A Framework for Evaluating and Improving Metadata Quality in the RADx Data Hub Using CEDAR Templates

Yan Cao, MS, Luna Baalbaki, MS, Crystal Han, MS, Matthew Horridge, PhD,
Marcos Martínez-Romero, PhD, Mark A. Musen, MD, PhD
Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

The Rapid Acceleration of Diagnostics (RADx) Data Hub is a centralized platform for curated, de-identified COVID-19 data, that enables researchers to discover, access, and analyze study data[1]. High-quality metadata are essential for ensuring that data are **FAIR (Findable, Accessible, Interoperable, and Reusable)**, but evaluation of metadata quality is a time-intensive task. This poster presents a semi-automated framework that is being used to find and resolve metadata issues in the Data Hub.

Example of "Subject Identifier" metadata in JSON-LD.

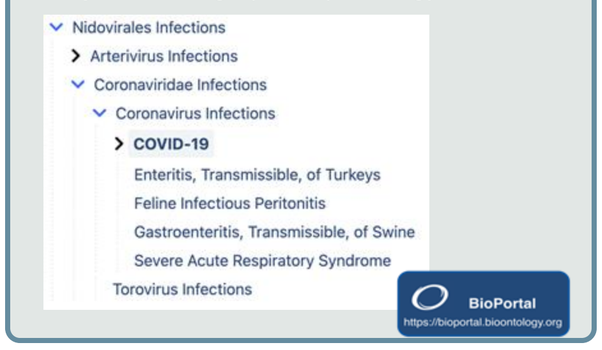
```
"Subject Identifier": {
  "@id": "http://purl.bioontology.org/ontology/MESH/D000086382",
  "rdfs:label": "COVID-19"
}
```

Visualization in the RADx Data Hub.

Our framework uses **CEDAR**[2] **templates**, which provide a precise specification of the expected metadata structure and values. The **Metadata Evaluator** first evaluates metadata against CEDAR template constraints, including controlled terms verification via **BioPortal**.

```
"Subject Identifier": {
  "type": "array",
  "minItems": 0,
  "items": {
    "_ui": {
      "inputType": "textfield"
    },
    "_valueConstraints": {
      "ontologies": [
        {
          "uri": "https://bioportal.bioontology.org/ontologies/MESH",
          "acronym": "MESH",
          "name": "MESH"
        }
      ],
      "valueSets": [],
      "classes": [],
      "branches": [],
      "requiredValue": false,
      "recommendedValue": true
    }
  }
}
```

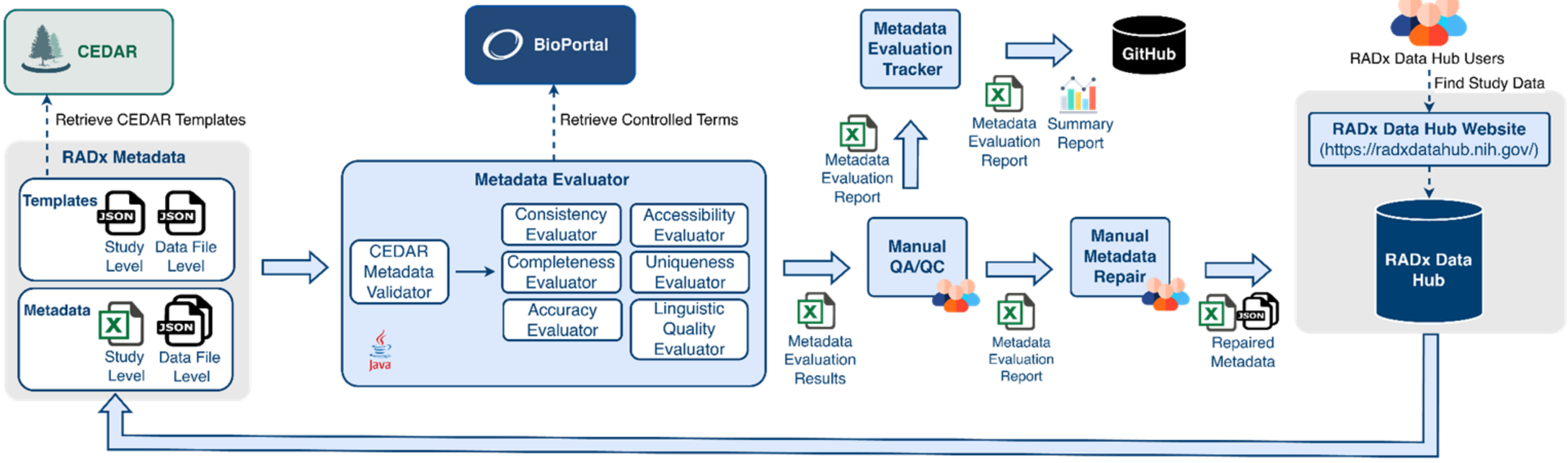
Retrieve ontology terms from **BioPortal**. ***COVID-19*** is a valid term in the Medical Subject Headings (MeSH) ontology.



Assess RADx Data Hub metadata across multiple evaluation criteria.

- ❑ **Validity**
- ❑ **Completeness**
- ❑ **Consistency**
- ❑ **Accuracy**
- ❑ **Accessibility**
- ❑ **Uniqueness**
- ❑ **Linguistic Quality**

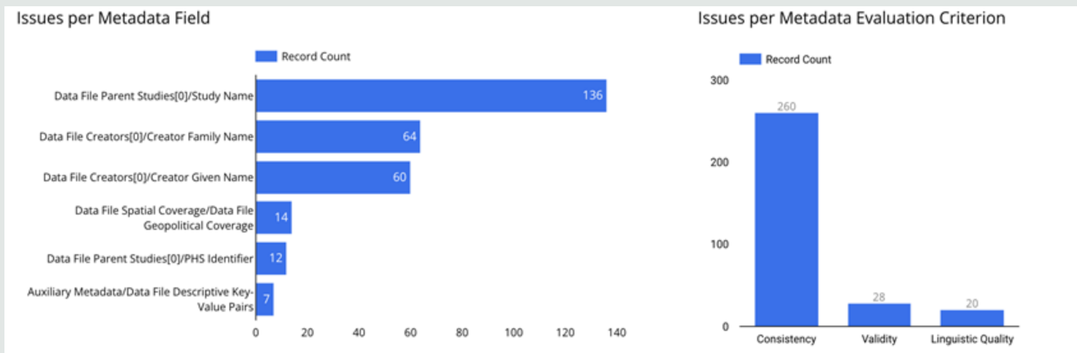
Metadata Evaluation and Improvement Schematic



Between December 2024 and January 2025, we evaluated metadata from 181 studies, identifying 586 issues currently being addressed.

Example of a metadata evaluation summary report (data file metadata):

308 issues across 172 files from 79 studies. Most relate to **Consistency**, including 134 mismatches between *Study Name* and study metadata, and 124 inconsistencies between *Creator Family/Given Name* and *Creator Full Name*. Smaller portions involve **Validity** (28 schema validation errors) and **Linguistic Quality** (20 formatting issues such as extra spaces).



Issue Count: 308
Distinct Data File Metadata Files: 172
Distinct Studies: 79

Metadata Issue Types

Issue Description Type	Criterion	Record Count
Study Name does not match the study metadata	Consistency	134
Creator Family Name is inconsistent with the provided full name	Consistency	64
Creator Given Name is inconsistent with the provided full name	Consistency	60
Extra space(s) found	Linguistic Quality	20
An empty string was detected in the Attribute Value field, which is not permitted.	Validity	14
Object instance has properties which are not allowed by the schema	Validity	14
PHS Identifier does not match the study metadata	Consistency	2

We developed a **semi-automated metadata evaluation framework** to access metadata quality across datasets. The framework integrates **pluggable evaluation criteria** with custom **CEDAR metadata templates, enabling flexible, reusable, and domain-adaptable assessments.** By systematically identifying issues across validity, completeness, consistency, accuracy, accessibility, uniqueness, and linguistic quality, the framework enhances the overall quality and reliability of metadata within the RADx Data Hub. In addition, **the framework allows for easy integration of new evaluation criteria and supports continuous refinement as metadata templates evolve.**

Future work will focus on enhancing automation, advancing AI-based validation, and strengthening support for customizable evaluation criteria.

Acknowledgments

This work was supported by the NIH under Other Transactions Authority award 1OT2DB000009.

References

1. Martinez-Romero M, et al. A Cloud-Based Platform for Harmonized COVID-19 Data: Design and Implementation of the Rapid Acceleration of Diagnostics (RADx) Data Hub. JMIR Public Health and Surveillance. 2025;11:e72677. doi:https://doi.org/10.2196/72677.

2. Musen MA, et al. Modeling community standards for metadata as templates makes data FAIR. Sci Data. 2022;9(1):696. doi:10.1038/s41597-022-01815-3.