D2.8 First Data policy recommendations

Lead Partner:	EPCC
Version:	1
Status:	Under EC review
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/document/3419

Deliverable Abstract
Building on current best practice, notably the EOSCpilot policy recommendations and the EC Expert Group report on FAIR data, we recommend 22 practical steps bridging general policy recommendations and future technical implementation of data sharing within the EOSC-hub service ecosystem.

OSC-HUB RECEIVES FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER RANT AGREEMENT NO. 777536.

COPYRIGHT NOTICE

DELIVERY SLIP

Date	***Name***	***Partner/Activity***	***Date***
From:	Rob Baxter	EPCC/WP2	21/01/2019
Moderated by:	Małgorzata Krakowian	EGI Foundation/WP1	
Reviewed by:	Alex Vermeulen	ICOS-RI	13/01/2019
Approved by:	AMB		

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author***
V0.1	17/09/2018	First outline and table of contents	R Baxter (EPCC)
V0.2	11/10/2018	First drafts of Introduction, FAIRness and Openness chapters	R Baxter
V0.3	26/10/2018	Drafts of Data Sharing and Information Governance chapters	R Baxter
V0.4	19/11/2018	Revisions after review at Amsterdam workshop	R Baxter, F Huigen (DANS), Y Chen (EGI.eu), S Varma (EMBL-EBI)
V0.5	27/11/2018	Revised section on DataTags; extensions to FAIR chapter	F Huigen, S Varma
V0.6	04/12/2018	Reviews of chapters 1-4, 6	Y Chen
V0.7	10/12/2018	Final version for internal working group review (missing Exec Summary and Appendix)	R Baxter
V0.8	13/12/2018	Completed for internal project review	R Baxter
V1	21/01/2019	Incorporated internal review feedback	R Baxter

TERMINOLOGY

https://wiki.eosc-hub.eu/display/EOSC/EOSC-hub+Glossary

Terminology/Acronym	***Definition***
Accession number	In bioinformatics, a unique identifier given to a DNA or protein sequence record in a single data repository. Used as a common form of *compact (persistent) identifier* in life science data management.
DataCite	https://datacite.org/
DataTags	A system of human-readable and machine-actionable labels that express conditions under which datasets can be stored, transmitted, or used; https://techscience.org/a/2015101601/
DOI	Digital Object Identifier, a well-recognised form of PID (qv); http://www.doi.org/
EOSCpilot	https://eoscpilot.eu/
FAIR principles	Principles of best practice in open research data management, an acronym of findability, accessibility, interoperability and reusability; https://www.force11.org/group/fairgroup/fairprinciples
FREYA	https://www.project-freya.eu/
OpenAIRE	https://www.openaire.eu/
PID	Persistent identifier, for example a DOI or accession number.
schema.org	"A collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet [and] on web pages"; https://schema.org/
Sensitive data	Data which, for whatever reason, cannot be openly shared without the risk of disclosure of legally or ethically sensitive information.

WORKING GROUP

Rob Baxter (EPCC); Ilona von Stein, Frans Huigen (DANS); Yin Chen, Yannick Legre (EGI.eu); Susheel Varma (EMBL-EBI); Serena Battaglia, Christian Ohmann (ECRIN); Michaela Th. Mayrhofer (BBMRI).

Contents

Executive summary

Facilitating access to research data is the central principle of the European Open Science Cloud, and a common policy framework for data sharing is an important ingredient in realising that principle. EOSC-hub is laying the foundations of a common layer of services for EOSC, building on preparatory work in existing e-infrastructures, research infrastructures and the EOSCpilot project.

In this report we adopt 11 key recommendations from EOSCpilot and translate them into 22 practical suggestions on the sharing of data across EOSC-hub. We cover the spectrum of potential research data, from the open and public to the highly controlled and "sensitive". These are preliminary recommendations for consideration by strategists and technical system integrators in EOSC-hub.

The recommendations fall under three broad headings:–

Implement FAIR.

We recommend a "Web first" approach to implementing the FAIR principles, with data objects published on the Web in open, non-proprietary, machine-readable formats, well described and referenced by resolvable persistent identifiers (PIDs).

We recommend widespread adoption of the current best practice of PIDs resolving to HTML landing pages which include (at least) minimal "discoverability metadata" based on the OpenAIRE and DataCite guidelines, either encoded in the page, or by content negotiation as JSON-LD metadata, following the schema.org approach.

Where data can be shared openly we recommend use of the Creative Commons 4.0 licensing scheme, in particular CC-BY, CC-BY-SA and CC0.

We recommend technical effort be invested in tracking the FREYA project; in building on the Elixir Beacon approach to sensitive metadata; and in direct retrieval of data objects by PID.

Build technical expertise in safe data and safe settings.

EOSC-hub (and EOSC more widely) provides an opportunity to lead the world on making sensitive data safely available for research. EOSC-hub should adopt the "Five Safes" principles (safe data, safe settings, safe projects, safe people and safe outputs) and work towards enabling continent-wide research that follows them.

Data objects should be tagged with a metadata tag indicating sensitivity, and managed according to the DataTags principles. Data objects flagged as non-open should not be distributed freely as a matter of course.

Working with sensitive data properly requires two things: services which can provide the necessary safe settings; and independent information governance. We recommend that EOSC-hub develop a technical design framework for safe settings in which researchers can work with sensitive data (*Safe Havens*), and work alongside wider governance activities in the EOSC ecosystem.

Support the wider development of ethical and information governance frameworks.

EOSC-hub should engage with a broader set of stakeholders, including social science and statistical data service providers, and the emerging EOSC governance function, to build a strong consensus and strong processes for cross-border research using sensitive data.

In support of research reproducibility we recommend EOSC-hub invest technical research and development effort in recording and tracking data provenance across the EOSC-hub service ecosystem.

For the research use of sensitive data, and personal data in particular, the General Data Protection Regulation recommends that communities develop Codes of Conduct to standardise ethical norms and practices. We recommend that EOSC-hub consider the feasibility and desirability of extending this report into just such a Code of Conduct for cross-border research with sensitive data.

1 Introduction

"This vision cannot be realised without specifications and standards for common components to enable interoperability across the FAIR data ecosystem" [4].

EOSC-hub brings together both users and providers of existing research computing and data services with the goal of harmonising interactions between service providers, thereby laying the foundations of a common set of e-infrastructure services for European research – EOSC, the European Open Science Cloud. Access to research data is the central principle of EOSC, and a common policy framework for data sharing is an important ingredient in realising that principle.

EOSC, of course, is composed of many different parts – services, research infrastructures, e-infrastructures – and many, if not all, of these constituents have drawn up data sharing policies in recent years, with particular regard to the protection of personal data under the General Data Protection Regulation, GDPR. The EUDAT2020 project created a series of recommendations for data service providers in late 2017 [1]; the Human Brain Project has a wealth of documentation on legal compliance in a complex biomedical ecosystem [2]. More recently, an overall policy framework for EOSC has been developed by the EOSCpilot project, captured in their report *D3.3: Draft Policy Recommendations* and four related whitepapers [3] and including a number of recommendations around data sharing. In terms of the FAIR data principles for finding and accessing research data which are interoperable and reusable, the recent EC Expert Group report *Turning FAIR into reality* [4] provides a definitive guide.

Rather than reinvent wheels, EOSC-hub has reviewed these key sources (and more) and evaluated in particular the EOSCpilot recommendations with an eye to practical implementation steps and guidance for service providers in EOSC-hub and beyond. EOSCpilot's recommendations touch on data sharing in numerous ways; our approach in this work has been to identify where these recommendations might be turned into practical guidance for service designers and integrators, and, by drawing on our working group's expertise in EUDAT, ELIXIR, ECRIN, BBMRI and other major data infrastructures, recommending in turn a series of steps towards effective data sharing policies for EOSC service providers.

We target these policy recommendations at data that are "ready for sharing". Research data move through a lifecycle, from dynamic "research objects" to formal, static datasets-of-record. Adopting the curation continuum model suggested by the Australian National Data Service [5] we recognise three main phases in the data lifecycle: private, dynamic research data that may be shared within a lab or between close collaborators; relatively stable data objects that can be shared more broadly within a research community; and static, published data-of-record, perhaps associated with one or more publications and potentially deposited in a long-term data repository. Our recommendations are geared towards the "published" and "broadly shared" categories and not necessarily the private, dynamic "research objects" (although we would recommend that the final state of research data always be considered in advance, and steps towards meeting the recommendations for shared data be included naturally in relevant data management plans).

While providing a common environment for the sharing of open research data is a guiding principle, we are first to acknowledge that not all research data can be shared openly. Research with “sensitive” data – medical records, genetic data, clinical trials data, statistical microdata, administrative or government data – offers the potential for tremendous public benefit but must be conducted in such a way that the legitimate rights and privacy expectations of data subjects are respected and balanced fairly. Our working group includes experts from a number of “sensitive data research services” across Europe, and practical considerations for the safe sharing of sensitive data for research form a key part of this report.

1.1 Scope and principles

We set out to define data sharing policies that should be adopted by data and service providers within the EOSC-hub consortium (“the EOSC-hub ecosystem”) but which could very easily (one might say naturally) be adopted by all such providers participating in EOSC generally.

In discussing EOSC we assume no particular organisational or governance form but rather cast EOSC as a network of independent legal entities working together to achieve common aims in open science. We thus seek to avoid statements or recommendations that assert, assume or require any form of “EOSC governance” beyond the model implicit in today's World-Wide Web.

1.1.1 The meaning of “users”

In this report we frame policies with two generalised classes of user in mind, where by “user” we mean anyone seeking to find and process shared research data objects within the EOSC-hub ecosystem, for any purpose. These two classes of user are: a human agent with a standard, interactive Web browser; and an autonomous program making http requests for data objects – a “script”. Scripts may be as simple as a Unix-style command line tool such as *curl* or *wget*, or as sophisticated as a large scientific workflow built from Python, C or Java.

1.2 Structure of this report

Chapter 2 introduces the key recommendations from EOSCpilot that have a bearing on data sharing; these form our starting point, and we map recommendations forwards to later chapters and sections of this report. Chapter 3 considers openness, principally with a view towards the openness or otherwise of personal or sensitive data for research. Chapter 4 considers the implementation of FAIRness for data across the EOSC-hub ecosystem: the properties of findability, accessibility, interoperability and reusability. Chapter 5 touches very briefly on reproducibility, concluding mainly that “more research is needed”, and Chapter 6 looks at frameworks for sharing sensitive data within the ideas of the “Five Safes” principles. Chapter 7 draws the report’s recommendations together and suggests time horizons for implementation; it also raises (but does not answer) the question of developing this report into a formal code of conduct for sensitive data research services (“formal” in the sense of registered with the European Data Protection Board).

2 Building on EOSCpilot

In August 2018 the EOSCpilot project published a report, *D3.3: Draft Policy Recommendations,* and four related whitepapers [3] which look at the policy landscape for EOSC. The EOSCpilot recommendations fall into four categories: Open Science and Open Scholarship (designated 'OS' in [3]); Data Protection ('DP'); Procurement ('P'); and Ethics ('L'). There are some 40 in total, more than half falling under Open Science and Open Scholarship. In analysing these we have settled on 11 that touch on data sharing and are most readily amenable to practical implementation within the horizon of the EOSC-hub project, and most likely to provide early value to data sharing within EOSC. For each recommendation the EOSCpilot report authors summarise possible implications for a number of stakeholder groups; in our extract below we have included the suggested implications for 'Research Infrastructures' as being most pertinent here. Where appropriate we provide forward references to sections of this report which describe our recommended next steps for EOSC-hub.

Id	EOSCpilot Recommendation	Implications for Research Infrastructures	See...
OS2	Adopt the AARC framework for enabling an interoperable AAI infrastructure.	-	§2.1
OS3	Adopt a minimum metadata schema and limited number of APIs to be considered as standard for services, infrastructures and other resources in the EOSC Service Catalogue.	RIs will need to adopt the approved set of minimum metadata and APIs for greater interoperability of RIs and services, if they wish to participate in the EOSC.	§4.2
OS6	Adopt a minimal set of standards for data/metadata and exchange protocols.	Develop and deploy standardisation tools and testing processes.	§4.2, §4.3, §4.4
OS9	Encourage the development of an EOSC TDM (Text and Data Mining) Policy Framework.	RIs can support the principles and expectations around openness which RIs and users should meet.	§4.4, §4.5
OS10	Develop principles for long-term data stewardship, enabling curation, provenance and quality.	RIs can support the data stewardship standards which RIs and users should meet.	§2.1
OS13	Make DMPs [data management plans] a	Support all usage applications of DMPs.	§2.1

Id	EOSCpilot Recommendation	Implications for Research Infrastructures	See...
	requirement and develop consistent (i.e. aligned) requirements for DMPs		
OS14	Encourage the use of unique and persistent digital identifiers.	Research outputs produced using RIs should be assigned unique and persistent digital identifiers. Supports research outputs to be open, FAIR and citable.	§4.3
OS18	Have proper IPR documentation when releasing or accessing a research resource.	Only host research content that contains IPR documentation. Provide tools and guidelines for clearing content. Ensure that IPR clearance takes place before any resource is shared through the infrastructure and only host IPR cleared material.	§4.2, §4.5
DP1	Legal basis for data protection: consent and legitimate interest of controller. For data processed through the EOSC: i) Explain the purpose of all data recording and processing; ii) Apply a concept of tiered consent (in compliance with “broad consent” of the GDPR); iii) Adapt privacy-by-design and privacy-by-default solutions (providing data subjects with a technological solution for consent withdrawal).	RIs need to provide (and if necessary develop) privacy-by- design/privacy-by-default systems and processes.	§6
DP3	Developing a user-friendly EOSC data protection policy a) Introduction of a special tag for the processing of data in the EOSC (as already done by some stakeholders). We recommend at least a differentiation between - personal data; - special categories of personal data;	Introduction of a tag that (at a minimum) differentiates between i) Personal data ii) Special categories of personal data iii) Data to be processed under special conditions. [To provide] Support for users via identification of respective regulations.	§3, §6

Id	EOSCpilot Recommendation	Implications for Research Infrastructures	See...
	- data to be processed under special conditions (e.g. the data of minors). b) Introduction of special regimes to classify data according to the level of data protection constraints.		
L0B	Metadata is managed and monitored to support research integrity (provenance, credit, status etc.).	Support of consistent application of the provenance and discovery metadata will be required. Tools to support correct metadata application will be required, will need development and funding.	§5

2.1 A note on out-of-scope recommendations

Some EOSCpilot recommendations are essential for the realisation of frictionless data sharing within the EOSC-hub ecosystem but are considered out of scope for this report.

OS2, enabling single sign-on, is the key to making EOSC work, and will be addressed by EOSC-hub working in concert with AARC.

OS10, long-term data stewardship, and OS13, data management plans, are important dimensions of the wider FAIR ecosystem but are, strictly speaking, out of scope for this report.

3 Policies for Openness

'As open as possible, as closed as necessary." [4]

Underlining the distinction drawn in the recent Expert Group FAIR report [4], we make a clear distinction between "open" data and "FAIR" data. FAIR data is a term coined originally for the FORCE11 FAIR Principles [6] and now widely used as a measure of the findability, accessibility, interoperability and reusability of research (or other) data. In the context of developing practical policies for data sharing within the EOSC-hub ecosystem, it is not an oversimplification to argue that the FAIR principles apply tests for the availability of research data that are principally *technical*: data should be finable, accessible, interoperable and reusable by technical means. In contrast, the openness, or otherwise, of data is principally a *legal* or *ethical* issue: a person's electronic medical records can easily be made openly available in machine-readable form on the public Internet, but *should not*!

EOSC-hub subscribes to the principles of data which are both open and FAIR. Policies for FAIRness are addressed in the next chapter; here we recommend approaches to openness. We can think of three principal barriers to making research data open: they are constrained by intellectual property law; they involve a data subject and thus may well contain personal data; or there are ethical sensitivities around their open publication. We consider each of these areas briefly.

3.1 Intellectual property restrictions

The application of intellectual property law to data relies first and foremost on the recognition of data as property; this is by no means a settled issue in Member State Law [6]. "Data" representing creative works will be covered by copyright; commercially sensitive data may be classed as trade secrets, or protected by patents; databases or data collections may be protected by *sui generis* database rights; other data may have no legal protection at all. In fields of research it is customary for research organisations, or funders, or grant Principal Investigators to claim rights over data generated through research, and to license these rights for the reuse of data under a variety of licensing scheme, some open, some not. There is thus a case to be made that the opening up through licensing of research data is more a cultural issue or a question of ethical norms than it is a legal one. This is the approach we shall take in this report, addressing the licensing of data under reproducibility in Chapter 4 on FAIR data. Publishing data freely and openly under a Creative Commons CC0 rights waiver is an excellent illustration of good research practice, and something to be encouraged; whether the researcher might actually possess any rights to waive in the first place is, strangely enough, a secondary question.

3.2 Personal data restrictions

Where a dataset contains personal information about one or more data subjects, the law on openness is largely clear. The EU General Data Protection Regulation [7] replaces the provisions of the earlier 95/46/EC Data Protection Directive and sets out the principles and conditions for processing personal data across the EU, as well as the rights of data subjects and the obligations of

data controllers and data processors. From 25 May 2018 all organisations within the EU (and in other countries "where Member State law applies by virtue of public international law") are subject to the GDPR.

The legal clarity around personal data means that much of our work on openness has focused on the GDPR and handling data of this kind. In particular, GDPR considerations have driven work in developing the idea of data tags as a means of ensuring that data objects are handled according to their level of sensitivity.

3.3 Ethical data restrictions

Personal data are by no means the only data which need sensitive handling. A 2014 report from the EU RECODE project[1] on *Legal and ethical issues in open access and data dissemination and preservation* classifies ethical concerns about the openness of data under five headings: unintended secondary uses and misappropriation; dual use; violations of privacy and confidentiality; unequal distribution of research results; commercialization, and; restriction of scientific freedom. Chapter 3 of that report offers good examples of each of these areas of concern, and notes that the principal mechanisms for handling data in such circumstances arise from the associated scientific or research communities and published codes of conduct.

As a common example of ethical norms in action here, research culture is quite comfortable with the idea of withholding data from full openness for a period of time to allow the originators of the data to publish first works with them – the "embargo" period. Embargo durations can and do vary according to specific policy provisions (e.g. as allowed by a funding agency) but usually take the form of a relative time span (e.g. 18 months) with a start date that, again, varies according to policy provisions. In practical terms, the OpenAIRE/DataCite basic metadata application profile (cf. Appendix I) supports this concept through the `<date dateType="Available">` tag.

3.4 DataTags: a common approach to handling

A DataTag is a label indicating a level of protection to be applied to the processing of the tagged data object. The concept was developed originally at Harvard University [9] and based on US legislation. During the EUDAT2020 project (2015-2018) DANS in the Netherlands developed a pilot version of the DataTags system based on the GDPR (see below). The idea of a DataTag follows directly from EOSCpilot recommendation DP3.

Following the original Harvard description, a DataTags repository is a repository of files held for data sharing that satisfies the following conditions:

1. A DataTag is a set of security features and access requirements for file handling. A DataTags repository has a finite, partially ordered set of DataTags, where the strictness and strength of DataTags' security features and access requirements dictate the ordering. A repository must have more than one DataTag.

[1] See http://recodeproject.eu/

2. All files in the repository must have a DataTag, and each file in the repository has one and only one DataTag. A file may optionally have additional handling requirements, such as an audit trail log or an expiration date. A file may optionally require additional terms for a data use agreement or additional terms of access by a recipient of the file from the repository. A file may have attributes that further describe it for reporting purposes. None of the optional requirements may weaken or replace the security requirements for the file's assigned DataTag, and none may adjust a DataTag's security requirements to be the same as another DataTag or stronger than a more restrictive DataTag.
3. A recipient who receives a file from the repository must satisfy the file's associated access requirements, produce sufficient credentials as requested, and agree to any terms of use required to acquire a copy of the file.
4. Technological guarantees exist that the requirements in 1 and 2 are satisfied for all files in the repository and for all accesses to those files from the repository. This imposes auditing obligations on transactions in the repository.

Security features and access credentials are independent components of a DataTag and at least one must be ordered to satisfy the first condition.

The use of DataTags this way enables the codifying and enforcement of privacy protection on personal or other sensitive data for optimal sharing within and across boundaries. DataTags offer the technical infrastructure for providing tiered access to data and for affixing machine-readable policies (and enforcing them in an auditable manner) in data sharing transactions. It combines the necessary legal requirements with a technical infrastructure.

3.4.1 A DataTags prototype for EOSC

At DANS, work on DataTags for GDPR is based on the original work at Harvard by Sweeney, Crosas & Bar Sinai (2015). The first prototype was developed under the EUDAT2020 project using the Zingtree decision tree application[2] to support researchers in complying with the GDPR [10]. The authors codified relevant GDPR Articles into more user-friendly questions to be answered by the data subject. Through a series of such questions, the tree results in DataTags that serve as advice for compliance. This questionnaire is viewed as a tool to be used at the start of the dataset deposit process, a guide for the conversation between depositor and repository, to help determine a dataset's content and assess any possible non-compliance issues.

The first prototype of the DANS DataTags approach garnered attention from across The Netherlands and more broadly from research institutions and universities. This interest spurred further development, with the idea of turning it into a framework that can be used universally in Europe. Certain adaptations needed to be made to the first prototype to make it suitable for this universal use. An example of this can be illustrated by GDPR Recital 52, concerning exceptions of prohibition for processing of special categories of personal data, provided for in individual EU Member States. If the to-be-deposited dataset contains genetic information about a data subject, appropriate technical and organizational measures should be in place to ensure lawful further

[2] See https://zingtree.com

processing of the dataset. To be able to give useful advice to the person answering the questions, it is imperative to determine conditions for processing, like informed consent and data minimisation.

Building on this first prototype through the EOSC-hub project, DANS has enhanced both the decision tree and the conceptual framework, refining questions and routing to help guide a researcher with the best possible advice given the answers (see Figure 1). Improvements have been made in reducing the complexity of the questions for researchers, and the decision tree has been remodelled to meet more closely the demands of the GDPR. This result in four distinct DataTags (see Table 1), where a blue tag refers to datasets containing no personal data at all, and a green tag to anonymised data. Anonymised data are, strictly speaking, outside the scope of the GDPR; nevertheless, in a risk-based model the chance of re-identification (through, for example, linkage of multiple datasets) is non-zero compared to datasets which never included personal data. In any other case, regardless of any specific facts in the dataset, the outcomes will be orange or red. Whenever identifiable (even pseudonymised!) personal information is part of the dataset, orange will be the resulting tag colour. A final result of red will occur when "special categories of personal data" (GDPR article 9) are involved. Note that Figure 1 captures a snapshot of the decision tree at time of writing; further improvements are planned.

Ideally this decision tree would become a generic tool for implementation across universities and research institutions; however, the nuances introduced by national variations in research data handling under GDPR mean that the tool needs to be used interactively between depositors and repositories, rather than purely automatically: questions and answers must lead the researcher to advice that is as specific as possible. The ultimate goal of the instrument is to improve understanding of this privacy regulation among researchers and research communities. We recommend this as a common approach to EOSC data services.

The DANS work is a concrete implementation of a DataTags system for GDPR. DataTags are not, though, bound to legislation or types of data: a DataTag is a token that advises how a given data object should be handled, irrespective of the underlying nature of the data. DataTags can equally be applied to non-personal sensitive data – detailed geo-locations of endangered species, for example – and can be adopted by data repositories as a general flag for the risk level of the associated data object (where ‘risk’ here should be interpreted as ‘disclosure risk’ or ‘risk of serious consequences [to data subject, researcher, repository or all three] should these data leak into the public domain’). Consequently, the four DataTags identified through the GDPR work are entirely general, risk-based measures of data sensitivity, and we recommend they be adopted across the EOSC-hub ecosystem. We also recommend that working groups be set up between communities in which different kinds of data sensitivity arise (e.g. biodiversity) and data curation professionals to develop similar decision tree-based approaches to assessing and tagging sensitive data objects.

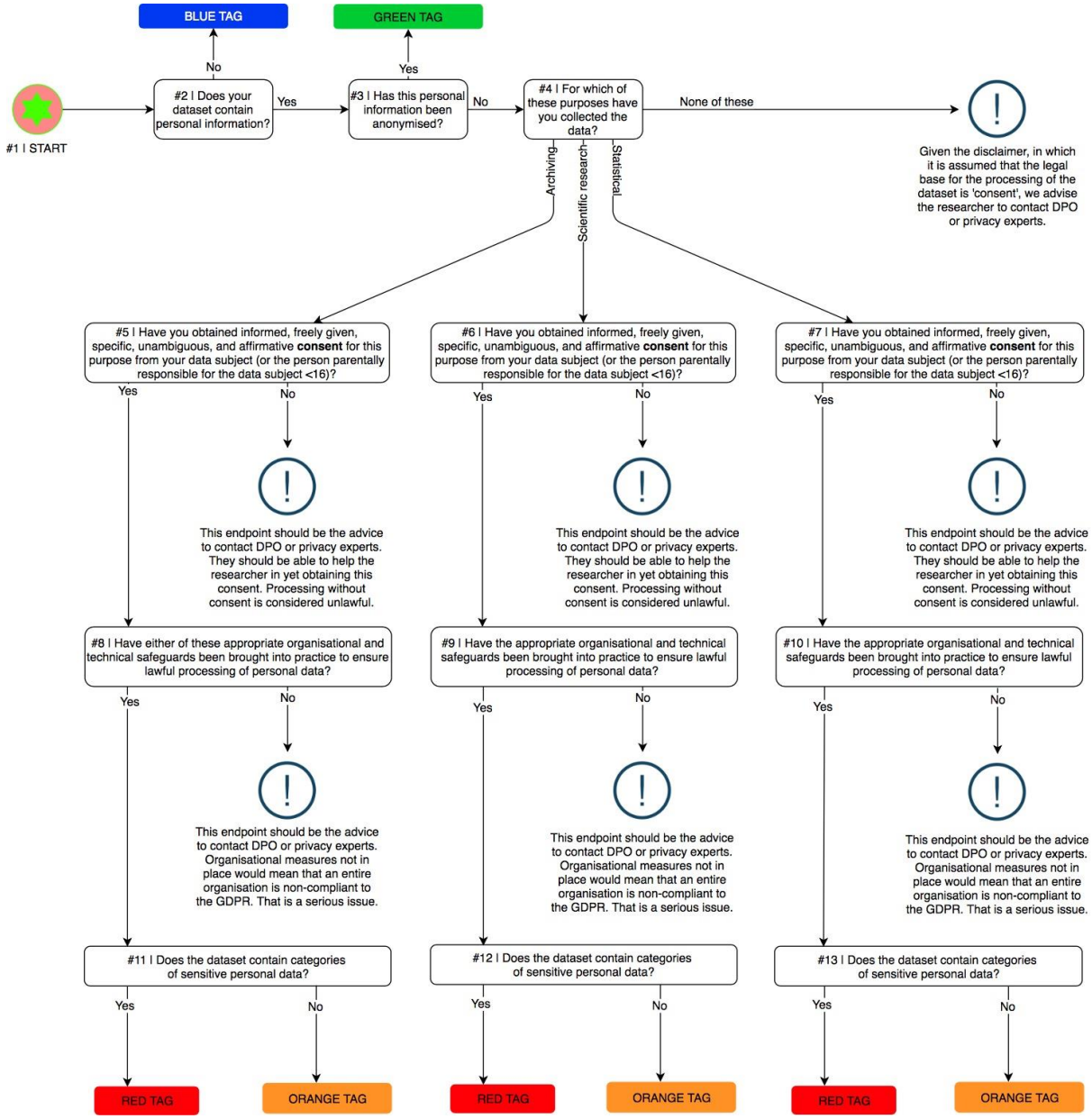

Figure 1. An example GDPR DataTags decision tree. Note that this is work in progress.

Table 1. Recommended DataTags arising from considerations of personal data under GDPR.

Risk Class	Technical and organisational measures	GDPR decision tree outcome	Description and desired message to the person that answered the ZingTree questionnaire
0 - Public	None.	Non-personal data.	Dataset contains no information that refers to any identified or identifiable living individual.
1 - Basic	Although anonymised data are out of scope of the GDPR, protection and authentication are desirable since de-identification is always possible. In addition, in aggregation with other datasets, the data could be traced back to original. Registration necessary, processing agreement is required for DANS, resulting in demonstrable accountability.	Anonymised personal data.	The dataset does contain personal information, but the researcher has made sure that this data is anonymised. Principles of anonymisation have been followed accordingly.

II - Increased	Examples include, but are not limited to:
	• Processing agreement
	• Data minimisation
	• Pseudonymisation
	• Authentication access policy:
	  ○ registered users only
	  ○ mandatory identification
	  ○ depositor approval
		Personal data. Consent obtained, including child's consent.	Dataset contains personal data. This data is collected in a lawful manner on the basis of obtained consent. This consent is obtained in compliance with articles 5, 6, and 7, and 8 in case of data subjects below 16. Message: “continue, but make sure appropriate safeguards are in place.”

III - High	Examples include, but are not limited to: • Processing agreement • Data minimisation • Pseudonymisation	Special categories of personal data. Consent	Given the answers provided, special categories of personal data are

Risk Class	Technical and organisational measures	GDPR decision tree outcome	Description and desired message to the person that answered the ZingTree questionnaire

	• Encryption
	• Two- or multi-factor authentication
	• Authentication access policy (depositor approval):
	  ○ registered users only
	  ○ protected environment access (special permission only)
	  ○ mandatory identification
	  ○ depositor approval
		obtained, including child's consent.	expected to be processed. This data is collected in a lawful manner on the basis of obtained consent. Since the GDPR provides multiple articles dedicated to these categories, additional prudence is advised. Message: “continue, but make sure appropriate safeguards are in place.”

3.5 Policy recommendations on openness

Rec 1 ESOC-hub should adopt a DataTag system for data objects based on at least four risk levels: blue, green, orange and red.

Rec 2 EOSC-hub should promote the use and development of decision tree-based tools for tagging data objects. A GDPR tool should be based on existing work at DANS; other tools should be developed in collaboration with relevant scientific and research communities.

4 Policies for FAIRness

Significant work on the implementation of FAIR policies has been carried out in the last twelve months, notably by the Expert Group on FAIR data [4]. Our task has been to identify the practical next steps for EOSC-hub service integrators, drawing on relevant community developments.

4.1 The FAIR Digital Object

The EC Expert Group on FAIR data introduces the concept of the FAIR Digital Object (Figure 8, page 35 of [4], reproduced below).

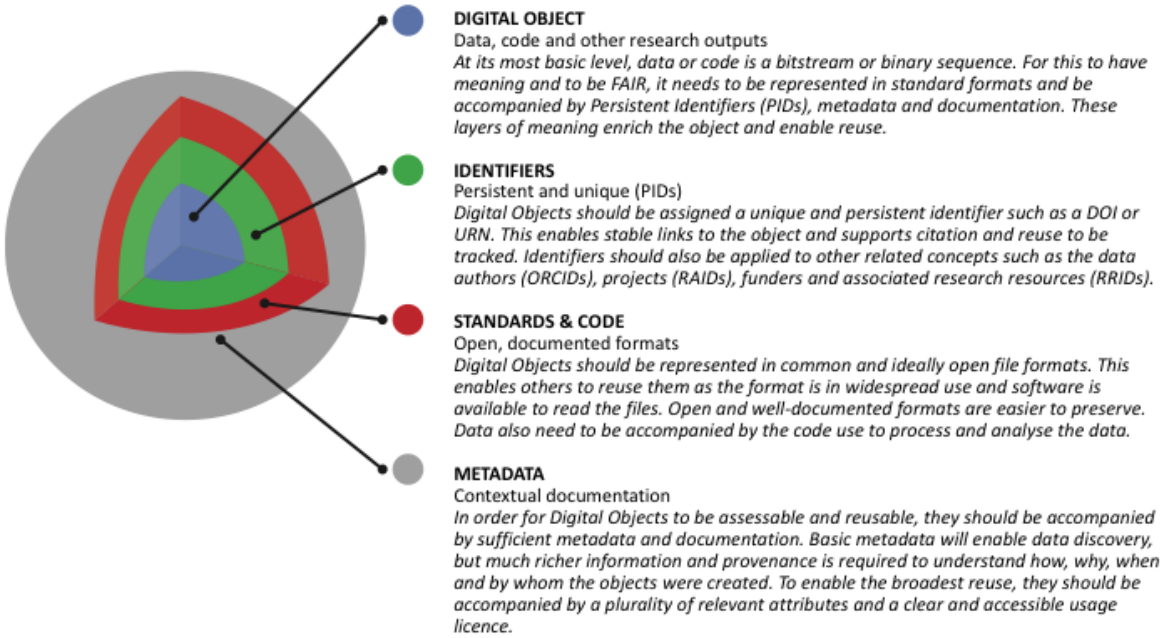

Figure 2. The FAIR Digital Object (reproduced from [4]).

We recommend that all shareable data objects in the EOSC-hub ecosystem should be FAIR Digital Objects.

The rest of this chapter explores practical steps – and areas for further technical research – for EOSC-hub service integrators in realising this policy goal.

4.1.1 Policy recommendations for FAIR data objects

Rec 3 All shareable data objects in the EOSC-hub ecosystem should be FAIR Digital Objects.

4.2 Findability

While findability is not explicitly mentioned in the EOSCpilot recommendations noted in the Introduction, we can assert that it is implicit and fundamental to many (e.g. OS3, minimal metadata schema). The EOSC builds on the concept of a rich layer of reusable research data; if

those data cannot be found, none of the higher value-added services matters very much. Findability is thus a key topic for EOSC-hub; we can easily argue it is the "killer app" for EOSC.

We assert that findability requires that the existence of a data object, and some basic facts about it (basic metadata), be discoverable on the public Web. Discoverability on the Web may be achieved by two methods:

1. data objects may be registered in public Web catalogues (either discipline specific or general purpose);
2. data objects may just be published on the Web without an accompanying catalogue entry.

The first method is more likely to arise for "published, final" data; the latter may perhaps be common for data shared within a community. In both cases data objects will be found by searching, the former within a catalogue, the latter using Web search engines; in both cases the searches need well-defined descriptive terms to target. Data objects and repositories of data objects in the EOSC-hub ecosystem should support both search methods.

The same could be said for the findability of data by scripts although, while Web search engines are generally accessible programmatically in a variety of ways, this is not always the case for individual data catalogues.

In drawing up practical recommendations for findability we take inspiration and guidance from the ELIXIR initiative. ELIXIR has a number of approaches to improving the findability of life-science resources across the ELIXIR node federation. The baseline ELIXIR approach is to provide core deposition databases[3] and core data resources[4] (as recommended by the ELIXIR Data Platform[5]) to allow ELIXIR member to deposit and access data in a well-known data repository determined by the community.

The ELIXIR Interoperability Platform[6] further provides a number of recommended services to improve discovery and findability of resources:

- Identifiers.org is an identifier resolving service that allows users to reference data independently of its repository location in the ELIXIR network (as multiple repositories may have metadata claims to the same identifier.
- FairSharing.org is a high-level educational catalogue of datasets, standards and policies used within the ELIXIR federation.
- Bioschemas.org is a schema.org extension (see Section 4.3 below) tailored for life-science concepts. It defines minimal metadata markup for each concept that improves the findability by search engines.

The ELIXIR Beacon project[7] pares the findability approach to its absolute minimum. It provides a standardised framework which gives only a binary yes/no answer (yes – have information, no – no

[3] See https://www.elixir-europe.org/platforms/data/elixir-deposition-databases
[4] See https://www.elixir-europe.org/platforms/data/core-data-resources
[5] See https://www.elixir-europe.org/platforms/data
[6] See https://www.elixir-europe.org/platforms/interoperability
[7] See https://beacon-project.io/

information) to specific genomic data collection queries. Each binary query response from a beacon (genomic data collection) can be additionally layered to increase the level of metadata revealed about the collection and the query. As a way to manage potential sensitivity in metadata records, this mechanism is worthy of further exploration within the EOSC-hub project.

Setting out to identify the perfect metadata schema for findability is dangerous; discussions about metadata very quickly often prove the adage that “the perfect is the enemy of the good”. We recommend that EOSC-hub follow the guidelines set out by OpenAIRE for data repositories [11], augmented by additional properties including a data tag. Similar guidelines have long been recommended by the EUDAT collaborative data infrastructure, and themselves derive from the DataCite guidelines associated with acquiring digital object identifiers (DOIs).

4.2.1 Policy recommendations for Findability

Rec 4 Data objects should be minimally described by a metadata record that follows the recommended schema in Appendix I.

Rec 5 A data object’s metadata record should be the minimally required description of it in a data catalogue.

4.3 Accessibility

In the FAIR Digital Object model, access to a data object is principally through a persistent identifier (PID) of some kind. The landscape of PIDs is surprisingly rich, although two main models stand out: the Compact Identifiers of n2t.net and identifiers.org used widely in the life sciences; and DOIs (and their underlying Handles) everywhere else [12]. The reality of multiple PID systems is what it is. While initiatives like FREYA are working towards closer alignment between PID resolver systems it is likely that EOSC-hub and EOSC more widely will have to support many types of PID for many years.

What matters more than any particular syntactic form for a PID is that, when it is presented to a user, it is presented as a properly formed and resolvable URL on the public Web, and that no matter what resolver system it uses (handle.net, identifiers.org, doi.org, ...) the user agent, whether Web browser or script, is able to understand the results. For PIDs, semantic – rather than syntactic – interoperability is the key to accessibility.

Current best practice for PID resolution is to return an HTML landing page to the requester, ideally encapsulating metadata about the data object [13]. Wimalaratne & Fenner in [13] also recommend that PID resolvers support http content negotiation: “Resolver services can support content negotiation so that users are not redirected to the landing page for a resource, but instead receive metadata in a standard, machine-readable format.”

As noted above (and in [13]), standardising metadata formats is no easy task. Wimalaratne & Fenner do, however, note the emergence of schema.org as a structured approach to metadata designed to be embedded in HTML pages. They note “schema.org is a collaborative initiative founded by the search providers Google, Bing, Yahoo and Yandex to markup metadata about web pages. There is an active community behind schema.org and it is being widely adopted by other communities such as life sciences” (as we note above). A recent dataset report [14] lists 32

repositories that support schema.org markup in JSON-LD format, including DataCite, Dryad, Pangaea, PDBe and UniProt. We recommend that EOSC-hub adopts this approach as good practice.

Why schema.org? The main reason is the support already in place from the big Web search engines (only Baidu is missing from the founding organisations). For data objects described and registered in public catalogues this is less important, but for data objects "just" published on the Web, embedding a metadata record following the schema.org scheme mobilises the power of the Internet search giants as a second catalogue provider.

Current best practice for PIDs focuses on resolving and returning metadata. What about getting hold of the actual data object itself? Being able to "get" a data object URL and receive a bitstream of the object would be a potential boon for scripted workflows and other automated user agents, but there are inherent risks: "getting" a data object of 1 TB in size will have significant resource implications and could well lead to workflow failure, deadlock or other unforeseen consequences for scripts not expecting something of that magnitude. Wimalaratne & Fenner note that "The current best practice to have a persistent identifier point to a landing page, or provide metadata via content negotiation, means that the persistent identifier should probably never resolve to the content itself, directly or via content negotiation."

Suggested mechanisms for handling this include adding a second "content URL" to the metadata record for an object or using the *10320/loc* field in a Handle record to point to a copy of the object. Schema.org supports a Dataset type which has a *download* property of type DataDownload; DataDownload includes the *contentUrl* property. Nevertheless, these mechanisms were designed for the Web, where data objects do not often reach the scale of some scientific data objects; applying them in the EOSC-hub ecosystem is feasible but needs further design and technical research.

4.3.1 Policy recommendations for Accessibility

Rec 6 A data object should have a unique persistent identifier.

Rec 7 A data object's persistent identifier must form part of its metadata record.

Rec 8 An http GET request on a data object's persistent identifier should return an HTML landing page that can be rendered in a standard Web browser.

Rec 9 A data object's HTML landing page should encode its metadata record according to the schema.org approach.

Rec 10 An http GET request on a data object's persistent identifier accepting a different return format (e.g. XML or JSON) should return the data object's metadata record in that format (content negotiation).

Rec 11 EOSC-hub should track the work of the FREYA project and adopt best practices in PID resolution as they emerge.

Rec 12 EOSC-hub should initiate a programme of technical research for metadata discovery that builds on the Elixir Beacon approach for cases where metadata records may themselves contain sensitive data.

Rec 13 EOSC-hub should initiate a programme of technical research for direct retrieval of data objects by PID.

4.4 Interoperability

The Expert Group report highlights three key dimensions to interoperability: good metadata, compatible licensing and open data formats.

In practical terms, while EOSC-hub can take concrete steps on basic "findability" or "discoverability" metadata (as noted above), the wider ocean of discipline-relevant metadata is beyond our scope. EOSC-hub does, of course, support the Expert Group in encouraging communities to develop and implement rich metadata descriptions for all shared and shareable data objects; this is a natural part of the wider FAIR EOSC ecosystem.

Licensing we treat below under reusability. The remaining dimension, open data formats, is perhaps the easiest aspect of interoperability to promote. As a touchstone we draw on Tim Berners-Lee's "5-star Open Data" model[8] (see also [15]) and recommend that EOSC-hub should support and promote the sharing and publication of data objects that are "good 3-stars" – on the Web, machine-readable, in non-proprietary formats.

We note that the Expert Group report offers two particular recommendations in this area:

- Rec. 8: Facilitate automated processing

which good 3-star data achieves, and

- Rec. 7: Support semantic technologies

which might suggest a leaning towards 4-star data or above (introducing the ideas of linked data and steps towards a full 5-star Semantic Web). However, the Expert Group report goes on to note (p. 41):

> "Many ontologies have been developed but they remain dramatically underused in current practice for a variety of reasons, relating to the diversity of ontologies available, the challenge of establishing mappings between different expressions of a concept, the need to update concepts as domains evolve, incompatible licensing terms and the relative lack in many domains of coordinated community approaches to semantics. There remains a need for concerted efforts from research communities to establish and implement more effective processes for community development, endorsement and adoption of ontologies and vocabularies."

EOSC-hub endorses this view and fully supports continuing community efforts to develop and adopt ontologies and vocabularies, but – again deploying the argument of practicality – recommends concrete efforts first and foremost to ensure data objects are on the Web, machine-

[8] See https://5stardata.info/en/

readable, and in non-proprietary formats. Given the breadth of data and data services in the EOSC-hub ecosystem, EOSC-hub makes no particular recommendations on data formats beyond this. Many research disciplines have, and always will have, their own preferred data formats; the EOSC-hub ecosystem must embrace them all. Only where legacy community data formats are not open should EOSC-hub intercede with a view to encouraging a change in practice and culture.

4.4.1 Policy recommendations for Interoperability

Rec 14 Data objects should be published on the Web in an open, non-proprietary format chosen to suit its content or subject.

4.5 Reusability

The technical reusability of data is fully addressed by the F, A and I recommendations noted so far. The final hurdle to fully reusability is legal – as a research user, do I have the *rights* to re-use this dataset the way I want to? As an automated script, am I able to interpret the response from a PID resolver in ways which tell me whether I can legally link the dataset I want with the data I already have?

The adoption of the EU *sui generis* database right into version 4.0 of the Creative Commons licence suite[9] has accelerated adoption what was already becoming a favoured scheme for licensing data (the EUDAT CDI has recommended CC 4.0 for a number of years now). Another extremely attractive feature of the CC suite is the machine-readable form of each licence; attaching a machine-readable licence statement to a data object's metadata record (as recommended in the OpenAIRE metadata guidelines) is a key step on the road to realising automated interoperability and reusability.

Thus we recommend that all data providers in the EOSC-hub ecosystem be encouraged to adopt a licence – open if possible – from the Creative Commons suite version 4.x. For openly shareable data, these would be

- CC-BY 4.0 (attribution);
- CC-BY-SA 4.0 (attribution with onward propagation);
- CC0 (public domain or rights waiver).

4.5.1 Policy recommendations for Reusability

Rec 15 Data objects in the EOSC-hub ecosystem should adopt licences from the Creative Commons 4.0 licence suite. Where data are openly shareable, these should be one of:

- CC BY 4.0 (attribution);
- CC BY SA 4.0 (attribution with onward propagation);
- CC0 (public domain or rights waiver).

[9] See Creative Commons, http://creativecommons.org/

5 Policies for Reproducibility

An unintended consequence of the open access and open data movements is the rise of the fake journal. The now near-legendary case of Australian computer scientist Peter Vamplew's 2014 resubmission of David Mazières and Eddie Kohler's classic 2005 paper to the International Journal of Advanced Computer Technology is just one high-profile example of the risks of "wild West" scientific publishing to the wider integrity of science[10].

EOSCpilot's recommendation L0B raises this ethical question and suggests provenance metadata as one possible approach to enhancing the reproducibility of science and guarding against wider fakery. Maintaining a traceable provenance chain from research paper back through processing steps, software and workflows, to base datasets is an excellent goal and one that could, in principle, be within the scope of EOSC-hub. It is, however, extremely difficult.

Scientific provenance is a field of active research. Recent work in the field has looked at extending standard provenance modelling frameworks to include "workflow" structures [16] and applying such ideas to particular scientific workflow environments [17][18]. How to apply such ideas in a broad, distributed ecosystem like EOSC-hub is very much an open question. There is no easy example of good practice to point to and recommend. Logging of user interactions with large computing resources is standard practice, and http requests to web servers are routinely logged; logging is an excellent mechanism for creating basic chains of trust across complex systems[11]. This might provide a starting point for service integrators in EOSC-hub, but there is a lot of design work still to do.

5.1 Policy recommendations for reproducibility

Rec 16 EOSC-hub should consider the logging and tracking of scientific provenance data as an element of service integration design.

Rec 17 EOSC-hub should consider convening a technical working group on the topic of recording provenance across the EOSC-hub service ecosystem.

[10] See https://www.vox.com/2014/11/21/7259207/scientific-paper-scam and https://en.wikipedia.org/wiki/International_Journal_of_Advanced_Computer_Technology

[11] See, for example, the discussion in Chapter 10 of R Anderson, *Security Engineering*, Second Edition, 2008 (Wiley), available at https://www.cl.cam.ac.uk/~rja14/book.html

6 Policies for Data Sharing

Where data are open – unencumbered by legal or ethical constraints on their use – data sharing is largely a technical issue to be facilitated by following the FAIR principles. Where data cannot be openly shared, but might be shareable for well-defined, approved research purposes under controlled conditions, questions of information security and governance arise. We look at both models for data sharing below, as viewed as a technical or management problem.

6.1 Distributing versus not distributing

Unlike physical assets, digital data can be trivially reproduced and thus distributed very easily by making a copy and sending it to a collaborator (we ignore issues of size for these purposes). Sensitive data can be protected by encryption before they are shared, their handling requirements codified in a data tag. However, no matter the security measures taken when a sensitive data object is copied and transmitted, once it is beyond the administrative scope of the originator, the risks of data leakage will simply increase. Maintaining oversight of easily "copyable" and "forwardable" digital assets in the wild becomes quickly infeasible. Legal protection, while possibly enabling an originator to recoup damages from a data leaker, provides no guard against the damage to a data subject, or to the wider public trust, that the loss of a sensitive data object might cause. Sharing copies of sensitive data is fundamentally risky. We recommend against it.

The alternative to sending sensitive data to an approved researcher is to allow an approved researcher to come to the data. Providing *in situ* access to sensitive or precious assets is nothing new in the world of physical data objects but is a relatively novel concept for digital datasets. So-called *Safe Haven* environments, carefully controlled against data leakage, offer researchers computational environments within which they can work with sensitive data – medical records, government administrative data – under a supervisory regime designed to maintain public trust in the ethical conduct of such sensitive research. Such approaches become particularly important when researchers are allowed to link multiple datasets together. Even sensitive data that have been de-identified carry a residual risk (accordingly they should be tagged 'green', not 'blue' – cf. Section 3.4), and linking datasets together increases the risk of an analytics query returning a "cohort of one" answer, leading to a high chance of re-identification.

This type of research cross-connecting multiple research datasets, is, of course, exactly the type of research EOSC aims to facilitate. EOSCpilot recommendations DP1 and DP3 highlight the principles of privacy-by-default and -by-design, of data tagging, and the development of "special regimes" to classify and identify data subject to processing restrictions. To support these principles, and to facilitate supervised research using sensitive data within the EOSC-hub ecosystem, we recommend that EOSC-hub focus on supporting the remote use of Safe Haven services, coupling this with the development of suitable information governance processes (see below).

6.2 The Five Safes

The concept of "Five Safes" arose in the early 1990's as a way to characterise models and methods for the safe sharing of statistical microdata (individual survey responses, for instance). From their origins in national statistical services, the ideas of "Five Safes" have begun to find their way into design and thinking around medical and health data services[12] and beyond. As such they provide an excellent conceptual starting point for safe data sharing services in EOSC-hub.

The five safes are typically written as:

Safe projects	Is this use of the data appropriate?
Safe people	Can the users be trusted to use it in an appropriate manner?
Safe settings	Does the access facility limit unauthorised use?
Safe data	Is there a disclosure risk in the data itself?
Safe outputs	Are the statistical results non-disclosive?

It is noteworthy that only one of these – safe settings – concerns the environment (technical or physical) in which data might be shared; the others concern people and procedures that need to exist around the safe environment in order to manage the risks inherent in working with personal or sensitive data.

In recent years, statistics providers across Europe have made strides in allowing cross-border access to national statistical microdata. The *Data without Boundaries* project[13], active from 2012 to 2015, and its current follow-on the *International Data Access Network*[14], bring together national statistical services from France, Germany, the Netherlands and the UK in a series of joint data sharing agreements, with remote access to, for instance, French data from the CASD (Centre d'Accès Sécurisé aux Données, the secure data access centre) now possible from GESIS (the Leibniz Institute for the Social Sciences) in Germany. IDAN is a good illustration of the possible – and of the challenges involved in providing cross-border access to sensitive data between countries with, nominally, the same personal data protection laws. Wider national regulations on disclosure control, statistical research, etc. require that frameworks of legal equivalence need to be negotiated and put in place before remote data sharing becomes feasible, and these are currently done case-by-case and point-to-point.

This is not something EOSC-hub can expect to achieve alone. EOSC-hub is principally about service interoperability and data interchange. As such, in the Five Safes model EOSC-hub might only reasonably expect to be able to address safe settings and, perhaps, safe data. Nevertheless, adopting good practice like the Five Safes approach to data handling and "Safe Haven" service

[12] See https://en.wikipedia.org/wiki/Five_safes for a good discussion of the principles and current applications.
[13] See http://www.dwbproject.org/
[14] See https://idan.network/

specifications will be a valuable contribution to the broader picture of research using sensitive data.

6.3 Safe Haven services

As "safe settings" for sensitive data research, Safe Haven services (SHSs) can be viewed a little like "digital fume cupboards": instead of a sealed glass box for dangerous chemicals, manipulated through built-in gloves, a SHS provides a digital environment for working with sensitive data. The SHS is designed not to let anything leak out, and enables researchers to manipulate and analyse, but not remove, data – either input data or research outputs – directly.

A number of EOSC-hub partners already operate Safe Haven environments in national contexts: TSD, the Sensitive Data Service, is developed and operated by the University of Oslo under the umbrella of UNINETT Sigma2, Norway[15]; ePouta, a secure cloud environment, is developed and offered by CSC, Finland[16]; the Scottish National Safe Haven is developed and operated by EPCC at the University of Edinburgh[17]. These services run in different ways under different regimes but share a number of common design features.

6.3.1 Required features of a Safe Haven service – an example

Common features across the EOSC-hub services serve as a useful starting point for characterising a possible standard approach to general secure Safe Haven services in EOSC-hub; we sketch an example below using the language of requirements specifications[18], identifying three independent roles: User – an approved researcher making use of the SHS for a specific, pre-authorised purpose (e.g. an approved research project); SHS Administration – the team responsible for running the SHS systems and hardware; and Information Governance – the necessary supervisory and approval functions which govern the operation of the SHS.

- A SHS MUST operate within a broader Information Governance process (see below).
- A SHS MUST be isolated from the Internet by a dedicated network firewall under the direct management of SHS Administration.
- Network access to a SHS MUST be through a Virtual Private Network ('VPN') connection.
- Network access to a SHS MAY be restricted to connections from a limited number of known IP addresses ('whitelisting' or 'safe rooms').
- A SHS MUST log all User connections and actions within the SHS environment.
- Authentication of Users to a SHS MUST involve more than one factor ('2FA+').
- Users of a SHS MUST be authorised in advance to access only specific data objects.
- Users of a SHS MUST NOT have access to any data objects for which they are not specifically authorised ('default deny').
- Authorisation for users MUST be a function of Information Governance, not of SHS Administration.

[15] See https://www.uio.no/english/services/it/research/sensitive-data/
[16] See https://research.csc.fi/epouta/
[17] See https://www.epcc.ed.ac.uk/projects-portfolio/nhs-national-services-scotland-nss-national-safe-haven
[18] See https://www.ietf.org/rfc/rfc2119.txt

- User workspaces in the SHS MUST be isolated from one another.
- User workspaces in the SHS MUST be isolated from the broader SHS environment.
- User workspaces in the SHS MUST NOT allow outgoing network connections.
- Users SHOULD NOT be able to introduce data or software into their workspaces.
- Extraction of data and research outputs from the SHS MUST be done only by Information Governance.
- Users MUST NOT be able to extract any data or research outputs from the SHS.

This list is not exhaustive, nor at this stage definitive. We would recommend further development of these ideas by EOSC-hub to support the 'safe settings' principle of sensitive data access.

6.4 Information governance

Two of the Five Safes relate to physical or technical aspects of working with sensitive data – safe data and safe settings. The other three relate to the people and procedures needed around the data and Safe Haven environments to manage the inherent risk of data leakage. As noted in Section 3.4, data which were once personal but have been "de-identified" nevertheless carry a residual risk of re-identification, especially in the context of data linkage (see below). Thus, Safe Havens by themselves are insufficient in providing the necessary level of comfort for safe research with sensitive data: the Five Safes approach demands a system of procedures and approvals (and thus an approving body or bodies) also be in place. This is the "Information Governance" role referred to above.

A full information governance framework is beyond the immediate scope of EOSC-hub. However, we recommend that EOSC-hub engage with a broader complement of stakeholders (in particular, stakeholders from the statistical microdata and social science fields) in helping to develop information governance ideas for the whole EOSC.

6.4.1 Information governance for linked data – an example

One of the goals of the EOSC model is to facilitate greater sharing of research data, not least to enable greater *linkage* of related data in cross-disciplinary fields. Data linkage creates a special case for the handling of sensitive data; linking multiple de-identified data sets together, and combining them with publicly available data (such as social media posts, and clinic GPS locations and opening times) magnify risks of accidental (or malicious) disclosure and data leakage. The Scottish National Safe Haven was designed to support data linkage from across health and local government, and operates in the information governance framework generalised here as an example of what may be needed in EOSC-hub and beyond. This example is a slightly simplified view of the information governance process used currently to oversee research using both unconsented clinical register data and government administrative data within the National Safe Haven.

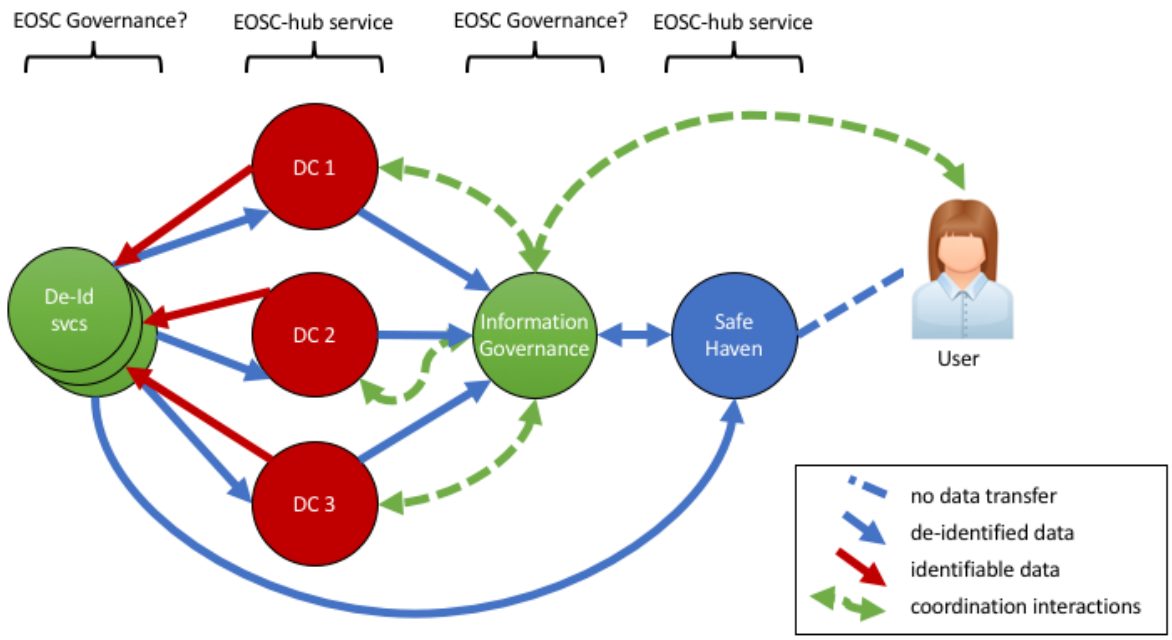

Figure 3. Generalised information governance for linked data in a Safe Haven environment.

Figure 3 sketches the interactions of a research User who wishes to perform research using a linked version of three potentially sensitive datasets, each of which is under the control of a different Data Controller (DC1 to DC3). We assume each of these datasets contains un-consented personally identifiable information (PII), the use of which for research purposes is permitted under a public benefit/public trust argument. The research will ultimately be conducted in the Safe Haven service, but authorisation, approval and coordination are all managed by a separate Information Governance body. In this model, the research project would proceed as follows.

1. The User submits a proposal for the three-data-set study to the Information Governance (IG) body.
2. The IG body seeks ethical approval as it needs to (possibly by escalating to a higher-level authority). We assume approval is granted.
3. The IG body sends individual requests to each of the Data Controllers (DC), requesting a de-identified copy of the dataset in question.
4. Each DC sends its identifiable dataset (red arrows) to a “de-identification service” ('De-Id svc'), requesting they remove the PII and replace it with arbitrary tokens.
5. The De-Id service does this for each dataset, using different tokens in different datasets for the same PII.
    - One complication is that the Researcher wants to be able to link the three datasets together by individual (e.g. all John Smith's data from the three datasets on one line). Where there is one De-Id service they are able to build up a “master index file” that connects the tokens used for “John Smith” in each of the three datasets; where there are multiple De-Id services they will need to coordinate.

6. The De-Id service returns each de-identified dataset (blue arrows) to the respective DC.
7. Each DC forwards their de-identified dataset (blue arrows) to the IG body.
8. The IG body sends the de-identified datasets (blue arrows) to the Safe Haven.
9. Independently, the De-Id service sends the "master index file" to the Safe Haven.
10. The Safe Haven uses the "master index file" to link the three datasets into one, places that linked dataset into a workspace, and notifies the IG body.
11. The IG body notifies the User, and monitors her research activities in the Safe Haven.
12. The User conducts her analysis and, upon completion, notifies the IG body about which results she would like to publish.
13. The IG body performs disclosure control, evaluating the safety of the research results; they may deny release if the results carry significant risk of the re-identification of individuals or the leakage of other sensitive data.

From this example, the importance of the Information Governance role is clear. This role approves research studies; coordinates data acquisition, preparation and de-identification; puts data into, and extracts data from, the Safe Haven; and gatekeeps any research results that arise. In this view, the Safe Haven service itself does very little but provide a secure computing environment.

In EOSC terms, we can think of the three data controllers as data or service providers, and the researcher as a user with a web browser. The Information Governance body is clearly part of an ethical oversight function within the envelope of "EOSC governance"; indeed, it is certainly from the same stable as the *Ethics and Legal Advisory Board (ELAB)* recommended by EOSCpilot (Ethics recommendations L2A and L1), if not indeed the same animal.

The role of the De-Identification services is potentially challenging. Given their role in working with PII they are most likely to be statutory or governmental bodies (this is certainly the case in Scotland in the UK). Note, however, that this role would only come in to play where multiple datasets need to be linked together by a "PII key". For single datasets, the data controller can (in principle) perform the necessary de-identification before passing their dataset to the IG body (although they may not have the full range of skills and tools available to a specialist service).

6.5 Policy recommendations for data sharing

Rec 18 Data objects with tags of green or higher (de-identified personal data and above) should not be freely copied or distributed within the EOSC-hub ecosystem.

Rec 19 EOSC-hub should adopt the Five Safes principles as guidance for the management and handling of sensitive data in the EOSC-hub ecosystem.

Rec 20 EOSC-hub should develop a technical design framework for Safe Haven services to support the safe data and safe settings dimensions of the Five Safes principles.

Rec 21 EOSC-hub should engage with a broader set of stakeholders, including social science and statistical data service providers, in supporting the design of a Europe-wide framework for research with sensitive data.

7 Conclusions and Next Steps

The twenty-one recommendations arising from this report are intended as practical steps towards implementing key policy recommendations from the EOSCpilot project on data sharing within the EOSC ecosystem. Broadly speaking they can be summarised under three headings:

1. Implement FAIR.
2. Build technical expertise in 'safe data' and 'safe settings'.
3. Support the wider development of ethical and information governance frameworks.

We tabulate the recommendations accordingly below.

7.1 Towards a Code of Conduct for research with sensitive data?

Codes of conduct are recognised under the GDPR – particularly in a research context – as useful elements of "soft law" within individual research disciplines. GDPR Article 40 notes:

Article 40 (1): "The Member States, the supervisory authorities, the Board and the Commission shall encourage the drawing up of codes of conduct intended to contribute to the proper application of this Regulation, taking account of the specific features of the various processing sectors and the specific needs of micro, small and medium-sized enterprises."

Article 40 (2): "Associations and other bodies representing categories of controllers or processors may prepare codes of conduct, or amend or extend such codes, for the purpose of specifying the application of this Regulation…"

Further, Recital (98) notes: "In particular, such codes of conduct could calibrate the obligations of controllers and processors, taking into account the risk likely to result from the processing for the rights and freedoms of natural persons" (emphasising the principle of "balance-of-risk" prevalent throughout the GDPR).

At time of writing there is ongoing work on a code of conduct for health data, coordinated by BBMRI[19] and two finalised EU cloud codes of conduct, the Cloud Infrastructure Providers Europe (CISPE) Code[20] and the EU Cloud Code of Conduct[21], of potential relevance to EOSC-hub. As of December 2018 CISPE lists 104 compliant services in its public register; the EUCOC website lists publicly only two organisations as offering "adherent services", and those at only "preliminary" level. Note that both cloud codes are pitched at business to business interactions between cloud service providers in data controller and data processor roles and, in terms of personal data, cover mostly transfers of consented personal data for business – rather than research – purposes.

Is there scope to develop the ideas captured in this report into a "code of conduct for sensitive data research services"? We add this question as a final recommendation:

[19] See https://code-of-conduct-for-health-research.eu/
[20] See https://cispe.cloud/
[21] See https://eucoc.cloud/en/home.html

Rec 22 EOSC-hub should consider the feasibility and desirability of developing a code of conduct for sensitive data research.

7.2 Collected recommendations

For each recommendation we give an indication of its principal scope: technical recommendations that could be implemented by individual service providers (TS); technical recommendations that require a more coordinated approach across EOSC-hub (TC); or policy recommendations (P). We also offer an estimate of time horizon (Short, Medium or Long term, deliberately vague) as an indicator of perceived complexity or maturity.

		Scope	Horizon
Implement FAIR			
Rec 3	All shareable data objects in the EOSC-hub ecosystem should be FAIR Digital Objects.	TS	M
Rec 4	Data objects should be minimally described by a metadata record that follows the recommended schema in Appendix I.	TS	S
Rec 5	A data object's metadata record should be the minimally required description of it in a data catalogue.	TS	S
Rec 6	A data object should have a unique persistent identifier.	TS	S
Rec 7	A data object's persistent identifier must form part of its metadata record.	TS	S
Rec 8	An http GET request on a data object's persistent identifier should return an HTML landing page that can be rendered in a standard Web browser.	TS	S
Rec 9	A data object's HTML landing page should encode its metadata record according to the schema.org approach.	TS	M
Rec 10	An http GET request on a data object's persistent identifier accepting a different return format (e.g. XML or JSON) should return the data object's metadata record in that format (content negotiation).	TS	M
Rec 11	EOSC-hub should track the work of the FREYA project and adopt best practices in PID resolution as they emerge.	TC	M
Rec 12	EOSC-hub should initiate a programme of technical research for metadata discovery that builds on the Elixir Beacon approach for cases where metadata records may themselves contain sensitive data.	TC	M
Rec 13	EOSC-hub should initiate a programme of technical research for direct retrieval of data objects by PID.	TC	M
Rec 14	Data objects should be published on the Web in an open, non-proprietary format chosen to suit its content or subject.	TS	S
Rec 15	Data objects in the EOSC-hub ecosystem should adopt licences	TS	S

	from the Creative Commons 4.0 licence suite. Where data are openly shareable, these should be one of: • CC BY 4.0 (attribution); • CC BY SA 4.0 (attribution with onward propagation); • CC0 (public domain or rights waiver).		
Build technical expertise in 'safe data' and 'safe settings'			
Rec 1	ESOC-hub should adopt a DataTag system for data objects based on at least four risk levels: blue, green, orange and red.	TS	M
Rec 2	EOSC-hub should promote the use and development of decision tree-based tools for tagging data objects. A GDPR tool should be based on existing work at DANS; other tools should be developed in collaboration with relevant scientific and research communities.	TC	M
Rec 18	Data objects with tags of green or higher (de-identified personal data and above) should not be freely copied or distributed within the EOSC-hub ecosystem.	P	S
Rec 19	EOSC-hub should adopt the Five Safes principles as guidance for the management and handling of sensitive data in the EOSC-hub ecosystem.	P	S
Rec 20	EOSC-hub should develop a technical design framework for Safe Haven services to support the safe data and safe settings dimensions of the Five Safes principles.	TC	L
Support the wider development of ethical and information governance frameworks			
Rec 16	EOSC-hub should consider the logging and tracking of scientific provenance data as an element of service integration design.	TC	L
Rec 17	EOSC-hub should consider convening a technical working group on the topic of recording provenance across the EOSC-hub service ecosystem.	TC	L
Rec 21	EOSC-hub should engage with a broader set of stakeholders, including social science and statistical data service providers, in supporting the design of a Europe-wide framework for research with sensitive data.	P	L
Rec 22	EOSC-hub should consider the feasibility and desirability of developing a code of conduct for sensitive data research.	P	M

8 References

[1] R Baxter *et al*, *D2.8: Guidelines on Open Access and Restricted Data (final)*, EUDAT2020, September 2017, http://doi.org/10.23728/b2share.cd52606e3c21493bb6343bcafc3f5eb2

[2] J Bjaalie, K McGillivray, B Stahl, T Fothergill, *Human Brain Project: Data Policy, Quick Guide*, HBP, March 2018, https://sos-ch-dk-2.exo.io/public-website-production/filer_public/25/2a/252ac9d0-d408-41fc-8fae-bab89dae5ee6/hbp_data_policy_quick_guide.pdf

[3] S Battaglia *et al*, *D3.3: Draft Policy Recommendations*, EOSCpilot, August 2018, https://eoscpilot.eu/content/d33-Draft-Policy-Recommendations

[4] S Hodson, S Jones *et al*, *Turning FAIR into reality*, European Commission Expert Group on FAIR Data, November 2018, https://doi.org/10.2777/1524

[5] ANDS, *Curation Continuum,* ANDS Guide, February 2016, https://www.ands.org.au/guides/curation-continuum

[6] M Wilkinson *et al*, (2016), *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data 3:160018, https://doi.org/10.1038/sdata.2016.18

[7] Osborne Clarke LLP, *Legal study on ownership and access to data*, EU DG-CNECT, 2016, https://doi.org/10.2759/299944

[8] EU, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, April 2016, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[9] L. Sweeney, M. Crosas, M. Bar-Sinai, *Sharing Sensitive Data with Confidence: The Datatags System.* Technology Science [Internet], 2015, http://techscience.org/a/2015101601/

[10] P Doorn & E Thomas, *Tagging Privacy-Sensitive Data According to the New European Privacy Legislation: GDPR DataTags - a Prototype,* International Digital Curation Conference (IDCC), 2018, Barcelona.

[11] OpenAIRE Guidelines for Data Archives, https://guidelines.openaire.eu/en/latest/data/index.html

[12] C Ferguson *et al*, *D3.1 Survey of Current PID Services Landscape*, FREYA project, July 2018, https://doi.org/10.5281/zenodo.1324296

[13] S Wimalaratne & M Fenner, *D2.1 PID Resolution Services Best Practices*, FREYA project, June 2018, https://doi.org/10.5281/zenodo.1324300

[14] M Fenner, M Crosas *et al*, *Listing of data repositories that embed schema.org metadata in dataset landing pages*, March 2018, https://zenodo.org/record/1263942

[15] T Berners-Lee, *Linked Data*, W3C design note, 2009, https://www.w3.org/DesignIssues/LinkedData.html

[16] P Missier *et al*, *D-PROV: extending the PROV provenance model with workflow structure*. In: TaPP; 2013, https://www.usenix.org/system/files/conference/tapp13/tapp13-final3.pdf

[17] T Guedes *et al*, *PROV-Df model and application to Apache Spark (SAMbA)*, 13th Workshop on Workflows in Support of Large-Scale, Dallas, November

2018, https://sc18.supercomputing.org/proceedings/workshops/workshop_files/ws_works111s2-file1.pdf

[18] A Spinuso *et al*, *S-PROV developed in VERCE:* https://www.knmi.nl/kennis-en-datacentrum/project/s-provflow and https://github.com/aspinuso/s-provenance

Appendix I. EOSC-hub Application Profile (draft)

This proposed metadata profile for findability in EOSC-hub differs from the OpenAIRE Application Profile (see https://guidelines.openaire.eu/en/latest/data/application_profile.html) only in the introduction of a new DataTag property. The OpenAIRE guidelines build on the DataCite Metadata Schema v3.1; following OpenAIRE's approach we list the full schema below but only describe the differences recommended for EOSC-hub.

This is not a definitive application profile: a number of recommendations, notably Rec. 12, 13 and 14, suggest further work on metadata properties for EOSC-hub. Technical recommendations from these activities will feed in due course into this application profile.

We follow the same terminology as OpenAIRE:

- **Mandatory (M)** = the field must always be present in the metadata record. An empty element is not allowed.
- **Mandatory when applicable (MA)** = when the value of the field can be obtained it must be present in the metadata record.
- **Recommended (R)** = the use of the field is recommended.
- **Optional (O)** = the property may be used to provide complementary information about the resource.

Property	**Comment**
1. Identifier (M)	
1.1 identifierType (M)	
2. Creator (M)	
2.1 creatorName (M)	
2.2 nameIdentifier (R)	
2.2.1 nameIdentifierScheme (R)	
2.2.2 schemeURI (R)	
2.3 affiliation (R)	
3. Title (M)	
3.1 titleType (O)	
4. Publisher (M)	
5. PublicationYear (M)	
6. Subject (R)	
6.1 subjectScheme (O)	
6.2 schemeURI (O)	

7. Contributor (MA/O)	
7.1 contributorType (MA/O)	
7.2 contributorName (MA/O)	
7.3 nameIdentifier (MA/O)	
7.3.1 nameIdentifierScheme (MA/O)	
7.3.2 schemeURI (O)	
7.4 affiliation (O)	
8. Date (M)	
8.1 dateType (M)	
9. Language (R)	
10. ResourceType (R)	
10.1 resourceTypeGeneral (R)	
11. AlternateIdentifier (O)	
11.1 alternateIdentifierType (O)	
12. RelatedIdentifier (MA)	
12.1 relatedIdentifierType (M)	
12.2 relationType (M)	
12.3 relatedMetadataScheme (O)	
12.1 schemeURI (O)	
12.1 schemeType (O)	
13. Size (O)	
14. Format (O)	
15. Version (O)	
16. Rights (MA)	
16.1 rightsURI (MA)	
17. Description (MA)	
17.1 descriptionType (MA)	
18. GeoLocation (O)	
18.1 geoLocationPoint (O)	
18.2 geoLocationBox (O)	
18.3 geoLocationPlace (O)	
19. DataTag (R)	A tag representing the sensitivity and handling requirements of the data object.

19. DataTag (R)

The definitions proposed here are for consultation purposes and should not at this stage be regarded as normative.

A DataTag is a label indicating a level of protection to be applied to the processing of the tagged data object (occurrences: 0-1). Note that, if a data object has a DataTag, it has one and one only, regardless of how many possible classification bases might be applied (e.g. for an object subject to multiple regulatory frameworks). Which tag the object should carry may be a matter of policy, but the path of greatest risk reduction suggests that the strictest tag suggested should be the one used.

19.1 code (MA)

The colour code of the DataTag (occurrences: 1).

Allowed values, examples, other constraints

Controlled List Values:

- blue
- green
- orange
- red

19.2 basis (O)

The legal, ethical or other security basis on which the data object has acquired this DataTag (occurrences: 0-1).

Allowed values, examples, other constraints

Free text, eg. GDPR

19.3 Handling (MA)

An indication of the technical treatment the tagged data object should receive in terms of storage, transmission or access authorisation (occurrences: 0-1).

19.3.1 storage (MA)

An indication of how the tagged object should be handled when "at rest" (occurrences: 1).

Allowed values, examples, other constraints

If 19.3 Handling is used, 19.3.1 storage is mandatory.

Controlled List Values:

- clear
- encrypt

19.3.2 transit (MA)

An indication of how the tagged object should be handled when being transmitted (either electronically across a network, or physically on a portable storage device) (occurrences: 1).

Allowed values, examples, other constraints

If 19.3 Handling is used, 19.3.2 transit is mandatory.

Controlled List Values:

- clear
- encrypt

19.3.3 auth (MA)

An indication of what type of access authorisation is required for the tagged object (occurrences: 1).

Allowed values, examples, other constraints

If 19.3 Handling is used, 19.3.3 auth is mandatory.

Controlled List Values:

- none
- password
- oauth
- signed

19.4 DUA (O)

A data object may optionally require additional terms for a data use agreement or additional terms of access by a recipient of the object from a repository. It may have attributes that further describe it for reporting purposes.

Currently undefined.

Example

```
1<DataTag code="red" basis="GDPR">
2  <Handling storage="encrypt" transit="encrypt" auth="signed">
3</DataTag>
```