About Relationships in Data Lakes

Ahlame Diouan[1,2][0009−0003−1481−4662], Eric Ferey[1], Jérôme Darmont[2][0000−0003−1491−384X], and Sabine Loudcher[2][0000−0002−0494−0169]

[1] Université Lumière Lyon 2, UR ERIC, Lyon, France
{a.diouan,sabine.loudcher,jerome.darmont}@univ-lyon2.fr
[2] BIAL-X, Lyon, France
{ahlame.diouan,eric.ferey}@bial-x.fr

Abstract. In the era of Big Data, managing voluminous and heterogeneous data presents significant challenges for organizations. To tackle these challenges, the concept of a data lake has emerged as a promising solution, allowing the storage of raw data from diverse sources in their original format. An efficient metadata management system plays a crucial role in preventing data lake to turn into an unusable data swamp by providing a structured framework for organizing, categorizing and establishing relationships between data entities.
In this paper, identify the various relationships from diverse domains found in the literature. Then, we categorize the types of relationships and propose a relationship typology that classes relationships by similarity, containment, grouping and provenance. Eventually, we also aim to check whether goldMEDAL, a state-of-the-art generic metadata management model, adequately supports all such relationships. This evaluation is particularly relevant for Bial-X, which seeks to implement a robust metadata management system based on goldMEDAL's concepts.

Keywords: Data lakes · Data discovery· Semantic relationships· Big data.

1 Introduction

In recent years, there has been a huge increase in global data production and organizations' decision-making processes have been revolutionized by the availability of large volumes of heterogeneous data, known as Big Data. This exponential growth not only presents real opportunities, but also challenges related to data volume, velocity and variety that exceed the capabilities of traditional data storage and management systems [18].

To address this issue, James Dixon proposes the concept of a data lake as a practical solution [6]. A data lake allows storing raw data from heterogeneous sources in their original format. In the absence of a data schema, the presence of a robust metadata system is crucial for enabling data queries and thus preventing the data lake from becoming a data swamp, i.e., an unusable data lake. Moreover, an efficient metadata system provides users with a unified interface to search,

explore, and understand the available data entities and the relationships between them.

Bial-X's customers require a metadata management system to effectively manage a data lake and establish semantic relationships between data entities. Note that there are many terms similar to relationship, e.g., relation, link, linkage and connection. However, after reviewing the literature, relationship appears to be the most frequent term. Finding relationships provides users with a global view of metadata, through which they can interpret said relationships and gain valuable context into how various data entities are interconnected, facilitating a deeper understanding of their significance within the data lake. Since the data lake literature seems unanimous about the importance of a metadata system, we benchmarked state-of-the-art metadata management systems, i.e., DataGalaxy[3], Atlas[4], Open Data Discovery[5] and OpenMetadata[6]. These tools offer various forms of relationships, including operational and structural relationships, e.g., "entity In" and "aggregation"), but also lineage (provenance) relationships, which are important for understanding the origins and transformations of data entities.

Lineage relationships belong to so-called semantic relationships, but there are still other semantic relationships that metadata management systems do not support. Semantic relationships are defined as "any form of hierarchical, generic or predefined semantic relationships (semantic connections between data sets, e.g., for provenance or governance)" [14].

Eventually, our contribution is threefold.

1. We survey the various relationships between data entities found in the literature, notably aiming to pinpoint all semantic relationships that meet Bial-X's specific needs.
2. We categorize the types of relationships and propose a relationship typology that classes relationships by similarity, containment, grouping and provenance.
3. We hypothesize and check that goldMEDAL, a state-of-the-art generic metadata management model [24], can adequately support all the relationships we identify in our survey. This evaluation is crucial for Bial-X, as the company funded a PhD thesis that was part of goldMEDAL's design.

In the remainder of this paper, we first explicit our survey methodology and present the metadata metamodel goldMEDAL that we use throughout this paper (Section 2). Next, we present and discuss our relationship typology, i.e., similarity, containment, grouping and provenance relationships (Section 3). Finally, we conclude this paper and hint at future works (Section 4).

[3] https://www.datagalaxy.com
[4] https://atlas.apache.org
[5] https://opendatadiscovery.org
[6] https://open-metadata.org

2 Preliminaries

2.1 Survey Methodology

We conduct a systematic literature review to analyze relevant articles addressing specific questions related to relationships between data entities within data lakes. We adopt the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) protocol [19] to ensure rigor and transparency throughout the whole process.

Research Questions First, we identify several key questions to guide our literature review.

- How are relationships between data entities defined in the literature?
- What types of relationships exist between data entities?
- Are there any semantic relationships between data entities?
- How relationships can enhance metadata management models or systems?
- Can goldMEDAL's concepts support all identified types of relationships?

Sources To address our research questions, we conduct extensive searches across several academic databases, i.e., the ACM Digital Library, SpringerLink and IEEE Xplore. Moreover, Google Scholar is occasionally used to access papers not available in the above databases. Yet, Google Scholar is not a primary source, because the results it yields includes numerous non-peer-reviewed papers.

Search Strategy Our search strategy involves selecting key terms designed to capture the full breadth of literature related to data lakes, relationships between data entities, and dataset discovery. Our search query below incorporates a range of terms and Boolean combinations to cover all relevant facets of the topic.

```
"Data lakes" AND
("relationships" OR "Semantic relationships" OR "Dataset
    discovery") AND
("data entities" OR "datasets" OR "tables") AND
(("Similarity" OR "Related" OR "Proximity") OR ("Containment"
    OR "Inclusion" OR "Encapsulation") OR ("Provenance"   OR
    "Lineage" OR "Tracking") OR ("Grouping" OR "Clustering"
    OR "Categorization"))
```

These results yield about 500 papers retrieved from ACM (118), IEEE (11), and Springer (371) databases.

Filtering Applying a date filter onto articles published between 2016 and 2024 reduces the set to 315 papers. Next, we filter by publication type, narrowing down to 80 from ACM, 9 from IEEE and 199 from Springer. Then, we conduct an initial screening based on titles and abstracts, focusing on relevant keywords

from our search query and assessing the alignment of abstracts with the scope of our study. This process leaves us with 50 articles from ACM, 6 from IEEE and 70 from Springer for further evaluation. Eventually, we apply inclusion and exclusion criteria (Table 1). The final result yields 10 papers from ACM, 2 from IEEE and 7 from Springer, for a total of 19 papers in the final review.

Table 1. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Publication type: conference/journal	Brief papers or limited in scope
Publication date: from 2016 to 2024	Language: non-English
Article type: research/survey	Named relationship without definition
Relevant keywords	Accessibility: paper cannot downloaded
	Duplicates

2.2 goldMEDAL and Relationships

goldMEDAL is a generic metadata model that bears a high level of abstraction to be very flexible for any data lake use case [24]. It encompasses three levels of modeling (conceptual, logical and physical) and is built upon four core concepts: data entity, grouping, link and process (Figure 1).

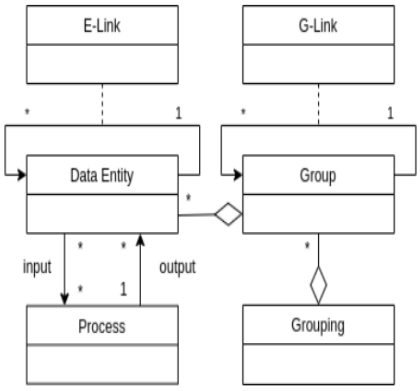

Fig. 1. goldMEDAL conceptual metadata model [24]

Data entities are the basic units of the metadata model. For example, a data entity can represent both raw data and transformed data. It might be a spreadsheet file, a textual file, a semi-structured document, an image, a database table,

a tuple or an entire database. This flexibility enables goldMEDAL to seamlessly handle data at various levels of granularity. Furthermore, the introduction of any new element such as file, document, image, etc., into the data lake triggers the creation of a new data entity.

A grouping involves organizing data entities into sets denoted as groups based on common properties. For instance, within data lake architectures, the raw and preprocessed data zones constitute two groups within a zone grouping. Another example is a grouping of textual documents according to the language of writing.

Links associate either data entities with each other or groups of data entities with each other. Such links may be either directed or undirected.

A process refers to any transformation or update applied to one or several data entities to produce a new data entity. It is used to track the relationships between data entities. Each process connects one or more "parent" data entities to "children" data entities. Yet, unlike links, processes represent the execution context of a transformation or modification operation (user, script, etc.).

By leveraging these core concepts, goldMEDAL facilitates the exploration of relationships between data entities within a data lake ecosystem. For example, processes can be employed to trace the lineage and parenthood relationships among data entities, while groupings aid in structuring related data entities, resulting in the establishment of containment relationships between different groups within the same grouping. Links can depict hierarchical or semantic connections between data entities or groups.

However, the practical application of these concepts needs to be assessed. While goldMEDAL offers a high level of abstraction and flexibility, it is essential to determine whether its core concepts can effectively support the implementation of all relationship types, including similarity, containment, grouping and provenance.

Ensuring that goldMEDAL's theoretical framework can guide the development of a metadata management system is vital for our future works. Bial-X especially needs a tailored and robust metadata management system.

3 Relationship Typology

We architecture our typology in four types of relationships: similarity (Section 3.1), containment (Section 3.2), grouping (Section 3.3) and provenance (Section 3.4) relationships. Each type of relationships is synthesized in a table with four columns:

- *Relationship name* as per the source article;
- bibliographical *Ref.*;

- a quote by the authors characterizing the relationship (*Authors' quote*);
- the goldMEDAL concept associated with the relationship (*gM concept*).

The relationships identified in the literature are categorized by the authors' definitions and descriptions within each paper. Some papers address multiple types of relationships. However, we do not include in our study named relationships that are not sufficiently defined.

Each identified relationship is classified according to our typology. Finally, we thoroughly analyze relationship definitions to determine what goldMEDAL concept can (or not) implement a particular relationship. We now present the selected papers and discuss them.

3.1 Similarity Relationships

In Table 2, we observe that the similarity relationships found in the literature may be based on different aspects such as content, structure or both a combination of content and structure. By classifying relationships into these three types, we can discern both differences and commonalities between them.

Content similarity Content similarity refers to resemblance or overlap in the information contained within data entities, particularly in terms of attributes, values or content. It indicates how closely related or similar two data entities are based on the content they store. Such similarity can be measured through various metrics, such as shared attributes, common values or semantic overlaps.

For Halevy et al., content similarity focuses on checksums and Locality Sensitive Hashing (LSH) values as indicators of content similarity [13]. Alserafi et al. emphasize the overall similarity of real-world objects or concepts stored in datasets [2]. On the other hand, Ravat and Zhao introduce the concept of partial overlap, suggesting that content similarity can exist even when data entities overlap rather than bearing strictly identical attributes [21]. Additionally, Eltabakh et al. define content similarity as establishing connections between a text document and table based on various criteria such as overlapping values, semantic similarity or metadata similarity. Moreover, they assess relatedness by assigning a score to each relationship between the document and table columns [8].

Furthermore, Kaminsky et al. expand the concept of content similarity by introducing joinability, which refers to the ability to link two similar columns from the same domain [16].

Structural similarity Structural similarity refers to the resemblance between data entities based on their structural aspects. It evaluates how data entities are similar based on their structure, including factors such as the types of variables, attributes names and data constraints. For Hai et al., a a structural similarity involves clustering similar schemas together and selecting the core of each cluster as its representation.This process is based on the similarity between schemas

Table 2. Similarity relationships

Relationship name	Ref.	Authors' quote	gM concept
Content similarity	[13]	... find datasets with content that is similar or identical to the given dataset, or columns from other datasets that are similar or identical to columns in the current dataset.	Link
Schema grouping	[12]	... clusters the schemas and picks up the core of each cluster as its presentation. The necessity of grouping depends on the schema similarity calculated over the imported data sources.	Link
Related	[2]	... Related pairs of datasets describe similar real-world objects or concepts from the same domain of interest. These datasets store similar information in (some of) their attributes.	Link
Proximity	[1]	... We utilise a novel proximity mining approach to assess the similarity of datasets.	Link
Similarity relationship	[5]	It's used to present that an object is "similarTo" another object.	Link
Relationship constraint	[11]	... the analyst may be interested in finding similar datasets to the ones found so far to make sure no information is missing (content similarity). Or, having already found a handful of relevant datasets, the analyst may want to find a join path to join them together (a primary-key/foreign-key (PK/FK) candidate).	Link
Property constraints	[11]	... selecting columns with unique values, or columns with a string in the schema name, which are all properties of the data. For instance, the analyst who is building the stock change prediction model may start with a search for tables that include metrics of relevance... (schema similarity).	Link
Content similarity	[20]	... Which means that different datasets share the same attributes	Link
Partial overlap	[20]	... Partial overlap which means that some attributes with corresponding data in different datasets overlap.	Link
Similarity Link	[23]	... Reflect the strength of the similarity between two objects. Unlike object groupings, similarity relationships refer to the intrinsic properties of objects, such as their content or structure.	Link
Schema matching	[1]	... It seeks to identify schematic overlaps between datasets. This involves detecting related objects (instances or attributes) and matching instances between two different schemata.	Link
Union	[9]	... Table union search aims to find all tables that are unionable with the query table. To determine whether two tables are unionable, existing solutions first identify all pairs of unionable columns from the two tables based on column representations, such as bag of tokens or bag of word embeddings.	Link
Doc to Table (From Document to Tables)	[8]	... A Table T with column set A is related to a text document D if there exists $A_i \in A$ such that D and A_i are related via overlapping values, semantic similarity, or metadata similarity, each with a relatedness score.	Link
Table j Table (Joinable Tables)	[8]	... Table T with column set A is joinable to Table T' with column set A' if there exists $A_i \in A$ and some $A'_j \in A'$ such that: 1. A_i and A'_j have value overlap suggesting syntactic join, or 2. A_i and A'_j have semantic overlap suggesting semantic join.	Link
Table U Table (Unionable Tables)	[8]	... Table T with column set A is unionable to Table T' with column set A' if a one-to-one mapping $H : A \rightarrow A'$ exists wherein there exists $h \in H$ such that the column pair given by h exhibits name, value, or semantic similarity.	Link
Joinability relationship	[16] [26]	... Joinability means that two columns can be linked together because they contain similar data from the same domain	Link

calculated over imported data entities [12]. Alserafi et al. emphasize on identifying overlaps between data entity schemas by detecting related objects or attributes and matching instances between different schemas [1]. Moreover, the use of proximity mining adds another criterion by employing proximity scores to identify similar data entities with respect to structural similarity [1]. Eltabakh et al. propose joinable tables based on syntactic or semantic overlaps between columns [8]. Finally, Fernandez et al. introduce a property constraint that focus on selecting data entities according to specific properties, such as unique values or schema names, which represent inherent structural features [11].

Hybrid similarity Hybrid similarity combines elements from both content and structure, offering a more comprehensive perspective on finding relevant and similar data entities. For Diamantini et al. [5] and Sawadogo et al. [23] similarity relationship focus on establishing a relationship between data entities, considering both their content and structural characteristics. Moreover, Fernandez et al. introduce constraints based on specific properties. such as columns with unique values or columns with a particular string in the schema name. These constraints aim to select data entities by integrating both content and structural aspects [11]. Eventually, Fan et al. [9] and Eltabakh et al. [8], focus on identifying unionable tables, considering both content and structural overlaps between their columns to determine similarity.

Despite the different terms used to describe similarity relationships, they share the common understanding that content or structural similarity is determined by shared attributes, values, semantic overlaps or structural aspects, using different methodologies and metrics. The comparison highlights that goldMEDAL's concepts can handle diverse similarity relationships. For example, the notion of data entities corresponds well with content similarity, focusing on the similarity or overlap in the data they contain. Furthermore, goldMEDAL's link concept facilitates the implementation of similarity relationships between data entities. After analyzing the various similarity relationships outlined in Table 2, it becomes evident that goldMEDAL is a robust metamodel to implement these relationships within a data lake.

3.2 Containment Relationships

In the context of data lake management, containment relationship refers to the hierarchical structure of data entities (Table 3). This relationship indicates how a data entity can be encapsulated or nested within another, such as sub-data entities or sub-tables within its structure. It illustrates how data entities are grouped or organized in a hierarchical manner, with some data entities being contained within others.

According to Halevy's et al., a containment relationship refers to how data entities may contain other data entities, such as bigtable column families [13]. Deng et al. focus on the subsumption relationship, providing a function to identify data entities or groups of data entities that are contained in or contain other

Table 3. Containment relationships

Relationship name	Ref.	Authors' quote	gM concept
Dataset containment	[13]	... Some datasets may contain other datasets.	Link
Subsumption relationship	[4]	... a list of tables or groups of tables that have some form of subsumption relationship (i.e., are contained in or contain) with respect to the reference table;	Link
Structural relationship	[5]	... Which is used to present that an object "contains" another object.	Link
Granularity Indicator	[7]	... collecting metadata on different granular levels. These levels are closely tied to some kind of structure in the data.	Link
Containment relationship	[15]	... containment relationship, i.e., a parent entity T may contain another child entity Ti.	Link
Containment fraction	[25]	... If A and B are schemas, $n(B)$ refers to the length of the flattened schema set in B, and $\|A \cap B\|$ refers to the length of the intersection between the flattened schema sets. If they are tables, $n(B)$ refers to the number of rows in B and $\|A \cap B\|$ refers to the number of rows common to both tables.	Link
Inclusion dependency	[10]	... $T_u.A_v \subset_{level} T_q.A_r$, where each T_i is a table, each A_j is an attribute, and level is the fraction of the values in $T_u.A_v$ that are contained in $T_q.A_r$. When the level = 1, there is a full inclusion dependency, and when the level < 1, there is a partial inclusion dependency.	Link

data entities [4]. Additionally, Diamantini et al. present structural relationship, indicating how an object contains another object, like a relational database containing tables and the same way tables contain attributes [5].

Eichler et al. introduce the granularity indicator entity, enabling the collection of metadata on multiple granularity levels, closely tied to some kind of structure in the data, such as object instances or key-value pairs within a JSON document [7].

Huang et al. describes containment relationships as a parenthood relationship between two data entities, i.e., a parent entity contains a child entity. Shah et al. quantify how much a data entity is contained in another, either in terms of their structure or content [25]. Finally, Fernandes et al. focus on inclusion dependency, where data values from one data entity are contained within another, either fully or partially [10].

Overall, these authors agree in their interpretation, collectively defining containment relationship as a hiearchical link between data entities within data lake, which align closely with goldMEDAL's Link concept. Each relationship, focusing on different aspects of containment within data entities, finds resonance in goldMEDAL's approach of using links to represent hierarchical and structural connections between data entities. Consequently, goldMEDAL's Link concept effectively captures the essence of these relationships, facilitating their implementation within data lake.

3.3 Grouping Relationships

Grouping relationships signify how data entities are grouped and classified together, based on various criteria as outlined in Table 4, enabling a more effective data management, discovery and analysis within data lakes.

For Hai et al., grouping focuses on clustering schemas based on common attributes and similarities [12]. Halevy et al. [13] and Ravat et al. [20] group data entities from the same domain or with similar attributes due to duplication.

Table 4. Grouping relationships

Relationship name	Ref.	Authors' quote	gM concept
Schema grouping	[12]	... clusters the schemas and picks up the core of each cluster as its presentation. The necessity of grouping depends on the schema similarity calculated over the imported data sources.	Grouping
Logical cluster	[13]	... We identify datasets that belong to the same logical cluster.	Grouping
Logical cluster	[20]	... Which means that some datasets are from the same domain (different versions, duplication etc.).	Grouping
Objects groupings	[23]	... Organize objects into collections, each object being able to belong simultaneously to several collections.	Grouping
Categorization	[7]	... The categorization entity is a label assigned according to the metadata element's context.	Grouping
ZoneIndicator	[7]	... The zoneIndicator entity is a label on the data entity supplying information on the location of the data element in the data lake's zone architecture.	Grouping
Outlier datasets	[1]	... Which have no similarity with any other dataset in the DL (i.e., no similar attributes in the DL).	Grouping

Sawadogo et al. propose to organize data entities into collections, where each data entity can belong to several collections simultaneously. These groups are generated automatically based on semantic metadata, including tags and business categories [23]. Eichler et al. introduce two types of grouping relationships [7]. The first one aims to categorize data entities using their metadata elements with labels based on their context. For example, operational labels for metadata elements storing access information. The second one aims to assigning labels to data entities to indicate their location within the data lake's zone architecture. In both Categorization and ZoneIndicator, the grouping is only based on metadata.

Moreover, Al-serafi et al. propose another approach which involves identifying data entities exhibit no similarity with any other data entities. This lack of similarity can be used as a criteria to categorize data entities that belong to no grouping of shared attributes [1]. Despite the diversity in approaches, these authors converge to the same goal : categorizing data entities into groups or collections within data lake using different criteria. Furthermore, these different approaches are not mutually exclusive, rather, they can complement each other, contributing to an efficient data lake management.

As a conclusion, goldMEDAL's grouping concept effectively aligns with these various relationships by providing mechanisms to organize and categorize data entities into clusters or collections based on their similarities, context, or other criteria within the data lake.

3.4 Provenance Relationships

Provenance relationships refer to the lineage or origin of data entities, tracing their evolution within a data lake. Table 5 shows these relationships, which offer a comprehensive understanding of data origins and transformations, and provide insights into data entities' history.

In the context of identifying relationships between data entities, Halevy et al. suggest content similarity that aims in identifying data entities with similar

Table 5. Provenance relationships

Relationship name	Ref.	Authors' quote	gM concept
Content similarity	[13]	...Content similarity—both at the level of dataset as a whole and at the level of individual columns—is another graph relationship that we extract.... we rely on approximate techniques to determine which datasets are replicas of each other and which have different content.	Process
Logical cluster	[13]	...Datasets that are versions of the same logical dataset and that are being generated on a regular basis;datasets that are replicated across different data centers; or datasets that are sharded into smaller datasets for faster loading.	Process
Duplicated	[2]	...Duplicate pairs of datasets describe the same concepts. They convey the same information in most of their attributes, but such information can be stored using differences in data.	Process
Provenance	[13]	...For each dataset, we maintain the provenance of how the dataset is produced, how it is consumed, what datasets this dataset depends on, and what other datasets depend on this dataset.	Process
Tracing and Provenance	[3]	...Collect and aggregate tracing metadata (including descriptive, administrative and temporal metadata and build a provenance graph) for both data and the contextualized data.	Process
Logical cluster	[20]	...Which means that some datasets are from the same domain (different versions, duplication etc.).	Process
Parenthood relationship	[23]	...Reflect the fact that an object can be the result of joining several others. There is a "parenthood" relationship between the combined objects and the resulting object, and a "co-parenthood" relationship between the merged objects.	Process
Versions	[23]	...Raw data in the lake are often modified through updates that result in the creation of new versions of the initial data, which can be considered as metadata.	Process
Representations	[23]	...Raw data (especially unstructured data) can be reformatted for a specific use, inducing the creation of new representations of an object	Process

content [13]. Content similarity indirectly contributes to data provenance by highlighting data entities that may have originated from the same source or undergone similar transformations. For Halevy et al. and for Ravat et al., logical cluster helps in organizing related data entities within the data lake, particularly those with shared attributes or versions [13, 20]. This organization facilitates the tracing of data lineage by grouping together data entities that are likely to have similar origins or same transformations.

Alserafi et al. highlight the importance of recognizing duplicated data entities, as they may reveal common sources or transformations [2]. Which can helps trace back to the original sources and gain insights into the data entities's history and transformations.

Provenance, as emphasized by Halevy et al., involves tracking the production, consumption and dependencies of datasets, offering direct insights into their lineage and origins [13]. This explicit documentation provides information on how datasets are created and used, aiding in understanding their provenance within the data lake. Beheshti et al. advocate aggregating tracing metadata to build a thorough provenance graph, facilitating the reconstruction of data lineage within the data lake [3]. Sawadogo et al. describe the parenthood relationship, which reflects the connections between combined data entities and their resulting data entities, offering insights into their dependencies and lineage [23]. Additionally, Sawadogo et al. highlight the importance of tracking data entities versions and representations, which provide insights into data evolution and transformations over time [23]. Versioning and representation tracking contribute to data prove-

nance and data lineage by documenting changes to data entities and their structures, allowing for a comprehensive understanding of their history and evolution.

Provenance of data entities holds significant importance in the context of managing data lakes. goldMEDAL's process, enables the tracking of data entities changes over time, their origin, usage, status in the life cycle, aligning well with the notion of documenting data origins advocated in the literature. Despite the existence of various approaches, goldMEDAL's process concept demonstrates flexibility in implementing different provenance relationships outlined from literature.

4 Conclusion and Perspectives

One key challenge in data lakes is to find and discover relationships between different data entities, which facilitate the process of data integration, discovery and analysis. While various metadata management systems exist, they often do not address relationships and particularly semantic relationships.

Our primary contribution is an extensive literature review and analysis, where we identify and categorize relationships based on their underlying characteristics and implications for data management. The outcome is a relationship typology that shed light on the diverse semantic relationships between data entities within data lakes.

Furthermore, we had hypothesized that goldMEDAL could support all the relationships found in our survey. Tables 2–5 show that goldMEDAL's concepts cover all the types of relationships identified in our survey. It is somehow a validation that goldMEDAL's conceptual model provides a flexible and comprehensive framework for metadata management and a promising solution for enhancing data discovery, exploration and analysis in data lake environments.

In future research, we plan to design a metadata management system that not only supports operational and structural relationships, but also semantic relationships. As of today, we have not ruled between:

1. contribute to the open source metadata management systems available, i.e., Open Data Discovery and OpenMetadata, and extend one of them to support semantic relationships;
2. build a metadata management system from scratch, based on the goldMEDAL metadata metamodel.

Furthermore, there are explicit relationships that are easy to spot, e.g., when designing a database schema. Yet, there are also implicit relationships that are hidden, especially in data lakes with highly heterogeneous data. Such high-potential relationships, e.g., similarity relationships, can be mined by machine learning or Large Language Models (LLMs). The ultimate goal is to interlink data entities so as to navigate and search data within a whole data lake.

Eventually, we lately identified additional relationships, i.e., causality [17] and correlation [22]. Of course, they are definitely different, so we need to investigate these relationships further.

Acknowledgments. Ahlame Diouan's PhD is funded by BIAL-X. The authors thank the anonymous reviewers for their useful comments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Serafi, A.M.M.: Dataset proximity mining for supporting schema matching and data lake governance. Ph.D. thesis, Universitat Politècnica de Catalunya (2021)
2. Alserafi, A., Calders, T., Abelló, A., Romero, O.: Ds-prox: Dataset proximity mining for governing the data lake. In: Similarity Search and Applications: 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings 10. pp. 284–299. Springer (2017)
3. Beheshti, A., Benatallah, B., Nouri, R., Tabebordbar, A.: Corekg: a knowledge lake service. Proceedings of the VLDB Endowment **11**(12), 1942–1945 (2018)
4. Deng, D., Fernandez, R.C., Abedjan, Z., Wang, S., Stonebraker, M., Elmagarmid, A.K., Ilyas, I.F., Madden, S., Ouzzani, M., Tang, N.: The data civilizer system. In: Cidr (2017)
5. Diamantini, C., Giudice, P.L., Musarella, L., Potena, D., Storti, E., Ursino, D.: A new metadata model to uniformly handle heterogeneous data lake sources. In: New Trends in Databases and Information Systems: ADBIS 2018 Short Papers and Workshops, AI* QA, BIGPMED, CSACDB, M2U, BigDataMAPS, ISTREND, DC, Budapest, Hungary, September, 2-5, 2018, Proceedings 22. pp. 165–177. Springer (2018)
6. Dixon, J.: Pentaho, Hadoop, and Data Lakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (October 2010)
7. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., Mitschang, B.: HANDLE – A generic metadata model for data lakes. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 73–88 (2020)
8. Eltabakh, M.Y., Kunjir, M., Elmagarmid, A., Ahmad, M.S.: Cross Modal Data Discovery over Structured and Unstructured Data Lakes. PVLDB **16**(11), 3377–3390 (2023)
9. Fan, G., Wang, J., Li, Y., Zhang, D., Miller, R.: Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning. PVLDB **16**(7), 1726–1739 (2022)
10. Fernandes, A.A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N.W., Sakellariou, R.: Data preparation: A technological perspective and review. SN Computer Science **4**(4), 425 (2023)
11. Fernandez, R.C., Abedjan, Z., Koko, F., Yuan, G., Madden, S., Stonebraker, M.: Aurum: A data discovery system. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE). pp. 1001–1012. IEEE (2018)
12. Hai, R., Geisler, S., Quix, C.: Constance: An intelligent data lake system. In: Proceedings of the 2016 international conference on management of data. pp. 2097–2100 (2016)
13. Halevy, A.Y., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S., Whang, S.E.: Managing google's data lake: an overview of the goods system. IEEE Data Eng. Bull. **39**(3), 5–14 (2016)

14. Hoseini, S., Theissen-Lipp, J., Quix, C.: A survey on semantic data management as intersection of ontology-based data access, semantic modeling and data lakes. Journal of Web Semantics p. 100819 (2024)
15. Huang, R., Song, S., Lee, Y., Park, J., Kim, S.H., Yi, S.: Effective and efficient retrieval of structured entities. Proceedings of the VLDB Endowment **13**(6), 826–839 (2020)
16. Kaminsky, Y., Pena, E.H., Naumann, F.: Discovering similarity inclusion dependencies. Proceedings of the ACM on Management of Data **1**(1), 1–24 (2023)
17. Liu, J., Sun, S., Nargesian, F.: Causal dataset discovery with large language models. In: Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics. pp. 1–8 (2024)
18. Miloslavskaya, N., Tolstoy, A.: Big data, fast data and data lake concepts. Procedia Computer Science **88**, 300–305 (2016)
19. Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L.A., Group, P.P.: Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. Systematic reviews **4**, 1–9 (2015)
20. Ravat, F., Zhao, Y.: Data lakes: Trends and perspectives. In: Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30. pp. 304–313. Springer (2019)
21. Ravat, F., Zhao, Y.: Metadata management for data lakes. In: European Conference on Advances in Databases and Information Systems. CCIS, vol. 1064, pp. 37–44 (2019)
22. Santos, A., Bessa, A., Chirigati, F., Musco, C., Freire, J.: Correlation sketches for approximate join-correlation queries. In: Proceedings of the 2021 International Conference on Management of Data. pp. 1531–1544 (2021)
23. Sawadogo, P.N., Scholly, E., Favre, C., Ferey, E., Loudcher, S., Darmont, J.: Metadata Systems for Data Lakes: Models and Features. In: 1st International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019), Bled, Slovenia. Communications in Computer and Information Science, vol. 1064, pp. 440–451 (September 2019)
24. Scholly, E., Sawadogo, P.N., Liu, P., Espinosa-Oviedo, J.A., Favre, C., Loudcher, S., Darmont, J., Noûs, C.: Coining goldMEDAL: A New Contribution to Data Lake Generic Metadata Modeling. In: International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data. CEUR, vol. 2840, pp. 31–40 (March 2021)
25. Shah, R., Mukherjee, K., Tyagi, A., Karnam, S.K., Joshi, D., Bhosale, S.P., Mitra, S.: R2d2: Reducing redundancy and duplication in data lakes. Proceedings of the ACM on Management of Data **1**(4), 1–25 (2023)
26. Youngmann, B., Cafarella, M., Salimi, B., Zeng, A.: Causal Data Integration. Proceedings of the VLDB Endowment **16**(10), 2659–2665 (2023)