MetaData Repository (MDR) for clinical study objects

Sergei Gorianin

Updated: 25.02.2020

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

FYI, just to remind

Findable

Accessible

Interoperable

Reusable

The FAIR Data Principles are a set of guiding principles in order to make data findable, accessible, interoperable and reusable (Wilkinson et al., 2016).

These principles provide guidance for scientific data management and stewardship and are relevant to all stakeholders in the current digital ecosystem.

They directly address data producers and data publishers to promote maximum use of research data. Research libraries can use the FAIR Data Principles as a framework for fostering and extending research data services.

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

Problem – clinical trials

Findable ✗

Accessible ~

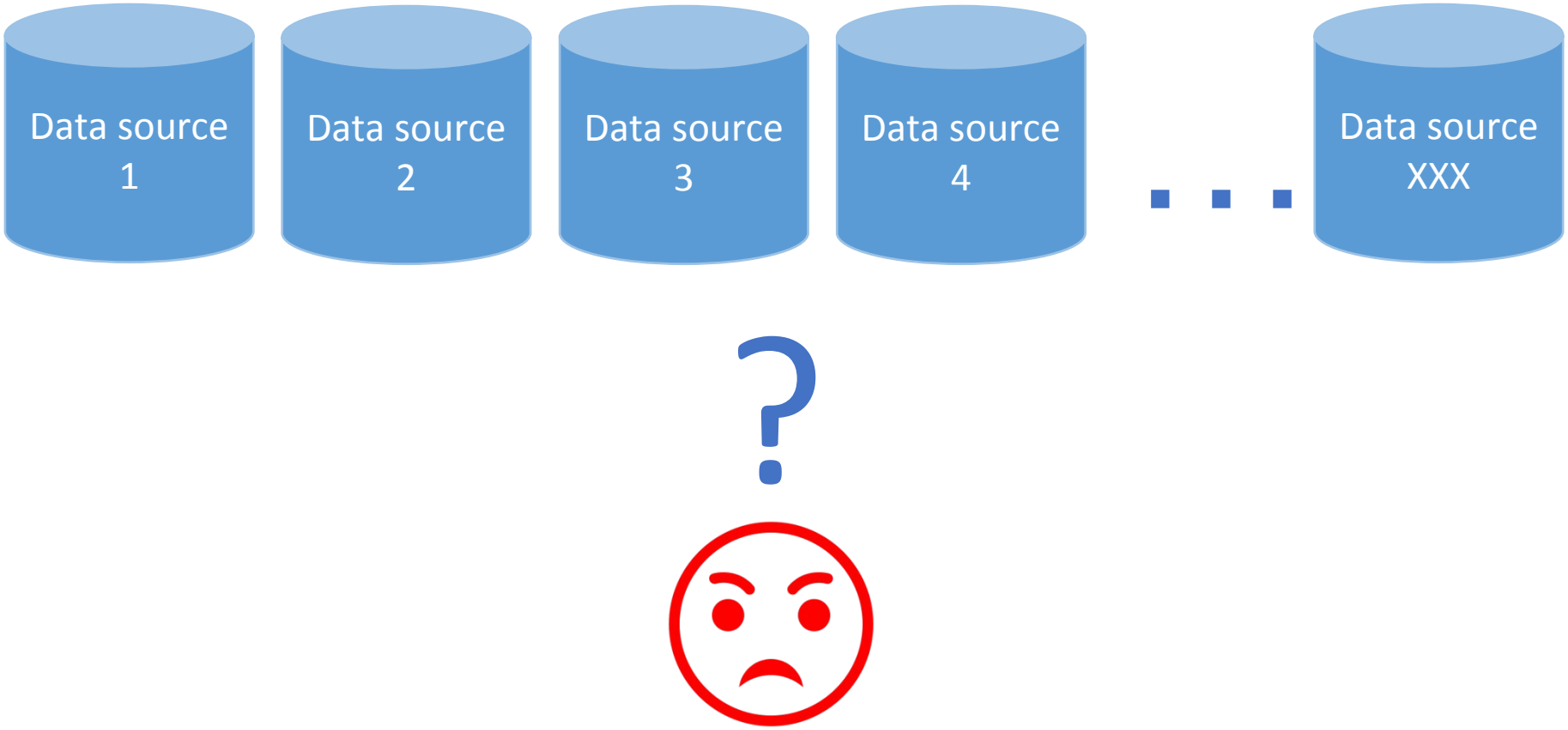
Interoperable ~

Reusable ~

- *Principles & Recommendations for clinical data sharing (Ohmann et al. BMJ Open 2017)*
- *Definition of processes (Ohman et al. F1000Research 2018)*
- *Quality criteria for repositories (Banzi et al. Trials 2019)*
- *Pilot repository (Tilki et al. Submitted)*
- *Recommendations for anonymisation technical solutions (ongoing)*

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

MDR – idea to improve the Findability ?



ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

Let’s summarize the problems

- Plenty of data sources
- Unstandardized data
- Unlinked data
- Unidentified data
- Collecting data is ~~very~~ quite problematic
- Possible duplicates
- etc…

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

How great would be to have all standardized and useful information in a single place…

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

MDR… Why?

To maximize the **discoverability** of all these data objects, it is necessary to collect the *metadata* about them, including object provenance, location and access details, into a single system.

To that end ECRIN is currently developing an **MDR** (MetaData Repository) to standardize, assemble and display the metadata about clinical studies and the data objects generated by them, and provide access to them through a single system, accessed via a web portal.

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

A ***study*** is any clinical research study with humans as study participants, and which is therefore subject to ethical approval, whether or not the study is interventional (a 'clinical trial'), or observational (including disease registry data), or a case study.

A ***data object*** is any information available in electronic form (a 'data stream') and may be a document (e.g. a pdf), a dataset in one of a variety of formats (spreadsheet, csv files, database file, XML etc.), a media (audio, video) file, an image (e.g. a conference poster) or simply a web page with useful text.

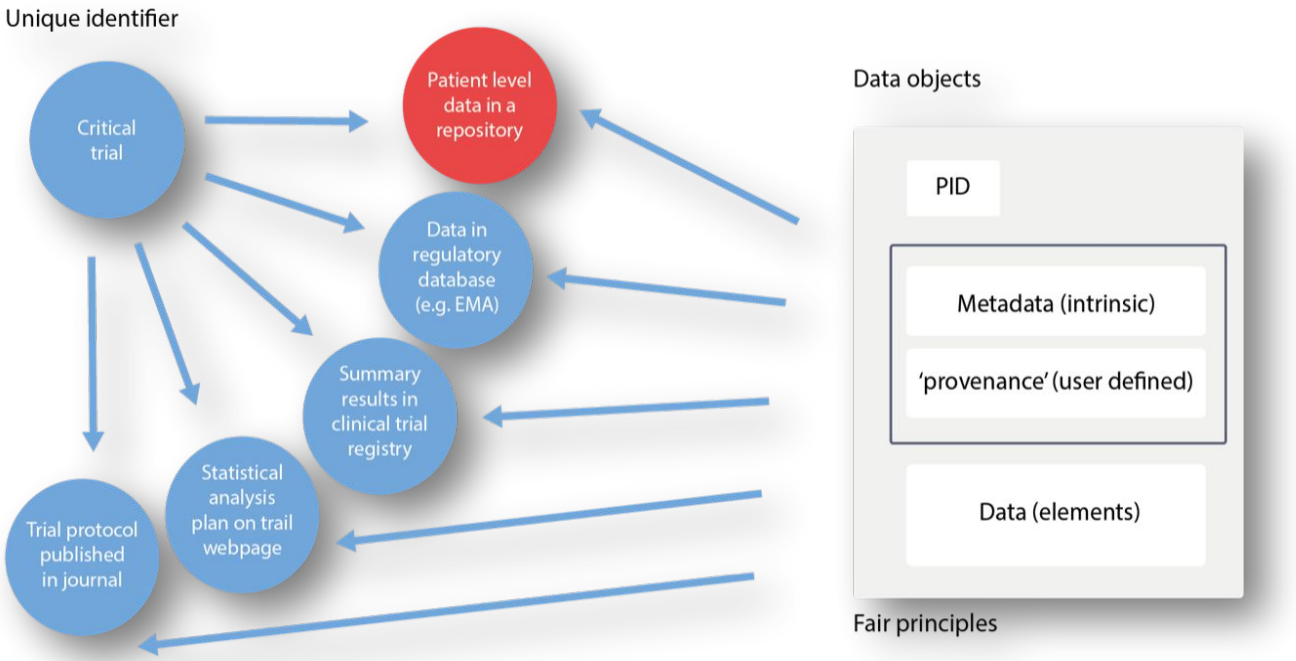


Fig. 1 – Structure of clinical trial (study, data object)

MDR… Implementation

The web portal is developed in collaboration with ONEDATA (Cyfronet, Poland); functionality for discoverability of studies and related data objects is provided by INFN (Italy)*. So far metadata from 4 data sources have been collected using different modalities (e.g. XMLs processing, APIs, scraping of web pages) and stored as JSON objects/relational DB form on the test bed server at INFN. Data will be then ingested, uploaded and made available for the users.

The ONEDATA solution will implement the metadata collection and transport from multiple OneProvider to the central OneZone service; it will also deal with the ACLs management to implement the data protection.

** Within H2020 XDC project*

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

eXtreme DataCloud

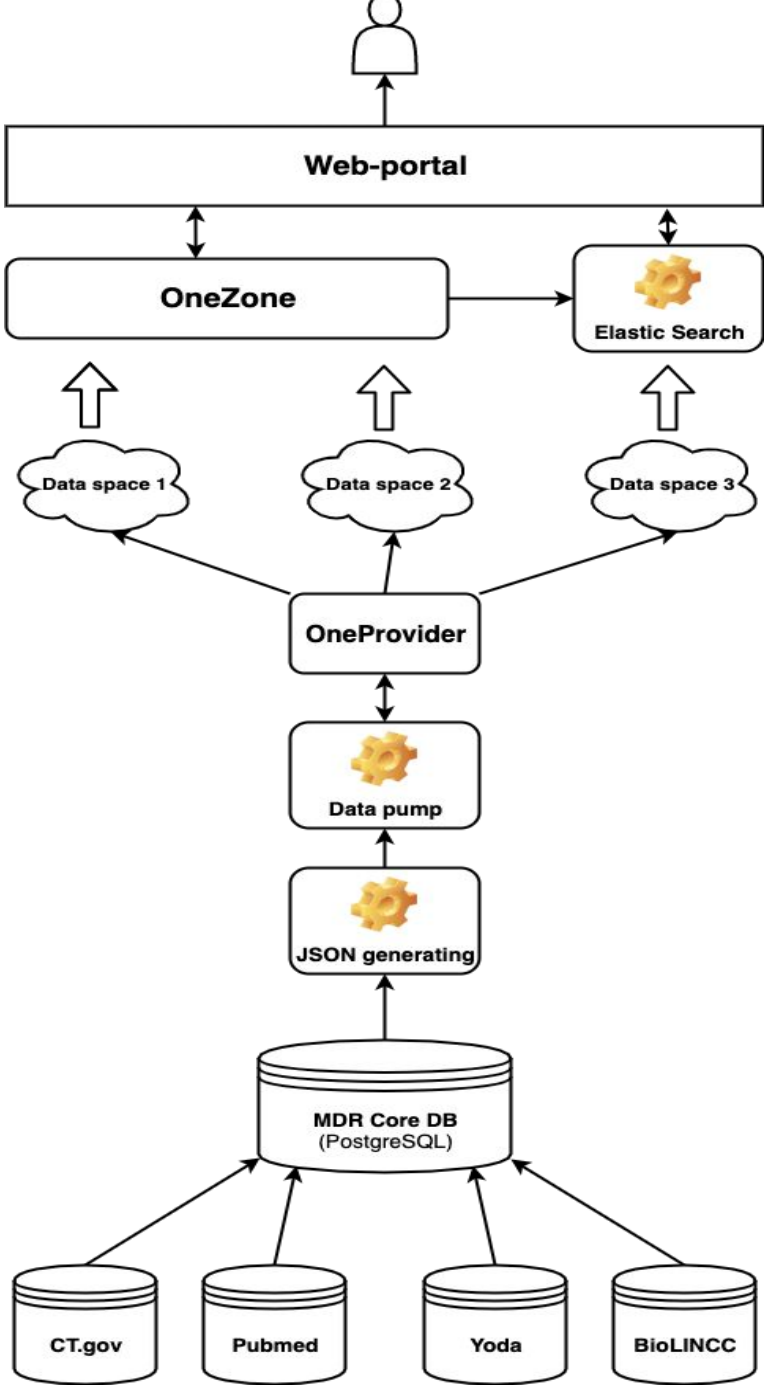
ONEDATA

ONEDATA

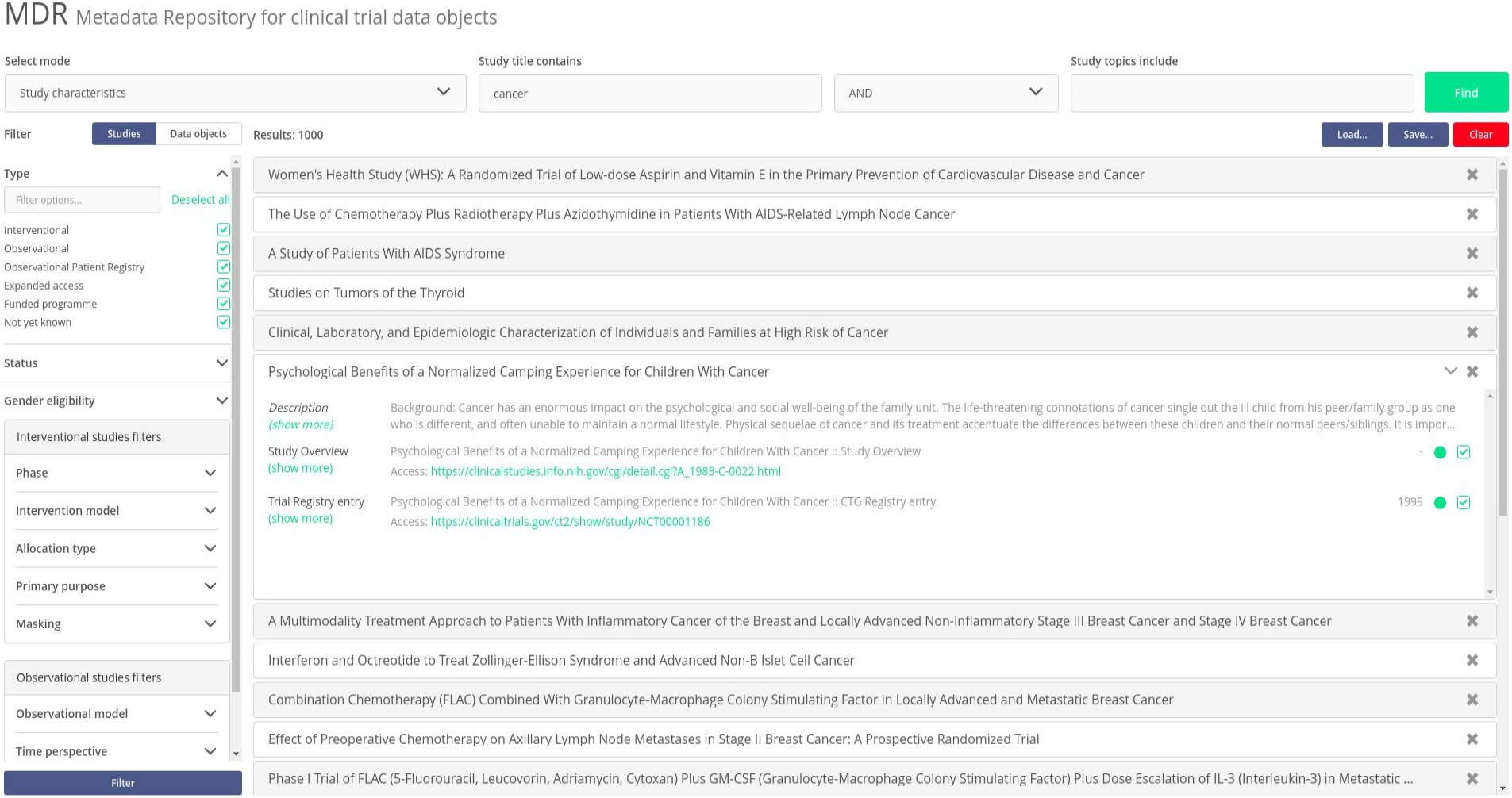
ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

INFN
Istituto Nazionale di Fisica Nucleare



User interface for Web portal (developed by ONEDATA)



ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

Current progress

- 4 data sources (PubMed, CT.gov, BioLINCC, Yoda) are connected
- Single and universal metadata schema was developed*
- about 800.000 studies and data objects uploaded

✔ JSON Schema - http://ecrin-mdr.online/index.php/JSON_Schemas

✔ ECRIN Clinical Research Metadata Schema https://zenodo.org/record/3562911

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

Next steps

- Refine current **metadata model** to increase usefulness to users, and further investigate alignment with other approaches, e. g. data tags.
- Modify **data extraction** to better handle periodic interrogation of the same source (i.e. only handle new or revised data).
- Get metadata from more major **studies registries** (about 6); establish links between new and already extracted studies and data objects.
- Extend extraction to other **data repositories** (about 10).
- Work on **search engine** (ElasticSearch or PostgreSQL alternative); preparation of API supporting data access.
- Develop and deploy **web-portal**.
- Look at developing support tools.
- **Containerization** of developed tools and components.
- Integration with **AAI**, developed and provided by EOSC-Life.
- **Publish as FAIR data resource in the EOSC**.

Acknowledgments

S. Battaglia, S. Canham, C. Ohmann

S. Nicotri, A. Italiano, G. Donvito

J. Kudzia, L. Dutka,
M. Orzechowski, M. Borzecki

More details?

Our wiki:

http://ecrin-mdr.online/index.php/Project_Overview

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURE NETWORK

Thank you!

Any questions?