A Comprehensive Study of Recent Metadata Models for Data Lake

Redha Benaissa*†‡, Omar Boussaid†, Aicha Mokhtari‡, and Farid Benhammadi§
*§ DBE Laboratory, Ecole Militaire Polytechnique, Bordj el Bahri, Algiers, Algeria
† ERIC, Universite de Lyon, Lyon 2, France
‡ RIIMA Laboratory, USTHB University, Algiers, Algeria
Email: *benaissa.redha@gmail.com, †omar.boussaid@univ-lyon2.fr, ‡amokhtari@usthb.dz, §fbenhammadi2008@gmail.com

***Abstract*—In the era of Big Data, an unprecedented 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 paper 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.**

Keywords–Metadata; Metadata models; Data Lakes; Big Data.

I. INTRODUCTION

In the era of Big Data, data has become more and more unstructured rendering traditional data storage models, such as Relational Data-Bases and their Management Systems (RDBMS) ill-adapted to meet these new needs. Indeed, traditional DBMS models are only suitable for applications having limited volume with relatively infrequent updates. Such systems, for instance, are unable to meet the exponentially growing data processing requirements for giant IT (Information Technology) companies, such as Google, Facebook, and Amazon. These limitations emphasize the need to review the methods of storing and processing this massive data and how to extract the relevant information. As a result, the concept of Data Lake (DL) has emerged. Within Data Lakes, massive heterogeneous data coming from different sources, is stored in its raw format without any transformation in order to accommodate multiple use-cases and applications. This, however, introduces new challenges to the management of these data with regards to discovery, storage, query, and construction of the catalog, as can be seen from Figure 1. To address such issues, metadata techniques are deployed within Data Lakes to reorganize and store data. The extraction of metadata from different heterogeneous data sources allows the construction of the DL catalog, which is essential for querying the DL. Within a Data Lake, a metadata catalog is a metadata management system, i.e., a formal system, which provides authority information on the structure and semantics of each element ingested within the Data Lake. It provides for each element the definition, the qualifiers associated with it, as well as the correspondences with equivalents in other languages or other diagrams, and finally the reference of the physical location of this element for retrieval data. Nevertheless, the extraction of the right metadata to build the catalog remains a challenging issue.

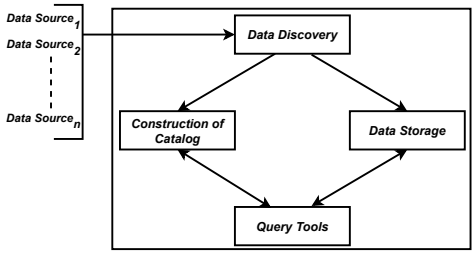


Figure 1. Data management process in a Data Lake.

This paper 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 point 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.

The remainder of this paper is organized as follows: in Section 2, we detail and discuss existing metadata management methods on Data Lake, which can be categorized into three categories namely level, typology, and content metadata. In Section 3, we study some metadata management models and systems that support the description and semantics of data ingested in the Data Lake. In Section 4, we present some limits of different metadata management existing models that have been identified and may be the subject of research directions to explore. The paper ends in Section 5 with conclusions and future directions.

II. METADATA IN DATA LAKE

With the emergence of Data Lakes, which refer to a massively scalable storage repository that contains a large amount of raw data [1], and for good management of heterogeneous data sources, only metadata can guarantee efficient management and effective interoperability of data sources [2]. However, until now, the representation and management of metadata on Data Lakes remains an open research area. In this section, we detail and discuss existing metadata management methods on Data Lake, which can be categorized into three categories namely level, typology, and content metadata, as can be seen from Figure 2.

 ISBN: 978-1-61208-816-7

A. Metadata Level

This category encompasses three models [3] [4] namely technical metadata, operational metadata, and business metadata, detailed below.

1) Technical metadata: Technical metadata [3] describes the technical aspects of data sets. It is used by the ingestion engine to determine the type of data encoding and to automatically convert the data sets into encodings according to the need or specification of the format and the type of encoding used in the ingestion target. It includes the type and format of the data (text, images, JSON, etc.) and the structure or the schema [5]. This latter reports the names of the sources, their data types, their lengths, and whether they can be empty or not.

2) Operational metadata: Operational metadata [5] contains information on the quality and origin of the data. It includes information about the source and target locations of the data, file size, number of records, and the number of records rejected during data preparation. Operational metadata can come in two forms [3]:

- Run-Time operational metadata: Reflects the state of the data sets each time a record is added, modified, or deleted.
- On-boarding metadata: Describes the cycle and life expectancy of data sets attributes provided by the ingestion phase.

3) Business metadata: Business metadata [3] 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. This type of metadata [5] covers management rules, such as setting an upper/lower limit on wages or determining the data that must be deleted from certain jobs for security and confidentiality reasons, for instance.

B. Metadata Typology

Defining a model for Data Lakes also involves identifying the data to be considered. Six key functionalities, expected by a metadata system from a Data Lake, have been identified by Sawadogo [6], which can be summarized as follows:

- *Semantic enrichment (SE)*, to generate a description of the context of the data, with interpretable and understandable tags based on ontologies.
- *Data indexing (DI)* consists in setting up a data structure essential for the recovery of data sets via specific characteristics (keywords or models). This requires the construction of forward or reverse indices.
- *Link generation and conservation (LG)* is the process of discovering similarity relationships or integrating relationships between data sets.
- *Data polymorphism (DP)* as storing multiple representations of the same data.
- *Data version (DV)* refers to the ability of the metadata system to handle data changes while keeping the previous states.
- *Usage tracking (UT)* saves the interactions between users and the Data Lake.

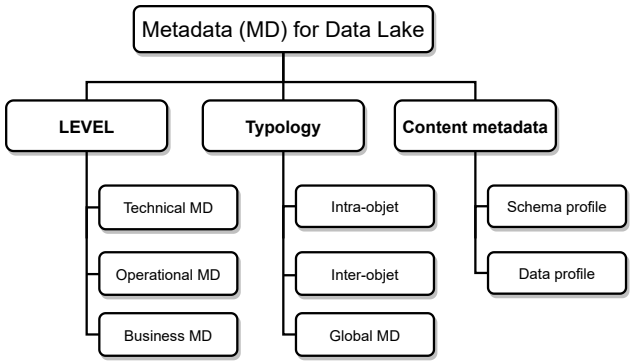


Figure 2. Metadata categories in Data Lake.

Besides, Sawadogo and al. [6] also proposed a typology of metadata, which categorizes it into intra-object, inter-object and global metadata. The following subsections detail each one of them.

1) Intra-object metadata: This category identifies *properties, summaries and previews, versions, and semantic metadata* associated with a given object. The properties provide a general description of an object, in the form of key/value pairs, obtained from the file system (the title of the object, size, date of the last modification, path, etc.). Summaries and previews provide an overview of the content or structure of an object. They can take the form of a data schema in the context of structured or semi-structured data, or a word cloud for text data. On the other hand, the creation of new versions of the initial data follows updates to the raw data in the Data Lake. Likewise, raw data (especially unstructured data) can be reformatted, inducing the creation of new representations of an object. Finally, semantic metadata are annotations that help to understand the meaning of the data (descriptive tags, text descriptions, or professional categories), useful for detecting object relationships.

2) Inter-object metadata: Inter-object metadata describes the relationships between at least two objects and has two main elements namely object groupings and similarity links. The former organizes objects into collections, which are derived from semantic metadata. Besides, properties like format or language can be used to group objects. The latter refers to the intrinsic properties of objects, such as their content or structure. It measures the compatibility of the diagrams of two structured or semi-structured objects, or other measures of common similarity.

3) Global metadata : Global metadata concern the entire Data Lake. They provide a contextual layer to the Data Lake that is essential for its analysis. Also, two new types of global metadata are presented. Semantic resources are essentially knowledge bases (ontologies, taxonomies, thesauri, dictionaries) used to generate other metadata and improve analyzes. Generally, they come from external sources such as ontologies. Furthermore, indexes are data structures that help find an object rapidly, and logs are used to track user interactions with the Data Lake.

C. Content metadata

According to [7], content metadata is the representation of all possible types of profiles in the Data Lake. Indeed, when analyzing raw data, the discovery of structural models and statistical distributions is based on the extraction and

 ISBN: 978-1-61208-816-7

profiling patterns of traditional data [8]. Ontology alignment techniques [9] are used to analyze the metadata and schema extracted. These techniques use *schema metadata and data profile metadata* to match different attributes of different datasets, generating the information profile. A *schema profile* describes the schema of datasets, e.g. the number of attributes, the names of the attributes and their data types [7]. The *data profile* describes the values of the dataset, i.e., the statistics values of single-attribute [7]. Information profiles are called metadata of relationships between datasets. Information profiles use the data profile models and schemas. For example, annotating attributes that can be linked based on the approximate similarity of data distributions and data types. The ingestion of data in the Data Lake allows the construction of the metadata catalog that offers added value to the enormous amount of data stored in DL. The metadata catalog describes this data and allows querying to extract hidden knowledge from data sources ingested in the Data Lake.

III. METADATA MANAGEMENT SYSTEMS IN A DATA LAKE

In a Data Lake, the extraction of knowledge is based and articulated on metadata, which describes the sources of ingested data. It may have other data from other sources that satisfy the request, and building semantic bridges between metadata will increase the performance of querying the DL. In this section, we study of recent metadata management models and systems that support the description and semantics of data ingested in the Data Lake.

A. Network-based model for Data Lake

In the model presented in [10], the aim is to offer an approach for the extraction of complex knowledge schemes from concepts belonging to structured, semi-structured and unstructured sources in a Data Lake. In [10], the term complex knowledge model is used to indicate a semantic relationship (specifically, a synonymy or part of a relationship), that focuses on the semantics of data sources, and, therefore, only business metadata is considered. They include the business names and descriptions assigned to the data fields. They also cover business rules, which can become integrity constraints for the corresponding data source. This model adopts a typical notation of XML, JSON and many other semi-structured models to represent business metadata. The proposed approach [10] is based on an appropriate network, which represents all the sources of Data Lakes. It builds a structured representation of keywords, generally flat, used to represent unstructured data sources. Formally, a complex knowledge model consists of a logical succession $x_1, x_2, ..., x_w$ of w objects. With this uniform and network-based representation of sources in the Data Lake, the extraction of complex knowledge models can be carried out by using tools based on graphs. It consists in constructing suitable paths going from the first node (ie, x_1) to the last node (ie, x_w) of the succession expressing the patterns. The proposed approach seeks an appropriate path (if it exists) connecting x_1 to x_w. Since x_1 and x_w can belong to different sources, the approach considers the possible presence of synonymies between concepts belonging to different sources, and should model these synonymies by means of an appropriate form of arcs (cross arcs or *c-arcs*), and should include both intra-source arcs (internal arcs or *i-arcs*) and c-arcs in the path connecting x_1 to x_w and representing the complex knowledge model of interest.

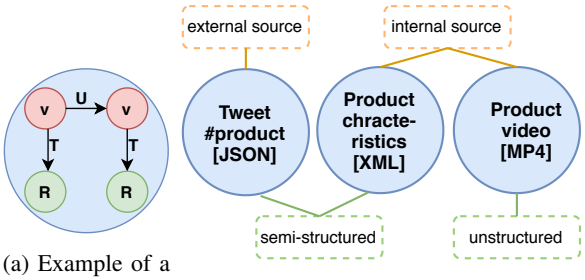
In addition, there are cases where synonymies are not sufficient to find a complex knowledge model from x_1 to x_w. In such cases, the proposed approach makes two other attempts in which it first tries to imply similarities in chains and, even if these properties are not sufficient, partial relationships. If neither the synonymies, the similarities of strings, nor the partial relations allow the construction of a path from x_1 to x_w, the proposed approach concludes that, in the Data Lake considered, a complex knowledge model of x_1 to x_w does not exist.

The biggest difficulty concerns unstructured data because a consistent flat representation by a simple element, for each keyword provided to designate the content of the source, is not recommended. In fact, this type of representation would make it very difficult to reconcile and next integrate an unstructured source with the other sources (semi-structured and structured) of the Data Lake. Therefore, it is necessary (at least partially) to "structure" unstructured data. To solve this problem, the proposed approach creates a complex element to represent the source as a whole and a simple element for each keyword. The approach exploits the lexical and string similarities. In particular, the lexical similarity is considered by declaring that there is an arc of the node n_{k1}, corresponding to the keyword $k1$, to the node n_{k2}, corresponding to the keyword $k2$ (and vice versa). It is possible if $k1$ and $k2$ have at least one common lemma in an appropriate thesaurus. To this end, the approach adopts the ontology or multilingual semantic network BabelNet [11]. The chain similarity is applied via an appropriate chain similarity metric on $k1$ and $k2$, is "sufficiently high". In this case, N-Grams [12] is used as a chain similarity metric.

B. MEDAL

MEtadata model for DAta Lakes (MEDAL) [6] adopts a logical representation of metadata based on a hypergraph, a nested graph and the concepts of assigned graph. An object is represented by a hypernode containing various elements (versions and representations, properties, etc.). The hypernodes can be located between them (similarity, groupings, etc.). Objects can take the form of structured data (relational database tables, CSV files, etc.), semi-structured (JSON, XML, YAML, etc.) and unstructured (images, text documents, videos, etc.). The metadata, obtained on the typology, are subdivided into three components: $M = (M_{intra}, M_{inter}, M_{glob})$, where M_{intra} is the set of intra-object metadata, M_{inter} is the set of inter-object metadata and M_{glob} is the set of global metadata.
Each hypernode contains *representations*, associated with an object. There is at least one representation per hypernode, corresponding to the raw data of the Data Lake. Other representations all derive from this initial representation. Each representation corresponds to a node carrying simple or complex attributes. The *transition* from one representation to another is done via a transformation, formalized by an oriented edge, which also carries attributes or properties describing the transformation process (complete script or description, in case of manual transformation). A hypernode may also contain *versions* associated with nodes with attributes to manage data revolution of the lake over time. Indeed, a hypernode contains a tree whose nodes are representations or versions and the oriented edges are transformations or updates. One representation (resp. Version) is derived from another by a transformation (resp. Update), as can be seen from Figure 3a.

 ISBN: 978-1-61208-816-7



(a) Example of a hypernode. (b) interconnected hypernodes.

Figure 3. Hypergraph representation of MEDAL [6].

So, the root of the tree is the initial raw representation of the hypernode and each version has its own representation subtree.

A group of objects is modeled by a set of undirected hyper-edges, i.e., edges, which can connect more than two (hyper) nodes. Each hyper-edge corresponds to a collection of objects. This grouping is performed on a hypernode attribute depicted in Figure 3b. A similarity link between two hypernodes is represented by an undirected edge with attributes: the value of the similarity metric, the type of the metric used, the date of the metric, etc. A hypernode can be derived from other hypernodes via a parental link. To translate this relation, a directed hyper-edge is used from all hypernodes ”parents” towards the hypernode ”child”. Hypernodes are grouped in relation to a given parameter (often an attribute) and by parental relations.

The global metadata gravitate hypernodes and operated, as required, that is to say almost always, especially logs and indexes.

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

In [13], metadata can help users find data that matches their needs, accelerates data access, verifies the origin of the data and treatment history to find relevant data and thereby enriches their analysis. The proposed metadata classification [13] has the advantage of integrating both intra-metadata and inter-metadata for all data sets or datasheets.
For Inter-metadata, the classification of [14] [See Section 2] is completed by subcategories. *Dataset Containment* indicates a containment relationship between the datasets. *Partial overlap* expresses the overlap of certain attributes in certain datasets. *Provenance* means that one dataset is the source of another dataset. *Logical clusters* mean that certain datasets are in the same domain. *Content similarity* finds common attributes shared between different datasets.
For Intra-metadata, the classification of [15] is extended to include access, quality and security. *Data characteristics* includes information such as the identification, name, size, type of structure and date of creation of the data sets. *Definition metadata* specifies the meaning of data sets. They are classified into semantic and schematic metadata. Semantically, structured and unstructured data sets can be described by text or by certain key words (vocabularies). Schematically, a structured dataset can be presented by a database schema. Other characteristics are described such as Navigation metadata, which relates to the location of data sets, Lineage, which presents the life cycle of the data, Access metadata, which presents access

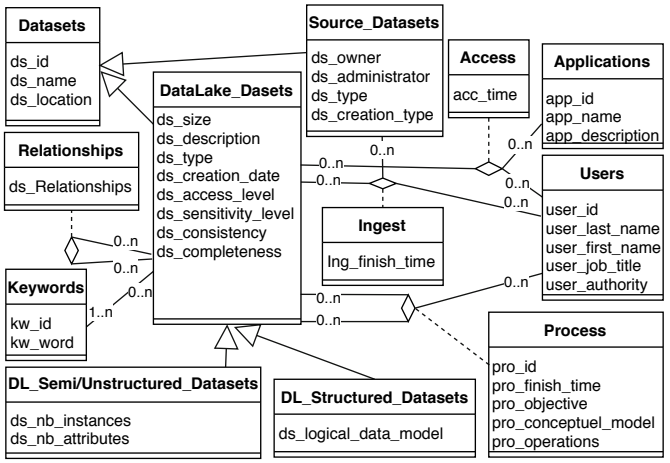


Figure 4. Scheme of the proposed conceptual metadata [13].

information, Quality metadata, which is the consistency and completeness of the data to ensure the reliability of the data sets and Security metadata, which includes data sensitivity and level of access.
Based on the classification in [13], a conceptual metadata schema, shown in Figure 4, is presented. A structured or unstructured dataset is ingested from one or more sources by one or more users. Datasets can be processed by users to transform into new datasets. Users can access datasets for their analyzes with certain tools. Datasets stored in a Data Lake can have relationships.

D. Model for integrating evolving heterogeneous data sources

In [16], a metadata model is proposed to describe the schemas and additional properties of datasets or datasets extracted from sources and transformed to obtain integrated data in order to perform the analysis in a flexible way. In addition, it keeps all changes that occur in the system. To collect metadata on the structure of data sources and to keep information on the changes that occur there, the conceptual model [16], presented in Figure 5, shows the metadata used.

In this section, we focus on the model classes that describe the schemas of the data sources and pipeline levels of data processing shown in Figure 5. The *Data Set class* is used to represent a collection of Data Items that are individual pieces of data. The *Data Set class* is divided into three subclasses according to the type and format. *Structured data Set* represents a relational database table where the data elements correspond to the columns of the table. *Semi-structured data* reflects the files in which the data elements are organized in a schema, which is not predefined. The *Type* attribute of a data element embedded in such a data source indicates its position in the schema. *Unstructured Data Sets* include data that has no recognized organization or schema, such as text files, images, other multimedia content, etc. A dataset can be obtained from a *Data Source* where it can be part of a *Highway Level* data processing pipeline level. In addition, information about the speed at, which data in the dataset is collected or updated by assigning one of the speed types and frequency attribute to the *Data Set class*.

In general, there are relationships between data elements in the same data set or between different data sets determined by the format of these data sets. These relationships are

 ISBN: 978-1-61208-816-7

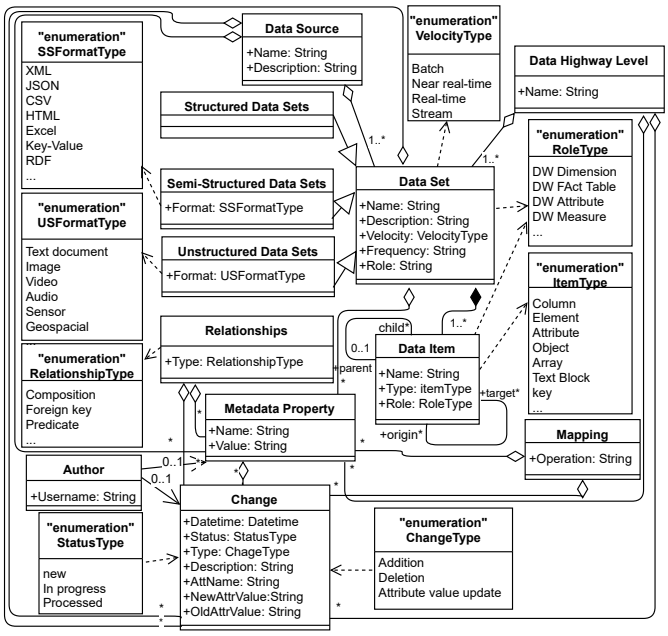


Figure 5. Conceptual diagram of the proposed metadata [16].

modeled by a *Relationship* association class that connects the child and parent data elements and assigns the corresponding relationship type. The *Equality* relationship type is assigned if two different dataset elements contain the same data.
To maintain the metadata of provenance of the data sets within the data processing pipeline and allow their lineage to be followed, a *Mapping* association-class has been introduced to define the way in which a target data element is derived from the elements of original data by a transformation that indicated in the *Operation* attribute of the *Mapping* class.
In the event that revolution is caused by a change in the value of an attribute of a model element, including the metadata property, the name of the attribute is saved as attribute $AttrName$ of the class *Change* and both the value before the modification ($OldAttrValue$ attribute) and after ($NewAttrValue$ attribute).

IV. Summary and open research issues

According to the study carried out on these different metadata management existing models, some limits have been identified and may be the subject of research directions to explore.

A. Limits of existing models

Concerning the work of Paolo et al. [10], who propose a model based on a network or graph to represent and manage the data sources of a Data Lake:

- In terms of lexical similarity between keywords describing the data ingested within the Data Lake, the approach adopts the BabelNet [11] multilingual semantic network or ontology. As a result, the choice of the ontology domain depends on ingested data within the Data Lake.
- The relevance of the similarity measure in relation to the choice of ontology impacts the semantic representation of the Data Lake data. The weighted aggregation, the measure of similarity, and the choice of ontology contribute to the improvement of the semantic representation.
- The approach exploits the lexical similarities of character strings to carry out the mapping between the attributes that describe the data sources or the N-Grams measurement is used. Other metrics can be used (Cosine, Minkowski distance, etc.).
- The extraction of knowledge in this work is to find an optimal path in the graph representing result. This graph is evaluated with the metric (average local coefficient, density and transitive). Other metrics (Betweenness centrality, Closeness centrality, etc.) can be used.

When it comes to MEDAL [6], according to the classification of metadata relating to the typology category:

- Compared to intra-object metadata:
  - The change in values is represented by transformation, however, the updates concerning the structure of the data ingested in the Data Lake is not supported.
  - The risk of repetition of descriptive tags between the different representations is true for structured data (update of BD), but not for semi or unstructured data. For a better analysis, it is necessary to save the history of the data ingested in the DL.
- Compared to inter-object metadata:
  - The grouping of hypernodes is based on functions and, therefore, the choice of the latter is essential and impacts the categorization of hypernodes (in this case, some criteria have been cited (the origin of the data source, the type of the latter (structured, semi or unstructured)).
  - 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, which will be based on similarity measures.
- Compared to global metadata:
  - Requirement of tools to identify semantic sources to add them to the DL metadata representation graph.

However, there is no system that automatically extracts inter or intra-metadata from different types (structures, semi-structures, non-structures) of datasets. With regard to the system proposed [13], based on a generic and extensible classification of metadata:

- Compared to intra-object metadata:
  - Unstructured data sources have no schema and, therefore, of according to the definition metadata, will not have schematic metadata.
  - Furthermore, under the proposed definition of metadata, semantic metadata is based solely on descriptive text and requires tools for the extraction of descriptive tags.
- Compared to inter-object metadata:

 ISBN: 978-1-61208-816-7

- The similarity of the content is based only on the same attributes shared by the different data sets, hence the need to measure the similarity between attributes to further expand this similarity.
- The conceptual schema of metadata offers [13] takes into account the structural aspect of data sources or datasets. It does not deal with the semantic aspect intra or inter datasets because it is limited to the identification of the same attributes that appear in these datasets.

Finally, in [16], the model for the integration of evolving heterogeneous data sources, used to store metadata, describing the schemas of the implied data sets and their changes, is one of the central components of the data warehouse architecture in the context of Big Data.

- Within this model, there may be a link between an unstructured data set and structured / semi-structured data. These relationships are modeled by an association class Relationship, which is limited to two types: Parent-children or Equality. Equality happens when two elements of different data sets contain the same data, but the case of synonymous or equivalent data elements is not considered.
- In 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, but there is no change in the scheme, ie. the structure remains unchanged.

B. Summary and challenges (Open research)

In this context of metadata management within a Data Lake and to overcome the limitations mentioned above, research work can be oriented to:

- 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 within the Data Lake.
- 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.

V. CONCLUSION

In this paper, we have presented a comprehensive study of recent metadata models for Data Lake that points out their rationales, strengths, and weaknesses. Specifically, we have provided a layered taxonomy of recent metadata models and their specifications. Afterward, 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 the main characteristics, strengths, and missing points is presented. Consequently, we have bridged the gap in the literature and identified open research issues that require community attention. As future work, we plan to propose a meta-metric that merges the results obtained according to several similarity measurement metrics to enrich the possible relationships between the concepts that describe the data sources ingested in Data Lakes.

REFERENCES

[1] N. Miloslavskaya and T. Alexander, "Big data, fast data and data lake concepts," Procedia Computer Science, vol. 88, 2016, pp. 300–305.

[2] I. Suriarachchi and B. Plale, "Crossing analytics systems: A case for integrated provenance in data lakes," in e-Science (e-Science), 2016 IEEE 12th International Conference on. IEEE, 2016, pp. 349–354.

[3] S. Badih, G. Gregory, and Y. Q. Herman, "Metadata-driven data management platform," Sep. 6 2018, uS Patent App. 15/909,833.

[4] C. Diamantini, G. Paolo, L. Musarella, P. Domenico, E. Storti, and D. Ursino, "A new metadata model to uniformly handle heterogeneous data lake sources," in European Conference on Advances in Databases and Information Systems. Springer, 2018, pp. 165–177.

[5] Oram, "Managing the data lake," in Managing the Data Lake OReilly, Sebastopol, CA, USA, 2015. IEEE, 2015.

[6] 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.

[7] A. Alserafi and O. R. A. Abello, "Towards information profiling: Data lake content metadata management," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE, 2016, pp. 178–185.

[8] F. Naumann, "Data profiling revisited," ACM SIGMOD Record, vol. 42, no. 4, 2014, pp. 40–49.

[9] R. Hauch, A. Miller, and R. Cardwell, "Information intelligence: metadata for information discovery, access, and integration," in Proceedings of the 2005 ACM SIGMOD international conference on Management of data, 2005, pp. 793–798.

[10] 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.

[11] R. Navigli and S. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," in Artificial Intelligence, vol. 193, IEEE. Elsevier, 2012, pp. 217–250.

[12] W. H. Gomaa, A. A. Fahmy et al., "A survey of text similarity approaches," vol. 68, no. 13. Citeseer, 2013, pp. 13–18.

[13] 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.

[14] H. Alon, K. Flip, N. N. Fridman, O. Christopher, P. Neoklis, R. Sudip, and W. S. Euijong, "Managing google's data lake: an overview of the goods system." IEEE Data Eng. Bull., vol. 39, no. 3, 2016, pp. 5–14.

[15] B. Bilalli, A. Abelló, T. Aluja-Banet, and R. Wrembel, "Towards intelligent data analysis: The metadata challenge." in IoTBD, 2016, pp. 331–338.

[16] 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.

 ISBN: 978-1-61208-816-7