Building an Ontology to FAIRify RWD Metadata – A proposal

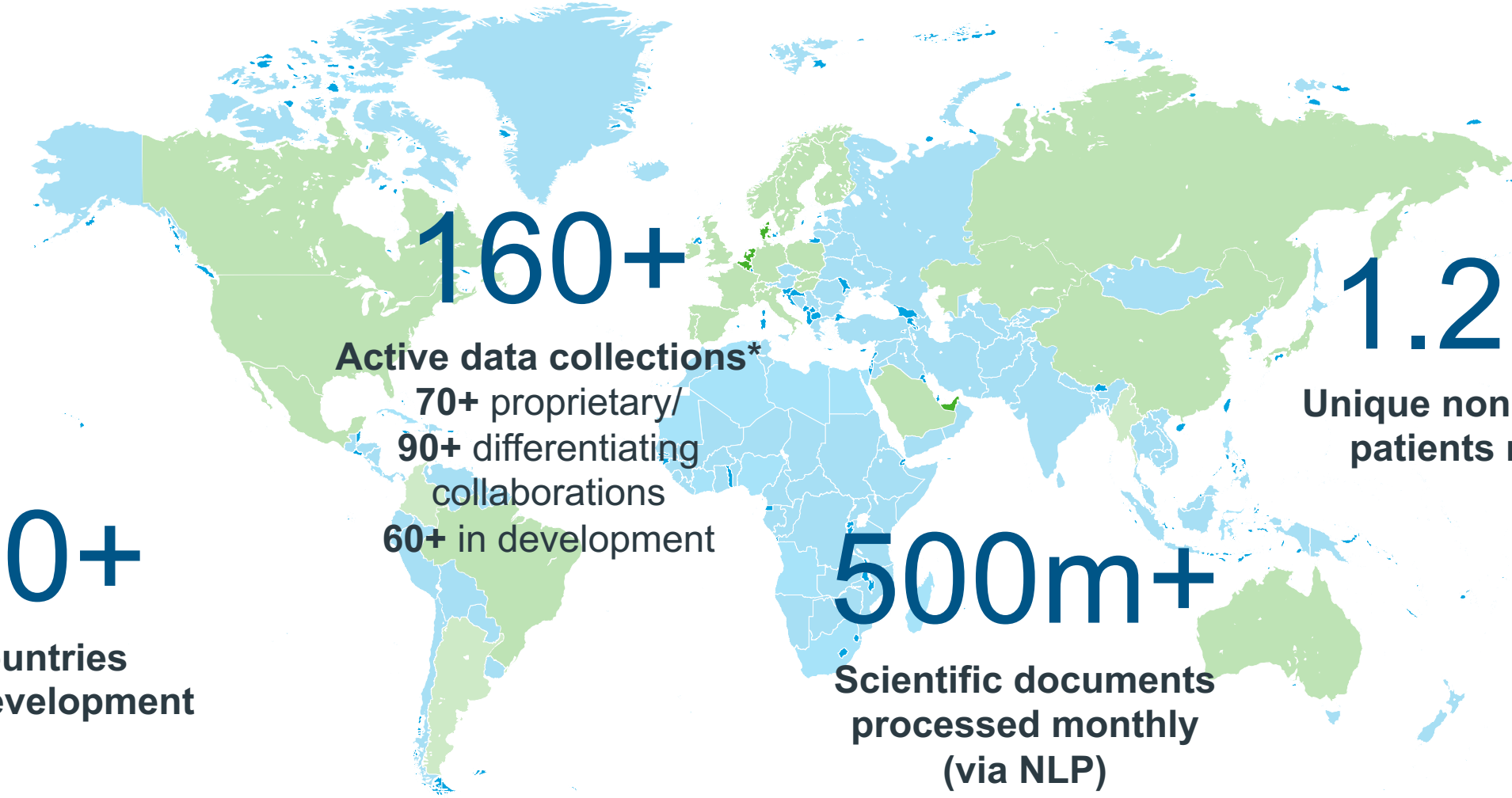
Pistoia Alliance 2024 London Conference

Emiliano Reynares
Andrea Splendiani
Michele Arnoe

© 2024. All rights reserved. IQVIA® is a registered trademark of IQVIA Inc. in the United States, the European Union, and various other countries.

RWD in IQVIA

IQVIA features an expansive Real World Data portfolio



160+

Active data collections*
70+ proprietary/
90+ differentiating
collaborations
60+ in development

1.2B+

Unique non-identified patients records

50+

Countries
5 in development

500m+

Scientific documents processed monthly (via NLP)

IQVIA health data catalogue

One stop shop to quickly identify the right data sources

Dataset content description
Patient demographics, Biometrics details, Diagnosis and treatment information, referential dictionaries

Data enhancement capabilities
Platform, Linkage, Enrichment

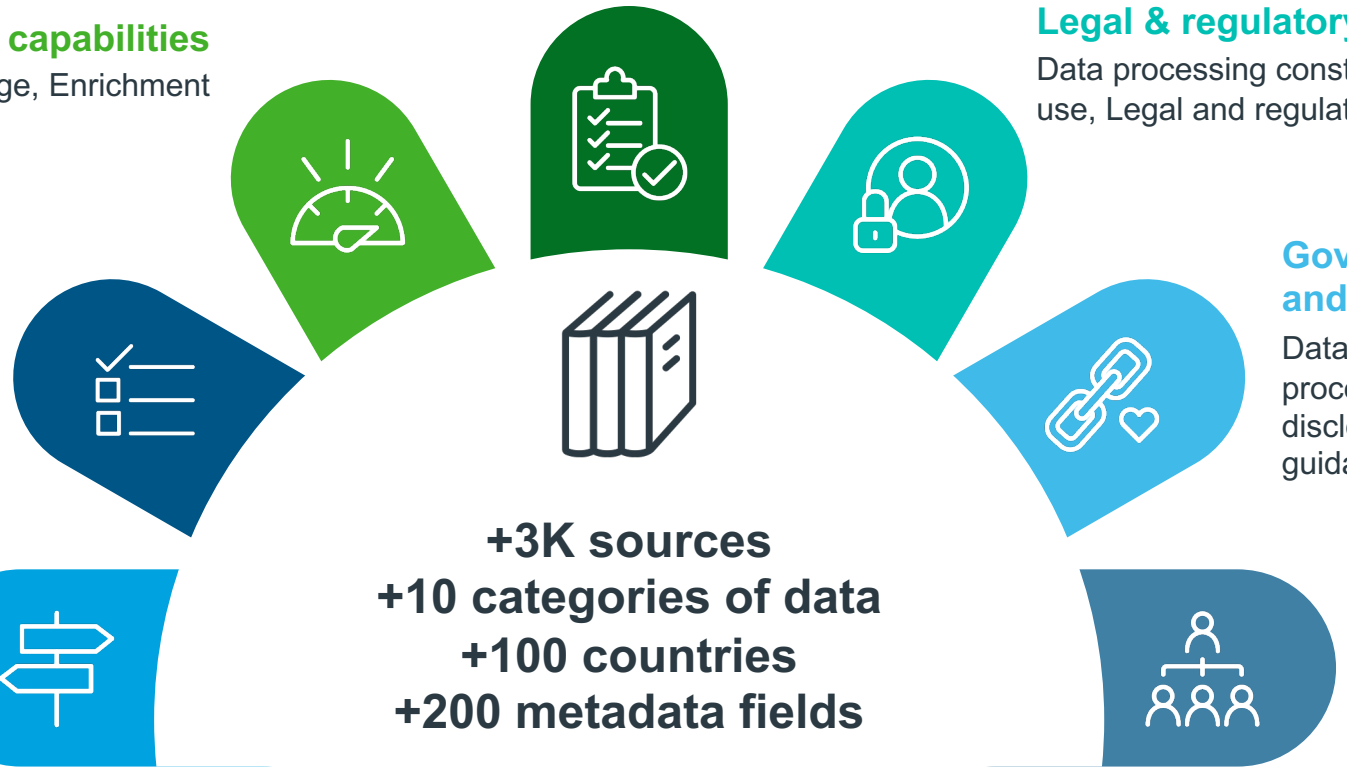
Legal & regulatory requirements
Data processing constraints, Prohibited use, Legal and regulatory requirements

Dataset characteristics
Data Type, collection methodology, Representativeness, Coverage, History

Governance processes and guidances
Data privacy and data governance processes, (Data access, data disclosure and data transfer guidance), Data security guidance

Data usage
Projection, granularity, appropriate use by TA and audience, Considerations (i.e. limitations)

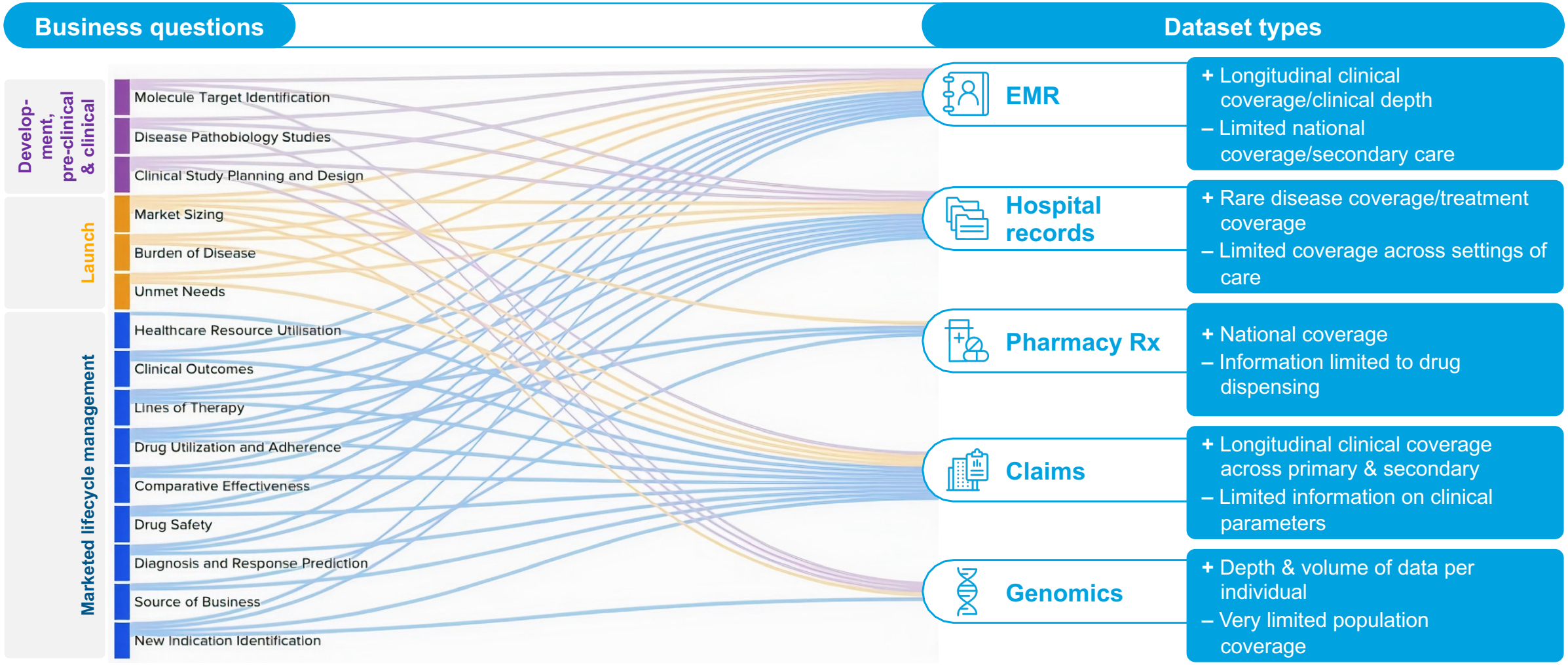
Data quality and validation
Data conformance, data completeness, Data Plausibility



+3K sources
+10 categories of data
+100 countries
+200 metadata fields

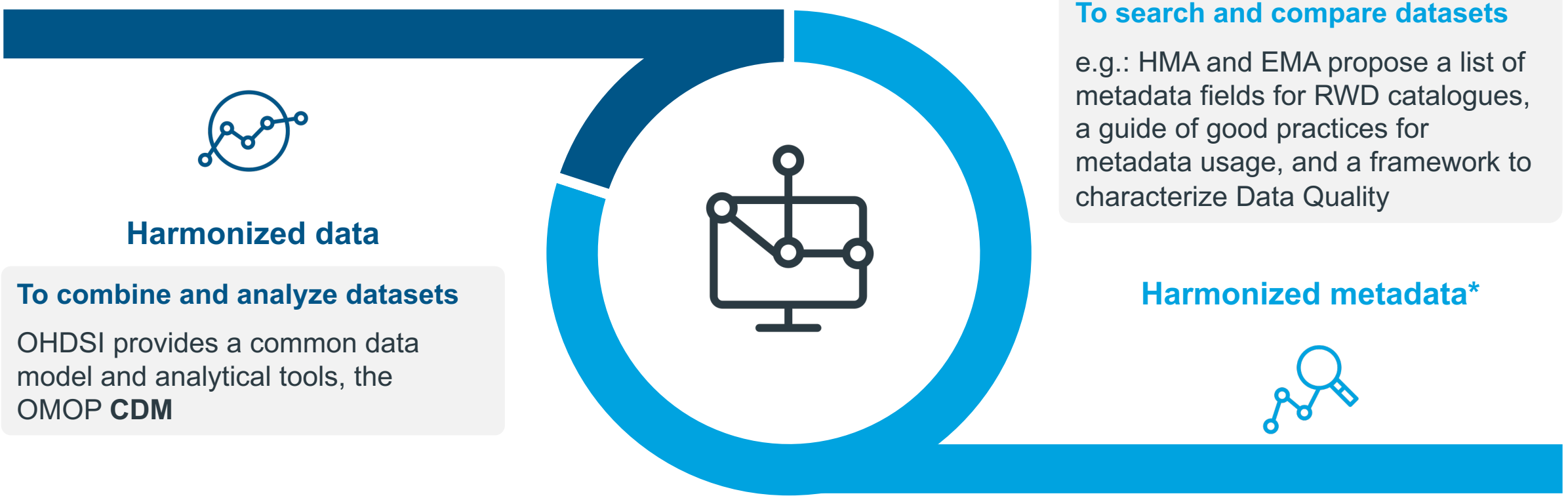
Real-World Data sources and research questions

For any researcher in the healthcare ecosystem who wants to leverage fit for purpose dataset



How to make use of this vast and diverse amount of RWD?

It's (also) about finding, comparing, selecting and understanding your assets



Harmonized data

To combine and analyze datasets

OHDSI provides a common data model and analytical tools, the OMOP **CDM**

To search and compare datasets

e.g.: HMA and EMA propose a list of metadata fields for RWD catalogues, a guide of good practices for metadata usage, and a framework to characterize Data Quality

Harmonized metadata*

*metadata – Descriptive data that characterize other data to make it understandable in terms of meaning and fit-for-purpose

What if RWD metadata would be FAIR?

What if we would have a way to annotate data represented as an ontology?

Findable

Metadata and data should be easy to find for both human and computers, supporting automated discovery and complex search

Interoperable

Metadata should be easily integrated with other (metadata) and directly usable by applications and workflows for data processing

Accessible

Metadata should be always available and clarify how and when data is accessible

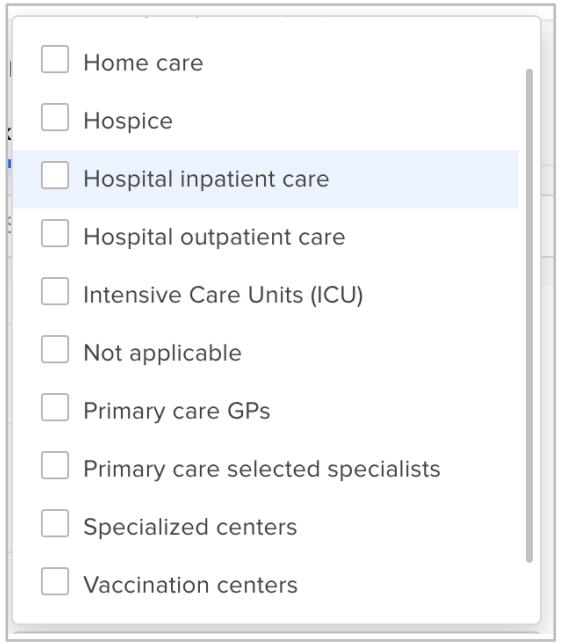
Reusable

Metadata should follow community standards and describe the data to understand its fit-for-purposedness

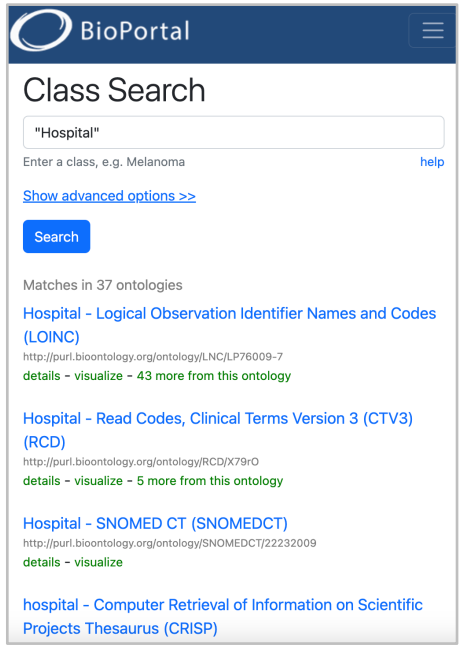
What is missing to make RWD Metadata FAIR? (the simple)

What identifiers should we use for metadata elements?

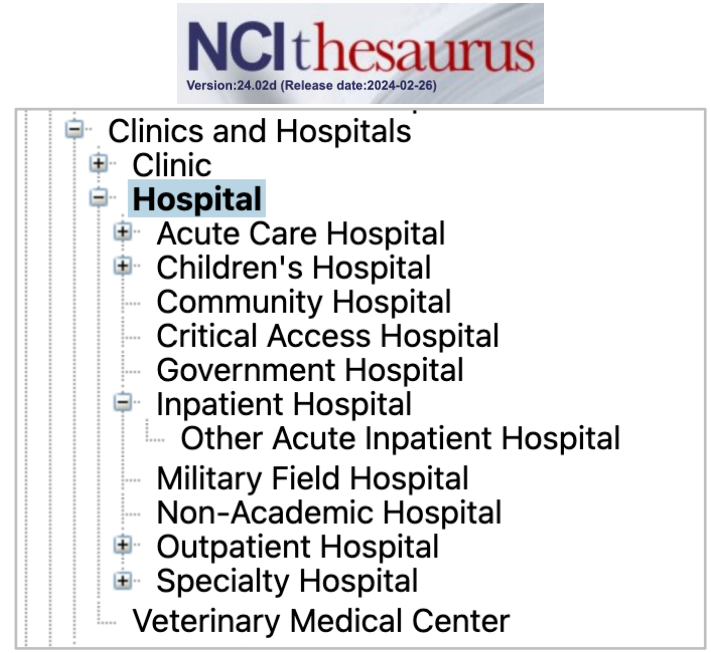
Making a list of provider types FAIR: What ontology?



Too many options



Not an ideal coverage



Definitions	A hospital in which patients are admitted for treatment and kept overnight or longer.
ID	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C201881
code	C201881

What is missing to make RWD Metadata FAIR? (the complex)

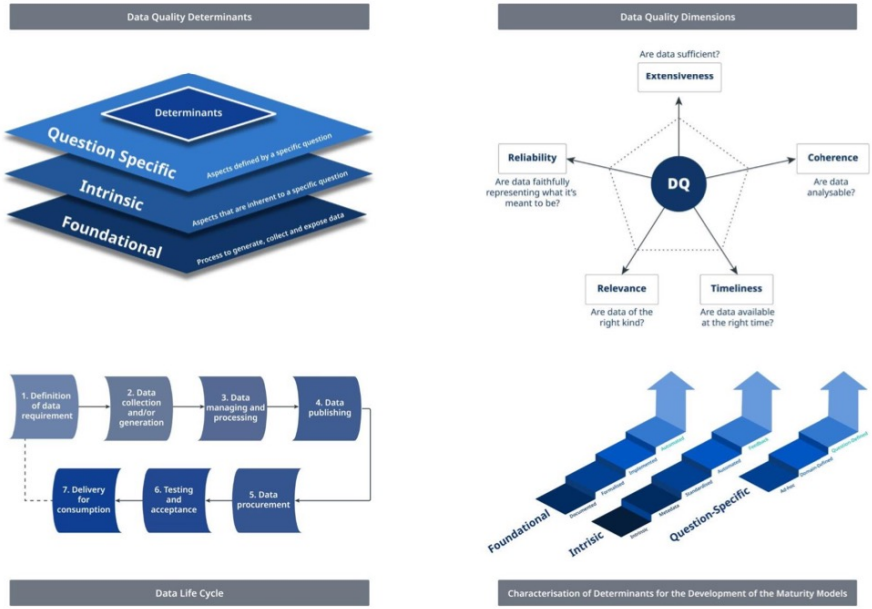
How can we characterize data quality?

HMA
Heads of Medicines Agencies

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

30 October 2023
Data Analytics and Methods Task Force
EMA/326985/2023

Data Quality Framework for EU medicines regulation



EDM Forum
EDM Forum Community

eGEMs (Generating Evidence & Methods to improve patient outcomes)

9-11-2016

A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data

TEHDAS
Towards European Health Data Space

Deliverable 6.3

Recommendations on a Data Quality Framework for the European Health Data Space for secondary use

26 September 2023

Data on the Web Best Practices: Data Quality Vocabulary

W3C Working Group Note 15 December 2016

This version:
https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/

Latest published version:
https://www.w3.org/TR/vocab-dqv/

IQVIA

Initial steps

Building an ontology for RWD Metadata

A Metadata Profile to Promote Real-World Data Discoverability

Emiliano Reynares[1], Andrea Splendiani[2] and Hanne van Ballegooijen[3]

[1] *IQVIA, Provença 392, Barcelona, Spain*
[2] *IQVIA, Kirschgartenstrasse 14, Basel, Switzerland*
[3] *IQVIA Solutions B.V., Herikerbergweg 314, Amsterdam, the Netherlands*

Abstract

Real-World Data (RWD) are patient level data routinely collected from a variety of sources other than traditional clinical trials. RWD is increasingly needed for purposes such as making drug discovery more efficient and supporting healthcare systems to become more sustainable. However, the specific characteristics of RWD pose integration and findability-related challenges that arise from the heterogeneity, dynamism, and voluminosity of their sources. While previous work has addressed these issues by defining common data models and frameworks to integrate RWD into specific processes, a machine-actionable representation is needed to promote data discoverability and fit-for-purpose assessment. This paper introduces a Data Catalog Vocabulary (DCAT)-based metadata profile for an RWD-specific metadata vocabulary. The profile aims to tackle the findability-related challenges by enabling querying and processing of metadata in a standardized fashion.

Proceedings of SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26-29, 2024, Leiden, The Netherlands.

What standard is out there?

DCAT or Schema.org -based? Both are are domain-agnostic but …

DCAT

Structured vocab to describe datasets and data services

Aimed to describe datasets and data services

Developed by the W3C

Comprises RDF –specified classes, properties and constraints

Reuses vocabularies like the Dublin Core Metadata Terms Recommendation

Extensible by adding classes, properties, constraints, other vocabularies and access mechanisms. Example: DCAT-AP, the DCAT profile for European data portals

Schema.org

Structured markup vocab to describe web resources

Aimed to describe web resources, e.g. books, events

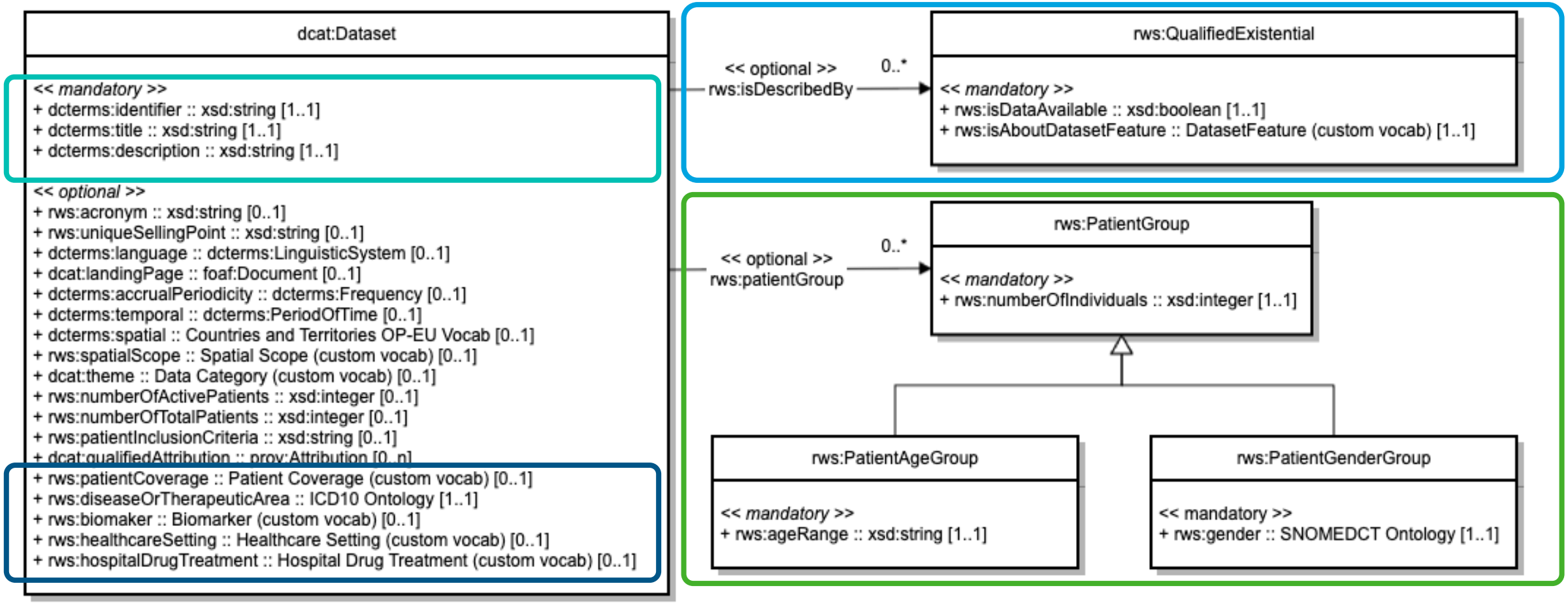
Launched by Google, Yahoo, Bing, and Yandex

Defines types of information and their properties

Extensible by adding types, properties, and constraints. Example: Bioschemas, the schema.org of the life sciences

Preliminary DCAT profile for RWD Metadata

Extension with new properties, classes and constructs to cover the specifics of RWD



If you want to go fast, go alone. If you want to go far, ***go together!***

We propose to build a RWD metadata ontology

A standard-based ontology to describe RWD assets aimed to...

1. **Increase the findability of data assets**
2. **Improve the understanding of their content & fit-for-purpose assessment**
3. **Enable comparison of different data sources**
4. **Support automation of data processes**

ontologies@iqvia.com