EOSC-Life

EOSC-LIFE: BUILDING A DIGITAL SPACE FOR THE LIFE SCIENCES

EOSC-LIFE WP14:
COVID-19 Repository Data Sharing Policy

WP14 – **Design, development, implementation and use of a repository for individual participant data from COVID-19 trials**

Lead Beneficiary: **ECRIN, UiO**
Date of publication: **20 September 2021**

Grant agreement no. 824087
Horizon 2020
H2020-INFRAEOSC-2018-2
Type of action: RIA

Authors: **Christian Ohmann, Mihaela Matei, Steve Canham**
With contributions from: **Maria Panagiotopoulou, Jacques Demotes, Gard Thomassen, Niklas Blomberg**

Table of Contents

Introduction

EOSC-Life (https://www.eosc-life.eu/) is a 4-year project funded under the European Union's Horizon 2020 programme. The project kicked off in March 2019 and aims to bring together the 13 Research Infrastructures in the Health and Food domain of the ESFRI Roadmap to create an open, collaborative digital space for life sciences in the European Open Science Cloud (EOSC). EOSC-Life will achieve this by publishing life science data and tools as FAIR Data Resources, linking reusable tools and workflows to standardised computing services in national life-science clouds, connecting users across Europe to a single login authentication and resource authorisation system, and developing the data policies needed to preserve and deepen the trust given by research participants and patients volunteering their data and biosamples.

In March 2020, the World Health Organisation (WHO) assessed that the COVID-19 outbreak can be characterised as a pandemic. In a rapid response to the COVID-19 pandemic, the European Commission announced a set of short-term coordinated research and innovation actions at EU level, the ERA vs. Corona Action Plan. The Action Plan set out key measures to coordinate, share, and increase support for research and innovation. For EOSC-Life, this resulted in the addition of two work packages:

- WP13 - Extension of COVID-19 Data Portal: core portal development and operationalisation of the extended portal through deep support and integration services

- WP14 - Design, development, implementation and use of a repository for individual participant data from COVID-19 trials

Many scientific organisations, funders and initiatives have expressed their commitment to more open scientific research. This "cultural shift" has been extended to also include clinical research and in particular clinical trials. The results of clinical trials are increasingly considered as a public good, and access to the individual participant data (IPD) generated by clinical trials for further research or healthcare is viewed by some as a fundamental right of the participants and respect to their contribution as it helps to avoid unnecessary repetition of studies. Especially in times of a pandemic crisis, there is an urge for timely and accurate collection, reporting and sharing of data within and between research communities, public health practitioners, clinicians and policymakers to inform assessment of the severity, spread and impact of a pandemic and to implement efficient and effective response strategies.

In EOSC-Life WP14, the European Clinical Research Infrastructure Network (ECRIN, https://ecrin.org/) has partnered with the University of Oslo (UiO, https://www.uio.no/english/) in order to commonly design, develop, implement and operate a repository for individual participant data from COVID-19 clinical research studies that is compliant with European regulations and in particular with the GDPR. For the storage of the IPD, the TSD infrastructure of the University of Oslo (https://www.uio.no/english/services/it/research/sensitive-data/) will be used, which is a multi-tenant remote access system with a strong set of built-in security measures. The repository will be part of the EU COVID-19 data portal (https://www.covid19dataportal.org/).

The following sections highlight the Data Sharing Policy of the COVID-19 repository under development in EOSC-Life WP14.

EOSC-Life WP14 COVID-19 Repository Data Sharing Policy

Purpose

This document describes the policy framework of the repository with respect to the management of data objects within it. It first clarifies the interpretation of several key terms, and then sets out the organising principles of the repository's approach. It continues with a high-level description of the options that will be made available for secondary use of data objects (or 'data sharing'), the ways in which data transfer agreements will underpin later data object use, and the data sharing process itself.

This policy document is designed to complement any specific contractual agreement between the repository and other organisations, whether for data object transfer or for secondary use. It is also designed to act as a basis for more detailed documents (e.g. Standard Operating Procedures) that provide information on how the policy should be interpreted in practice.

This version

1.0: 19th July 2021

Authors

Christian Ohmann, Mihaela Matei, Steve Canham

With contributions from: Maria Panagiotopoulou, Jacques Demotes, Gard Thomassen, Niklas Blomberg

Approval by the COVID-19 repository Steering Committee

Member	Date of approval
Jacques Demotes	15/07/2021
Gard Thomassen	21/04/2021
Mihaela Matei	19/04/2021
Niklas Blomberg	17/06/2021

A document revision history is available at the end of the document

Terminology

Many of the terms used in discussions of secondary use have multiple meanings depending on context. It is therefore necessary to clarify how such terms are used in this document. In particular:

1. ***A Data Object*** is any file available in electronic form, of *any type* (document, data, media etc.). The repository stores and makes available *data objects*, not simply 'data' (or datasets, as defined below), because it is designed to contain protocols, analysis plans, consent forms, and other documents associated with a clinical research study, as well as data, including individual participant data (IPD). Thus, throughout this document, reference is made to data object transfer, data object secondary use, data object storage etc.

2. To try and clarify things further the term ***Dataset*** is used when referring to a data object that contains only data – e.g., a spreadsheet, CSV, JSON or XML file, database dump, etc. In the context of the repository, 'dataset' (or datasets) will usually refer to the file or files of individual participant data derived from a clinical research study.

3. A ***Data Object Provider*** (or simply 'the Provider') is an organisation that provides data objects to the repository, i.e. that enters into a Data Transfer Agreement with the repository. Unless those data objects are already explicitly in the public domain, the Data Object Provider is assumed to have the legal power to enter into that agreement, for instance they would hold the copyright or intellectual property rights on data objects. For datasets of sensitive personal data, the Provider would be the Data Controller as defined under the GDPR.

4. The Data Object Provider must be a legal entity, and thus able to enter into agreements such as a Data Transfer Agreement or Data Use Agreement. The individuals who collected and analysed the data – the researchers – are here called the ***Data Object Generators***. The same distinction applies to secondary use – the individuals seeking to re-use data objects are called ***Data Object Secondary Users***, (or simply the 'secondary users') the organisation arranging such re-use on their behalf, normally their employer, is the ***Data Object Requester***.

5. The transfer of the data objects to the repository will be governed by a ***Data Transfer Agreement (DTA)***, with that document covering the transfer of ***all*** data objects, including but not limited to datasets. This agreement will reference the GDPR and associated legislation, but also intellectual property law. Each DTA will also have an appendix describing the data objects to be transferred to the repository, and – if datasets – whether they are categorised as anonymised or pseudonymised. Further details (e.g. the exact access arrangements desired for each object) will be required within object metadata. Note that a ***Data Use Agreement (DUA)*** (also known as a Data Access Agreement) will normally only apply to datasets. A model Data Transfer Agreement for further discussions with Data Object Providers is available on request (contact maria.panagiotopoulou@ecrin.org).

6. The term '***Data Sharing***' is used as a shorthand for the secondary use of any data object.

7. The term '***Repository'*** used in the policy refers to the repository as a whole, i.e. both the storage infrastructure and the systems in place to manage content, liaise with Data Object Providers, Requesters, etc. In the context of EOSC-Life it includes both ECRIN and TSD.

A. Principles

The following principles form the basis of the repository's operation:

1. The Data Object Providers retain control over all the data objects they transfer to the repository. In particular each Provider stipulates the access arrangements to be applied to their objects, identifying those that will be available without restraint, and those that will be under a controlled access regime. For objects under controlled access, the Provider can set fixed prerequisites for access, or reserve the right to review and grant access on a case by case basis. For datasets falling under the GDPR or related legislation the Data Object Provider will remain the Data Controller. For data objects that are explicitly put into the public domain the Data Object Provider should specify the licence under which such transfer is made.

2. The repository securely stores and processes data objects on behalf of the Data Provider but does not take independent decisions about making data objects that are under controlled access available to Requesters. If, however, the Data Object Provider has stipulated clear criteria for allowing access, the repository can check a request against those criteria and allow access if the criteria are met, in effect relieving the Data Object Provider / Generators of the burden of dealing with such requests.

3. Transfers of all data objects will be the subject of Data Transfer Agreements (DTAs), between the Data Object Provider and the repository, that will specify the data objects involved, specify the roles and responsibilities of the organisations involved, and also check the commitment of the Data Object Provider to provide the necessary metadata and de-identification for datasets. A model Data Transfer Agreement for further discussions with Data Object Providers is available on request (contact maria.panagiotopoulou@ecrin.org).

4. Controlled access to datasets will be managed using Data Use Agreements (DUAs) to ensure that suitable safeguards are in place. The repository will provide a standard or default Data Use Agreement, though individual Data Object Providers may negotiate amendments to that, to create their 'own' standard DUA. In the context of a specific request, the standard agreement for the relevant Data Object Provider will be used by the repository. The repository will not enter into negotiations to change this DUA - if the Data Requester is unhappy with the (Provider specific) DUA offered by the repository they will need to negotiate directly with the Provider.

5. The repository will hold both anonymised and pseudonymised data, *with the classification being made by the Data Object Provider*. For pseudonymised data the repository will *not* hold any linking or identifying data. From the point of view of the repository, and any Data Object Secondary Users, the pseudonymised data they hold or obtain, which will be de-identified as described below, will therefore be anonymous for practical purposes.

6. Datasets, whether legally classified as pseudonymised or anonymised, should be de-identified and a brief description of the de-identification process should be supplied. The level of de-identification required should, as a minimum, follow the guidelines provided by the US HIPAA rules, usually with the addition of date re-basing and the removal of narrative text fields. Although the repository will provide support services the final responsibility for implementing de-identification will be with Providers. Data Object Providers will also be encouraged to

provide the results of a re-analysis of the de-identified data to ensure that the conclusions of the original analysis are still supported.

7. To ensure FAIRness of data objects, all such objects should be linked to Discovery – Access – Provenance (DAP) metadata. The repository will provide support services, making use of the ECRIN DAP metadata scheme[1], to ensure such DAP metadata is present and then distributed (e.g. within ECRIN's Clinical Research Metadata Repository). Data Object Providers will also be encouraged to ensure originating studies, including observational studies, are registered, which will make the provision of DAP metadata much easier.

8. Datasets should be associated with detailed descriptive metadata, describing the individual data points. Ideally this would be in a standardised format (e.g. CDISC define XML) but may be a simple spreadsheet based 'data dictionary'. The repository should reserve the right to insist on descriptive metadata being available – usually as a public data object, even if access to the dataset itself is controlled.

B. Options for secondary use

The repository will aim to provide a range of different access arrangements, to maximise the choice available to Data Object Generators / Providers, especially given the controlled access that many Providers prefer to apply to IPD datasets. The repository will therefore offer the following options:

1. **Public access, freely available data objects**
   Many data objects, documents in particular, are expected to be made freely available. No special procedures are necessary for these data objects – there should be a link to them from the metadata, and that link should lead to the document. Normally the document should be visible in the browser and downloaded from there.

2. **Public access, after the users have identified themselves and their purpose**
   In order to provide better feedback, to both the Data Object Generators / Providers and the Repository itself, users downloading public data objects are first asked to identify themselves – their names, emails, affiliations, and also briefly summarise the reasons they are using the material. Once such information is provided the data object is freely available.

3. **Data objects with controlled access, managed solely by the Data Provider**
   This may be applicable, for example, if Data Object Providers have and wish to use their own Data Access Committee. For these Data Objects, users wishing access would be told, via the associated DAP metadata, to contact the Data Object Provider directly, (normally, specifically, the Data Generators).

4. **Data objects with controlled access, managed by the Repository**
   This is expected to be the most common option for managed access objects, though the details will depend on the exact stipulations of the Data Object Provider. A Data Object secondary user would first contact the repository, which checks if the prerequisites stipulated

[1] Canham, Steve. (2020). ECRIN Metadata Schemas for Clinical Research Data Objects Version 5.0. Zenodo. doi: 10.5281/zenodo.4133889. Available at https://zenodo.org/record/4133889

by the Data Provider have been met (e.g. there is a protocol describing the proposed use of the datasets).

If requested by the Data Object Provider within the original Data Transfer Agreement, and assuming the prerequisites were clear enough for their fulfilment to be unambiguous, the repository could then grant access to the data object(s). Alternatively, it could pass the request back to the Data Object Provider for their final decision, perhaps with a recommendation from a repository based Data Access Committee.

The next two options can be offered in combination with any of the four above (and / or with each other).

5. **Imposition of an embargo period**
   Data Object Providers will be offered the option of setting an embargo period. This could extend either to a fixed date, or until a pre-specified condition for lifting the embargo is met – for example once the primary paper is published. In either case *the embargo period should not exceed two years from the date the data objects were transferred to the Repository.* After the embargo is lifted the data objects would be subject to one or other of the access methods described above.

6. **On screen access only (for datasets, with analysis tools being available)**
   While the preferred option (because it is easier for users to process the data, and requires less work from the repository) is for data objects to be downloaded by secondary users, some Data Object Providers may insist on datasets only being viewed and processed within the TSD infrastructure, i.e. with file downloads prevented. Any analysis will have to be done *in situ*, using whatever statistical and data tools are available.

C. Data object transfer

The Data Object Transfer Process, and the associated Agreement(s), are key to both maintaining the quality of the repository's contents and in establishing the data object access regime for each data object transferred. To support this process, in outline:

1. Any initial enquiries need to be met with a clear explanation of the Data Object Transfer procedure, including the need for the provision of metadata.

2. The collection of DAP metadata for each data object (and for the generating study or studies) is a necessary first step in the Data Object transfer process, as it characterises the objects to be transferred.

3. For each object, the access regime to be applied should be clearly indicated. For controlled access regimes, the nature of any prerequisites to be checked by the repository should be made clear.

4. For datasets with personal data, the legal status of the datasets, (anonymised or pseudonymised) should be made clear by the Data Object Provider. In either case, such

datasets should be de-identified and their de-identification described. The repository will organise support, as necessary, for the de-identification process.

5. The data object transfer should be governed by a Data Transfer Agreement between the repository and the Provider, that will cover all data objects (not just the datasets). Even when data objects have been explicitly placed in the public domain the Data Transfer Agreement should indicate the specific licence that applies to them. The specific access regime and other details described above should be included within supporting appendices to each Data Transfer Agreement. A model Data Transfer Agreement for further discussions with Data Object Providers is available on request (contact maria.panagiotopoulou@ecrin.org).

6. The details of the data transfer process and the access requirements for each data object need to be stored in the Repository Management System and available to all repository staff. This system is used to direct and record the granting of the necessary access to upload data objects and metadata.

7. Once uploaded, the Data Objects will need to be checked by repository staff to ensure that they comply with the DAP metadata description, and that datasets have sufficient descriptive metadata and appear to be de-identified as described. Once those checks are successfully passed the DAP metadata for the data objects can be made public.

D. Secondary use

The details of the secondary use process will depend on the stipulations of the Data Object Provider, as written within the Data Transfer Agreement. Public access and download should be recorded by the repository, as part of its ongoing monitoring of use. For managed access where the repository is involved:

1. The repository will check if the stipulated prerequisites have been fulfilled by the Data Requester / Secondary Users. If previously instructed to do so, it will also pass the request to a Data Access Committee for their recommendation. If (but only if) the fulfilment of the Data Object Provider's requirements are clearly met the repository can make the requested data objects immediately available to the Secondary Users. If not, the request will have to be relayed to the Data Object Provider for their final decision.

2. Once access is agreed, the repository will offer a Data Use Agreement governing the secondary use of the datasets. This will, amongst other things, impose restrictions on usage – normally only for the task explicitly described by the Data Object Secondary Users – and on any further dissemination. To save time and effort, the repository will offer only limited flexibility in the wording of the Data Use Agreement – the intention is to develop and use a standard agreement, for all Data Object Requester organisations.

3. The progress of the secondary use process needs to be stored in the Repository Management System and available to all repository staff. This system is used to direct and record the granting of the necessary rights to access and / or download data objects.

4. The results of any secondary use should be actively followed up by the repository, in part to ensure compliance with the Data Use Agreement, in part to contribute to the ongoing

monitoring of repository usage and impact. Secondary users will be expected to report any publication resulting from their re-use of the data objects.

E. Document history

V0.1: 12/04/20201
Initial version in this format, based on earlier papers by Christian Ohmann and Mihaela Matei

V0.2 13/04/2021
Some rewording to try to clarify in response to feedback. Data Object users renamed to Data Object Secondary Users, or just secondary users. Reference to possible costs removed.

V0.3 13/04/2021
Comments removed and changes accepted where queries resolved. Bullet points replaced by numbered paragraphs in each section. A few queries remain to be resolved.

V0.4 13/04/2021
Remaining comments deleted and all changes accepted for a clean document.

V0.5 16/04/2021
Final revisions to policy text following internal discussion. A5 and D5 rewritten to reflect revised terminology for transfer agreements. A7 added to clarify 'repository'. Additional Reference added for the ECRIN metadata scheme.

V0.6 19/04/2021
Terminology made consistent throughout the document. Typographic corrections.

V1.0 19/07/2021
Initial agreed version, after approval. (Data object) Depositor replaced in all instances by (Data object) Provider, to bring the DSP in line with the DTA.