Outstanding Challenges in Federated Data Reuse: Data Quality in FAIR Data Points

By
Priyanka Ojha https://orcid.org/0000-0002-6844-6493
Amsterdam UMC

14/01/2025

What is Metadata

Metadata has been defined as “data describing the context, content and structure of records.[1]

Metadata includes information about technical and business processes, data rules and constraints, and logical and physical data structures. It describes the data itself, the concepts the data represents , and the connections between the data and concepts. [2]

1

https://committee.iso.org/files/live/sites/tc46sc11/files/documents/N800R1%20Where%20to%20start-advice%20on%20creating%20a%20metadata%20schema.pdf

2 DAMA DMBOK 2nd Edition, Page # 393

Types of MetaData

In Library / Information science, following categories of metadata exists :

1. Descriptive Metadata (e.g. Title, Author, DOI etc) describes a resource and enables identification and retrieval.
2. Structured Metadata describes relationships within and among resources and their components i.e. using standardized schemas, vocabularies, and formats. E.g. DCAT, JSON-LD etc.
3. Administrative Metadata(e.g. version numbers, archival dates etc) is used to manage resources over their life cycles.

What is Fair Data Point

- FAIR Data Points are the building blocks of a federated data ecosystem, enabling data sharing, reuse, and analysis across different domains and organizations.

- A FAIR Data Point ultimately stores information about data sets, which is the definition of metadata. And just like the webserver in the WWW in the beginning of the 1990s brought the power of publishing text to anyone, a FAIR data point aims to give anyone the power of putting their own data on the web. [1]

1: https://www.fairdatapoint.org/

What is Health Data Catalog

The National Health Data Catalogue enables sharing, finding and combining health data. The Catalogue provides an overview of data in the field of health and life sciences in the Netherlands. As a user, you can search and find data(sets). You can view more information (metadata) about each dataset and find contact details for requesting access to the data(sets) for research, policy and innovation. https://catalogus.healthdata.nl/

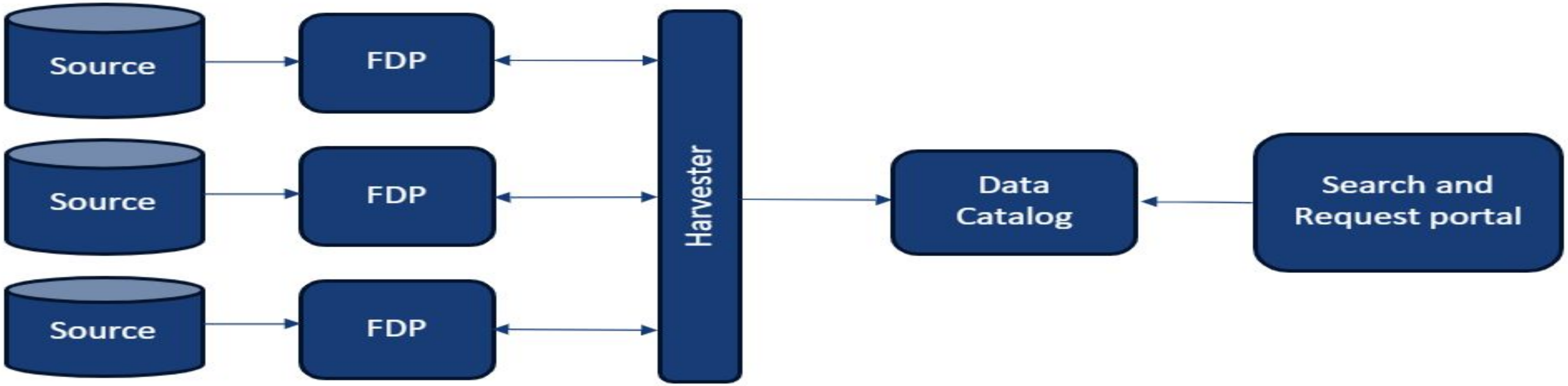

Image Source :

https://health-ri.atlassian.net/wiki/spaces/FSD/pages/279150593/Metadata+onboarding+on+the+National+Catalogue

Health DCAT-AP , DCAT-AP NL and HRI Metadata Schema

Health DCAT-AP (https://healthdcat-ap.github.io/)

- **Purpose:** An health-related extension of the Data Catalog application profile (DCAT-AP) specifically designed for sharing information about Catalogues containing Datasets and Data Services descriptions in Europe
- **Focus:** Covers a wide range of health-related data, including Omics, Imaging, Biobanks and collection.

DCAT-AP NL (https://docs.geostandaarden.nl/dcat/dcat-ap-nl30/)

- **Purpose:** A Dutch-language extension of DCAT-AP.
- **Focus:** Provides a Dutch-language vocabulary for describing data resources for government, semi government organisations.

HRI Metadata Schema (https://github.com/Health-RI/health-ri-metadata/tree/master)

- **Purpose:** It describes the minimum amount of information that should be used to describe resources across Health-RI nodes through the national directory, which is in line with what Plateau 1 offers. The schema can be changed or extended to meet the needs of different areas, and new versions will be released in the future.
- **Focus:** To make it easier to share, find and reuse data within Health-RI nodes via National Catalog and interoperable with other EU portals

Issues & Challenges for metadata mapping for FDP’s - 1

- **Data Inconsistency**:
    - Inconsistent Terminologies and conflicting definitions in source systems : E.g. Cohort / Study / Consortium/ Project etc.
    - Data discrepancies: The same data point may have different values across different sources, leading to confusion and inaccurate data . E.g. Publisher / Owner of Dataset
    - Inconsistent data formats, units, and naming conventions across sources hinder data integration and analysis.

Issues & Challenges for metadata mapping for FDP’s -2

- The presence of many mandatory fields in DCAT-AP poses a significant challenge when source systems lack the necessary data to populate them. https://github.com/Health-RI/health-ri-metadata/tree/develop_v2?tab=readme-ov-file
- Data Lineage challenges : It can be difficult to trace the origin, transformation, or usage of data assets across different systems for within an organisation and or external systems if things are done manually / without documentation.
- Lack of a single, reliable source of truth and curated datasets internally for exposure through FDPs
- Data providers often fail to prioritize persistent URLs for datasets and crucial metadata fields (e.g., Author, Organization, Funders).

Issues & Challenges for metadata mapping for FDP’s -3

  - Metadata schemas, SHACL versioning, and the challenges of schema validation and backwards compatibility for previously published datasets.
  - Metadata mapped incorrectly : e.g. https://github.com/FAIRDataTeam/FAIRDataPoint/issues/571
  - Lack of reliable schema validators and clear validation error messages hinders users from fixing dataset issues. E.g. : https://data.europa.eu/mqa/shacl-validator-ui/data-provision
  - Feedback for draft-v2 : https://github.com/Health-RI/health-ri-metadata/issues/169

  https://github.com/Health-RI/health-ri-metadata/issues

- How is same data across multiple organisations displayed, if they generate their internal PID’s and how to handle deduplication in this scenario.
- Deviation from the original Health DCAT-AP is a risk from my perspective and may hamper interoperability.
- Link rot issues.

Issues & Challenges for metadata mapping for FDP's : Demo

https://fdp.healthdata.nl/dataset/98ed818b-af83-4057-b5be-1e119993e143

- Multiple creators
- Unresolveable URL for Contact Point
- Publisher unknown (_g_L30C548) and different from Catalog, which has 2 publishers(_g_L16C1117, _g_L19C1221) https://fdp.healthdata.nl/catalog/a157fca9-2333-46ab-8ccc-b2197a4b2988
- No License specified.
- If you check the Turtle file, the authors are present, just mapped incorrectly. https://fdp.healthdata.nl/dataset/8b32b83f-a30f-47c4-baa0-b2c420bef851?format=ttl
- https://github.com/Health-RI/health-ri-metadata/blob/develop/Documents/Pre-release_metadata_CoreGenericHealth_p2.xlsx
- Good example : https://orphanet.fdps.ejprd.semlab-leiden.nl/

Possible suggestions /enhancements

- How frequently is metadata updated at the source systems (e.g., institutional repositories like Pure, Zenodo, DataVerse, Yoda)?
- Establish clear update schedules for harvesting and aggregating metadata from sources.
- Dataset expiry dates, if applicable.
- Publish the last processed date/time for each source to provide transparency.
- Document the rules and logic applied during metadata processing (e.g., data cleaning, transformation, validation).
- Maintain a history of metadata changes to track modifications and facilitate data provenance.
- Consider displaying a change log to users to increase transparency and trust.
- Map metadata fields from source systems (Pure, Zenodo, DataVerse, Yoda) to the target FDP schema (e.g., DCAT-AP).
- Define a core set of essential metadata fields for all sources to ensure consistency and data quality.
- Access rights , consents and Licences regarding data usage.
- Metrics for usage ,completeness, repository uptime etc (https://data.europa.eu/data/datasets/13015-gezonde-levensverwachting-geslacht-en-leeftijd/quality?locale=en)

Additional Reading

- https://www.health-ri.nl/nieuws/nederlands-metadata-profiel-dcat-ap-nl-officieel-vastgesteld
- https://www.health-ri.nl/health-ri-roadmap-plateauplanning
- https://healthdcat-ap.github.io/SENTIVE%20DATA%20HealthDCAT-AP%203.0.0.drawio.png
- https://github.com/Health-RI/health-ri-metadata/blob/develop_v2/Images/2.0_plateau2/HRI_metadata_p2.png
- https://health-ri.atlassian.net/wiki/spaces/FSD/pages/279150593/Metadata+onboarding+on+the+National+Catalogue
- https://zenodo.org/records/4697038
- https://research.vu.nl/ws/portalfiles/portal/220367454/Towards_a_standard_based_open_data_ecosystem_analysis_of_DCAT_AP_use_at_national_and_European_level.pdf
-

Acknowledgement & Thanks

I would like to thanks the colleagues I interacted with at Amsterdam UMC, Health-RI and others in relation to FDP.

Thank You !!!