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

## D1.2 – EOSC repository deployment for project demonstrators

WP1 – Publishing FAIR RI data Resource
Lead Beneficiary: Fraunhofer, EBI
WP leaders: Philip Gribbon and Helen Parkinson
Contributing partner(s): EMBL, BBMRI, EATRIS, ECRIN, EMBRC, EMPHASIS (FZJ), ERINHA, INFRAFRONTIER, UNIMIB, INRA, UNIVDUN, HMGU, CERBM, BSCRC, UOULU, CIRMMP, CSIC, VU, KNAW, UVEG, USMI, IMG, CNRI, UNIMAN, LUMC

Authors of this deliverable: **Helen Parkinson and Phil Gribbon, all authors listed below.**

Authors of this deliverable:

WP1: Helen Parkinson (EMBL-EBI, ELIXIR), Ugis Sarkans (EMBL-EBI, ELIXIR, Euro BioImaging), Phil Gribbon, Gesa Witt, Andrea Zaliani, Manfred Kohler (Fraunhofer, EU OPENSCREEN), Jason Swedlow, Jean-Marie Burel (UNIVDUN, EUBI), Morris Swertz, Esther van Enckevort (UMCG, BBMRI), Petr Holub (BBMRI ERIC), Marzia Massimi, Rafaele Matteoni (CNR, INFRAFRONTIER), Holger Maier(HMGU, INFRAFRONTIER), Reetta Hinttala, Anne Heikkinen (UOULU, INFRAFRONTIER), Philipp Gormanns (INFRAFRONTIER GMBH, INFRAFRONTIER), Laura del Cano, Laurent Vasseur, Sophie Leblanc, Yann Herault (CERBM-GIE, INFRAFRONTIER), Dimitris Kontoyiannis, Christina Chandras, Dimitra Panou (FLEMING, INFRAFRONTIER), José Miguel López Coronado, Rosa Aznar Novella (UVEG, MIRRI), Vincent Robert, Ammar Ben Hadj Amor (KNAW, MIRRI), Serge Casaregola, Jean-Luc Legras, Michel-Yves Mistou (INRA, MIRRI), Paolo Romano (USMI, MIRRI), Isabelle Perseil (INSERM, ERINHA), Romain David (ERINHA), Roland Pieruschka (FZJ, EMPHASIS), Katrina Exter, Marc Portier, Cedric Decruw (VLIZ, EMBRC), S. Canham, C. Ohmann, S. Goryanin (ECRIN-ERIC), Laura Del Cano (CSIC, Instruct), Maddalena Fratelli (IRFMN, EATRIS), Carole Goble, Stuart Owen, Stian Soiland-Reyes, Nick Juty (UNIMAN, ISBE), WP6: Susanna Sansone and Peter McQuilton (UOXF), Marco Roos (LUMC), Luiz Bonino (LUMC)
WP2: Carole Goble, Nick Juty (UNIMAN)

# Table of Contents

# Executive Summary

This deliverable addresses WP1's objectives related to publication of data and data resources in cloud repositories by making available resources in EOSC-Life's registries; the implementation of FAIR services and standards, by development of FAIR registries in collaboration with WP6; and evolution of repository infrastructure by making available registry entries deriving from WP3 demonstrators. Use cases for EOSC-Life registries have been defined for general and clinical research data use and the 'EOSC-Life' collection in FAIRsharing has been populated and released comprising more than 100 data resources and standards. A clinical MetaData Repository has also been developed to address challenges of clinical data sharing in Europe. We have engaged with EOSC, sister projects such as FAIRsFAIR and ENVRI to consider the features of registries, cross registry interoperability and use cases which link EOSC projects. We will continue to update the registries as the project progresses and as new dataset and resources are made available and will engage with pan-EOSC registry planning based on our work to date.

# Project Objectives

This deliverable has contributed directly to the following WP1 objectives:

- WP1 Objectives Assessment of cloud feasibility of data / data resources and publication of these repositories in EOSC for data reuse
- Implementation of FAIR services and WP1 standards for RI data resources and associated Demonstrator projects
- Advance the evolution of RI repository infrastructure for EOSC (sustainability) and the interfaces between the repositories for RI demonstrators and open calls

# Detailed Report on the Deliverable

## 1. Introduction

This work addresses the requirement for EOSC-Life data resources to be accessible through a data catalogue or registry (we are using these terms as synonyms, EOSC speaks of registries, but the term 'Catalog' is more common in the life sciences) for access by users. The work undertaken includes a landscape assessment of registries in EOSC-Life and an assessment of the wider EOSC landscape as well as a consideration of how registries can themselves be FAIR. Specifically: the criteria for choosing a registry and the sustainability and interoperability of registries which have informed our development choices. A major challenge for EOSC-Life is the breadth of the different research infrastructures (RIs) involved in EOSC-Life and the differences between these, highlighted in Figure 1. In the project proposal, we proposed the delivery of an instance of

OmicsDI for molecular data for EOSC-Life. However, on the development of the WP1 roadmap, it became clear that the data types constituting EOSC-Life were more diverse than can be supported by a registry which supports primarily molecular data types. Additionally, many of EOSC-Life's RIs already have a well developed work plan for their registries and, in some cases, these have been operational for many years, for example BBMRI's Sample Directory[1]. Further, that deep representation (e.g. down to assay types) would be challenging across the whole project and would also be impossible for the clinically focussed infrastructures (particularly ECRIN) to share data meaningfully at this granularity. We therefore started from the use cases shown (Table 1) and re-evaluated the implementation plan based on these.
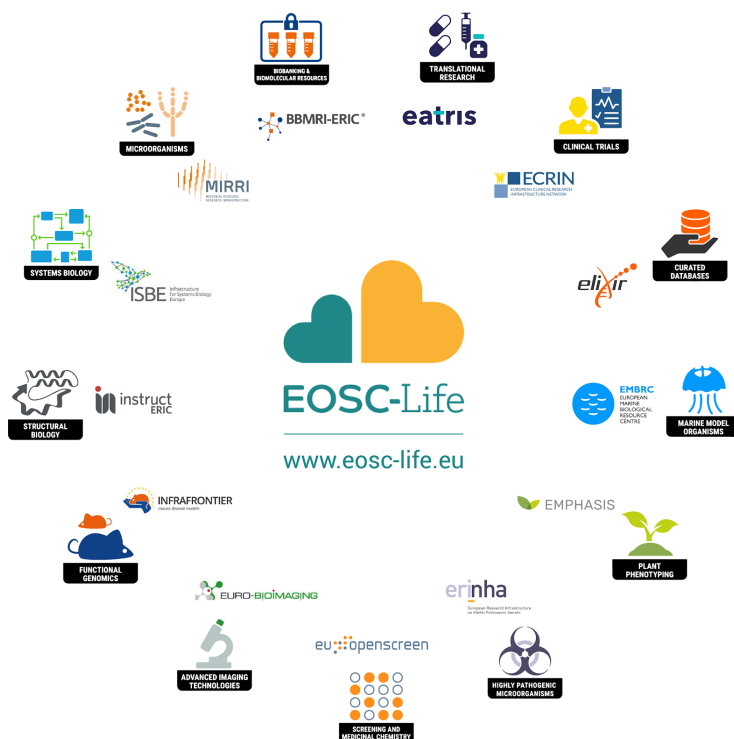


*Figure 1: A representation of the complexity and diversity of research infrastructures data types in EOSC-Life*

## 2. Description of Work

### 2.1. Landscape Analysis and Use Cases for EOSC-Life and wider EOSC

EOSC-Life is one of many EOSC projects and its unique feature is that it brings together delivery and deployment of cloud hosted data resources across many projects and domains aggregated via the RIs. It also delivers the eight EOSC-Life demonstrators, which bring together the RIs in a new way of working around a specific implementation scenario. EOSC-Life also delivers new projects

---

[1] https://www.bbmri-eric.eu/services/directory/

from WP1 and WP3 funding calls and all of these will bring new use cases and have variable stages of technical and content maturity. Therefore, in considering the use cases for this deliverable, we evaluated the diverse current and future needs as well as the sustainability of the work after EOSC-Life. The primary use cases and identified stakeholders for this deliverable are described in Table 1 below.

| Use case | Stakeholder |
|---|---|
| Provide cross EOSC-Life and cross EOSC data resource visibility | RIs, resource providers, funders, data owners |
| Provide support for data resources representing the diversity of EOSC-Life, including Clinical Data | EOSC-Life, data owners, resource providers, funders |
| Provide support for records of different type and different granularities | EOSC-Life, data owners, resource providers |
| Provide a trusted service with user support | EOSC-Life, data owners, resource providers |
| Provide a sustainable registry for long term use (>5 years) | Funders, registry developers, resource providers, data owners |
| Provide a simple and branded means of identifying EOSC-Life resources and data | EOSC-Life project, data resources, EOSC |
| Provide a strategy and means of interoperating across different repositories | Registry owners, EOSC projects, data owners, resource providers, standards bodies and developers (RDA, W3C etc) |
| Registry itself should constitute a FAIR resource | RIs, resource providers, funders, data owners |
| Application Programming Interface (API) access to data within the registry | Registry users, registry developers |
| Population should be simple, rapid and regularly updated during EOSC-Life | Resource providers and data owners, registry developers |

*Table 1: Primary use cases for D1.2 showing stakeholders*

| Use Case | FAIRSharing feature |
|---|---|
| Provide cross EOSC-Life and cross EOSC data resource visibility | A wide range of data types from life sciences and non life sciences domains are supported. FAIRsharing is an ELIXIR Recommended Interoperability Resource (RIR) and has visibility beyond EOSC-Life, its developers are active in international standards organisations including Research Data Alliance (RDA)[2]. |
| Provide support for data resources representing the diversity of EOSC-Life | Life sciences across all EOSC-Life domains are supported. |
| Provide support for records of different type and different granularities | FAIRsharing supports data resource records, ontologies, models, formats, identifier schema, reporting guidelines and data policies; counts and visualisations of these are available. |
| Provide a trusted service with user support | A user support email and helpdesk service is provided. |
| Provide a sustainable registry for long term use (>5 years) | FAIRsharing has existed since 2011 and we estimate that the resource has longevity as it is an ELIXIR Recommended Interoperability Resource (RIR) and content is continuing to grow[3] |
| Provide a simple and branded means of identifying EOSC-Life resources and data | FAIRSharing provides groups of records customised for projects and consortia |
| Provide a strategy and means of interoperating across different registries | FAIRsharing supports schema.org and BioSchemas[4] and is involved in interoperability discussions with other registries. |
| Registry itself should constitute a FAIR resource | Developers of FAIRsharing are partners in EOSC-Life WP6 and are implementing the FAIRassist.org and other FAIRification tools. |
| API access to data within the registry | API access is temporarily not provided by FAIRsharing as they move to a new API |

---

[2] https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases-wg/outcomes/fairsharing
[3] https://fairsharing.org/summary-statistics/?collection=all
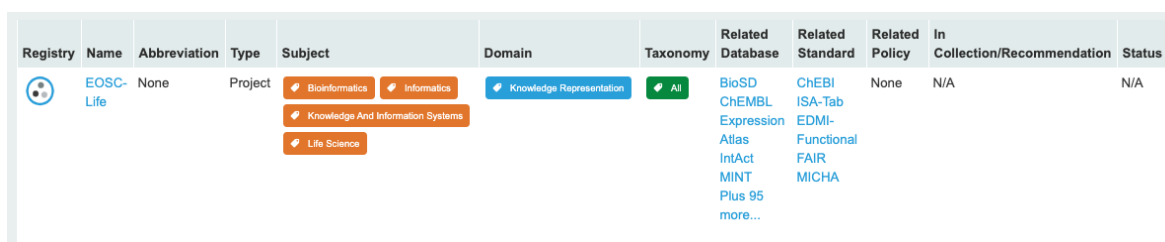[4] https://bioschemas.org/

| | |
|---|---|
| | system, however, interoperability cross registries mitigates this somewhat, though we note that an API would be useful for query. The customised batch loading processes provided for EOSC-Life mean that data submissions are already scalable and an API is not required for batch loading of data. |
| Population should be simple, rapid and regularly updated during EOSC-Life | Data submissions are via google sheet for EOSC-Life and data owners claim and can edit each record as needed[5]. |

## 2.2.   Implementation and population of the FAIRsharing registry

Given the use cases and the stakeholders and the availability of the FAIRsharing[6,7] resources within EOSC-Life (WP6), we evaluated the overhead of delivery of a new registry vs. reuse and therefore improved sustainability. This led to a decision to use FAIRsharing as the project's registry and we were supported in the delivery of the FAIRsharing EOSC-Life collection by WP6 colleagues.

Briefly, FAIRsharing is an ELIXIR Recommended Interoperability Resource[8] with a broad scope, including, but not limited to, Biomedicine and which was first launched in 2011 as BioSharing and rebranded (and expanded) as FAIRsharing in 2017. Due to its now significant content and breadth of coverage we decided to use it to fulfil WP1 registry needs. FAIRsharing  resource records are created and curated by in-house curators in collaboration with the resource's maintainer. To populate FAIRsharing for EOSC-Life, a Google-sheet based method was delivered by FAIRsharing, this was consistent with EOSC-Life's way of working, and enabling rapid and lightweight population of the registry. The EOSC-Life collection was first made available in September 2020 and currently has >100 data EOSC-Life resources registered, plus four related data standards (Figure 3).



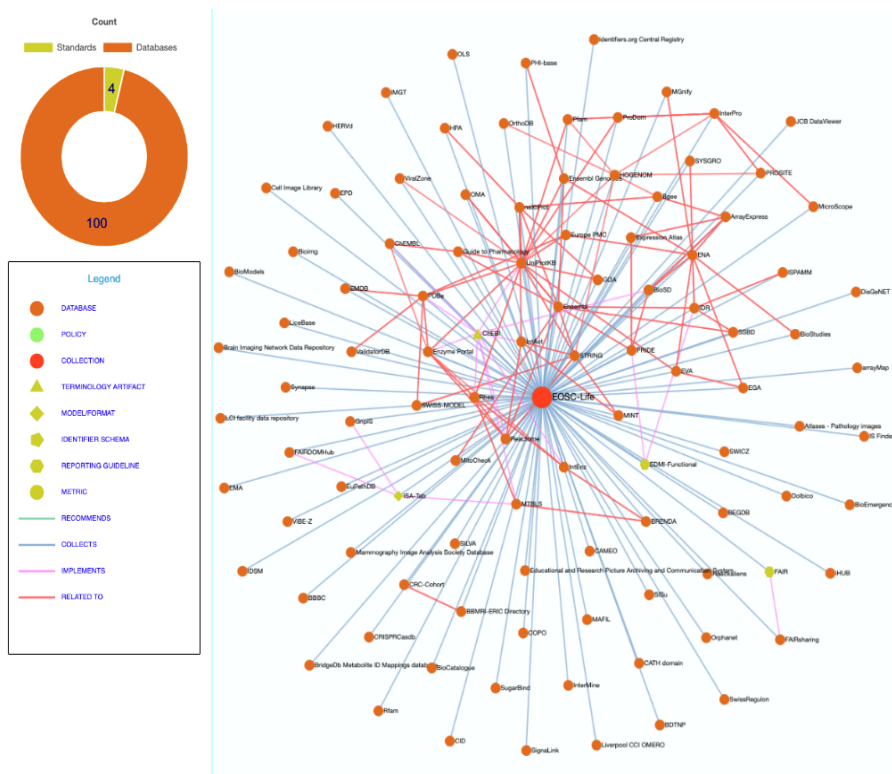*Figure 2: The EOSC-Life collection in FAIRsharing[9]*

*Figure 3: The EOSC-Life registry represented as network of 100 data resources and four related standards*

## 2.3.   Clinical research data use cases

One of the use cases for the EOSC-Life registry is the need to represent clinical trial data and related data. Some clinical research data is already integrated into FAIRsharing as a data set record, for example, BBMRI's Colorectal Cancer Dataset and BBMRI Sample Directory which includes preclinical models. However, there remains a challenging use case around clinical research data. Although more and more datasets and associated documents from clinical research are being made available (the data usually under controlled access), these various objects are often stored in a variety of places, can be difficult to find, and the mechanisms for gaining access to data are not always clear. ECRIN has therefore defined the need for, and is developing, a repository for metadata relating to clinical study data objects, the MetaData Repository (MDR)[10]. Comparable efforts are in progress in life sciences, for example EU Rare Disease Program's central meta data repository[11]. This provides a single, easy to search web portal for discovery, access and provenance metadata for all the data objects associated with clinical studies. The work plan for the MDR is provided in Appendix 1 and progress to the work plan is reported below. Note that the original Task 13 in the original work plan in the appendix has been incorporated into tasks 6 and 7, the current task 13 was added after the work plan was written.

---

[10] https://ecrin.org/clinical-research-metadata-repository
[11] https://www.google.com/url?q=https://eu-rd-platform.jrc.ec.europa.eu/erdri-description_en&sa=D&ust=1607933862291000&usg=AOvVaw36qHNMDBo2CvhRvmYoPHv4

| # | Task | Status | Comments |
|---|------|--------|----------|
| 1 | Re-examination and refinement of the current metadata model, to enhance searching and filtering capability. | **Completed:** Model now at version 5 for both studies and data objects. A variety of changes introduced, as listed in comments. | **Recent changes:** a) Provenance information introduced for both study and data object data (required to support source registries terms of use) b) EOSC risk and Data Use Ontology consent classification incorporated (see 2 below) c) DUO like scheme introduced for de-identification categories. d) Topic data re-organised to better support MeSH coding (Medical Subject Headings) e) Field names made more consistent. f) Simplification by removal of some unused fields. |
| 2 | Exploration of the inter-relationships and possible alignments with other ontologies and approaches to discoverability (e. g. OmicsDI, data tags, BioSchema). | **Ongoing:** a) DUO consent classification examined and incorporated. b) EOSC data tag for risk examined and incorporated. | **To do:** a) Explore and promote further compatibility with Persistent ID entifiers (PID) developments (e.g. from RDA PID forum), in particular Organisation ids (ROR) for contextual data. b) Investigation of possible mappings, to schema.org and DCAT schemas |
| 3 | Obtaining and integrating metadata from more (at least 6) major study registries. | **Completed:** a) All 18 WHO registries now serve as sources, though 15 of these are via WHO data. Creates about 580,000 study records. b) WHO data set processing now much improved – data from different registries split on initial download and then processed separately. | **Further work:** a) German, Dutch and Australian registries to be interrogated directly rather than through WHO data. b) EUPAS dataset to be added for additional observational studies |

| | | | |
|---|---|---|---|
| 4 | Extending extraction to other data repositories (at least 10). | **Ongoing:**<br>a) Only BioLINCC and Yoda being targeted at present, through web scraping.<br>b) Vivli data downloaded and analysed but appears too incomplete at the moment for use. | **To do:**<br>a) Examine suitability of DataDryad, Zenodo, CrossRef as potential data sources, and add each of these if possible.<br>b) Examine the possibility of using other NIH sponsored repositories, (i.e. similar to BioLINCC) as data sources.<br>c) Examine the possibility of using designated protocol documents, as published in Trials.<br>d) Examine possible contributions of one or two institutional repositories. |
| 5 | Establishing algorithms for identifying the links between new data and already extracted studies and data objects | **Completed:**<br>a) Mechanisms introduced into the system based around md5 hashes, to identify data objects and studies without PID.<br>b) A revised procedure for identifying links between studies now in place, based upon 'other study identifiers' listed in registries.<br>c) Management of cross-source study-study one-to-many relationships now added to the system. | **Further work:**<br>a) Study linkage based on title should be explored.<br>b) The possible use of text mining and ML techniques for establishing links and duplications needs to be explored. |
| 6 | Modification of data extraction to better handle periodic interrogation of the same source (i.e. only handle new or revised data). | **Almost completed:**<br>Data extraction and processing mechanisms now brought within a generic framework and scheduled operation introduced. Scheduling now being checked, logging of 'unsupervised' operations being improved. | **Recent changes:**<br>a) Download and processing mechanisms brought within a generic framework, for better control and monitoring.<br>b) Local data stores established for all sources.<br>c) Logging and tracking mechanisms introduced to identify correct candidate studies / objects for each process<br>d) Introduction of data download, processing and aggregation tasks as scheduled tasks (weekly at present). |

| 7 | Modularising data extraction architecture where possible, with a view to providing interchangeable components for uptake by other RIs. | **Ongoing:**<br>a) Different stages of the extraction process now separated into modules in order to better support modularisation and independent functioning. Documentation of the systems brought up to date in MDR wiki. | **To do:**<br>Assessment of possible usefulness to other RIs, but this is difficult until the full range of systems is developed, including APIs (see 13). |
|---|---|---|---|
| 8 | Developing ways of rationalising topics / keywords against a common UMLs based schema, to reduce duplication and enhance searchability. | **Ongoing:**<br>Me    SH codes selected as the best interim method of rationalising topic terms, and applied to the system. (Much source data is already Me    SH coded) | **To do:**<br>a) M    eSH coding, where possible, of uncoded terms against their Me    SH equivalents<br>b) Further exploration of UMLS systems and related services. Need to find as comprehensive a solution as possible. |
| 9 | Developing ways of processing names (of research, organisations, people) to better support matching and searching. | **Ongoing:**<br>Algorithms introduced for applying standardised versions of names during the import process, but not 100%. | **To do:**<br>a) Explore the text indexing capabilities of Postgres.<br>b) Explore how developments in PID management (e.g. from PID forum) can be applied to entities in the MDR in particular try to integrate with ROR organisation PIDs |
| 10 | Maintaining comprehensive documentation of all aspects of the system, including each extraction routine, within a project Wiki. | **Ongoing:**<br>a) Wiki re-organised and new material introduced for metadata and data extraction sections | **To do:**<br>a) Portal documentation needs bringing up to date<br>b) Shared 'to do' and issue tracking system needs to be introduced.<br>By its nature this task always 'ongoing' |
| 11 | Maintaining all extraction and data processing code in GitHub. | **Ongoing:**<br>a) Source code made more uniform and Github repository tidied up<br>b) Revised Readme files created for all 4 main data collection/extraction systems | **Comment:**<br>By its nature this task always 'ongoing' |

| 12 | Development of tests (including test data) for regular testing of extraction accuracy. | **Ongoing:** A strategy now outlined. Different types of tests required for different parts of the system. | **To do:** a) Initial selection of relevant test material (e.g. sample studies) for each source (or source type). b) Automated systems for comparing actual versus expected values required. |
|---|---|---|---|
| 13 | Publication of journal papers around the MDR. | **Begun:** Outline of initial paper circulated and agreed | **To do:** Text to be written in near future |
| 14 | Preparation and testing of a Restful (or possibly GraphQL) API for supporting data access. | **Not yet begun:** Other tasks have had to take priority up to now. | **Comment:** a) Characterisation of data demands from the portal interface need to be clarified. b) To explore the usefulness of GraphQL instead of or in addition to a RESTful API. |
| 15 | Developing a web-based support tool to help data generators more easily apply the metadata at source. | **Begun:** Initial design work being carried out (to support metadata capture in EOSC Life WP14) | **Comment:** Version 1 expected Spring 2021 |
| 16 | Integration with AAI developed and provided by EOSC-Life. Although the data itself will be public, access to development and data management systems will need to be controlled. | **Not yet begun** Not needed at the moment | **Comment:** Needs further details on how the portal will be integrated within EOSC hub and how development / production versions will be managed. |
| 17 | Contributing to an overall strategy around discoverability of data sources, within EOSC as a whole and within life science RIs in particular. | **Not yet begun** | **Comment:** Will be discussed with WP6 and wider EOSC in 2021. |

| 18 | Exploration of how data sharing can be improved by the MDR (i.e. demonstrations of usefulness). | **Not yet begun** | **Comment:**<br>The system needs to reach a certain degree of maturity before it can be properly evaluated. Once that is done a dialogue can be begun with users, both in general and with a designated test group. |
|----|----|----|----|

## 2.4.    Implementation in support of the Demonstrators

A detailed assessment of WP1 Demonstrator support activities and achievements was provided in the EOSC-LIFE Periodic report (see Table1_7[12]), where details on associated implementations can be found.

# 3.  Next Steps

## 3.1.    Continued population of the EOSC-Life collection

The EOSC-Life collection now forms the registry of data resources, and in some cases specific datasets from EOSC-Life's RIs. As the WP1 and WP3 newly funded projects commence, they will be requested to ensure their resources are included in FAIRsharing and in the EOSC-Life collection. This will allow us to track their progress and ensure their results are visible to the wider community. We will also request annual updates from the EOSC-Life WP1 partners to existing records to ensure these are appropriately maintained.

## 3.2.    Registry interoperability and the wider EOSC

There is a clear need to provide datasets (both private and public) direct to workflows implemented in WP2 of EOSC-Life and compatible with the supported workflow management systems. WP2 has implemented  WorkflowHub[13] a workflow registry to promote workflow sharing. A future direction will be to define best practice for provision of data to workflows and determine the use cases requiring dataset access for specific workflows from WP2 and WP1/WP3 use cases. We will use the WP3 demonstrators and open call funded projects to explore these use cases in addition to exploration of wider interoperability challenges.

During the first phase of EOSC-Life we have engaged in discussions with sister projects and also EOSC. There is a difference in approach within EOSC-Life, where we have not sought to deliver a new registry, but to sustain an existing registry and acknowledge the landscape for the life

---

[12] https://drive.google.com/drive/folders/1qHvz8zA3Vsl4janyJ00aE6r82SQLm49N
[13] https://about.workflowhub.eu/

sciences RIs is complex. This naturally leads to a requirement to interoperate across registries, both within EOSC-Life and across EOSC's registries. It is clear from the life sciences perspective that there are use cases related to e.g. clinical and bio molecular data linked to socioeconomic data, geophysical data and climate data, especially in relation to the COVID-19 pandemic. We are therefore working to understand how to interoperate our registries to answer these and emerging use cases. We note a recent ENVRI[14] project deliverable addresses strategies for interoperability across EOSC projects and we propose to explore these for EOSC-Life as the project progresses. Similarly, the EOSC Enhance deliverable D4.3 (Goble, Juty et al., UNIMAN) explores interoperability across its constituent Cluster projects and recommends that "*interoperability is always driven by a use case ... lightweight technologies are used* (e.g. schema.org, BioSchemas, DCAT)" and this is consistent both with EOSC-Life's landscape and expertise.

We have also engaged with the FAIRsFAIR project[15] via two workshops to address FAIRness related to metadata repositories and interoperability. This project will undertake an interoperability pilot using the standards such as the Data Documentation Initiative Cross Domain Integration (DDI-CDI) (and others). We will therefore monitor the outcomes of future workshops with FAIRsFAIR (and others) in collaboration with WP6 with the aim of ensuring cross registry interoperability for EOSC-Life and sister projects and to ensure standards development in this space meets our use cases.

In WP6, FAIR requirements for registries have been delineated (D6.1), and, for practical implementation, an extension of the W3C's Data Catalogue Vocabulary version 2 (DCAT2) is proposed to describe a FAIR registry in a machine actionable way. The DCAT description, implemented as a 'FAIR Data Point' (FDP, https://www.fairdatapoint.org/), provides a uniform FAIR access point for any type of underlying registry. It is relevant to highlight that the FDP approach, using DCAT, supports the registration of metadata of different types of digital objects. DCAT2 introduced the concept of Catalogued Resource as an abstract class that should be specialised to represent the type of resource, or digital object, that one would like to have metadata about such as dataset, workflow, website, registry, repository, etc. Demonstrating the use of DCAT-based FDPs to find and use registered FAIR resources (e.g. FAIR mappings or workflows) is planned as a next step and WP1 will collaborate in these efforts.

## 3.3.    Dissemination

The recent EOSC-Life mid term review discussed the need for visibility of EOSC-Life's resources to the wider research community. We will therefore work with WP10 to design materials to promote the registries (for example, https://www.eosc-life.eu/achievements/) and will populate the new LSRI website, designed to sustain visibility of WP1's resources after EOSC-Life ends. Dissemination is already underway within EOSC and via sister projects and we will continue to contribute to discussions via direct project engagement, attendance at workshops etc. Immediately, we will focus on supporting WP1 and WP3 call projects starting with a kick off workshop in January 2020. Our future KPIs will include the data/data resources available from relevant registries, use of data

---

[14] https://envri.eu/wp-content/uploads/2020/09/MS18-WP5-Design-of-the-service-catalogue-1.pdf
[15] https://zenodo.org/record/4134788 - .X8EaeOmg-M8

and data resources in WP1/WP3 collaborations via open calls and access statistics to FAIRsharing/MDR for the EOSC-Life relevant resources.

# Abbreviations

API - Application Programming Interface
DCAT2 - Data Catalogue Vocabulary version 2
DDI - Cross Domain Integration (DDI-CDI), a specification aimed at helping implementers integrate data across domain and institutional boundaries
DUO - Data Use Ontologym a Global Alliance for Genomics and Health standard
EOSC - European Open Science Cloud
FDP - FAIR Data Point
IPD - Individual Participant Data
MeSH - Medical Subject Headings
MDR - Metadata Repository
PID - Permanent ID    entifier
RDA - Research Data Alliance
RI - research infrastructures within EOSC-Life
RIR - Recommended Interoperability Resource

# Delivery and Schedule

The delivery is delayed from August 2020 to December 2020 due to the COVID-19 pandemic which prevented in person meetings and due to staff being assigned to COVID-19 projects temporarily delaying work on this deliverable. A Grant Agreement amendment was planned to be introduced before the end of the 1st reporting period, but was delayed due to an emergency amendment adding two COVID-19 related WPs, so the 2nd amendment will be progressed at the beginning of next year.

# Appendices

## Appendix 1. Work plan for delivery of the ECRIN Metadata Repository

### EOSC-Life WP1: A Metadata Repository (MDR) for clinical research as a data resource for the EOSC

Date: 11 March 2020

Authors: S. Canham, C. Ohmann, S. Goryanin, S. Battaglia (ECRIN)

## Background

In recent years there has been a growing acceptance that to accurately assess the results of trials and other clinical research studies, and in particular to combine the results from different trials in meta-analyses, it is necessary to have access to the original source data, the "Individual Participant Data" (IPD), as well as the result summaries found in published papers. In addition, to make sure that the IPD can be fully understood and properly analysed, a variety of other study documents (protocols, analysis plans, etc.) are required. As a result, under pressure from funders and journal editors, more and more researchers are making such material (generically, "clinical trial data objects") available for sharing with others. The datasets are rarely freely available - instead a variety of access mechanisms (e.g. individual request and review, membership of pre-authorized groups, or web based self-attestation), are used in combination with different access types (e.g. download versus in-situ perusal). Furthermore the various data objects are stored in a wide variety of different locations: a rapidly growing number of general and specialized data repositories, trial registries, publications, the original researchers' institutions, etc. The researcher or reviewer wishing to locate relevant data objects for a study is therefore faced with a bewildering mosaic of possible source locations and access mechanisms, and this problem of 'discoverability' will almost certainly become much worse in the future as more and more materials are made available for sharing.

## Aims of the Project

The principal aim of the project is to combat this discoverability problem, by making the data objects generated from clinical research easier to locate, and by describing how each of those data objects can be accessed, providing direct links to them where that is possible. The central idea is to develop systems that can collect the *metadata* about the data objects, including object provenance, location and access details, and aggregate it into a single **MetaData Repository** (or MDR). The MDR is therefore designed to assemble, and standardize the metadata about clinical studies and the data objects generated by them, and provide access to that metadata through one or more APIs.

## Status of implementation

Within the EU H2020 funded project eXtreme DataCloud (XDC, grant agreement 777367), a demonstrator for a MDR for clinical research has been designed and developed by **ECRIN** (the European Clinical Research Infrastructure Network), in collaboration with **ONEDATA** (Poland) and **INFN** (Istituto Nazionale di Fisica Nucleare, Bologna - Italy). For the demonstrator metadata from 4 external sources (CT.gov, Pubmed, Yoda, BioLINCC) have been collected, their metadata processed and stored in relational databases and mapped to a common metadata schema developed by ECRIN (http://ecrin-mdr.online/index.php/JSON_Schemas). In total, about 800,000 records (~325,000 of studies and ~472,000 of data objects) have been downloaded and links between studies and associated data objects have been established.

Within the XDC project, data from the MDR core database are converted into JSON files, which are then transferred and injected into the OneData environment. The JSON inside each file is attached to that file as metadata using PyFileSystem. A harvester built upon Elasticsearch was developed by INFN and implemented as a plugin to the OneData file system – it is able to read and index the metadata attached to the files. A web portal for searching studies and associated data objects was provided by OneData, using the Elasticsearch engine. The MDR demonstrator is currently under testing by ECRIN in the final phase of the XDC project. More detailed information is available in a Wiki: http://ecrin-mdr.online/index.php/Project_Overview.

## Proposals for the next stage of the project, in EOSC-Life WP1

The aim of the next stage of the project is to extend, complete and qualify the MDR through test and demonstration and finally to publish it in the EOSC, to promote FAIR data usage within clinical research. The current XDC project has successfully provided a positive proof of concept for the idea of a metadata repository in clinical research. The next stage will need to bring the MDR to maturity so that it can be configured as a resource within EOSC. To achieve this the following steps are planned within EOSC life:

1.  Re-examination and refinement of the current metadata model, to enhance searching and filtering capability.
2.  Exploration of the inter-relationships and possible alignments with other ontologies and approaches to discoverability (e. g. OmicsDI, data tags, BioSchema).
3.  Obtaining and integrating metadata from more (at least 6) major studies registries.
4.  Extending extraction to other data repositories (at least 10).
5.  Establishing algorithms for identifying the links between new data and already extracted studies and data objects.
6.  Modification of data extraction to better handle periodic interrogation of the same source (i.e. only handle new or revised data).
7.  Modularising data extraction architecture where possible, with a view to providing interchangeable components for uptake by other RIs.

8. Developing ways of rationalising topics / keywords against a common UMLS based schema, to reduce duplication and enhance searchability.

9. Developing ways of processing names (of research, organisations, people) to better support matching and searching.

10. Maintaining comprehensive documentation of all aspects of the system, including each extraction routine, within a project Wiki.

11. Maintaining all extraction and data processing code in GitHub.

12. Development of tests (including test data) for regular testing of extraction accuracy.

13. Creation of a coordinating    system for scheduling, triggering, monitoring and logging extraction activity, with a GUI for ease of use.

14. Preparation and testing of a Restful (or possibly GraphQL) API for supporting data access.

15. Developing a web-based support tool to help data generators more easily apply the metadata at source.

16. Integration with AAI, developed and provided by EOSC-Life. Although the data itself will be public, access to development and data management systems will need to be controlled.

17. Contributing to an overall strategy around discoverability of data sources, within EOSC as a whole and within life science RIs in particular.

18. Exploration of how data sharing can be improved by the MDR (i.e. demonstrations of usefulness).

Note that the creation and testing of an access portal, along with associated indexing, has been removed into a separate project – the focus of the work in EOSC life is the accurate assembly, processing and standardization of the data. Note also that the intermediate JSON file step is no longer required – it is intended to query the data directly via an API and / or via a pre-indexing mechanism such as Elasticsearch.

## Timelines

The project will start 1 April 2020 and will be finished 31 March 2022.

## Milestones

- Agreement on a revised metadata model

- Successful integration of 6 study registries

- Successful integration of 10 additional data repositories

- Creation of a co-ordination and logging system

- Creation and successful testing of an API

## Reporting

There will be regular reporting on updates of the MDR to EOSC-Life WP1, every 6 months.

| Deliverables | Dates |
|---|---|
| Enhanced metadata and software specification | 12 months |
| Code on GitHub | 12, 24 months |
| Co-ordination and logging system | 18 months |
| The completed project Wiki | 24 months |
| A fully documented and tested API | 24 months |

## Involvement of other partners

We would be very happy to work with all partners involved in supporting data discovery in their RI and interested in the architecture, development and application of the MDR.