ABSTRACT
We summarize here a paper published at the DOLAP 2021 international workshop, which was collocated with EDBT and ICDT. We introduce goldMEDAL, a generic metadata model for data lakes [5].

1 INTRODUCTION
The rise of big data has revolutionized data exploitation practices and led to the emergence of new concepts. Among them, data lakes are large heterogeneous data repositories that can be analyzed by various methods [1].

An efficient data lake requires a metadata system that addresses the many problems arising when dealing with big data. The study of data lake metadata models is currently an active research topic and many proposals have been made [2–4].

However, existing metadata models (including the most recent cited above) are either tailored for a specific use case or insufficiently generic to manage different types of data lakes. To address this issue, we propose goldMEDAL, a generalization of MEDAL, the MEtadata model for DAta Lakes [4]. This new metadata model is specified through a classical three-level modeling process, i.e., conceptual, logical and physical.

2 GOLDMEDAL CONCEPTUAL MODEL

GoldMEDAL’s conceptual model features four main concepts: data entity, grouping, link and process (Figure 1).

- **Data entities** are the basic units of our metadata model. They are flexible in terms of data granularity. For example, a data entity can represent a spreadsheet file, a textual or semi-structured document, an image, a database table, a tuple or an entire database. The introduction of any new element in the data lake leads to the creation of a new data entity.

- **Groupings** are a set of groups, with a grouping bringing together data entities based on common properties. For example, the raw and preprocessed data zones common in data lake architectures are the groups of a zone grouping. Another example is a grouping of textual documents according to the language of writing.

- **Links** are used to associate either data entities with each other or groups of data entities with each other. They can be oriented or not. They allow the expression of, e.g., simple similarity links between data entities or hierarchies between groups. For example, a temporal hierarchy month → quarter would have the months of January, February and March linked to the first quarter of a given year.

- **Process** refers to any transformation applied to a set of data entities that produces a new set of data entities.
3 GOLDMEDAL LOGICAL MODEL

At the logical level, goldMEDAL concepts are represented by a graph. Data entities translate into nodes, links translate into edges and groups and processes translate into hyperedges.

For example, in Figure 2, four data entities (say, textual documents) are represented by nodes $n_1$, $n_2$, $n_3$ and $n_4$. The set of hyperedges $H_1 = \{\theta_{11}, \theta_{12}\}$ represents a zone grouping, where $\theta_{11}$ and $\theta_{12}$ are hyperedges representing the groups Raw data zone and Processed data zone, respectively. Similarly, $H_2 = \{\theta_{21}, \theta_{22}\}$ represents a language grouping, where $\theta_{21}$ and $\theta_{22}$ represent the groups French and English, respectively.

Figure 3: Sample process graph logical model

Figure 2: Sample grouping graph logical model

Figure 3 represents the logical model of a process that merges two data entities (say, two relational tables) represented by nodes $n_7$ and $n_8$ into a new node $n_9$. Process $\Pi_1 = \{\Upsilon_1, \Omega_1\}$ is an oriented hyperedge, with $\Upsilon_1 = \{n_7, n_8\}$ and $\Omega_1 = \{n_9\}$ being the sets of input and output nodes of $\Pi_1$, respectively.

4 GOLDMEDAL PHYSICAL MODELS

At the physical level, we implemented goldMEDAL into three use-cases/data lakes dedicated to social housing, management sciences and archaeology, respectively. Metadata are managed with dedicated tools (the Neo4j\(^1\) graph database management system or the Apache Atlas\(^2\) data governance and metadata management) or a combination of such tools, i.e., Neo4j, the MongoDB\(^3\) document-oriented database management system (for handling non-atomic metadata) and Elasticsearch\(^4\) search engine (for indexing textual documents).

References


5 CONCLUSION

Through the three physical models implemented with goldMEDAL, we demonstrated the feasibility and flexibility of our metadata model. Moreover, we demonstrate that goldMEDAL’s concepts generalize those of the most recent models from the literature [2–4], which makes of goldMEDAL the most generic metadata model for data lakes as of today.

Another particularity of goldMEDAL is the explicit possibility of data lineage tracing with the concept of process. Thus, goldMEDAL can manage the dynamics of data, while the most recent metadata model, HANDLE [2], does not natively support it.

Future research and open issues include the “industrialization” of data lakes, i.e., providing a software layer, connected to the metadata system, which allows non-data or non-computer scientists to transform and analyze their own data in autonomy.

Furthermore, exploiting a data lake and its metadata system may contribute to open data and open science. A well-designed data lake should indeed readily enforce the four FAIR principles\(^5\).

Acknowledgements

Étienne Scholly (PhD), Pegdwendé Nicolas Sawadogo (PhD) and Pengfei Liu’s (postdoc) are funded by the BIAL-X company, the AURA Region and the IMU LabEx, respectively.

References


5 https://www.go-fair.org/fair-principles/