The Ninth International Conference on Data Analytics
DATA ANALYTICS 2020
October 25, 2020 to October 29, 2020 - Nice, France.

A Comprehensive Study of Recent Metadata Models for Data Lake.

Redha Benaissa, Omar Boussaid, Aicha Mokhtari, and Farid Benhammadi

Redha BENAISSA,
benaissa.redha@gmail.com

Abstract

In the era of Big Data, an unrepresented amount of heterogeneous and unstructured data is generated every day, which needs to be stored, managed, and processed to create new services and applications. This has brought new concepts in data management such as Data Lakes (DL) where the raw data is stored without any transformation. Successful DL systems deploy efficient metadata techniques in order to organize the DL. This work presents a comprehensive study of recent metadata models for Data Lake that points out their rationales, strengths, and weaknesses. More precisely, we provide a layered taxonomy of recent metadata models and their specifications. This is followed by a survey of recent works dealing with metadata management in DL, which can be categorized into level, typology, and content metadata. Based on such a study, an in-depth analysis of key features, strengths, and missing points is conducted. This, in turn, allowed to find the gap in the literature and identify open research issues that require the attention of the community.

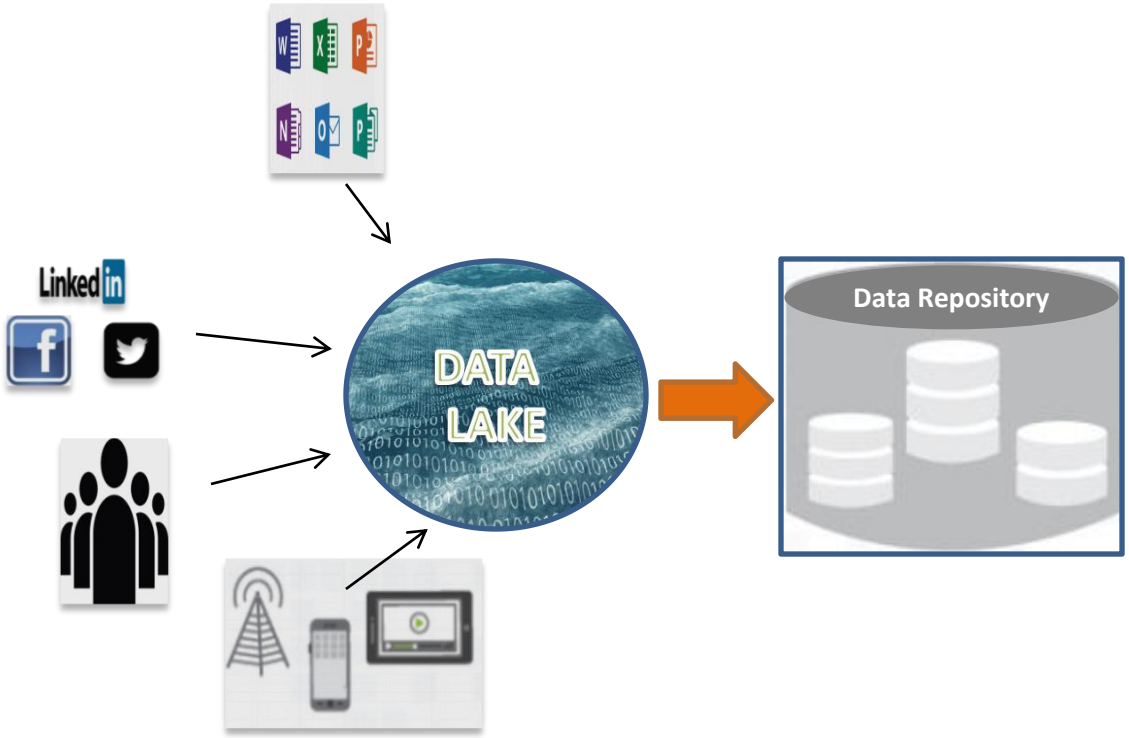
Topics of research interest

- Big Data Analytics.
- Metadata extraction in Data Lake.
- Scalable Machine Learning Algorithms using Apache Spark.
- Deploing scalable Machine Learning Algorithms for Data Lake.

Overview of Presentation

1. Introduction
2. Metadata in Data Lake (DL)
3. Metadata management systems in DL
4. Summary and open research issues
5. Conclusion and Future works

1. Introduction



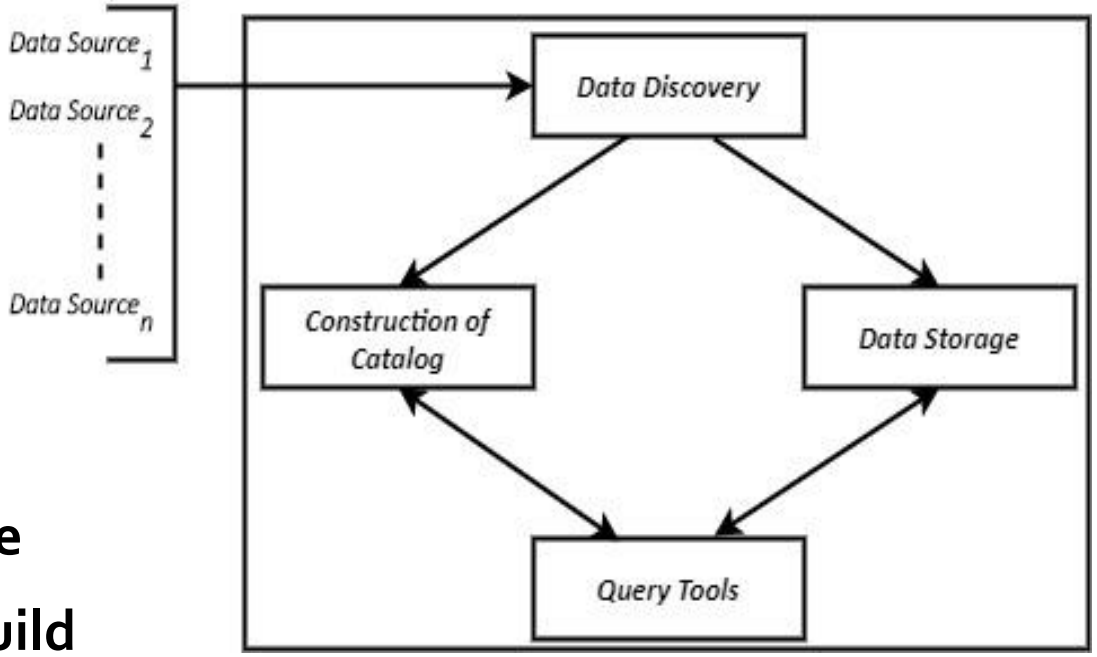
- ✓ A huge volume of data.
- ✓ Data in their raw format
- ✓ Heterogeneous data.

Some Data lake's capabilities:

- To capture and store raw data at scale for a low cost.
- To store many types of data in the same repository
- To perform transformations on the new data processing
- To define the structure of the data at the time it is used.

1. Introduction

- A **metadata catalog** provides the structure and semantics of each element ingested within the Data Lake.
- Challenge: extraction of the **right metadata** to build the catalog.

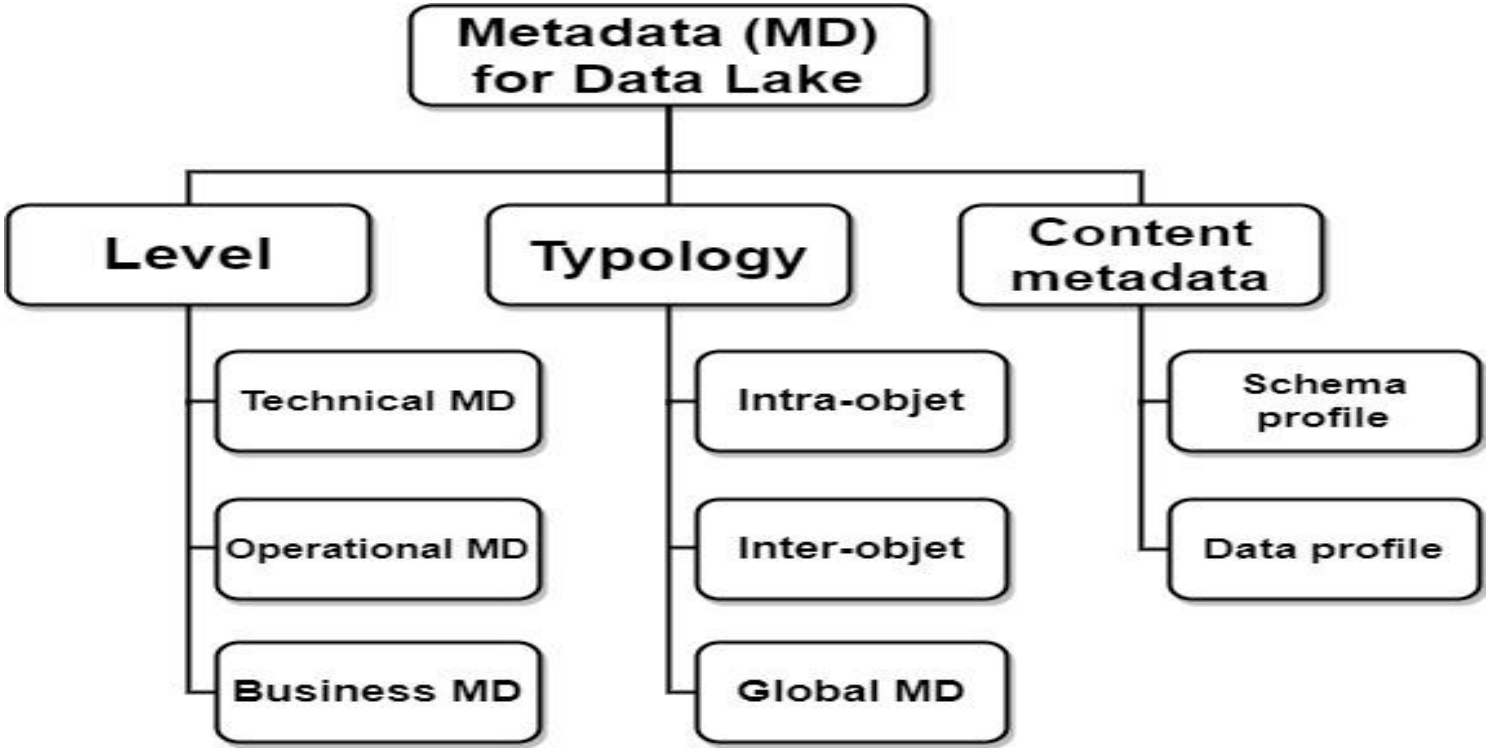


Data management process in a Data Lake.

- Goal: To present a comprehensive study of recent metadata models for Data Lake.

2.Metadata in Data Lake (DL)

- **3 Metadata categories in Data Lake.**



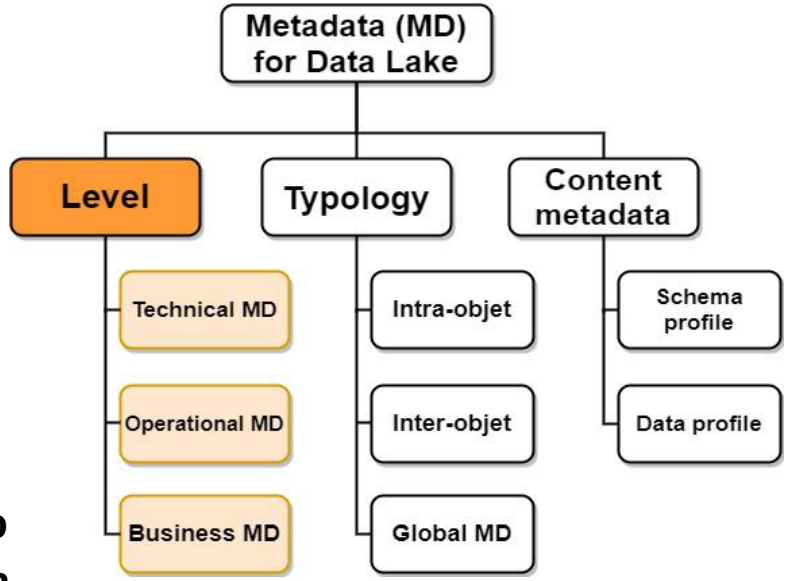
2.Metadata in Data Lake (DL)

A. Metadata (MD) Level

1) **Technical MD** describes the technical aspects of data sets. It is used to determine the *type of data encoding*.

2) **Operational MD** contains information on the *quality* and *origin* of the data. I

3) **Business MD** provides *meaning* and *semantics* to technical metadata to give more knowledge of the data sets. It provides information about the *data providers* and *source* systems.

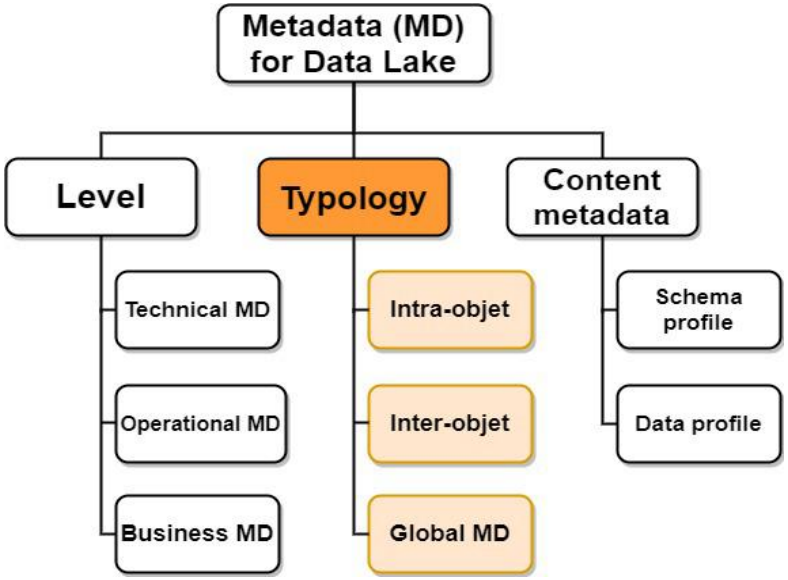


2.Metadata in Data Lake (DL)

A. Metadata (MD) Typology

1) **Intra-object MD** identifies *properties*, *summaries and previews, versions,* and *semantic* metadata associated with a given object.

2) **Inter-object MD** describes the *relationships* between at least two objects and has two main elements namely *object groupings* and *similarity links.*



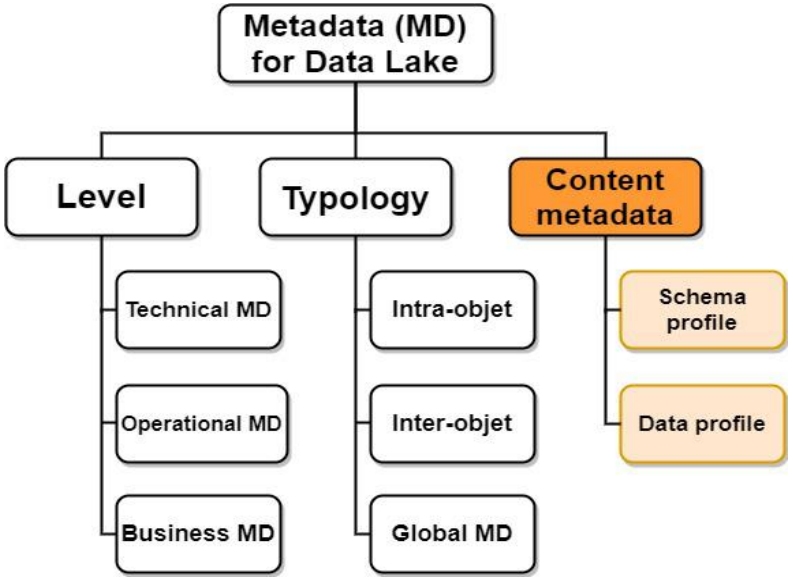
3) **Global MD** provide a contextual layer to the Data Lake that is essential for its analysis. Semantic resources are essentially knowledge bases (ontologies, taxonomies, thesauri, dictionaries) used to generate other metadata and improve analyzes.

2.Metadata in Data Lake (DL)

A. Content Metadata

1) **A schema profile** describes the schema of datasets (the number of attributes, the names of the attributes and their data types).

2) **The data profile** describes the values of the dataset, i.e., the statistics values of single-attribute. Information profiles are called *MD of relationships* between datasets.



3.Metadata management systems in DL

A. Network-based model for DL.

B. MEDAL

C. A generic and extensible classification of metadata-based System.

D. Model for integrating evolving heterogeneous data sources.

3.Metadata management systems in DL

A. Network-based model for DL

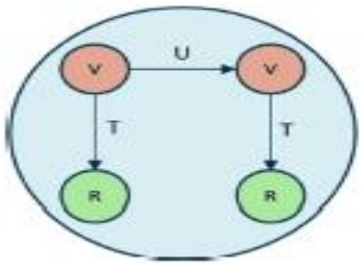
- The ***Complex knowledge model*** indicate a ***semantic relationship*** of data sources.
- A typical notation of ***XML, JSON*** is adopted to represent ***business metadata***.
- Based on ***an appropriate network*** to represent all the sources of DL.
- The extraction of complex knowledge model
  - → using tools based on graphs.
- Added an appropriate form of arcs
  - → The possible presence of synonymies between concepts belonging to different sources.
- Cases where synonymies are not sufficient to find a complex knowledge model
  - → An appropriate chain similarity metric is applied (N-Grams).

3.Metadata management systems in DL

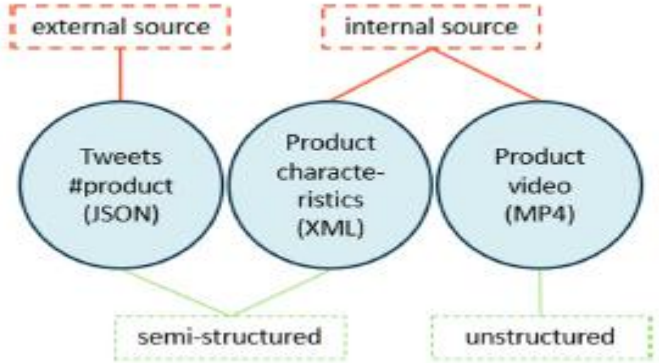
B. MEDAL

- A logical representation of metadata based on *a hypergraph.*
- *An object* is represented by a hypernode.
- *the forms of Objects* = Structured, Semi-structured and Unstructured data .
- *M* = (M_{intra};M_{inter};M_{glob}), where:

  M_{intra} is the set of intra-object metadata,
  M_{inter} is the set of inter-object metadata,
  M_{glob} is the set of global metadata.



Example of a hypernode



Interconnected hypernodes.

3.Metadata management systems in DL

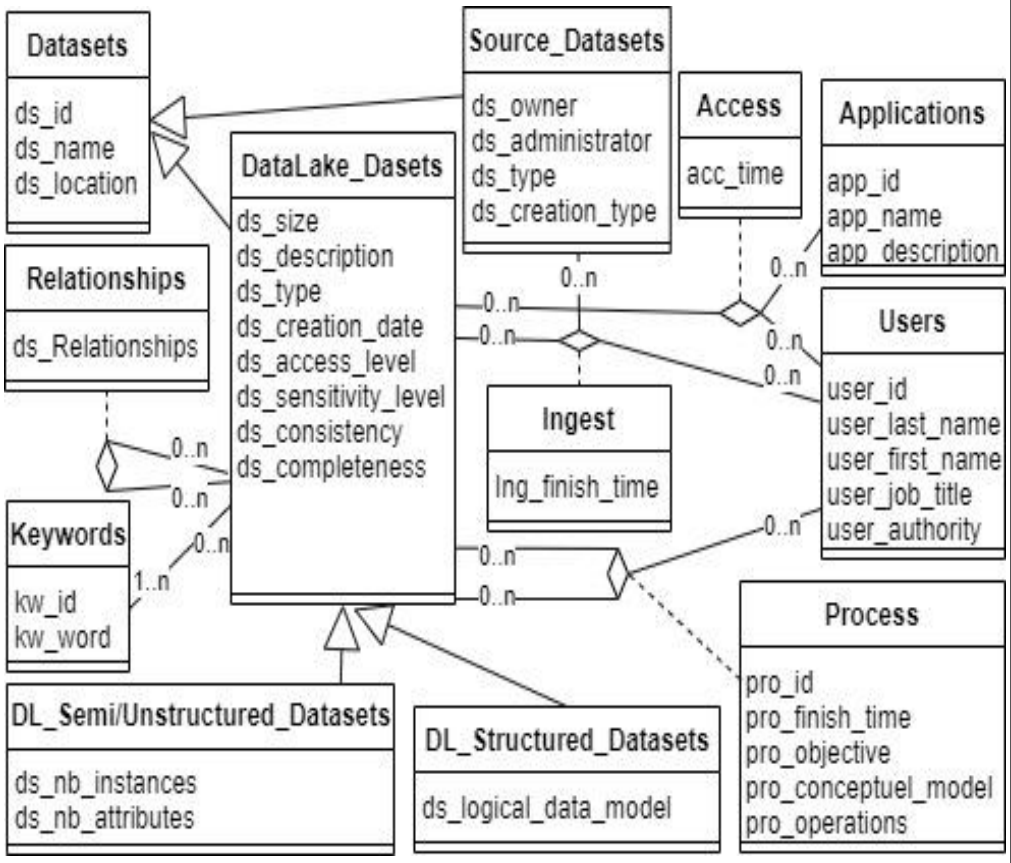
C. A generic and extensible classification of metadata-based System

- The proposed metadata classification integrates both **intra-metadata and inter-metadata** for all data sets or datasheets.
- **Inter-metadata**: the classification is completed by:

Dataset Containment, Partial overlap, Provenance, Logical clusters, Content similarity .

- **Intra-metadata**: the classification includes access, quality and security.

Data characteristics, Definition metadata (semantic and schematic metadata).

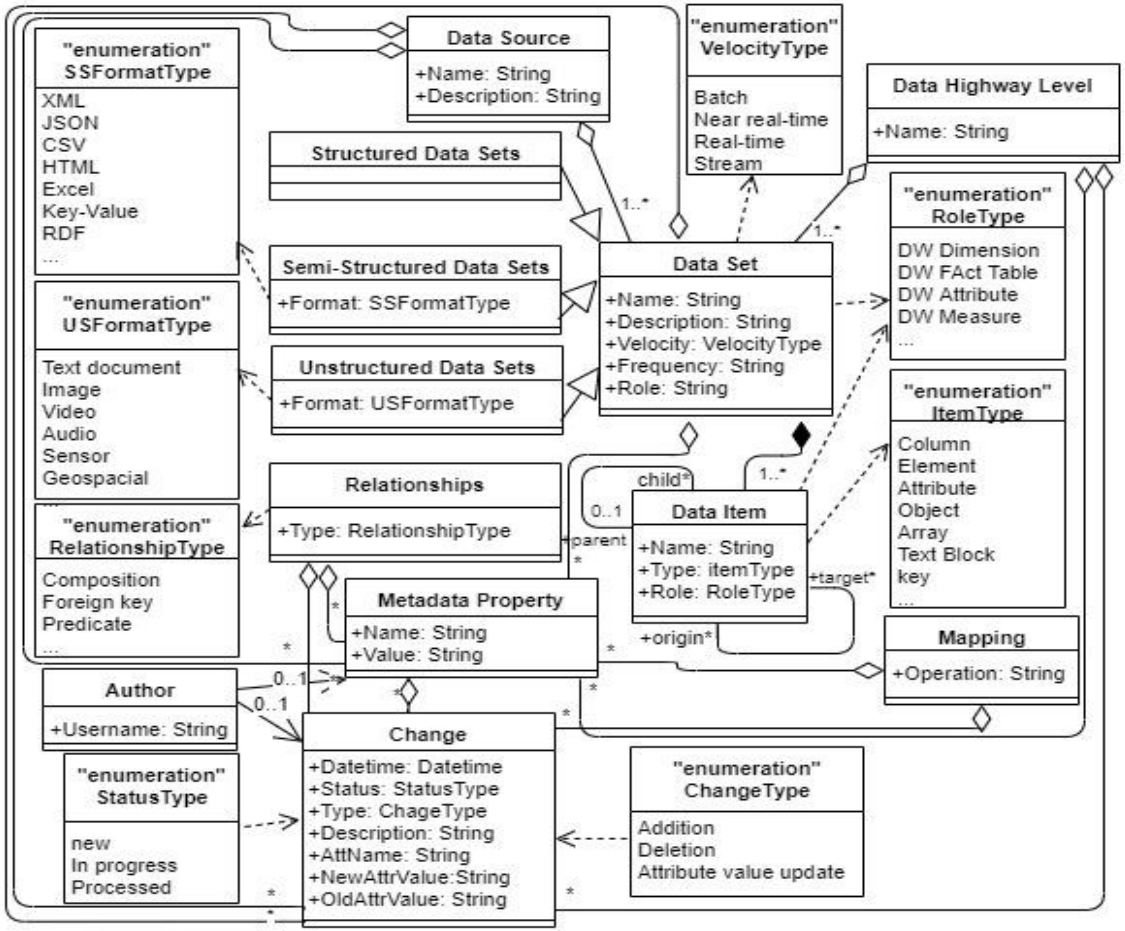


Scheme of the proposed conceptual metadata.

3.Metadata management systems in DL

D. Model for integrating evolving heterogeneous data sources

- To describe the schemas and additional properties of datasets or datasets.
- To collect metadata on the structure of data sources.
- To keep information on the changes that occur data sources



Conceptual diagram of the proposed metadata.

4.Summary and open research issues

A. Limits of existing models

A.1 Network-based model for DL

- **The BabelNet ontology** is adopted in terms of *lexical similarity between keywords* describing the data ingested within the DL.
  - → The choice of the ontology domain depends on ingested data.
- **The relevance of the similarity measure** in relation to the choice of ontology impacts the semantic representation of the DL data.
  - → The improvement of the semantic representation (The weighted aggregation, and the measure of similarity).
- **the N-Grams measurement** is used to carry out the mapping between the attributes that describe the data sources.
  - → Other metrics can be used (Cosine, Minkowski distance, etc.).
- The extraction of knowledge = to find an optimal path in the graph, which is evaluated with average local coefficient, density and transitive metrics.
  - → Other metrics (Betweenness centrality, Closeness centrality, etc.) can be used.

4.Summary and open research issues

A. Limits of existing models

Existing models	intra-object MD	inter-object MD
A.2 MEDAL	➢ The change in values is represented by *transformation*, however, *the updates* concerning the structure of the data ingested is not supported.	➢ The grouping of hypernodes is based on functions and, therefore, the choice of the latter is essential and impacts the categorization of hypernodes.
	➢ The risk of repetition of descriptive tags is true for structured data, but not for semi or unstructured data. it is necessary to save the history of the data ingested.	➢ The possible relationships between metadata are represented by parental type. Indeed, it is possible to extend this relationship by other types, such as include, friend, and equal.

4.Summary and open research issues

A. Limits of existing models

Existing models	intra-object MD	inter-object MD
A.3 A generic and extensible classification of metadata	➢ Unstructured data sources have no schema and, therefore, will not have schematic metadata. ➢ under the proposed definition of metadata, semantic metadata is based solely on descriptive text and requires tools for the extraction of descriptive tags.	➢ The similarity of the content is based only on the same attributes shared by the different data sets. ➢ The conceptual schema of metadata takes into account the structural aspect of data sources or datasets. It does not deal with the semantic aspect intra or inter datasets

4.Summary and open research issues

A. Limits of existing models

A.4 Model for integrating evolving heterogeneous data sources.

- Within this model, there may be a link between data sets. These relationships are modeled by an association class Relationship, which is limited to two types (Parent-children or Equality).

  → the case of synonymous or equivalent data elements is not considered.

- The case where the data set revolution is produced and is caused by a modification of the value of an attribute of an element of the model, the names of the old and new attributes are represented.

  → no change in the scheme, ie. the structure remains unchanged.

4.Summary and open research issues

B. Open research issues

- Enrich the possible relationships between the concepts that describe the data sources, based on the similarity measure score.

  → Several types of relations can be exploited, such as Include, Friend, Equal, Assigned, according to these scores (at intervals for each relationship type).

- Compared with textual descriptions of data sources, extracting relevant descriptive tags enhances the semantic representation of ingested data.

- Several similarity measures can be used to compare descriptive tags of the sources ingested within the Data Lake.

- Model a meta-metric that merges the results obtained according to several similarity measurement metrics.

Conclusion and Future works

- We have provided a layered taxonomy of recent metadata models and their specifications.

- A study of recent work dealing with DL metadata management models was conducted to classify metadata in 3 categories: by level, typology, and content.

- Based on such a study, an in-depth analysis of key features, strengths, and missing points is conducted.

- The future direction:

  To propose a meta-metric that merges the results obtained according to several similarity measurement metrics to enrich the possible relationships.

Main references

- P. N. Sawadogo, S. Etienne, F. Ccile, F. Eric, L. Sabine, and D. Jrme, "Metadata systems for data lakes: Models and features," in ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA,M2P, MADEISD and Doctoral Consortium Bled, Slovenia, September,811, 2019, Proceedings. IEEE, 2019, p. 440.

- P. L. Giudice, L. Musarella, G. Sofo, and D. Ursino, "An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake," vol. 478. Elsevier, 2019, pp. 606–626.

- F. Ravat and Y. Zhao, "Metadata management for data lakes," in European Conference on Advances in Databases and Information Systems. CCIS, vol. 1064. Springer, Cham., 2019, pp. 37–44.

- D. Solodovnikova, L. Niedrite, and A. Niedritis, "On metadata support for integrating evolving heterogeneous data sources," in European Conference on Advances in Databases and Information Systems. Springer, 2019, pp. 378–390.

- .....

Thank You!