D4.1 EOSC-Life Requirements for hosting/ distributing and access control for sensitive data

Eva García-Álvarez, Christian Ohmann, Gary Saunders, Harald Wagener, Ilaria Colussi, Jan-Willem Boiten, Maria Panagiotopoulou, Mónica Cano-Abadía, Romain David, Irene Schluender, et al.

EOSC-Life

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

D4.1 – Requirements for hosting/ distributing and access control for sensitive data

WP4 – Policies, specifications and tools for the management of data for biological and medical research
Lead Beneficiary: ECRIN-ERIC and BBMRI-ERIC
WP leader: Jacques Demotes and Michaela Th. Mayrhofer
Contributing partner(s): EATRIS, Charité, Lygature, ERINHA, EMBRC, INFRAFRONTIER, EuroBioImaging

Authors of this deliverable: **Eva García Álvarez** (BBMRI-ERIC), **Christian Ohmann** (ECRIN), **Gary Saunders** (EATRIS), **Harald Wagener** (Charité), **Ilaria Colussi** (BBMRI-ERIC), **Jan-Willem Boiten** (Lygature), **Maria Panagiotopoulou** (ECRIN), **Mónica Cano Abadía** (BBMRI-ERIC), **Romain David** (ERINHA), **Irene Schlünder** (BBMRI-ERIC), **Petr Holub** (BBMRI-ERIC), **Michaela Th. Mayrhofer** (BBMRI-ERIC)

Contractual delivery date: **31 July 2021**
Actual delivery date: **30 August 2022**
H2020-INFRAEOSC-2018-2

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

Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Executive summary

WP4 is concerned with policies, specifications and tools for secure management of sensitive data for research purposes and addresses this in general policies and guidelines for sensitive data storage, processing, sharing and reuse for research purposes across Research Infrastructures (RIs).

The particular objective of this deliverable is to establish the requirements for managing controlled access to sensitive data sets within the Life Science AAI (WP5) and to manage and analyse those sensitive data in clouds (WP7), in particular data coming from the life sciences Research Infrastructure (LS RI) demonstrators. These requirements may include GDPR-compliance, geographical location of the data centre, legal jurisdiction, institutional scope, etc.

T7.3 works with the WP7 cloud providers to identify their compliance with these sensitive data handling requirements and T5.1 integrates the sensitive data access controls into WP5 so that a single authoritative source of access can be used across the project. This requirements-gathering draws on technical expertise from the LS RIs in WP4 and from WP5 and WP7 to ensure both the requirements and implementation routes are shared and capture existing national requirements to host/distribute sensitive data (i.e. health and healthcare data, biomedical research data, classified data) and the tools used within the LS RIs to manage controlled access to such sensitive data sets.

Ultimately, this deliverable reports on the different sensitive data types that the LS RIs are working with and possibly hosting and providing access to. Regulations, best practices and standards that apply to all sensitive data types are collected. For completing this deliverable, a questionnaire on sensitive data was taken by each RI, based on the EOSC-Life toolbox data types[1], demonstrating the utility of its categorisation system.

Introduction: EOSC-Life project

The EOSC-Life project brings together the 13 Life Science Research Infrastructures (LS RIs) within the European Strategy Forum for Research Infrastructures (ESFRI)[2] to create a European Open Science Cloud (EOSC)[3] for the life sciences, an open collaborative space for digital life science. The project co-creates and integrates the EOSC federated core, while simultaneously creating, adapting and adopting the services and policies for Open Science that help researchers in the life sciences to manage, publish, analyse and reuse data.

The LS RIs[4] provide access to advanced instruments, research facilities and services helping researchers to describe biology from single molecules to ecosystems and long-term population cohorts. The LS RIs are all distributed organisations, each bringing together national facilities and centres into a connected European entity with harmonised access procedures, aligned quality assurance, and joint FAIR data management. Please see Annex 2 for a short description of each RI.

[1] https://zenodo.org/record/5507324#.YiYA3OjMKUk
[2] https://www.esfri.eu/
[3] https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en
[4] https://lifescience-ri.eu/home.html

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

The EOSC-Life project builds on the outcomes from previous cluster projects: in CORBEL[5], LS RIs established a foundation of collaborative scientific services that support users throughout the execution of scientific projects, from planning and grant applications through to the long-term sustainable management and exploitation of research data. EOSC-Life takes the next step: building on the EOSC foundation, projects are embedding the combined infrastructure capabilities into the scientific workflow of advanced users across the European research area.

The EOSC-Life vision of an open collaborative space for digital life science is central to the user support: scientists in advanced research projects need to access and integrate a broad range of data resources. The use of common metadata specifications, cataloguing and indexing in data catalogues will make LS RI datasets widely accessible across disciplines. The project has established mechanisms to continually align our strategy with those of other entities in the EOSC, other cluster projects (SSHOC[6], ESCAPE[7], ENVRI-FAIR[8], PaNOSC[9]), and the wider life science community (GA4GH[10], RDA[11]) while actively engaging to shape EOSC as a pan-European, interdisciplinary ecosystem.

It is the mission of EOSC-Life to make data resources from LS RIs 'FAIR' and publish them in the EOSC following guidelines and standards. Overall, this will drive the evolution of the RI repository infrastructure for EOSC and integration of the LS RI repositories. EOSC-Life is implementing workflows that cross disciplines and are addressing the needs of interdisciplinary science. Through open hackathons and bring-your-own-data events we will co-create the EOSC. EOSC-Life WP4 tackles the considerations raised by the sensitivity of the data. The sensitivity of the data might derive from its personal nature, from Intellectual Property considerations, Dual Use Research of Concern aspects, application of the Nagoya protocol, etc. In particular, this deliverable focuses on different sensitive data types that are relevant for LS RIs. It presents the different sensitive data types according to previous work done in this WP. In addition, a survey and requirements based on them are also reported here.

Detailed Report on the Deliverable

1. Sensitive data types

In order to provide information to researchers who wish to share and/or use sensitive data in a cloud environment in general, and the European Open Science Cloud in particular, an EOSC-Life WP4 Toolbox[12] has been conceptualised in EOSC-Life. The toolbox does not create new content, instead, it allows researchers to find existing resources that are relevant for sharing sensitive data across all participating research infrastructures (F in FAIR). The toolbox provides links to recommendations, procedures, and best practices, as well as to software (tools) to support data sharing and reuse. It is

[5] https://www.corbel-project.eu/home.html
[6] https://sshopencloud.eu/
[7] https://projectescape.eu/
[8] https://envri.eu/home-envri-fair/
[9] https://www.panosc.eu/
[10] https://www.ga4gh.org/
[11] https://www.rd-alliance.org/rda-europe
[12] https://zenodo.org/record/4483694#.YvUAFuzMLFp

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

based upon a tagging (categorisation) system, allowing consistent labelling and categorisation of resources[13]. The concept and terminology used in the toolbox documentation has been a multidisciplinary consensus building exercise involving senior experts from different RIs and scientific domains (ECRIN, BBMRI, EATRIS, EMBRC, ERINHA, EuroBioImaging).

In its final version, a new dimension was added to simplify the categorisation system. This dimension is "sensitive data type", which perfectly fits with the scope of this deliverable, highlighting the synergies between the different tasks of WP4. Similarly, here we keep the definition of sensitive data used in the toolbox and extracted from Templates for FAIRness evaluation criteria - RDA-SHARC ig[14] glossary. Concisely, it concerns "Information that is regulated by law due to possible risk for individuals and for public and private organisations".

Six different **sensitive data types** have been defined within EOSC-Life WP4:

Personal data

Personal data include information related to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership and data concerning the health or sex life of an individual. These data could be identifiable and potentially cause harm through their disclosure.

Environmental data

Environmental data cover environmental risks (nuclear or other sensitive installations, for example) or environmental preservation (habitats, protected fauna or flora, in particular).

Proprietary data

Proprietary data is generated data or documents that can contain technical or other types of information that is controlled by a firm or public institution to safeguard its competitive edge. It could potentially be protected by copyright, patent or trade secret laws. Under this term intellectual property and licences are also covered as well as embargo.

DURC data

Dual Use Research of Concern (DURC) is life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops and other plants, animals, the environment, materiel, or national security[15]. In case of environmental data, the term DURC should only be used, if there is an explicit reference to this type of life science research (e.g. in the title, abstract or keywords).

Classified information

Classified Information is material that a government body deems to be sensitive information that must be protected. Access is restricted by law or regulation to particular groups of people with the necessary security clearance and need to know and mishandling of the material can incur criminal

[13] https://zenodo.org/record/5507324#.YvUAK-zMLFp
[14] https://zenodo.org/record/3922069#.YvUATOzMLFp
[15] https://osp.od.nih.gov/biotechnology/dual-use-research-of-concern/

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

penalties[16]. The term classified information has been specified, for example, in EU[17] and US regulations[18].

Other sensitive data

Under "other sensitive data" all resources that do not fall under one of the above categories are allocated.

2. Life Science Research Infrastructures and Sensitive Data

In the framework of EOSC-Life, LS RIs work with some of the sensitive data types described above. Hence, to have an overview of which sensitive data type is managed by the different RIs, we leveraged the work done in EOSC-Life WP4 Toolbox and performed a survey, which was completed by six RIs (namely BBMRI, EMBRC, EATRIS, INFRAFRONTIER, ECRIN and EuroBioImaging). It provides information about how they manage sensitive data and what policies and best practices they have in place. Their feedback was critical to assess how their activities align with the sensitive data types defined in the toolbox, especially as the references in the toolbox are not identical to the work performed by the RIs. It should be noted that this deliverable is focused on sensitive data types the RIs are working with and not the ones they provide services for. The circulated questionnaire can be found in Annex 1 and the main results are summarised below and in Table 1:

- The following RIs indicated in the survey that they are working with personal data: BBMRI, EATRIS, ECRIN, ELIXIR, EuroBioImaging. In addition, INSTRUCT, ERINHA and ISBE are likely working with this kind of data, according to the description of their activities (see Annex 2). As it can be extracted from the survey, ECRIN hosts and is a controller for personal data but not for sensitive personal data. Similarly, BBMRI is the data controller of some datasets but, as a general procedure, directly identifying data never leave the biobanks. In the case of EATRIS and EuroBioImaging, data controllers are distributed across federated networks.
- EMBRC and ERINHA handle environmental data. Again, even though they are not reported in the survey, it can be extracted from their mission description (Annex 2) that Emphasis and MIRRI also work with this data type. In the case of EMBRC, it is responsible for sensitive environmental data within the EMO-BON project, while other data is controlled by its member institutions.
- EMBRC is also controlling proprietary data through the EMO-BON project. Other RIs, such as BBMRI or EATRIS, handle proprietary data coming from different sources, such as nodes, partners or users of their services. At the time of this document, ECRIN has no intellectual property considerations.

Question	BBMRI	EMBRC	EATRIS	INFRAFRON-TIER	ECRIN	Euro-BioImaging
Sensitive data types tags from WP4 toolbox						
Is this RI hosting, distributing or controlling access to Personal data?	Yes	No	No	No	ECRIN hosts and is controller for personal data (e.g. for the networks of Clinical Trials Units) but not for sensitive	Yes

[16] https://en.wikipedia.org/wiki/Classified_information
[17] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013D0488&from=EN
[18] https://csrc.nist.gov/glossary/term/classified_information

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

					personal data (e.g. Individual Patient Data).	
Is this RI hosting, distributing or controlling access to Environmental data?	No	Yes	No	No	No	(Microscopy images of environmental samples)
Is this RI hosting, distributing or controlling access to Proprietary data?	Yes	Yes	Yes	No	Yes	No
Is this RI hosting, distributing or controlling access to DURC data?	No	No	No	No	No	No
Is this RI hosting, distributing or controlling access to Classified information?	No	No	No	No	No	No
Is this RI hosting, distributing or controlling access to Other sensitive data?	No	No	No	No	No	

Table 1. Sensitive data types across different RIs.

As part of this deliverable, regulations, guidelines and best practices applying to the sensitive data types that the LS RIs are working with are collected with the aim to provide guidance. To set the requirements for hosting/distributing and access control for sensitive data, the survey collected relevant information from the responding LS RIs. The section below contains these requirements for sensitive data as derived from the survey responses and literature review.

3. Requirements

This deliverable collects recommendations that should be taken into account when hosting or controlling access to sensitive data. Of note, a set of technical capabilities that a cloud provider must have in place to be considered a trustworthy environment are being established in WP7[19]. As part of their work, a questionnaire has been created to be completed by cloud providers, to have an overview of which requirements they have in place. In addition to the general questionnaire, there is another one specific for health-related sensitive data[20]. The present document is meant to complement this work. Thus, to avoid duplication of requirements, only those not listed in the work of WP7 are going to be reflected here.

Hereafter we will provide some recommendations as to how to build a cloud service handling sensitive data. Before offering some guidance, we will briefly explore the existing 'sources' that can be considered, split by sensitive data types.

3.1. Sources

3.1.1. About CLOUD COMPUTING

A cloud service that handles sensitive data should be built on the basis of the following certifications for cloud computing, on which the above-mentioned questionnaires have been built as well:

[19] https://docs.google.com/document/d/1BJG_Df6wM-d3x8fV0v3hfCBWGM3r9bGZlMFP8EWObZA/edit
[20] https://docs.google.com/document/d/1ZMyv0VqQclBN_CxPGE6THOtsK9gwdQdNuGBzF5X_33U/edit#heading=h.gjdgxs

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- BSI C5 2016/2020 (the Cloud Computing Compliance Criteria Catalogue from the Federal Office for Information Security in Germany 'BSI').
- ISO standards (ISO 27001, 27017 and 27018 standards).

These standards are not compulsory for cloud services, but if followed they give a strong baseline security level for cloud computing.

3.1.2. About PERSONAL DATA

- From the legal viewpoint, the cornerstone of legally binding rules about personal data in the EU is the General Data Protection Regulation (GDPR)[21].
  The following parts of the GDPR are relevant for cloud computing:
  - Distinction between anonymised and pseudonymised data. According to GDPR definitions:
    - Anonymised data (Recital 26): "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.*"
    - Pseudonymised data (Article 4): "*means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.*"
    - Thus, only the latter falls under GDPR.
  - Data minimisation principle as embedded in Article 5: "*data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.*"
  - Purpose and storage limitations, both under Article 89.
  - Article 9, on the prohibition for processing special categories of particular sensitive data (Article 9: paragraph 1), excluding those data meeting the requirements exposed in Article 9: paragraph 2.
    - Article 46 about data sharing outside the EEA (European Economic Area), which is highly controversial, especially since the invalidation of the EU-US Privacy Shield as a result of the Schrems II case.
    - Article 89, which contains "Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes."
  - When dealing with sensitive personal data in the EU, it is also important to be compliant with domain-specific regulations. The Clinical Trials Regulation22 is an example, whose purposes are, among others, to improve sharing and to increase transparency of information on clinical trials. The Clinical Trial Regulation is legally binding and directly applicable to EU Member States, such as the GDPR.
- Other sources mentioning ethical requirements to be adopted when dealing with personal data are:

[21] http://data.europa.eu/eli/reg/2016/679/oj
[22] http://data.europa.eu/eli/reg/2014/536/oj

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

  - In the EU: the Charter of Fundamental Rights of the European Union (2012/C 326/02), which is legally binding among EU Member States, and it lists the right of privacy and freedom of expression, among the others (articles 8 and 11)[23].
  - At the international level (Council of Europe): the Convention for the Protection of Human Rights and Fundamental Freedoms[24], better known as the European Convention on Human Rights (mentioning the right to respect for private life at Art. 8); and the Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine, known as Convention on Human Rights and Biomedicine (Oviedo Convention)[25]. Both these sources are international treaties, legally binding for the States that have ratified and implemented them.
  - World Medical Association Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects[26], which is a statement of ethical principles addressed primarily to physicians. It is not legally binding but voluntarily followed.

- Beyond legally binding sources and ethical declarations, there are also standards to be considered with regard to the processing of personal data. A comprehensive collection of standards in clinical research and healthcare has been published in a previous report from this WP[27]. Apart from the standards mentioned in that work, others can be used, depending on the domain and also the data type. For instance, Data Catalog Vocabulary (DCAT)[28] facilitates interoperability between online Data Catalogues. More specific standards are also available, such as ECRIN's Data Centre Certification Programme[29], which implementation of this programme has a positive impact in the clinical trial community[30]; or the GA4GH Passports[31], a standard that allows the identification of researchers and their access permissions across different organisations and clouds. The standards are voluntarily followed.

3.1.3. About ENVIRONMENTAL DATA

- At the EU level, the legally binding Regulation referring to the management of biological resources and research results derived from them is the Access and benefit-sharing (ABS)[32] Regulation, implementing the international Nagoya Protocol. Indeed, the application of the ABS requirements is controlled by the Nagoya Protocol, which is a supplementary agreement to the legally binding Convention on Biological Diversity (CBD). In the EU, detailed rules for the implementation of ABS with regard to register of collections, the monitoring of user compliance, and to best practices can be found in the Implementing Regulation 2015/1866[33].
- In addition to legally binding sources, in the EU there are non-legally binding ones such as a report[34] and a guidance[35] on the provisions and implementation of the ABS. Noteworthy, the

[23] https://fra.europa.eu/en/eu-charter
[24] https://www.echr.coe.int/Documents/Convention_ENG.pdf
[25] https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=164
[26] https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
[27] https://zenodo.org/record/5810612#.Yr20d3ZBxPZ
[28] https://www.w3.org/TR/vocab-dcat-3/
[29] https://ecrin.org/activities/data-centre-certification
[30] https://www.sciencedirect.com/science/article/pii/S2451865416300825
[31] http://bit.ly/ga4gh-passport-v1#passport
[32] http://data.europa.eu/eli/reg/2014/511/oj
[33] http://data.europa.eu/eli/reg_impl/2015/1866/oj
[34] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1548339471740&uri=COM%3A2019%3A13%3AFIN

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

European Marine Biological Resource Centre (EMBRC), which is a European 'research infrastructure' that provides researchers and companies with access to marine organisms and the facilities to study them, has elaborated guidelines for ABS compliance[36].
Moreover, the European Environment Agency (EEA) Data Policy[37] provides guidelines to "*ensure that data is handled in a consistent and transparent manner. EEA aspires to promote the sharing of environmental data. In agreeing to share, data providers need to have assurance that their data are properly handled, disseminated and acknowledged following similar principles and rules across countries and stakeholders.*"

3.1.4. About PROPRIETARY DATA

- At the EU level, it comes into relevance the EU Directive[38] for the protection of copyright and related rights. The directive is legally binding but not directly applicable and its implementation is left to each Member State.
- As non-legally binding source, the EC communication entitled "*Making the most of the EU's innovative potential An intellectual property action plan to support the EU's recovery and resilience*"[39], with special focus on Intellectual Property, must be taken into consideration too.

3.1.5. About DURC DATA

- At the EU level[40], a Union legally binding regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items[41] is in place, thus cloud services dealing with DURC must be compliant with it.
- At the international level, there are international conventions such as the Biological and Toxin Weapons Convention (BWC)[42], the Chemical Weapons Convention[43], the Treaty on the Non-Proliferation of Nuclear Weapons (NPT)[44], and the UNSC Resolution 1540 (UNSCR 1540)[45], which depend on States' implementing legislation.
- In addition to formal international treaties and international law, groups of states may organise themselves to undertake tasks or coordinate policies regarding DURC data. Such States have created common control lists of items, including technology and software. These States' policies are voluntary and not legally binding[46].

---

[35] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.C_.2021.013.01.0001.01.ENG&toc=OJ%3AC%3A2021%3A013%3ATOC
[36] https://www.embrc.eu/embrc-guides-abs
[37] https://www.eea.europa.eu/legal/eea-data-policy#toc-0
[38] http://data.europa.eu/eli/dir/2019/790/oj
[39] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52020DC0760
[40] https://zenodo.org/record/4275835#.YnvNKOhByzW
[41] https://eur-lex.europa.eu/eli/reg/2021/821/oj
[42] https://www.un.org/disarmament/biological-weapons/
[43] https://www.opcw.org/chemical-weapons-convention
[44] https://www.un.org/disarmament/wmd/nuclear/npt/
[45] https://www.un.org/disarmament/wmd/sc1540/
[46] See for example https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/controllists.html, http://zanggercommittee.org/, https://www.wassenaar.org/.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

3.1.6. About CLASSIFIED INFORMATION

- As in the case of the other data types, an EU Regulation[47] could be taken into account in the first place, being a legally binding source. Then, specific binding acts apply to different topics, for instance those concerning Court documents[48] or historical European archives[49].

As each RI has its own structure (for instance, centralised or federated), here some general requirements are described, but each of them must have their own legal advisors to deal with domain-specific matters.

A similar situation is found when it comes to geographical location of the data centres. To address this topic, this WP has already published preliminary results of the mapping of national legal and ethical requirements[50].

3.1.7. Sources in progress

The Data Governance Act, which is currently in progress, aims to address the following situations, as reflected in the proposal text[51]:

- *Making public sector data available for re-use, in situations where such data is subject to rights of others.*
- *Sharing of data among businesses, against remuneration in any form.*
- *Allowing personal data to be used with the help of a 'personal data-sharing intermediary', designed to help individuals exercise their rights under the General Data Protection Regulation (GDPR).*
- *Allowing data use on altruistic grounds.*

The European Health Data Space (EHDS) regulation is also in progress. Even though it is not implemented yet, it aims to address health-specific challenges to electronic health data access and sharing. The main goals of this regulation, are collected in the proposal text[52] and summarised as follows:

- *EHDS will create a common space where natural persons can easily control their electronic health data. It will also make it possible for researchers, innovators and policy makers to use this electronic health data in a trusted and secure way that preserves privacy.*
- *It aims to ensure a legal framework consisting of trusted EU and Member State governance mechanisms and a secure processing environment.*
- *It aims to contribute to a genuine single market for digital health products and services, by harmonising rules, and so boost healthcare system efficiencies.*
- *The EHDS will also promote better exchange and access to different types of electronic health data, including electronic health records, genomics data, patient registries etc.*

3.2. Recommendations

[47] http://data.europa.eu/eli/reg/2018/1725/oj
[48] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32005D0012%2803%29
[49] http://data.europa.eu/eli/reg/1983/354/oj
[50] https://zenodo.org/record/3820390#.Yh4o6OjMKUk
[51] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767
[52] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197&qid=1656585315370

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

On the basis of the mentioned regulations, standards, guidance, policies and best practices and also considering the results based on previous projects[53] and emerging from significant publications[54,55], we can conclude that a cloud service that handles sensitive data should be built considering the following recommendations:

1. An institution handling sensitive data must have a clear governance plan. This implies having a concrete, consistent and detailed schema of the legal basis of the data workflow within the infrastructure, including the collection of data as well as data sharing. As part of this governance plan, different policies must be in place as living documents, being frequently updated. Examples of these policies include:

- Data management plan
- Data management policy
- Use and access policy

2. Some requirements must be met at the collection step for hosting sensitive data:

- Have a clear legal basis on data access and make sure that the contributor (data provider[56]) has one.
- A legal agreement between the data provider and the infrastructure must be in place.
- The data provider should make sure that the infrastructure has organisational and technical measures in place, according to the GDPR[57].

3. Once collected, some security measures must be in place for ensuring a safe storage of the data. Most of them, such us encryption[58], pseudonymisation and computer and network protection are collected in previous reports[59]. In addition, it is of main importance to document all the security measures that are in place.

4. For accessing the data

- An access policy must be in place. It must indicate the mode of access and set responsibilities regarding granting access (as mandated by the data controller in case of personal data).
    - Occasionally, a Data Access Committee, Ethics Advisory Board or Ethics Committee must be aware of the data access process and should supervise it if needed.
- In many cases, a DAA (Data Access Agreement) should be signed between the data user and the data provider. This agreement outlines the access conditions.
- Interoperability standards should be in place for a streamline access to resources. A previous report from this WP can be used as a guideline to improve data interoperability[60].
- The data providers must have rule- or committee- based access control.
- Restricted access should include a way to ensure the veracity of the users' identity. This can be done through different Levels of Assurance (LoA[61]). For data access, LoA higher or equal to 2 is

[53] https://documents.egi.eu/public/ShowDocument?docid=2677
[54] Shabani, Mahsa, Gauthier Chassang, and Luca Marelli. "The Impact of the GDPR on the Governance of Biobank Research." GDPR and Biobanking. Springer, Cham, 2021. 45-60.
[55] Shabani, Mahsa, Adrian Thorogood, and Madeleine Murtagh. "Data Access Governance." The Cambridge handbook of health research regulation. Cambridge University Press, 2021. 187-196.
[56] https://zenodo.org/record/6787119#.YtAeK3ZByzU
[57] http://data.europa.eu/eli/reg/2016/679/oj
[58] https://gdpr-info.eu/issues/encryption/
[59] https://documents.egi.eu/public/ShowDocument?docid=2677
[60] https://zenodo.org/record/5810612#.Yh48G-jMKUl

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

recommended, being higher if direct access or process to pseudonymized or raw data is provided.
- Different measures for ensuring accountability must be in place, such as:
    - Logs of accessed data must be kept and accessible if needed.
    - Users' accession can be tracked just for auditable purposes, and they must be always informed about this tracking process, which has to be clearly documented.
- Secure access must be provided. The EOSC Authentication and Authorization Infrastructure (AAI)[62] is a service that enables secure access to resources, establishing "*a common global ecosystem for identity and access control infrastructures for the EOSC*".

Abbreviations

- RIs: Research Infrastructures
- AAI: Authentication and Authorization Infrastructure
- GDPR: General Data Protection Regulation
- EOSC: European Open Science Cloud
- ESFRI: European Strategy Forum for Research Infrastructures
- FAIR: Findable, Accessible, Interoperable and Re-usable
- DURC: Dual Use Research of Concern
- LoA: Level of Assurance

Delivery and Schedule

D4.1 "Requirements for hosting/distributing and access control for sensitive data." was due for M29 (July 2021) as rescheduled with the 2nd Grant Agreement Amendment.

The delivery has deviated from this deadline due to additional COVID-19 related workload for the co-authors and difficulties that the partners encountered in recruiting additional human resources during the pandemic. As a result, the actual delivery date is M42 (August 2022).

Adjustments

Adjustments made:

None

[61] Burr, William E., et al. "Electronic Authentication Guideline: Recommendations of the National Institute of Standards and Technology-Special Publication 800-63-1." (2012).
[62] https://op.europa.eu/en/publication-detail/-/publication/d1bc3702-61e5-11eb-aeb5-01aa75ed71a1

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Appendices

Appendix 1. RI questionnaire

D4.1	Question	Response format	BBMRI	EMBRC	EATRIS	INFRA-FRON-TIER	ECRIN	Euro-BioImaging
Toolbox sensitive data type: Personal data	Is this RI hosting, distributing or controlling access to Personal data?	YES/NO reply	Yes	No	No	No	ECRIN hosts and is controller for personal data (e.g. for the networks of Clinical Trials Units) but not for sensitive personal data (e.g. Individual Patient Data).	Yes
	If yes, please provide some examples.	Open-Ended Response	Heath data: data controller of CRC cohort, biobanks data (notably, directly identifying data never leaves the biobanks).		Not directly. Data controllers are distributed across the federated network.		ECRIN is not the data controller for Individual Participant Data. Controllership normally remains with the clinical trial sponsor. The sponsors normally are the ones deciding upon access to data arrangements.	Various types of medical imaging data, usually collected, stored and controlled by the Nodes. The Nodes are responsible for collecting patients/volunteers' data sharing informed consent. Shared data are anonymized.
Toolbox sensitive data type: Environmen-tal data	Is this RI hosting, distributing or controlling access to Environmental data?	YES/NO reply	No	Yes	No	No	No	(Microscopy images of environmental samples)
	If yes, please provide some examples.	Open-Ended Response		EMO-BON (https://www.embrc.eu/emo-bon), other biodiversity data anyhow is controlled by EMBRC member institutions	Not directly. Data controllers are distributed across the federated network.		N/A	
Toolbox sensitive data type: Proprietary data	Is this RI hosting, distributing or controlling access to Proprietary data?	YES/NO reply	Yes	Yes	Yes	No	Yes.	No
	If yes, please provide some examples.	Open-Ended Response	Project proposals within	EMO-BON (https://www.embrc.eu/	Sensitive documents of various		ECRIN has not so far been concerned with intellectual property considerations. This	

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

			Negotiator	emo-bon)	kinds with Node and institute partners.		might change if ECRIN starts acting as a clinical trial sponsor.	
Toolbox sensitive data type: DURC data	Is this RI hosting, distributing or controlling access to DURC data?	YES/NO reply	No	No	No	No	No	No
	If yes, please provide some examples.	Open-Ended Response	N/A				N/A	
Toolbox sensitive data type: Classified information	Is this RI hosting, distributing or controlling access to Classified information?	YES/NO reply	No	No	No	No	No	No
	If yes, please provide some examples.	Open-Ended Response	N/A				N/A	
Toolbox sensitive data type: Other sensitive data	Is this RI hosting, distributing or controlling access to Other sensitive data?	YES/NO reply	No	No	No	No	No	
	If yes, please provide some examples.	Open-Ended Response	N/A				N/A	
	Please list here the regulations, guidance, best practices or policies relevant for the data types managed in the RI. If publicly available, kindly provide the web links where relevant documentation can be found?	Open-Ended Response	(1) Security and privacy architecture. (2) BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data (https://www.bbmri-eric.eu/wp-content/uploads/D3.2_Rev1-submitted.pdf)		https://eatris.eu/wp-content/uploads/2020/03/EATRIS_Privacy-Policy-website_updated-20200511.pdf		ECRIN is not directly concerned as it is not acting as a data controller for sensitive data and does not hold or provide access to sensitive data. Nevertheless for clinical trials in general the following would potentially apply: • Charter of Fundamental Rights of the European Union (2012/C 326/02); • Convention for the Protection of Human Rights and Fundamental Freedoms; • World Medical Association Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects; • Directive 2001/20/EC of the European Parliament and of the Council on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use; • Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the	

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

							Application of Biology and Medicine: Convention on Human Rights and Biomedicine. • EU regulation 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 (GDPR). • Council of Europe Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data, CM/Inf (2018)15-final • EU Regulation No 536/2014 of the European Parliament and of the Council on clinical trials on medicinal products for human use, repealing Directive 2001/20/EC (OJ L 158, 27.5.2014)	
	Is this RI involved in one of the EOSC-Life demonstrators that deal with sensitive data?	YES/NO reply	YES	No	Yes	No	No	No
	If yes, please provide the title.	Open-Ended Response	A1228 – Cloudification of BBMRI-ERIC CRC-Cohort and its Digital Pathology Imaging		https://www.eosc-life.eu/d1/		N/A	
	If yes, please provide the data types that are managed within the demonstrator.	Open-Ended Response	Personal data (CRC-Cohort data: https://www.bbmri-eric.eu/scientific-collaboration/colorectal-cancer-cohort/)		Data produced from drug sensitivity screens & protein/ ligands 3D structural interaction screens		N/A	

Appendix 2. RIs descriptions

1. ELIXIR, the European life-science infrastructure for biological information, is a unique initiative that consolidates Europe's national centres, services, and core bioinformatics resources into a single, coordinated infrastructure. ELIXIR brings together Europe's major life-science data archives and connects these with national bioinformatics infrastructures throughout ELIXIR's members, of which there are 22 across Europe. ELIXIR's organisational team ('Hub') is based in the UK, within the European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI). ELIXIR

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

supports users addressing the Grand Challenges in diverse domains ranging from marine research via agriculture to health research and medical sciences. By consolidating the expertise and outputs of such a broad range of institutes, ELIXIR offers researchers access to bioinformatics expertise, data, tools, compute resources and training for their research.

2. BBMRI-ERIC improves the accessibility and interoperability of the existing comprehensive collections, either population-based or clinical-oriented, of biological samples from different (sub)populations of Europe. With more than 20 member states and one international organisation, BBMRI-ERIC is one of the largest research infrastructures in Europe. Its Directory lists over 600 biobanks and over 100 million samples and datasets, accessible through the Negotiator tool. BBMRI-ERIC also supports members through its CS IT, Quality Management and ELSI services. In the context of EOSC-Life, legal and ethical experts support the project on identifying and appropriately addressing ethical requirements as well as to assess policy documents, legal and ethical requirements and conceptualise tools, recommendations, codes that support sensitive data handling, especially in relation to the open calls.

3. EATRIS, the European Infrastructure for Translational Research offers a new collaboration model for fostering innovation. The infrastructure plays a fundamental role in the advancement of knowledge and technology in translational research and drug development. With over 115 leading institutes, across 13 EU Member States, EATRIS provides access to the entire pipeline of academic translational infrastructure and expertise, and optimises the route from discovery to proof-of-concept in medicines development. It provides a new development pathway, open to researchers and companies in need of support for advancing biomedical innovations. EATRIS helps pooling and exploiting the translational academic capacities of the infrastructure in omics technologies to enable researchers to better address the scientific and societal challenges of Personalised Medicine.

4. ECRIN is a not-for-profit intergovernmental organisation that supports the conduct of multinational clinical trials in Europe. Multinational clinical trials provide greater access to patients, facilities and medical expertise; raise methodological standards; enable the sharing of costs, tools and procedures; increase the potential for broad implementation of research outcomes; and prevent duplication of research. However, various trial obstacles – from infrastructure interoperability to regulatory and ethical requirements and management and funding issues – deter many investigators from attempting multinational trials. This is especially the case for independent or academic trials which are more frequently conducted in a single country than industry-sponsored trials. This limited scope means reduced potential impact on global public health. ECRIN provides a means to overcome the above challenges by offering researchers support to prepare and implement multinational trials. Support areas include the preparation of applications for funding, protocol evaluation, trial management, quality assurance and more. The clinical research ecosystem is currently undergoing a digital revolution, with the possibility to reuse data already collected in the context of research projects (trials, cohorts, registries), or healthcare and health systems (electronic health records, hospital data warehouses, health databases) for research purposes.

5. The European Marine Biological Resource Centre (EMBRC-ERIC) is a pan-European Research Infrastructure for marine biology and ecology research. It provides researchers a variety of unique

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

resources that are custom to the demand of the user and is a driver in the development of blue biotechnologies, supporting both fundamental and applied research activities for sustainable solutions in the food, health and environmental sectors. EMBRC brings its expertise in FAIR data management, data integration, data policies, trainings and ABS principles and legislations.

7. European Research Infrastructure on Highly Pathogenic Agents AISBL (ERINHA-AISBL) is the Pan-European distributed Research Infrastructure dedicated to the study of highly infectious emerging and re-emerging diseases classified as Risk Group 4 (RG4). ERINHA brings together European high containment and complementary research facilities and expertise in the field of highly infectious diseases and preparedness to outbreaks of high consequences pathogens. It provides access to the rare resources – Biosafety Level 4 laboratories – to perform excellence-driven research (developing diagnostic capabilities, increasing the understanding of diseases, developing new interventions (e.g. therapeutics, vaccines), helping to translate interventions to the market). ERINHA contributes to the enhancement of the European and global capacity, capability and emergency preparedness in response to global outbreaks.

6. EMPHASIS is the European infrastructure for plant phenotyping. Currently in the preparatory phase, EMPHASIS will enable researchers to use facilities, resources and services for multi-scale plant phenotyping from 2021 onwards. In particular, EMPHASIS will facilitate research on plant performance in different agro-climatic scenarios in Europe to address the challenge of food security in a changing climate. Within the framework of EOSC-Life, EMPHASIS will enable interoperability of plant phenotyping data according to the FAIR principle, making them re-useable for the plant phenotyping community and beyond. In addition, EMPHASIS will develop data analysis tools to address specific scientific challenges related to plant sciences in general, plant breeding, and development of technologies including imaging, robotics and automation. In its activities, EMPHASIS will build on existing initiatives to make data management FAIR such as MIAPPE and BrAPI, which will directly contribute to the development of joint standards and ontologies as well as incorporation of both plant genetic and phenomic information.

8. EU-OPENSCREEN ERIC (EU-OS) is a non-profit research infrastructure, which operates on a global scale. We are funded by European countries (currently Czech Republic, Denmark, Finland, Latvia, Norway, Poland, Spain with Germany as host country) and offer access to academic high-throughput screening facilities and medicinal chemistry groups in these countries. Scientists from academia and industry can implement their screening projects at EU-OS partner laboratories using our European Chemical Biology Library (ECBL), which is composed of 100,000 commercial compounds plus a growing number of compounds collected from academic chemistry groups. Hit optimisation can be done at our medicinal chemistry sites. All compounds will be profiled in a panel of about 20 non-project specific assays, which will deliver extensive information on physico-chemical, cellular toxicity and anti-microbial properties. Compound structures and primary screening data will be made public in the European Chemical Biology Database (ECBD), with the possibility to request an embargo period of up to three years so that data can be patented or published. The primary activity of EU-OPENSCREEN is to provide access to its distributed research infrastructure for scientists seeking a better understanding on how fundamental molecular processes act to govern biological function at

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

the organismal, tissue, cellular and pathway levels. EU-OPENSCREEN will extend the application of Chemical Biology and will develop novel research 'tools' for all fields of the Life Sciences, incl. molecular, cell, plant, structural and micro-biology; synthetic and medicinal chemistry; pharmacology and early drug discovery. EU-OPENSCREEN will drive quality and consistency in the generation and analysis of biological data by creating consistency in the process of target validation through the application of agreed experimental and data analysis standards across its network.

9. Euro-BioImaging's mission is to allow imaging research to flow unhindered, thereby ensuring excellent research and development across the life sciences. In practical terms, Euro-BioImaging offers all scientists – regardless of affiliation, area of expertise, or field of activity – open access to imaging instruments, expertise, training opportunities, and data management services that they do not find at their home institutions or among their collaboration partners. Euro-BioImaging has 15 members hosting 21 Nodes, located across 9 countries and EMBL and offers access to over 40 different imaging technologies. In terms of data, Euro-BioImaging offers a wide range of image data services – the Nodes support their users in data management and quality control, primary data analysis and the transfer of large data sets to the home base. In addition, collaboration with the BioImage Archive (www.ebi.ac.uk/bioimage-archive/) which was launched as public central archive for biological and biomedical image data, which will make it easier for researchers around the world to store, share, access and analyse images. This wealth of scientific images can now start to be reused, reanalysed, and interconnected to create new knowledge. New computational tools for image analysis and processing accessible via the cloud will complement the offered services. With the advances in modern imaging technologies, the wealth of digital imaging data is rising exponentially. While this will lead to exciting breakthroughs, new challenges are also presenting themselves such as image data storage and interoperability, which are a key-focus of Euro-BioImaging.

10. INFRAFRONTIER is the European Research Infrastructure for the development, phenotyping, archiving and distribution of genetically modified mouse models for basic biomedical research. It is our mission to provide unique scientific resources, services, and expertise to advance the understanding, prevention and treatment of human diseases using rodent models (mouse and rat). In EOSC-Life, INFRAFRONTIER is cooperating with many partners on the FAIRification and cloud deployment of animal model resources (e.g. IMPC data). Seven of our core consortium partners are involved in publishing FAIR RI resources (WP1). These include heterogeneous types of data and workflows which enable applications in various fields of work within and beyond the INFRAFRONTIER research infrastructure. One of the key tasks is to enhance the usability of RI resources for a broad user community and to create links to research areas with human data. Therefore, we are developing user driven analysis workflows in WP2.

11. Instruct-ERIC is the single point of access to technology and expertise for structural biology research. Through its specialist research centres in Europe, Instruct-ERIC offers funded research visits, training, internships and R&D awards. By promoting integrative methods, Instruct-ERIC enables excellent science and technological development for the benefit of all life scientists. Instruct-ERIC is utilising its cross-research infrastructure support tool, ARIA, to support the project

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

activities and enhance integration with other research infrastructures in the life-sciences domain. As co-leads of WP5, Instruct-ERIC will also be supporting the delivery of Life Science Login, a single sign-on technology that will allow for seamless data sharing across multiple life-science domains.

12. ISBE – Infrastructure for Systems Biology Europe – is a coordination effort to interconnect the best experimental and modelling facilities for Systems Biology in Europe. ISBE provides stewardship and insight into biological data and their acquisition. It is composed by one pillar, FAIRDOM, and three national candidate nodes: ISBE.NL, ISBE.SI, ISBE.IT. Each node provides different services to the user community, and has its own budget, but ISBE Italy is in charge of the interim coordination of the ESFRI project ISBE (ISBE Europe). FAIRDOM is the Data and Model Stewardship pillar of ISBE prep phase (http://fair-dom.org). FAIRDOM platform and resources support Systems Biology to make Data (models, data, SOPs, samples, workflows) FAIR, and generalisable to any kind of FAIR management. The FAIRDOM has been integrated with national data e-Infrastructures in Norway (NeLS) and Slovenia (p-ISA).

13. MIRRI / CIRM (Centre International de Ressources Microbiennes) is a network of five Biological Resource Centres of the microbial domain (mBRCs). This network belongs to the Institut national de la recherche agronomique (INRA), the leading European agricultural research institute and the world's second largest in agricultural research; INRA's research is dedicated to food, nutrition, agriculture and the environment. CIRM preserves the microbial resources generated by a large number of projects conducted by INRA since its creation in 1946. As a distributed infrastructure, CIRM is composed of five thematic mBRCs: food bacteria, human and animal pathogenic bacteria, plant-associated bacteria, filamentous fungi and yeasts of biotechnological interest. It currently preserves and distributes more than 22,000 isolates. The majority of these isolates are natural, but CIRM also maintains mutants and reference strains for research. This unique set of biological resources with which many data are associated (date of access, geographical origin, substrate of isolation, phenotypic, molecular and taxonomic characteristics, etc.) has a socio-economic impact in fields as varied as food fermentations, white and green biotechnologies as well as the control of plant pathogens and bacterial infections in animals. As a supplier of authenticated and reliable strains with their associated data, CIRM participates in research partnerships with university and private scientists. In addition to the conservation and distribution of strains, CIRM offers a wide range of services to the scientific community (identification and characterization of strains, molecular typing, confidential security deposit, field collection, etc.). The expertise of CIRM's five mBRCs covers taxonomy, genomics, biotechnology, etc. CIRM operates according to ISO 9001 certified processes to (i) ensure the quality of strains, associated data and services provided and (ii) provide assurance that the regulatory and legal framework is respected.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.