

Metadata traceability and reuse in research projects

Alberto Ballesteros Rodríguez, Miguel Ángel Sicilia
Biomedical Data Science and Engineering Unit at IRYCIS

Background

Why metadata traceability and reuse matter?

- Why it matters in biomedical research?
 - Biomedical research generates huge amount of sensitive and complex data
 - Metadata traceability improves transparency, compliance, and reproducibility
 - Faster discoveries and less effort due to reusability
- What is a DMP?
 - A Data Management Plan (DMP) defines how research data is collected, processed, stored, and shared throughout a project.
 - Comply with FAIR principles (Findable, Accessible, Interoperable, Reusable)
- Machine-actionable DMPs (maDMPs) Python library: <https://madmpy.readthedocs.io>
 - Structured and standardized DMPs that systems can automatically process
 - Dynamic updates, validation, and integration with other research tools
 - RDA DMP Common Standard as a widely adopted specification

Electronic Data Capture (EDCs) software in research projects

- Structured tools used for capturing research data (e.g., REDCap)
 - Widely used in hospitals, clinical trials, and observational studies

Data fragmentation

- Each project stores data in isolated silos
- Disconnected from DMPs
 - No direct link between DMPs and the data actually collected
 - Compliance and planning can't be verified or automated

Consequences

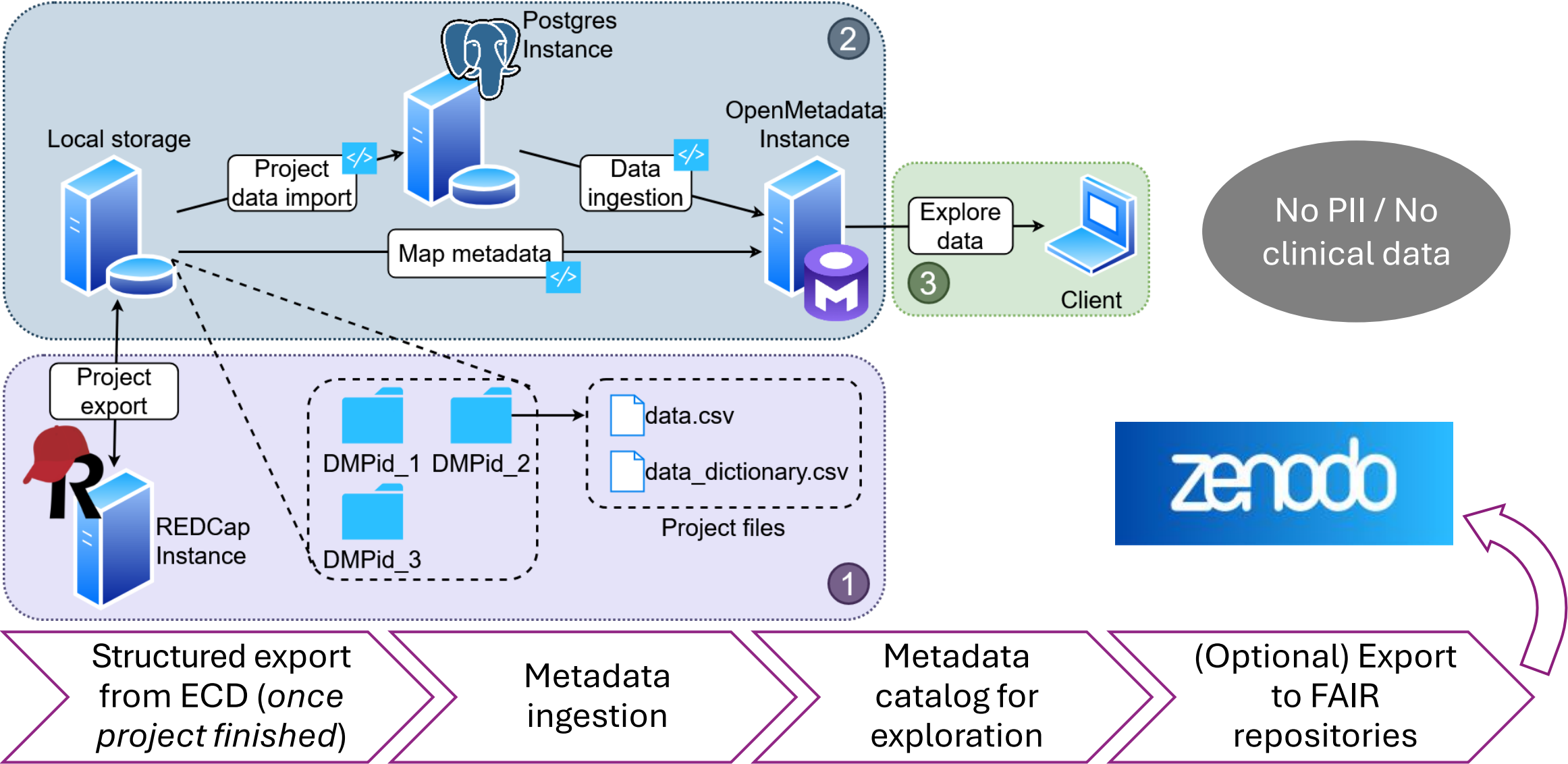
- Hard to know what data exists and where it was used
- Limits interoperability, reproducibility, or secondary use
- Reduced transparency and oversight at institutional level

Record ID record_id	Data Access Group redcap_data_access_group	Sexo: sexo	Fecha de nacimiento: fecha_nacimiento	Edad edad	Peso: peso	Altura: altura	IMC: imc	HTA hta	Tabaquismo tabaquismo	Dislipemia dislipemia	Diabetes mellitus diabetes_mellitus	Tratamiento: tratamiento_diabetes	Fa fi
<u>1</u>	HU Ramon y Cajal							Sí (1)	Sí (1)	No (0)	No (0)		S
<u>2</u>	HU Ramon y Cajal							Sí (1)	No (0)	Sí (1)	No (0)		N
<u>3</u>	HU Ramon y Cajal							Sí (1)	Sí (1)	Sí (1)	No (0)		S
<u>4</u>	HU Ramon y Cajal							Sí (1)	No (0)	No (0)	No (0)		N

Proposal

From disconnected silos to metadata-centric integration

- Metadata-driven integration
 - Centralize project metadata in a unified catalog
 - Automatically ingest structured exports from ECD systems
 - Enrich metadata with terminologies (e.g., SNOMED CT, LOINC, ...)
- Bridge between planning and reality
 - Link DMPs (planning) to real data (ECDs) using shared identifiers
 - Allow validation of compliance between expected and collected data
- Toward reuse and interoperability
 - Identify common metadata across projects
 - Support metadata reuse for new studies
 - Facilitate integration with CRIS systems and standards



- Mapping process
 - Metadata enriched with standard clinical terminologies (in the ECD)
 - Vocabularies, curated by the OHDSI Athena repository, provide terminology mappings
 - During ingestion, codes extracted and converted into semantic tags in OpenMetadata
- Why semantic tagging?
 - Cross-project metadata harmonization
 - Supports integration with research networks and standard terminologies
 - Discoverability and filtering by concept or domain
 - Reuse semantic alignment across systems
- Future extensions
 - Expand to additional vocabularies
 - Support institution-specific terminologies or controlled vocabularies
 - Compatible with CERIF classification schemes through mapping layers

- What researchers and data stewards can do
 - Browse all project schemas and metadata in a unified catalog
 - Filter variables by name, data type, or terminology tag
 - Identify overlapping attributes across studies (e.g., "BMI", "Age", ...)
 - Analyze collection patterns and metadata reuse potential
- Results
 - Improved discoverability of existing data assets
 - Ensures semantic consistency across projects
 - Easier integration with external systems and analysis tools

Project structure in OpenMetadata

Database Services / eurocris / fevar

public

No Domain | No Owners | No Tier

Tables 4

Stored Procedures 0

Activity Feeds 0

Custom Properties

Description

standard public schema

Deleted

NAME	DESCRIPTION
eventos_articos_mayores	No description
perioperatoria	No description
postoperatoria	No description
preoperatoria	No description

Data Products

--

Tags

+ Add

Glossary Term

+ Add

All

Search for Data Assets

ctrl K

All Domains EN ? admin Default

0

0

0.1

*Each table represents a structured part of the clinical study

Metadata traceability and reuse in research projects – euroCRIS MM 2025

Biomedical Data Science and Engineering (IRYCIS)

May 14, 2025

9 / 15

Metadata structure and semantic tagging in OpenMetadata



All ▾

Search for Data Assets

ctrl K

🔍

All Domains ▾ EN ▾ 🔔 ? 



eurocris / fevar / public

 **preoperatoria**

No Domain  | No Owners  | No Tier  | Retention Period: --  | Type: Regular | Usage: 0th pctile

Schema

Activity Feeds & Tasks 24Sample DataQueries 0Data ObservabilityLineageCustom Properties

NAME	TYPE	DESCRIPTION	TAGS	GLOSSARY TERMS	DATA QUALITY
sexo	character varying	No Description	 263495000	+ Add	--
fecha_nacimiento	character varying	No Description	 184099003	+ Add	--
edad	double precision	No Description	 SNOMED.184099003 Fecha de nacimiento / Date of birth669002	+ Add	--
peso	character varying	No Description	 43664005	+ Add	--
altura	character varying	No Description	 309534003	+ Add	--

localhost:8585/tag/SNOMED.184099003

**Semantic tag from SNOMED CT: 184099003 (Date of birth)*

Terminology navigation in OpenMetadata

The screenshot shows the OpenMetadata interface with the following components:

- Top Bar:** Search bar with 'All' and 'SNOMED' filters. User profile 'admin' with 'Default' role.
- Left Sidebar:** Navigation icons for Data Assets, Tags (10 items), and other features.
- Main Content Area:** A list of SNOMED terms with their IDs and descriptions.
 - 184099003:** Fecha de nacimiento / Date of birth. No Domain • No Owners.
 - 263495000:** Género / Gender. No Domain • No Owners.
 - 397669002:** Edad / Age. No Domain • No Owners.
 - 309534003:** Altura normal / Normal height. No Domain • No Owners.
- Right Panel:** Detailed view of the term '184099003'. It shows the usage 'preoperatoria' and a list of associated terms: 397669002, 263495000, and 184099003.

**Users can explore clinical terminologies directly from the metadata catalog.*

Concept-based discovery across projects

All

Search for Data Assets

ctrl K

All Domains

EN

admin

Default

Classifications / SNOMED

184099003

184099003

Overview

Assets 1

Description

Fecha de nacimiento / Date of birth

Add Assets

<>

Domain

**Where has this concept been used across projects?*

Overview

Assets 1

Search Assets

☐ Select All

☐ eurocris / fevar / public / preoperatoria

preoperatoria

No description

No Domain • No Owners • Usage 0th pctile

preoperatoria

Type Regular

Queries No Queries

Columns 39

Incidents 0

Profiler & Data Quality

Profiler is not enabled for this table.

- Centralized metadata as a source
 - OpenMetadata acts as a central metadata layer
 - Unified structure enables cross-project visibility
 - Metadata is enriched, structured, and semantically tagged
- Integration with other standards
 - Metadata can be exposed via APIs or exported in standard formats (JSON, XML, CSV)
 - Mappings can be created to align OpenMetadata objects with CERIF entities:

CERIF Entity	Mapped concept in proposal
cfResProd	Datasets or individual variables
cfProj	ECD projects (e.g., REDCap projects)
cfPers / cfOrgUnit	Project creators, PIs, departments
cfClassScheme	Domain-specific terminologies

- Benefits
 - Unified view across projects
 - Simplifies management & discovery
 - Compliance with FAIR & RDA-DMP standards
 - Interoperability and reusability
 - Adaptable to different standards, institutions, research environments
- Limitations
 - Requires institutional IT integration & infrastructure support
 - Metadata quality depends on input
 - Poor annotation reduces value and discoverability

- Conclusions
 - ECD data is often disconnected from DMPs and institutional metadata systems
 - Semantic tagging enhances interoperability and reuse
 - Model compatible with different vocabularies
 - Model scalable across institutions
 - FAIR-aligned metadata can be exported to public repositories
- Outlook
 - Integration with institutional CRIS
 - Extend tagging support to other vocabularies
 - Publish reusable tools (e.g., ingestion scripts, tag mappers, ...)
 - Link to external DMP systems for automated validation and synchronization



euroCRIS Membership Meeting 2025
Leuven, Belgium

Thank you!

Alberto Ballesteros Rodríguez, Miguel Ángel Sicilia
Biomedical Data Science and Engineering Unit at IRYCIS
alberto.ballesterosr@uah.es



MDG-23-11625