POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

A Metadata Model for Data Lake in Industry 4.0: the MADE Experience

TESI MAGISTRALE IN MANAGEMENT ENGINEERING – INGEGNERIA GESTIONALE

AUTHORS: FILIPPO TUBINO, EDOARDO TONETTI

ADVISORS: prof. PIERLUIGI PLEBANI, prof. CAPPIELLO CINZIA

ACADEMIC YEAR: 2021-2022

1. Introduction

The increasing amount of data collected and used for analysis required a change from traditional data warehouses to data lakes. Despite this, data lakes are still a relatively new technology and a defined approach for their implementation is lacking. Crucial to the management of this system is the management of metadata, through which data can be easily found in the repository once saved. Several researchers have proposed metadata models: frameworks for metadata management that offer different features whose utility depends on the context of use [1]. To date, it has not yet been defined which metadata and metadata model are effective for successfully implementing a data lake in Industry 4.0 (I4.0) environments.

The goal of this dissertation is therefore to define a metadata classification useful for I4.0 context, validate it in a case study, understand the link between metadata and data lake features, and finally to understand which data lake features are useful for I4.0 organizations. This will allow selecting the most tailored metadata model that prevents the entire system from becoming a "data swamp", a repository of data in which data analysts cannot find what is of interest.

2. Adopted method

To select the most appropriate metadata model, information about the use of data in I4.0 contexts is required. To obtain this information, we did a literature state-of-the-art analysis and then we validated the work in a case study. To do this, we turn to MADE, an I4.0 competence center of which Politecnico di Milano is one of the founders. This collaboration allowed us to interview different area managers, each one specialized in a different field of I4.0. The aim is to identify the needs and requirements when implementing a data lake in I4.0 environments.

The results validation with MADE requires two sets of interviews. The first one is cognitive, in which questions inquiring about the as-is situation in the area of competence, thus allowing the gathering of information on how the organization currently works. The second phase of interviews instead, is aimed at investigating the usefulness of

various metadata categories that have been selected as relevant from the literature.

2.1 Sawadogo et al. metadata classification

Considering the importance of metadata, it is necessary to define the type of metadata that can be found within a data lake and how these are organized. After an analysis of the literature, we identified the most cited and comprehensive Sawadogo et al. metadata classification, which organizes data based on the structural metadata types. This metadata categorization considers an extended metadata typology that categorizes metadata into intra-object, inter-object and global metadata with new types of inter-object (relationships) and global (index, event logs) metadata [2]. According to this classification we distinguish between Intra-object metadata, Inter-object metadata, and Global metadata as described in the following.

2.1.1 Intra-object metadata

These are the metadata associated with characteristics related to a single object within the repository. They are subdivided into:

- **Properties (PR)**: provide a general description of an object. Provide details such as object title, file name, size…
- **Summaries and previews (S&P)**: give a general explanation of an object's structure or content.
- **Semantic metadata (SM)**: provide a textual description that makes it easier to understand the content of the data.

2.1.2 Inter-object metadata

It represents the relationships existing between the different data in the system. These links between different objects can be between 2 or more elements. We distinguish:

- **Object groupings (OG)**: allows objects to be organised in groups. These can be generated automatically on the basis of certain intra-object metadata, such as semantic metadata.
- **Similarity links (SL)**: these metadata reflect the similarity between two objects based on their intrinsic characteristics.
- **Parenthood links (PL)**: this category of metadata indicates the relationships between objects that have been generated by the transformation of other data.

2.1.3 Global metadata

Global metadata are data structures intended to give a contextual layer to the entire data lake. These are therefore not information attributable to individual objects but to the entire data lake. Here we identify:

- **Semantic resources (SR)**: indicate knowledge bases which allow once that a metadata has been associated with an object, to associate also tags with comparable semantic descriptions to it.
- **Indexes (IX)**: are data structures that facilitate data retrieval. These allow the user to query the data lake with word-based queries and find metadata with similar meanings.
- **Logs (LS)**: these metadata make it possible to record data access by different users.

2.2 Metadata classification validation in I4.0 context

Interviews with actors involved in MADE gave the possibility to validate the proposed metadata model and to collect significant feedback on the role of metadata in an Industry 4.0 setting as described below.

2.2.1 Intra-object metadata results

Intra-object metadata are the starting point on which both inter-object and global metadata are built. These are important because if in a data lake all objects are well-constructed analysis is easier and the other two macro-categories are unimportant if this is not done correctly. Properties prove to be the most essential metadata since they provide useful information for understanding what can be figured out from that data and this is often a starting point for analyses. Summaries and previews are very useful in a context where there are large amounts of unstructured data collected, in order to have an overview of their content. With few unstructured data used they lose usefulness. Finally, Semantic metadata has proven useful for areas where the semantics used is standardized, while it can cause problems and misunderstandings when this is not regulated.

2.2.2 Inter-object metadata results

The usefulness of the inter-object metadata strictly depends on the goodness of the intra-object metadata. These metadata are useful for improving search and facilitating the discovery of data clusters for analysis. The importance of these metadata is directly proportional to the amount of data collected. Object grouping metadata are very used since associating metadata based on semantics turns out to be a widely used way of searching data, if not even necessary. Similarity links metadata can be considered an advanced function compared to the other two inter-object categories, and useful to automatically extract insights from data. Last, Parenthood links metadata show a very high degree of usefulness due to the way they actually work. The grouping of elements that belong to the same product, process or source data makes it much easier to understand the data.

2.2.3 Global metadata results

The usefulness of these metadata is recognized especially when the amount of data increases and when there are exchanges of information between areas or with third parties. Semantic resources and Indexes metadata allow to easily find data in a similar but opposite way. Semantic resources allow additional tags similar to those already present, to be attached to the data. On the other hand, indexes, at the moment in which a user queries the data lake with a textual query, the system will also search for semantic tags similar to those entered by the user. This facilitates the identification of data of interest. Finally, logs were found to be unimportant since data security has not been declared a priority in the use case.

2.3 Metadata in I4.0 environments

Starting from the results obtained we group these metadata into 3 macro utility groups, based on the mark that every single metadata received by respondents during the interviews. We can distinguish between essential, useful, and advanced metadata as shown in Table 1.

- **Essentials metadata**: with these metadata, it is possible to describe an object with basic metadata, group it based on these attributes, and keep track of the relationships and hierarchical structure of the data. The usefulness of this metadata is that many area managers work using this information.
- **Useful metadata**: The semantics used for metadata is often causing confusion and slowing down the analysis process. These functions, if not essential for the proper functioning of the data lake, prove to be very useful especially when there are no policies for semantic standardization within the company.
- **Advanced metadata**: with "advanced" we mean those functions that are considered something more than basic metadata management operations. They are useful metadata when large amounts of data are collected since they are automatically assigned metadata tags and allow automatic identification of relevant data clusters.

Table 1 Essential, useful, and Advanced metadata

Essential metadata	Properties
	Object groupings
	Parenthood relationships
Useful metadata	Summaries and previews
	Semantic metadata
	Indexes
Advanced metadata	Similarity links
	Semantic resources
	Logs

3. Data Lake Features and Enabling Metadata

The analysis of the fundamental features of a data lake have been inspired by the work done by Sawadogo et al. [2] and R. Eichler et al. [3].

After having merged the two classifications, considering that some functionalities of Sawadogo overlapped logically with some of Eichler and vice versa, it is noticed that there is a grey area not covered by any functionality: **Data provenance (PV)**. Indeed, is needed a function that stores and allows us to trace back any kind of transformation that the data has undergone. This same function

must also allow us to trace the physical, and geographic provenance and the path the data has taken to get to the destination where we find it.

In the end, the final merged set of functionalities is:

- **Semantic enrichment (SE):** it enables adding textual descriptions to data, describing their content and making the data more comprehensible.
- **Data Indexing (DI):** this is the search engine function of a data lake. It allows searching for data using keywords or patterns making it easier to search for data within the data lake.
- **Link Generation (LG):** this function makes it possible to generate links between different data in order to facilitate searching. These links can be generated manually or automatically by the tools.
- **Data Polymorphism (DP):** if data is transformed to be adapted to a new context, there must be a reverse function that allows going back to the original state. In this way, multiple representations of the same data are allowed at different levels of detail or structure.
- **Data Versioning (DV):** this functionality automatically relates two or more data, one of which is the latest updated or modified version of the other.
- **Usage Tracking (UT):** allows the recording of iterations (creation, access, and update of data) between users and the data lake.
- **Granularity Levels (GL):** it is necessary to consider the fact that data can be aggregated hierarchically following the various dimensions, according to the context where this data belongs.
- **Data Provenance (PV):** it allows to identify the provenance of data and it allows to trace back any kind of transformation that the data has undergone.

Now that data lake functionalities are known we have to link metadata to data lake features to turn the information on metadata usefulness into feature usefulness as summarized in Table 2. This information will then allow us to select the most appropriate metamodel based on the features required in I4.0 environments. Metadata is seen as the enabler for those functionalities. It is the input for the proper functioning of data lake features. Analyzing which metadata enabled certain features, we found that there were missing some metadata categories not present in the Sawadogo et al. classification.

Therefore, it is necessary to add three new metadata categories to the classification presented in chapter 2 and evaluate their usefulness using interview responses:

- **Data Version (DV):** this metadata enables data versioning. This metadata reflects the version and reasons for a data update. It falls under inter-object metadata as it relates to two distinct data elements within the data lake. It is considered an advanced metadata since it would be useful in a dynamic context, where different versions of the same data are often created, so very far from the MADE use case.
- **Difference Links (DL):** this one enables data provenance. This metadata contains quantitative information regarding what distinguishes the actual version of a data object from the previous ones. It is an inter-object metadata and it falls under advanced metadata category for the same reasons of DV.
- **Link Indicator (LI):** The link stores the information about the source from which the data was imported into the zone as well as the appropriate timestamp. This one must be an intra-object metadata. It is considered by MADE area manager as essential information for as-is analysis since it gives the possibility to trace the provenance and contextualize it. It is therefore considered an essential metadata.

Table 2 Metadata as data lake feature enabler

Function	Metadata1	Metadata2	Metadata3
SE	PR	SM	SR
DI	**IX**	**SL**	**SR**
LG	OG	SL	
DP	**S&P**		
DV	DV		
UT	**LS**		
GL	PL		
PV	**IL**	**DL**	

4. Metadata as input for data lakes features

To transfer the utility of metadata to the feature of a data lake they enable, a directly proportional relationship between the utility of metadata and the respective feature enabled is found to exist. The average or the median of the enabling metadata utility are not suitable for this purpose. This is because features with more enabling metadata are penalized if there is a lot of unhelpful metadata among them. This is wrong because it only takes a few very useful metadata enabled by a feature to make it very important. This problem leads to lower ratings of features that are enabled by multiple metadata simultaneously. The method that allows the utility of the features to be more closely aligned with the opinions gleaned from the interviews is to assign each feature the higher utility of the metadata that uniquely enables the function.

$$DL\ feature\ utility = MAX(\ enabling\ metadata\ utility)$$

According to the results, functions are divided into two categories based on their utilities. We can distinguish between **advanced** features and **basic** features as shown in Table 3. The basic ones are functions that are indispensable for the operation of a data lake in I4.0 industries, functions from which one cannot disregard; without them, the data lake would turn into a data swamp. Advanced features, on the other hand, increase in usefulness as the complexity of the lake increases. Complexity is related to a large number of data extracted from numerous and heterogeneous sources both structured and unstructured, or a dynamic context, in which data are updated frequently.

Table 3 Basic and Advanced data lake features

Basic	Advanced
SE	DP
LG	**DV**
GL	UT
PV	
IX	

5. Metadata model selection

Once understood the importance of each data lake feature in industry 4.0 context, the focus shifts to assessing which is the optimal metadata model for managing a data lake in this environment.
In assessing the most suitable metamodel the presence of basic features is prioritized. As can be seen from Table 4, the metamodel with the highest number of basic features satisfied is goldMedal [4].

Table 4 Metadata features availability

Metadata models	SE	LG	GL	PV	IX	DP	DV	UT
goldMedal	X	X	X	Y	X	X	X	X
Medal	X	X		Y	X	X	X	X
Handle	X	X	X		X	X		X
Ravat & Zhao	X	X			X	X	X	X
GEMMS	X		X		X			
Ground	X				X		X	X
Diamantini	X	X	X		X	X		
GOODS	X	X		X	X		X	X
CoreKG	X	X			X	X		X
			Basic				Advanced	

Another driver to consider in addition to the number of basic features supported is the complexity of implementation due to the presence of advanced features. The more advanced features are present in a metamodel, the higher the implementation complexity, meaning that it would be more difficult to develop the metamodel in the data lake. Here Diamantini [5], looks like the winner, that even if is not a metamodel designed ad-hoc for data lakes, is a candidate for being an optimal metamodel for data lakes in I 4.0.
The last driver to consider in the choice is the disclosure of details about the metamodel functioning. Actually, GOODS R. [6] and CoreKG [7] are considered black boxes, since are proprietary metamodels. Especially GOODS is developed and used by Google to manage its database. On the other hand, E. Scholly et al. the goldMedal developers, state that they take great care in following users in the appropriation of the know-how needed to exploit the metamodel at its maximum. GoldMedal is also designed in order to ensure understandability to non-technical users.

In the end, priority needs to be given to the metamodel that satisfies the most basic features, since without them the data lake would be unusable. Moreover, considering that goldMedal is well disclosed in terms of functionalities and implementation methods, it looks like the best candidate to be chosen as the optimal metamodel for I4.0 environment.

6. Conclusions

The requirements and needs for implementing a data lake change depending on the industry or context in which they are used. Data management within the data lake is still the main challenge for an effective implementation. A poor choice for metadata management can turn the entire data lake into a so-called "data swamp", a repository of data in which it is difficult to find what you are looking for. The selection of the most appropriate metadata model depends on many factors such as the data used and the analyses that are done.

Through the work done in this dissertation, it has been possible to define the utility of different metadata categories and data lake features useful in Industry 4.0 environments, selecting goldMedal as the most appropriate metadata model in these contexts. As a result of the work done so far, the fundamentals for designing a data lake in I4.0 and IoT-related fields have been laid. It is to be considered as a starting point for the applicative development of a data lake. Once the metadata model is clear, it will then be necessary to build and program the data lake considering the metadata organization identified. Doing so will require having appropriate skills and selecting the right tools to build the final solution.

Although the literature and knowledge regarding data lakes are constantly evolving there are few real-world examples of metadata model implementations. This is to be expected as data lake technology is relatively new and the literature is constantly evolving. Since this is a theoretical work all the analyses done on metadata, features and metadata model are perfectly implementable on most open-source libraries (such as Apache Hadoop). How to do this remains an open issue and it must be a subject of analysis in future research. The implementation of most of the proposed features requires the integration of several additional modules into Apache Hadoop. With different levels of complexity, it is needed to work to be able to develop the technology needed to implement the different functions.

Another open issue remains the integration of the metadata system with the data catalog. This tool allows non-data analyst users to easily navigate the data lake, making the system easier to navigate and increasing the number of users who can interact with it. The goodness of the data catalog strictly depends on how well the metadata model used is integrated with it.

The industrial value of having a data repository in which you cannot find what you are looking for is zero. Some analyses may also not be feasible given the inability to find the data of interest. Due to the work done this risk decreases by selecting the most tailored metadata model for the context of use. This will optimize and speed up data searching.

References

[1] S. B. Alice LaPlante, in *Architecting data lakes: Data management architectures for advanced business use cases*, O'Reilly Media, 2018, pp. 1-55.

[2] P. N. S. E. F. C. F. E. L. S. &. D. J. Sawadogo, "Metadata systems for data lakes: Models and features," *European Conference on Advances in Databases and Information Systems,* vol. 1, pp. 440-451, 2019.

[3] R. G. C. G. C. S. H. &. M. B. Eichler, "Handle-a generic metadata model for data lakes," in *International Conference on Big Data Analytics and Knowledge Discovery*, 2020, pp. 73-88.

[4] E. S. P. L. P. E.-O. J. A. F. C. L. S. e. a. Scholly, "Coining goldMEDAL: A new contribution to data lake generic metadata modeling," 2021.

[5] P. L. G. L. M. D. P. E. S. D. U. C. Diamantini, "A New Metadata Model to Uniformly Handle Heterogeneous Data Lake Sources.," *Springer International Publishing,* p. p. 165, 2018.

[6] K. F. N. N. O. C. P. N. R. S. W. S. A.Y Halevy, "Goods: organizing google's datasets. In Proceedings of the 2016 international conference on management of data," *SIGMOD 2016,* 2016.

[7] B. B. R. N. A. T. A. Beheshti, "CoreKG: A knowledge lake service.Proceedings of the VLDB Endowment.," p. 11(12), 2018.