruce Alberts. "Making data maximally available." Science 331, no. 6018 (2011): 649-649.

Co-funded by the Horizon 2020 Framework Programme of the European Union

[HEFCE et al 2016] Higher Education Funding Council for England, Research Councils UK, Universities UK, Wellcome Trust, Concordat on Open Research Data, 28 July 2016, http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/

[Hey 2009] Hey, Tony, Stewart Tansley, and Kristin M. Tolle. The fourth paradigm: data-intensive scientific discovery. Vol. 1. Redmond, WA: Microsoft research, 2009.

[H2020 ORD 2016a] H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[H2020 ORD 2016b] Guidelines on FAIR Data Management in Horizon 2020, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, retrieved 7 Feb 2017.

[ISO 15836-1:2017] [ISO 15836-1:2017. The Dublin Core metadata element set. https://www.iso.org/standard/71339.html

[ISO 26324: 2012] ISO 26324:2012(en) Information and documentation — Digital object identifier system. https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en

[Metadata Standards Directory WG] Metadata Standards Directory Working Group http://rd-alliance.github.io/metadata-directory/standards/

[National Science Board 2005] National Science Board. (2005). Long-lived digital data collections. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

[Nosek 2015] Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G. and Contestabile, M., 2015. Promoting an open research culture. Science, 348(6242), pp.1422-1425.

[OAI-ORE] Open Archives Initiative Object Reuse and Exchange - ORE Specifications and User Guides. http://www.openarchives.org/ore/1.0/toc.

[OAI-PMH] The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14, Document Version 2015-01-08. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm. Retrieved 2017-06-08.

[OECD 2015a], "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/5jrs2f963zs1-en

[OECD 2015b] Carthage Smith and Giulia Ajmone Marsan, Directorate for Science, Technology and Innovation, OECD Open Science – the Policy Challenges, https://jipsti.jst.go.jp/rda/common/data/pdf/lecture/Smith_Symposium.pdf

[OGC] Open Geospatial Consortium - Observations and Measurements http://www.opengeospatial.org/standards/om

[OpenAIRE] The OpenAIRE open access network, https://www.openaire.eu, retrieved 2017-02-08.

[OpenDOAR] The Directory of Open Access Repositories. http://www.opendoar.org/. Retrieved 2017-06-08.

[Open Definition] The Open Definition http://opendefinition.org/, retrieved 2017-05-23.

[Open Definition Licenses] Open Definition - Conformant Licenses http://opendefinition.org/licenses/

[Paskin 2010] Paskin, Norman. "Digital object identifier (DOI) system." Encyclopedia of library and information sciences 3 (2010): 1586-1592.

Co-funded by the Horizon 2020 Framework Programme of the European Union

[PREMIS] PREMIS Data Dictionary for Preservation Metadata - http://www.loc.gov/standards/premis/

[PREMIS Dict] The PREMIS Data Dictionary (full document), Version 3.0. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[PREMIS Figs] http://www.loc.gov/standards/premis/v3/premis-3-0-figures.pdf

[RFC 2119] Key words for use in RFCs to Indicate Requirement Levels. https://www.ietf.org/rfc/rfc2119.txt. Retrieved 2017-06-08.

[Re3Data] Re3Data.org, Registry of Research Data Repositories, http://www.re3data.org/, retrieved 2017-02-08

[Royal Society Science 2.0 2014] Royal Society - Consultation response: Science 2.0, https://royalsociety.org/topics-policy/publications/2014/consultation-response-science-2-0/

[Russell 2012] Russell, R., 2012. Adoption of CERIF in higher education institutions in the UK: A landscape study.

[Taylor 2016] Taylor, Stephen and Boniface, Michael (2016) Techniques for increasing trust in federated experimental platforms. At European Conference on Networks and Communications 2016, Athens, GR, 27 - 30 Jun 2016.

[Wilkinson 2016] 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 (2016). http://dx.doi.org/10.1038/sdata.2016.18

[Zenodo] The Zenodo open research data repository, https://www.zenodo.org/, retrieved 2017-02-08.

[Zenodo FAQ] The Zenodo FAQ. http://help.zenodo.org/. Retrieved 2017-06-09.

[Zenodo Policies] Zenodo Polices. http://about.zenodo.org/policies/. Retrieved 2017-06-09.

[Zenodo Principles] Zenodo Principles http://about.zenodo.org/principles/, retrieved 2017-05-23.

[Zenodo Terms] Zenodo Terms of Use v1.0. http://about.zenodo.org/terms/. Retrieved 2017-06-09.

Co-funded by the Horizon 2020 Framework Programme of the European Union

7 BIBLIOGRAPHY

[Ball 2012] Alex Ball. (2012). Minimum Mandatory Metadata Set for RAIDmap (version 1.0). REDmMED Project Document redm1rep111124ab10. Bath, UK: University of Bath. https://core.ac.uk/download/pdf/8795253.pdf

[Ball 2014] Ball, A. (2014) How to License Research Data. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf

[Ball 2015] Ball, A., & Duke, M. (2015). 'How to Cite Datasets and Link to Publications'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

[Boniface 2013] Boniface M., Inglesant P., Papay J. (2013) Counting the Cost of FIRE. In: Galis A., Gavras A. (eds) The Future Internet. FIA 2013. Lecture Notes in Computer Science, vol 7858. Springer, Berlin, Heidelberg

[Borgman 2012] Borgman, C.L., 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), pp.1059-1078.

[BPMN 2011] Object Management Group, "Business Process Model and Notation (BPMN) Version 2.0," http://www.omg.org/spec/BPMN/2.0/PDF/ , retrieved 2017-02-13.

[Cerf 2015] Google's Vint Cerf warns of 'digital Dark Age', 13 February 2015, http://www.bbc.co.uk/news/science-environment-31450389

[CERIF] http://www.eurocris.org/cerif/main-features-cerif

[Creative Commons] Creative Commons: About The Licenses, https://creativecommons.org/licenses/, retrieved 2017-02-13.

[DataCite] DataCite https://www.datacite.org/

[DataCite About] About DataCite. https://www.datacite.org/assets/datacite.pdf. Retrieved 2017-05-23.

[DataCite Metadata Kernel_v4.0] DataCite Metadata Working Group. (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. DataCite e.V.http://doi.org/10.5438/0012.

[DataCite Tech] DataCite Technical Introduction. https://www.datacite.org/assets/tech_intro.pdf. Retrieved 2017-05-23.

[Data Packages] Data Packages, http://frictionlessdata.io/data-packages/

[DCAT] Data Catalog Volcabulary (DCAT) W3C Recommendation 16 January 2014, https://www.w3.org/TR/vocab-dcat/

[DCC] Digital Curation Centre (DCC), http://www.dcc.ac.uk/

[DCES 1.1] Dublin Core Metadata Element Set, Version 1.1 http://dublincore.org/documents/dces/

[Dechamp 2016] Jean-François Dechamp, Open by default: the challenges of research data in Europe, European Commission, Directorate-General for Research & Innovation, @OpenAccessEC, 3rd LEARN Workshop, 28 June 2016, Helsinki

[DOI] Digital Object Identifier, https://www.doi.org/index.html

[Dryad] The Dryad Digital Repository. http://datadryad.org/. Retrieved 2017-06-08.

Co-funded by the Horizon 2020 Framework Programme of the European Union

[Dryad Policies] Dryad Digital Repository – Policies. http://datadryad.org/pages/policies/. Retrieved 2017-06-09.

[Dublin Core] The Dublin Core Metadata Initiative (DCMI). http://dublincore.org/specifications/

[EC DG RI 2016] European Commission, Directorate-General for Research and Innovation, Open innovation, open science, open to the world - A vision for Europe. ISBN: 978-92-79-57346-0 DOI: 10.2777/061652. Publication year: 2016

[EC ORD Infographic 2016] European Commission, OPEN RESEARCH DATA IN HORIZON 2020 https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

[EC Open Science Cloud 2016] European Open Science Cloud - first report from the High Level Expert Group, https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[EC Science 2.0 2014] European Commission, Validating the results of the Public Consultation on Science 2.0: Science in Transition, 2014, https://ec.europa.eu/research/consultations/science-2.0/science_2_0_final_report.pdf

[Engen 2015] Engen, Vegard, Veres, Galina, Crowle, Simon, Bashevoy, Maxim, Walland, Paul and Hall-May, Martin (2015) A Semantic Risk Management Framework for Digital Audio-Visual Media Preservation. In, The Tenth International Conference on Internet and Web Applications and Services (ICIW), June 21 - 26, 2015, Brussels, Belgium, Brussels, Belgium, IARIA

[FreeBSD] The FreeBSD Copyright. https://www.freebsd.org/copyright/freebsd-license.html

[GPL v3] GNU General Public License, version 3.0. https://www.gnu.org/licenses/gpl-3.0.en.html

[Hanson 2011] Hanson, Brooks, Andrew Sugden, and Bruce Alberts. "Making data maximally available." Science 331, no. 6018 (2011): 649-649.

[HEFCE et al 2016] Higher Education Funding Council for England, Research Councils UK, Universities UK, Wellcome Trust, Concordat on Open Research Data, 28 July 2016, http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/

[Hey 2009] Hey, Tony, Stewart Tansley, and Kristin M. Tolle. The fourth paradigm: data-intensive scientific discovery. Vol. 1. Redmond, WA: Microsoft research, 2009.

[H2020 ORD 2016a] H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[H2020 ORD 2016b] Guidelines on FAIR Data Management in Horizon 2020, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, retrieved 7 Feb 2017.

[ISO 15836-1:2017] [ISO 15836-1:2017. The Dublin Core metadata element set. https://www.iso.org/standard/71339.html

[ISO 26324: 2012] ISO 26324:2012(en) Information and documentation — Digital object identifier system. https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en

[ISO 31000 2009] ISO/IEC, 31000:2009 Risk management - Principles and guidelines, ISO Std., 2009.

[Lavoie 2014] Technology Watch Report 14-02: The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition) by Brian Lavoie 2014. http://dx.doi.org/10.7207/twr14-02

Co-funded by the Horizon 2020 Framework Programme of the European Union

[Li et al 2014] Li, C. and Sugimoto, S., 2014, October. Provenance Description of Metadata using PROV with PREMIS for Long-term Use of Metadata. In International Conference on Dublin Core and Metadata Applications (pp. 147-156).

[Metadata Standards Directory WG] Metadata Standards Directory Working Group http://rd-alliance.github.io/metadata-directory/standards/

[Mitcham et al 2015] Mitcham, Jenny; Chris Awre; Julie Allinson; Richard Green; Simon Wilson (2015): Filling the Digital Preservation Gap. A Jisc Research Data Spring project. Phase One report - July 2015. figshare. https://doi.org/10.6084/m9.figshare.1481170.v1 Retrieved: 14 07, May 03, 2017 (GMT)

[National Science Board 2005] National Science Board. (2005). Long-lived digital data collections. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

[Nosek 2015] Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G. and Contestabile, M., 2015. Promoting an open research culture. Science, 348(6242), pp.1422-1425.

[OAI-ORE] Open Archives Initiative Object Reuse and Exchange - ORE Specifications and User Guides. http://www.openarchives.org/ore/1.0/toc.

[OAI-PMH] The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14, Document Version 2015-01-08. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm. Retrieved 2017-06-08.

[OAIS ISO 2012] Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model. ISO 14721:2012 https://www.iso.org/standard/57284.html

[OAIS 2012] The Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Magenta Book. Issue 2. June 2012, CCSDS 650.0-M-2. https://public.ccsds.org/Pubs/650x0m2.pdf

[OECD 2015a], "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/5jrs2f963zs1-en

[OECD 2015b] Carthage Smith and Giulia Ajmone Marsan, Directorate for Science, Technology and Innovation, OECD Open Science – the Policy Challenges, https://jipsti.jst.go.jp/rda/common/data/pdf/lecture/Smith_Symposium.pdf

[OGC] Open Geospatial Consortium - Observations and Measurements http://www.opengeospatial.org/standards/om

[OpenAIRE] The OpenAIRE open access network, https://www.openaire.eu, retrieved 2017-02-08.

[OpenAire RDM 2017] – OpenAIRE Research Data Management Briefing paper https://www.openaire.eu/briefpaper-rdm-infonoads

[OpenDOAR] The Directory of Open Access Repositories. http://www.opendoar.org/. Retrieved 2017-06-08.

[Open Definition] The Open Definition http://opendefinition.org/, retrieved 2017-05-23.

[Open Definition Licenses] Open Definition - Conformant Licenses http://opendefinition.org/licenses/

[Paskin 2010] Paskin, Norman. "Digital object identifier (DOI) system." Encyclopedia of library and information sciences 3 (2010): 1586-1592.

[PREMIS] PREMIS Data Dictionary for Preservation Metadata - http://www.loc.gov/standards/premis/

Co-funded by the Horizon 2020 Framework Programme of the European Union

[PREMIS Dict] The PREMIS Data Dictionary (full document), Version 3.0. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[PREMIS Figs] http://www.loc.gov/standards/premis/v3/premis-3-0-figures.pdf

[RFC 2119] Key words for use in RFCs to Indicate Requirement Levels. https://www.ietf.org/rfc/rfc2119.txt. Retrieved 2017-06-08.

[Re3Data] Re3Data.org, Registry of Research Data Repositories, http://www.re3data.org/, retrieved 2017-02-08

[Royal Society Science 2.0 2014] Royal Society - Consultation response: Science 2.0, https://royalsociety.org/topics-policy/publications/2014/consultation-response-science-2-0/

[Russell 2012] Russell, R., 2012. Adoption of CERIF in higher education institutions in the UK: A landscape study.

[Taylor 2016] Taylor, Stephen and Boniface, Michael (2016) Techniques for increasing trust in federated experimental platforms. At European Conference on Networks and Communications 2016, Athens, GR, 27 - 30 Jun 2016.

[Van den Eynden 2011] Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton. Managing and Sharing Data. UK Data Archive, ISBN: 1-904059-78-3, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

[Whyte 2015] Whyte, A. (2015). 'Where to keep research data: DCC checklist for evaluating data repositories' v.1.1 Edinburgh: Digital Curation Centre. Available online: www.dcc.ac.uk/resources/how-guides

[Wilkinson 2016] 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 (2016). http://dx.doi.org/10.1038/sdata.2016.18

[Zenodo] The Zenodo open research data repository, https://www.zenodo.org/, retrieved 2017-02-08.

[Zenodo FAQ] The Zenodo FAQ. http://help.zenodo.org/. Retrieved 2017-06-09.

[Zenodo Policies] Zenodo Polices. http://about.zenodo.org/policies/. Retrieved 2017-06-09.

[Zenodo Principles] Zenodo Principles http://about.zenodo.org/principles/, retrieved 2017-05-23.

[Zenodo Terms] Zenodo Terms of Use v1.0. http://about.zenodo.org/terms/. Retrieved 2017-06-09.

Co-funded by the Horizon 2020 Framework Programme of the European Union

8 APPENDIX 1 – ZENODO PRINCIPLES

These principles are retrieved from Zeondo [Zenodo Principles] and included here for reference.

8.1 BEST EFFORT PRINCIPLES

Zenodo does not sign SLAs. This is not a weakness, it is by design and marks a philosophy that we believe is most appropriate for Science. Instead, Zenodo is run by leading practitioners according to best practices.

What Science needs is inherent reliability, or more accurately demonstrated reliability based on open best practices. Furthermore the users should be able to influence these best practices. In the long-term, a service which is trusted is much more valuable than one for which assurances must be bought.

Service failure can never be undone. Enforcing an SLA means being prepared to litigate against the contract, which means compensation, frequently assessed on the basis of loss of revenue… but none of these concepts have any place or relevance in the free exchange of research results!

Living by these principles, Zenodo strives to make available architecture, implementation, practices and statistics. Please see for example the infrastructure page. We are also aiming to have these certified.

8.2 FAIR PRINCIPLES

FAIR Principles definition as referenced from: *Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: **10.1038/sdata.2016.18** (2016).*

8.2.1 To be Findable:

- **F1**: (meta)data are assigned a globally unique and persistent identifier
  - A DOI is issued to every published record on Zenodo.
- **F2**: data are described with rich metadata (defined by R1 below)
  - Zenodo's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichements.
- **F3**: metadata clearly and explicitly include the identifier of the data it describes
  - The DOI is a top-level and a mandatory field in the metadata of each record.
- **F4**: (meta)data are registered or indexed in a searchable resource
  - Metadata of each record is indexed and searchable directly in Zenodo's search engine immediately after publishing.
  - Metadata of each record is sent to DataCite servers during DOI registration and indexed there.

8.2.2 To be Accessible:

- **A1**: (meta)data are retrievable by their identifier using a standardized communications protocol
  - Metadata for individual records as well as record collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name.

Co-funded by the Horizon 2020 Framework Programme of the European Union

  - Metadata is also retrievable through the public REST API.
- **A1.1**: the protocol is open, free, and universally implementable
  - See point A1. OAI-PMH and REST are open, free and univesal protocols for information retrieval on the web.
- **A1.2**: the protocol allows for an authentication and authorization procedure, where necessary
  - Metadata are publicly accessible and licensed under public domain. No authorization is ever necessary to retrieve it.
- **A2**: metadata are accessible, even when the data are no longer available
  - Data and metadta will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
  - Metadata are stored in high-availability database servers at CERN, which are separate to the data itself.

8.2.3 To be Interoperable:

- **I1**: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML.
- **I2**: (meta)data use vocabularies that follow FAIR principles
  - For certain terms we refer to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE).
- **I3**: (meta)data include qualified references to other (meta)data
  - Each referrenced external piece of metadata is qualified by a resolvable URL.

8.2.4 To be Reusable:

- **R1**: (meta)data are richly described with a plurality of accurate and relevant attributes
  - Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and Zenodo's enrichments.
- **R1.1**: (meta)data are released with a clear and accessible data usage license
  - License is one of the mandatory terms in Zenodo's metadata, and is referring to a Open Definition license.
  - Data downloaded by the users is subject to the license specified in the metadata by the uploader.
- **R1.2**: (meta)data are associated with detailed provenance
  - All data and metadata uploaded is tracable to a registered Zenodo user.
  - Metadata can optionally describe the original authors of the published work.
- **R1.3**: (meta)data meet domain-relevant community standards
  - Zenodo is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

Co-funded by the Horizon 2020 Framework Programme of the European Union