Project acronym: EOSC4CANCER
Grant Agreement Number: 101058427
Project full title: EOSC4CANCER
Call identifier: HORIZON-INFRA-2021-EOSC-01

# D1.3 Initial data resources map (DEC Report)

| | |
|---|---|
| Version: | 1 |
| Status: | Final |
| Dissemination Level: | Public |
| Deliverable Type: | DEC — Websites, patent filings, videos, etc |
| Due date of deliverable: | 22.05.2023 |
| Actual submission date: | 17.05.2023 |
| Work Package: | WP 1 Federated data spaces to enable accessing, using, reusing and sharing cancer-related data |
| Lead partner for this deliverable: | UMCG |
| Partner(s) contributing: | UMCG |


Main author(s):

| | |
|---|---|
| Morris Swertz | UMCG |
| Gerieke Been | UMCG |

Other author(s):

| | |
|---|---|
| Joeri van der Velde | UMCG |
| Ype Zijlstra | UMCG |



Funded by the European Union

## Revision History

| Version | Date | Changes made | Author(s) |
|---|---|---|---|
| 0.1 | 12/05/2023 | Several suggestions | Mariska Bierkens (NKI) |
| 1 | 17/05/2023 | Final adjustments for submission | UMCG |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

**Statement of originality**

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

# TABLE OF CONTENTS



# LIST OF FIGURES

# Background

The complex nature of Cancer requires integration of advanced research data across national boundaries to enable progress. The Europe mission board for cancer has identified access to data, knowledge and digital services - accessible across European research. In EOSC4Cancer the mission is to make cancer genomics, imaging, medical, clinical, environmental and socio-economics data accessible; using and enhancing existing federated and interoperable systems for securely identifying, sharing, processing and reusing FAIR cancer data across borders. EOSC4Cancer use-cases will cover the patient journey from cancer prevention to diagnosis to treatment, laying the foundation of data trajectories and workflows for future cancer mission projects. Curated and FAIR datasets will be essential for advanced analytics and computational methods to be reproducible and robust, including machine learning and artificial intelligence approaches.

The overall goal of WP1 is to enhance access to cancer-related data by increasing the FAIRness of the participating resources that contribute data to the EOSC4Cancer metadata catalogue. The resources are data owners both from within the consortium as well as data owners from institutes not involved in EOSC4Cancer, who will contribute their relevant study metadata and their data variables. We will develop a queryable directory of the participating resources to enable discovery (collaboration with WP2 Beacon). This allows researchers and clinicians to discover relevant resources and datasets containing variables including phenotypes of interest. Within WP1, the first steps focus on creating the EOSC4Cancer metadata catalogue infrastructure, which is an initial data resources map based on FAIR principles to allow data collection and sharing across borders.

This deliverable aims to provide FAIR cataloguing tools to map existing resources metadata for discovery (building on experience and standardised templates developed in previous projects, such as BBMRI-ERIC), request and as reference during research, and to support the process of data harmonisation and versioning of the variable mappings in EOSC4Cancer towards joint data analysis and study across study comparisons (i.e. promoting the FAIR principles).

# Initial data resources map development

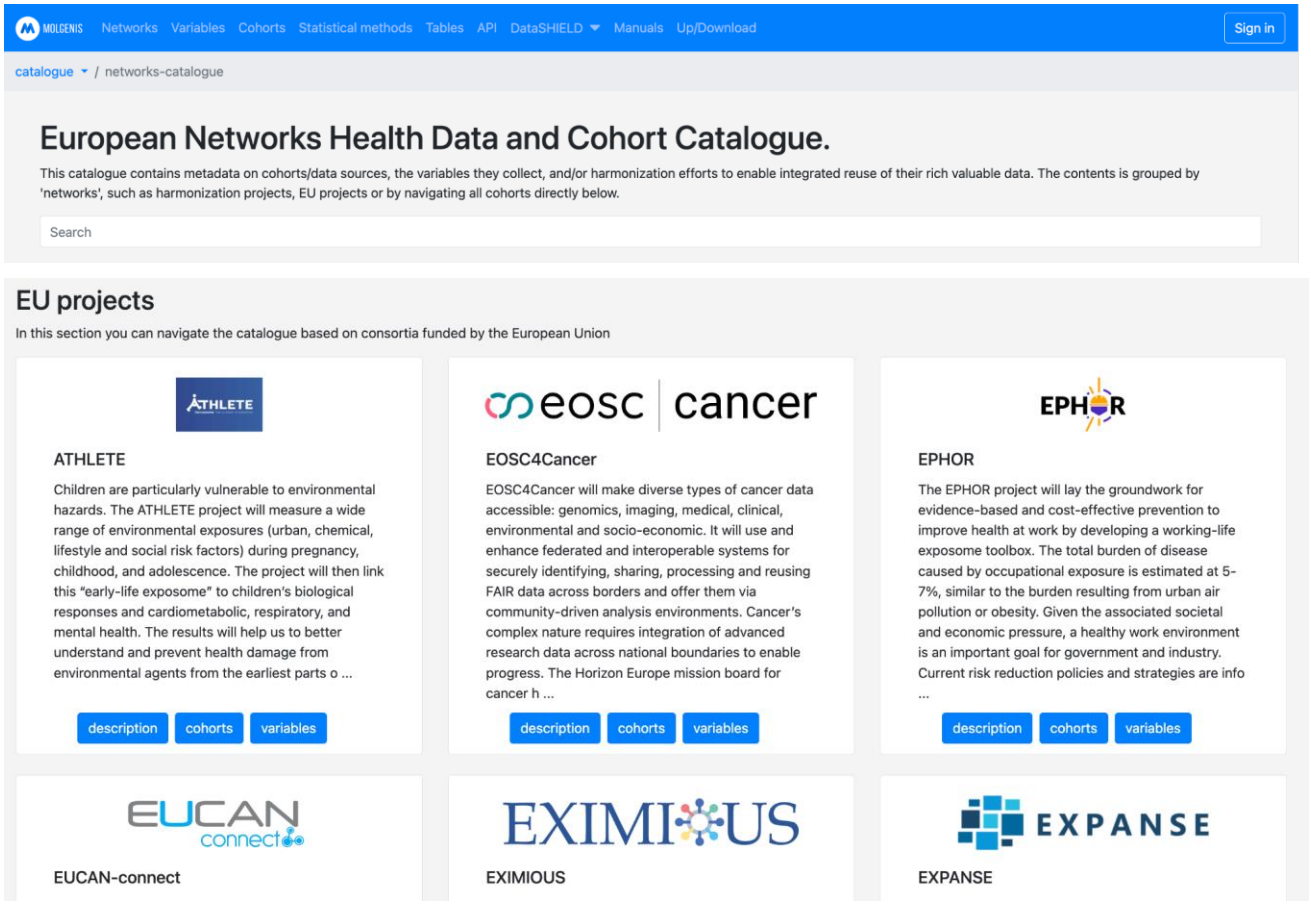
The EOSC4Cancer consortium has access to established cancer networks and European research infrastructures for cancer data, including cancer genomics, imaging, medical, clinical, environmental, socio-economics data and biobanks. A **metadata catalogue** has been initialised to enable the EOSC4Cancer researchers to (i) assess all these data sets, (ii) assess the suitability of the data sets to answer specific research questions, and (iii) to facilitate the deployment of the WP4 on federated multi-institutional research use-cases.

EOSC4Cancer partners, the University Medical Centre Groningen (UMCG), have committed to build the data resources map, a FAIR metadata cataloguing platform. As per project planning, the catalogue has been joined with the existing MOLGENIS based European Networks Health Data and Cohort Catalogue, which is implemented using the existing MOLGENIS data catalogue platform software. See figure 1. MOLGENIS is a unified framework that enables European-wide sharing of data catalogues as described in Swertz et al (2022)[1]. Our ambition is to connect the EOSC4Cancer data to a large resource of other EU project metadata rather than developing a separate catalogue, which will ease reuse of catalogue records, sharing of data harmonisation protocols and ease sustainability beyond EOSC4Cancer. For EOSC4Cancer, we have refined the catalogue data model and have used this model in several new releases of the software and polished the user interface, and most importantly started with population of the resources map.

The aim was to create  a publicly findable metadata catalogue to make the relevant cancer related datasets collected findable, comprising the relevant cancer related data sets collected both within EOSC4Cancer and beyond the project consortium. Specifically, for EOSC4Cancer an EU consortium metadata catalogue has been equipped and added to the European Networks Health Data and Cohort Catalogue.

[1] Swertz et al (2022) Towards an Interoperable Ecosystem of Research Cohort and Real-world Data Catalogues Enabling Multi-center Studies. Yearb Med Inform. https://pubmed.ncbi.nlm.nih.gov/36463884/

**Figure 1**: European Networks Health Data and Cohort catalogue. https://data-catalogue.molgeniscloud.org/



# Catalogue Functions and Features

The catalogue comprises:

- A **findability function** enabling users to access *rich metadata* about the data sets including, for example, the type of data set (e.g., cohort), the population, number of participants, and the data variables; See figure 2.
- A **data model** that allows networks and EU projects to capture the required variables, the data model is accessible via the MOLGENIS EMX2 Github repository https://github.com/molgenis/molgenis-emx2/blob/master/data/datacatalogue/molgenis.csv
- A **harmonisation mapping system** that enables cataloguing on the 'variable level', documenting harmonised 'standard' variables and providing interoperability maps showing if and how collected variables from the data sets have been mapped to these standards; See figure 3.
- **Manuals** to explain to users how to use the catalogue; and
- **Standard operating procedures** for deployment, upgrading and maintenance of the software application and database.

**Figure 2:** Rich metadata describing the datasets, EOSC4Cancer can be found in the networks section

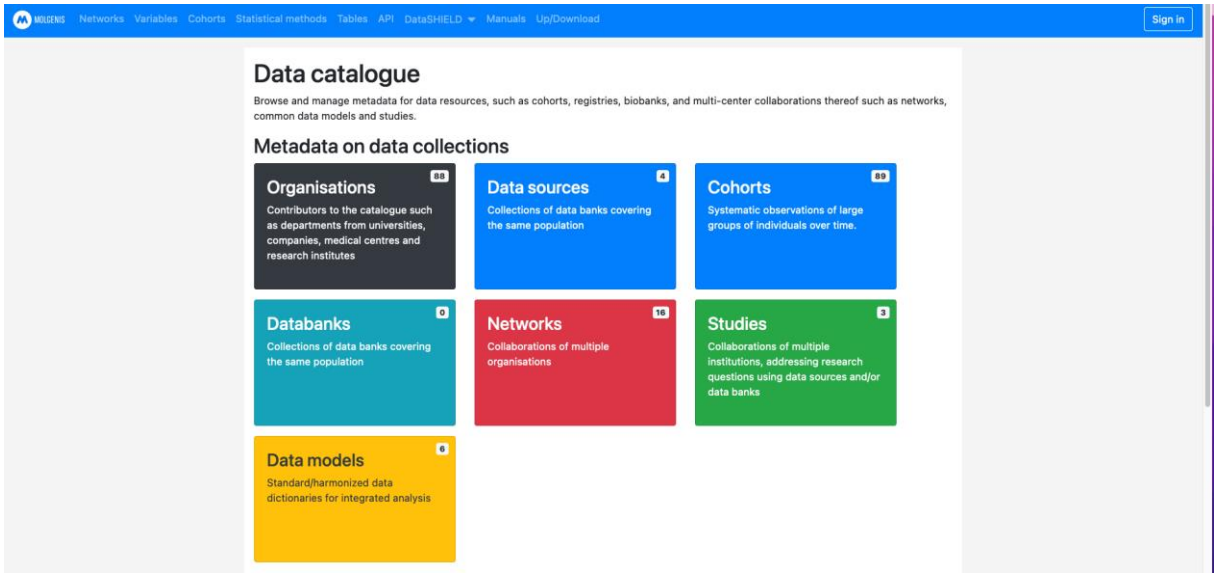
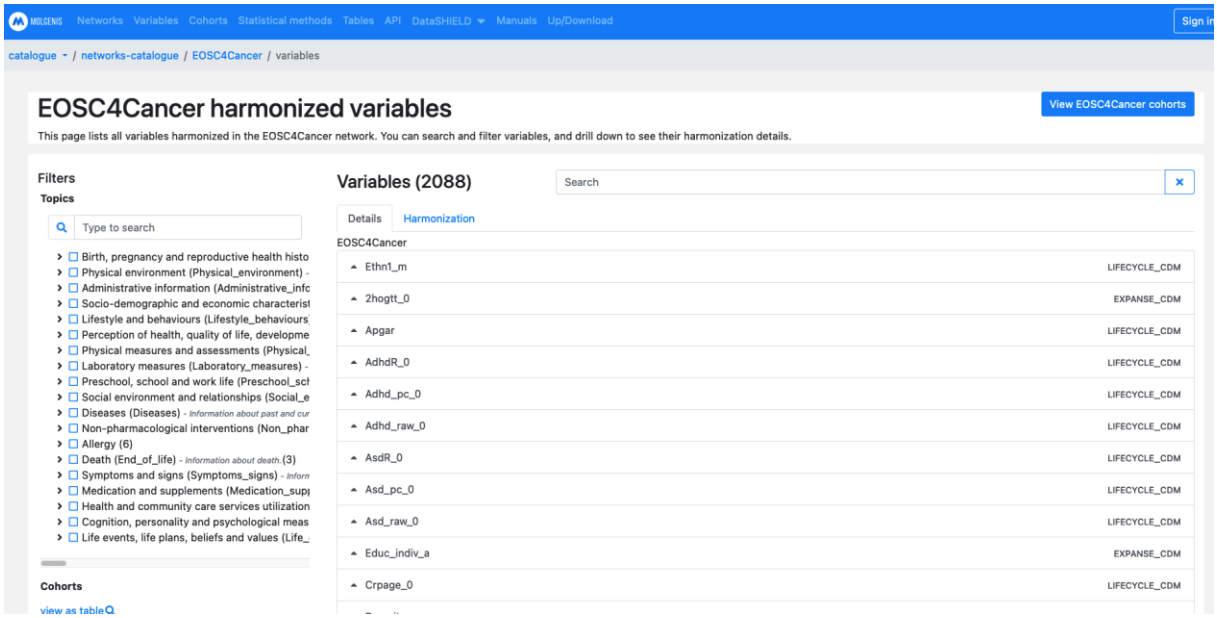


**Figure 3:** Harmonisation mapping system. Currently, 2088 harmonised variables are included in the European Networks Health Data and Cohort Catalogue which will be mapped to EOSC4Cancer variables in the future by WP2 to achieve metadata interoperability. The EOSC4Cancer variables are grouped and can be selected separately from all variables.




The catalogue only contains metadata (in simple terms data about data). The actual data within the data sets is held locally or is stored in a centralised system and accessed by the researcher independently of the metadata catalogue.

The catalogue is open access, web-based and can be accessed at [https://data-catalogue.molgeniscloud.org](https://data-catalogue.molgeniscloud.org). The catalogue software is built into the MOLGENIS FAIR data platform and is available for free as open source for reuse by other networks at [http://github.com/molgenis/molgenis-emx2](http://github.com/molgenis/molgenis-emx2). A screenshot of the current metadata catalogue landing page is in Figure 1 above. Initial documentation can be found at https://molgenis.github.io/molgenis-emx2/#/, with description of both MOLGENIS underlying platform as well as documentation specific to the cataloguing tools.

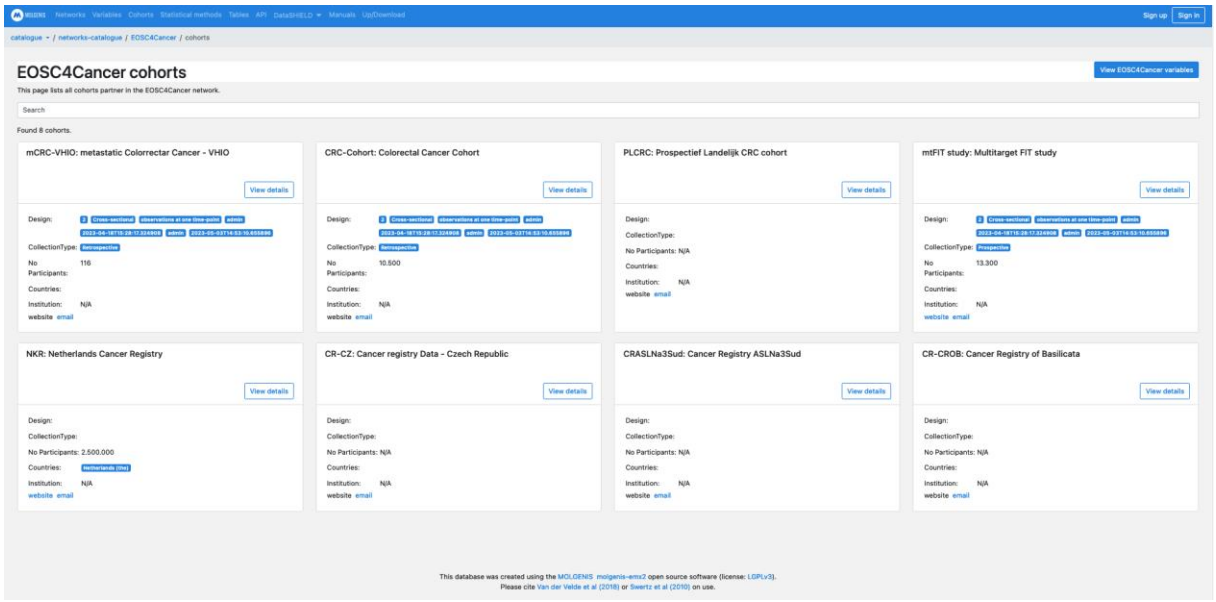
# Population of the data resources map

As an initial proof of concept data managers of the EOSC4Cancer partner institutions provided information about their institutions and data sets, this 'rich metadata' was added to the catalogue. To achieve this, for each use case in WP4 (and the other WP's), people were asked to fill out a Google Form to get a first overview of the available data sets.

Over the past months the first meetings with these partners have taken place with the aim to give them information in the form of a demonstration of the catalogue and giving them the needed support to add their dataset(s) to the catalogue. For the preparation of the cohorts metadata to be added to the catalogue a staging area is set up for each partner. A demo staging area can be found at https://data-catalogue.molgeniscloud.org/testCohort/tables/#/

Currently the catalogue is populated with a number of selected datasets,of partners who initially filled out the shared Google form and cancer related datasets from the BBMRI biobank, as shown in Figure 3. Besides the datasets that are publicly available in the metadata catalogue, several partners are in the process of adding their datasets in the staging area and meetings have been scheduled with the partners who will be onboarding next.

**Figure 4:** Screenshot of the datasets currently populating the EOSC4Cancer initial data resources map



# Dissemination

The metadata catalogue is one of the key EOSC4Cancer outputs and the information about it will be widely disseminated to maximise the utilisation by both EOSC4Cancer researchers and other potential users.

Dissemination activities include:

- presentation of the catalogue to EOSC4Cancer partners at project meetings
- the EOSC4Cancer web site features a page about WP1, which will contain information and links to the metadata catalogue[2] describing what it is and its purpose and includes explanation of the terminology used for a non-scientific audience. People will be given the opportunity to subscribe to receive progress updates and contact us to collaborate.
- the catalogue is connected to other networks and EU projects via the European Networks Health Data and Cohort Catalogue[1]

[2] https://data-catalogue.molgeniscloud.org/catalogue/catalogue/#/networks-catalogue

# Future Development

The catalogue and supporting materials are ready to start the support of the EOSC4Cancer harmonisation and federated analysis work. It is expected that, driven by the needs of the researchers and data managers, the catalogue system and standard operating procedures for deployment, upgrading and maintenance of the software application and database will continue to be iteratively improved, however most effort will be applied in further population of the catalogue in collaboration with WP4. It is hoped that the approach to create a sustainable catalogue, usable beyond EOSC4Cancer, will inspire more partners and data owners beyond the project to participate. For further development and to support the outreach of the catalogue a new version of the EOSC4Cancer metadata catalogue landing page will be designed.

The main next steps include:

- **Populating the catalogue** with EOSC4Cancer resources, linking and integrating cataloguing efforts from partners and from the data sources directly
- **Evaluating the functionality** of the catalogue against available use cases, in particular in close collaboration with WP2 and WP4 to work on delivery of standardised, harmonised codebook variables and a method to transform data to this format.
- **Refinement of the data catalogue metadata model.** While the current catalogue already includes many best practices, we see options for extensions for particular domains such as imaging and genomics (e.g. using image catalogue efforts from EUCAIM partners and FAIR genomes[3] recommendations that are also now developed in GDI).
- **Optimise the user experience**. We have positioned EOSC4Cancer within the larger Health Data Catalogue initiative and we want to develop the user interface to fit both the needs of the EOSC4Cancer use cases as well as interested other users, including researchers and clinicians. (See Figure 1).
- And finally, in close collaboration with WP2, explore options for **federated discovery** using MOLGENIS integrated beacon functionality.

[3] Van der Velde et al (2022) FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. https://pubmed.ncbi.nlm.nih.gov/35418585/