Towards Augmenting Metadata Management by Machine Learning

Christopher Julian Kern [1], Thomas Schäffer [2], Dirk Stelzer [3]

Abstract: Managing metadata is an important section of master data management. It is a complex, comprehensive and labor-intensive task. This paper explores whether and how metadata management can be augmented by machine learning. We deduce requirements for managing metadata from the literature and from expert interviews. We also identify features of machine learning algorithms. We assess 15 machine learning algorithms to determine their contribution to meeting the requirements and the extent to which they can support metadata management. Supervised and unsupervised learning algorithms as well as neural networks have the greatest potential to support metadata management effectively. Reinforcement learning, however, does not seem to be well suited to augment metadata management. Using Support Vector Machines and identification of metadata as an example, we show how machine learning algorithms can support metadata management.

Keywords: Artificial Intelligence, Machine Learning, Master Data, Metadata Management

1 Introduction

Digital transformation in companies, government agencies and our society as a whole requires appropriate data management. In a corporate context, managing master data is a key element of data management. The term master data refers to critical business objects of an organization. Master data management aims at planning, evaluating and controlling master data so that they can be used effectively and efficiently [Lo08]. Managing metadata is a central element of master data management [Lo08, HOÖ11]. Metadata is often referred to as data about content data [BNP99]. Hüner et al. define metadata [HOÖ11] as structured data referring to other data. Metadata focus either on technical characteristics [To99] or on features of applying content data [BNP99]. On the one hand, the challenges for metadata management are growing with increasing digitization, on the other hand, managing metadata is a complex, comprehensive and labor-intensive task [Lo08]. Many organizations will therefore need support for metadata management. One solution to this issue can be automation or at least automated support.

[1] EBS Universität für Wirtschaft und Recht, Management Group, Gustav-Stresemann-Ring 3, Wiesbaden, 65189, christopher.kern@ebs.edu, https://orcid.org/0000-0002-6864-2259
[2] Hochschule Heilbronn, Institut für Wirtschaftsinformatik, Max-Planck-Str. 39 74081 Heilbronn, thomas.schaeffer@hs-heilbronn.de, , https://orcid.org/0000-0001-8097-286X
[3] Technische Universität Ilmenau, Fachgebiet Informations- und Wissensmanagement, Postfach 100565, Ilmenau, 98693, dirk.stelzer@tu-ilmenau.de, https://orcid.org/0000-0002-6757-9411

Machine learning, a section of artificial intelligence, describes the ability of a machine or software to learn the execution of specific tasks. Machine learning is a powerful tool for analyzing large volumes of data [Mu12]. It is therefore reasonable to study to what extent machine learning is suitable to support metadata management. In this paper, we explore how machine learning algorithms may help to fully or partially automate selected tasks in metadata management. Using artificial intelligence for master data management has received considerable attention. The Gartner Group has labelled it Augmented MDM [JW20]. This paper focusses on assessing whether and how machine learning algorithms may be used to augment metadata management.

Prior research has only sparsely assessed the field of automating operations performed on metadata. Ganesan et al. have explored automatic identification of entities and their relationships which were then visualized based on metadata [Ga20]. Murthy et al. [Mu10] have considered automated extraction of metadata from a set of master data objects. This research was rather prototypical and the results are more than ten years old. Therefore, this topic deserves a fresh look and a more detailed analysis.

The objective of this paper is to identify to what extent selected machine learning algorithms may support metadata management tasks. We aim to answer two questions:

- RQ1: What are key requirements for augmenting metadata management?
- RQ2: Which machine learning algorithms support implementing these requirements?

2 Methodology

Figure 1 illustrates our research approach. The first step is based on the functional architecture for master data management proposed by Otto and Hüner [OH09]. This architecture refers to master data management as a whole. It consists of numerous sub-architectures, of which metadata management is one. We also conducted a literature review to elicit functional requirements for tools supporting metadata management. We followed the approach proposed by Fettke [Fe06] and analyzed the results according to Webster and Watson [WW02]. To identify relevant literature we used Google Scholar, Web of Science, the AIS eLibrary, IEEE Xplore and Springer Link. We used the search term [„metadata management“ OR „Metadatenmanagement”] and analyzed each article’s keywords and abstract. We considered 29 publications as depicted in Fig. 1. We included articles on metadata management in the context of master data management. From the findings of the review, we derived a set of requirements via a requirements elicitation process according to ISO 29148 [IS18]. We identified 37 requirements and summarized them in seven categories.

In the second step, we discussed the practical relevance of each requirement and prioritized the requirements with 15 participants of the master data management roundtable at the Heilbronn University of Applied Sciences. The participants are involved in strategic or operative management of master data in eight German companies. We also conducted four expert interviews. Each of the experts is responsible for master data management and works either in digital sales, process management or R&D. Our questionnaire consists of 25 closed-ended questions as well as six open-ended questions.

We evaluated answers to the closed-ended questions using a qualitative content analysis proposed by Mayring [Ma15]. We present our findings in form of a concept matrix based on Webster and Watson which is depicted in table 1 [WW02].

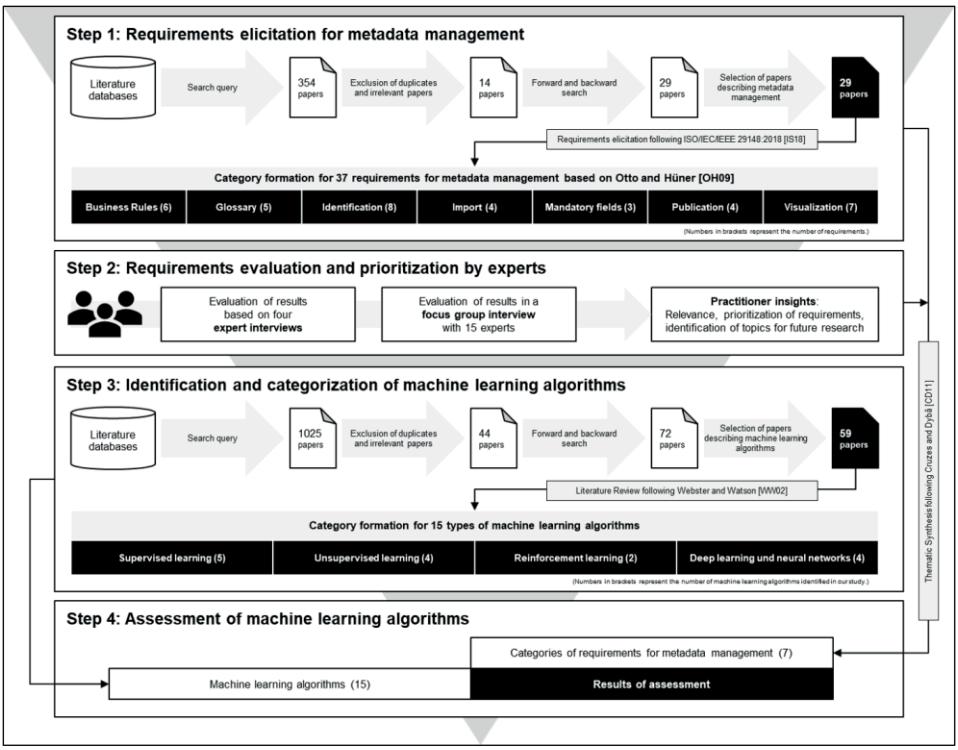

Fig. 1: Research approach

In the third step, we performed a second literature review to identify and categorize machine learning algorithms. We followed a similar approach as in the first review described in step 1. We used the search term ([„master data management“ OR „metadata management“ OR “Stammdatenmanagement” OR “Metadatenmanagement” OR “algorithms” OR “Algorithmen”] AND [„machine learning“ OR „maschinelles Lernen“]). After analyzing keywords and abstracts, we considered 59 publications. We explored papers published between 2005 and 2020. We included articles on machine learning and machine learning algorithms, respectively. We studied 15 types of machine learning algorithms and grouped them into four categories, namely *supervised learning*, *unsupervised learning*, *reinforcement learning* and *neural networks*. We identified functions and features, potential fields of application as well as benefits and downsides when applying these algorithms.

The fourth step consists of assessing selected machine learning algorithms. We assessed the algorithms based on an argumentative-deductive approach using the 37 requirements identified in step 1. We used the approach presented by Cruzes and Dybå [CD11] and

extracted a model of higher-order themes from our results, merging sub-requirements into requirements. We used a four-eye-principle as proposed by Peffers et al. [Pe12]. We present our findings in section 3.4.

3 Findings

3.1 Requirements for Metadata Management

Subsequent to the first literature review, we performed a requirements elicitation process following ISO 29184 [IS18]. This resulted in 37 requirements, which we grouped in seven categories. Six of these categories were originally proposed by Otto and Hüner [OH09]. We added the category *identification* of metadata as this is regarded also as an important area of metadata management in the literature. In the following passage, we briefly explain the seven categories of metadata management tasks (see figure 1).

- Documentation of *business rules* seeks to document process related guidelines, partly formalized rules and individual knowledge of staff members. Documented business rules are essential for process automation [OÖ16]. These rules can be formed with the help of metadata definitions. They refer to individual master data elements or to a set of these data.
- Creation of a *glossary* or dictionary is a prerequisite for identifying, documenting and maintaining master data. A glossary can be created and updated using metadata. It supports documentation of and communication on business processes for which master data is required [Lo08].
- *Identification* means creating metadata for already existing master data. These metadata must be continuously updated afterwards. Collecting all definitions of master data in an organization creates a basis for metadata management [Lo08]. These definitions can be documented in a data catalogue or data dictionary [Hi21].
- *Metadata import* helps to consolidate metadata stored in different formats into a coherent system. Import of metadata from heterogeneous systems is an element of integration processes as described by Loshin [Lo08].
- *Mandatory fields* must contain valid values if master data is inserted into a database or updated [Hi18]. Managing mandatory fields is a prerequisite to ensure adequate quality of master data. Metadata is used to define mandatory or optional fields.
- *Metadata publication* denotes the process of making metadata available for operational information systems. This service supports a coherent use of master data. Glossaries and data dictionaries help to maintain a consistent use of definitions and formats when integrating information systems [Lo08].
- The aim of *visualization* is to display master data in a graphical form. Visualization uses metadata to represent data objects graphically, e. g., with Entity-Relationship- or UML-Diagrams.

3.2 Evaluated and Prioritized Requirements

As already mentioned, we conducted expert interviews to evaluate and prioritize the results of our requirement elicitation process. In the following focus group interviews, we summarized and discussed the results of the interviews with a larger group of experts. We asked them to prioritize the seven metadata management tasks described above. The experts assigned priority one to the following tasks: documentation of *business rules*, *identification* of metadata and *visualization* of master data. Priority two was assigned to definition of *mandatory fields* and metadata *import* and priority three to creation of a *glossary* and metadata *publication*. These priorities provide valuable guidance for the selection of practical projects to support metadata management by machine learning.

3.3 Selected Machine Learning Algorithms

In step 3, we identified and categorized four types of machine learning algorithms. In the following step, we assess how they can be used to augment metadata management.
Supervised learning denotes the ability of algorithms to allocate a given set of inputs to a set of outputs [GGA20]. We further evaluated four subtypes of supervised learning, presented by Sindhu et al. [SS20], namely ‘support vector machines‘, ‚naïve bayes‘, ‚linear regression‘ and ‚decision trees‘. We also subsume ‘semi-supervised learning’ as another sub-category of supervised learning.
Unsupervised learning seeks to find patterns in a given set of data [GAM20]. Nock and Nielsen [NN06] describe four types of algorithms as key examples: ‘K-means-clustering’, ‘fuzzy c-means’, ‘gaussian-expectation-maximization’ and ‘harmonic k-means-clustering’.
Reinforcement learning trains algorithms with a set of reinforcements, ‘rewards’ and/or ‘punishments’ [Ru16]. Weber describes ‘q-learning’, ‘temporal difference learning’ and ‘generative adversarial networks’ as instances of reinforcement learning [We20a].
Neural networks are based on the assumption that mental activity consists primarily of the electrochemical activity of networks of neurons [Ru16]. In a neural network, ‘artificial neurons’ - arranged in a network - perform the computation. The basic architecture consists of three layers of artificial neurons: an input layer, a hidden layer and an output layer [Ha09]. There are numerous variations of neural networks [Ru16]. We decided to address ‘auto-encoders’, ‘convolutional neural networks’ and ‘recurrent neural networks’.

3.4 Selected Findings of Assessments

In step 4 of our research process, we assessed the extent to which the algorithms meet the requirements. Table 1 shows the results of the assessment. For reasons of space, we cannot show all results in detail here, but only at the level of the requirement categories. The original presentation showing all 37 requirements is available from the authors upon request. We illustrate the results using Harvey-Balls. A full black ball represents a strong match between the requirements and the functionality of an algorithm, a white ball expresses that an algorithm does not match the requirements at all.

Machine learning algorithms	Requirements: Business rules	Glossary	Identification	Import	Mandatory fields	Publication	Visualization
Supervised learning							
Support vector machines	◔	●	●	◕	◑	●	●
Naïve bayes	◕	●	◕	◕	◑	●	◕
Decision trees	◑	◕	◑	◑	◑	◑	◑
Semi-supervised learning	◑	◕	◕	◑	◕	◕	●
Linear regression	○	○	○	○	○	○	○
Unsupervised learning							
K-means-clustering	◑	◕	◑	◑	◔	◕	●
Fuzzy-c-means	◑	◕	◑	◑	◔	◕	●
Gaussian-EM	◑	◕	◕	◑	◔	◕	●
Harmonic-k-means-clustering	◑	◕	◑	◑	◔	◕	●
Reinforcement learning							
Q-learning	○	○	○	○	○	○	○
Temporal difference learning	○	○	○	○	○	○	○
Generative adversarial networks	○	○	○	○	○	○	○
Neural networks							
Auto encoders	◑	●	◑	◕	●	◕	●
Convolutional neural networks	◔	◑	◕	◔	◔	◑	◑
Recurrent neural networks	◑	◕	◕	◔	◑	◕	◑

Tab. 1: Assessment of machine learning algorithms with metadata management requirements

Four types of algorithms were assessed as not suitable for meeting any requirement. 'Linear regression' primarily aims at predicting future states. This does not correspond with any of the requirements. The same applies to *Reinforcement Learning Algorithms*, namely 'q-learning' and 'temporal difference learning'. These algorithms are primarily suitable for problems that can be solved with the help of Markov chains. 'Generative adversarial networks' are particularly fit for generating sets of data, e.g., for manipulating or generating images.

The experts have ranked the identification of metadata as one of the most important functional requirements. We have assessed 'support vector machines' (SVM) as the algorithm with the highest potential to support identifying metadata. Therefore, we focus on this task to illustrate how machine learning algorithms can support metadata management.

SVM are among the most popular supervised learning algorithms. Algorithms are trained to draw boundaries through a given set of data and to group them into classes [Ru16].

Training SVM needs a large number of data of sufficient quality [We20b]. Bahmani et al. [BBV17] use SVM for data cleansing and for the detection of duplicates to be merged into a single data object. Singh et al. [STS16] name text classification as the main application of SVM. Advantages are high accuracy of the algorithms, avoidance of overfitting and a good suitability for generalizing facts. Disadvantages are the complexity of SVM, the extensive training effort needed and that performance depends on the choice of appropriate parameters. In the following paragraphs, we address the eight requirements that we have summarized in figure 1 and table 1 under the heading *identification* of metadata.

In the context of SVM, we primarily assessed classification-, entity detection- and entity resolution problems. Classification problems connect a given set of inputs with a discrete number of outputs [Mu12]. Text classification problems use certain words or word combinations to characterize an object [STS16]. Duplicates belong to certain groups of objects which can be characterized by their attributes [BBV17]. Conflicts resulting from duplicates can be resolved by either replacing them with a single object [BBV17] or by connecting the two objects – which also aids in representing semantic aspects [HOÖ11, Lo08].

The first requirement is ‘The algorithm shall list all definitions of master data objects that exist within an organization in the form of metadata’. This is feasible because SVM can use text classification to recognize data objects and definitions. However, this requires considerable training.

The second requirement, 'The algorithm shall recognize and list similarities and differences of master data definitions in the form of metadata', can easily be satisfied, as it is a duplicate detection problem in combination with text recognition. The amount of training required will probably also be high.

The third requirement is 'The algorithm shall standardize and harmonize similar definitions of master data objects in the form of metadata'. The implementation is feasible since text recognition is the main task here as well. Training algorithms will probably be time-consuming. The results may need to be improved by experts.

The fourth requirement is 'If similar definitions of master data objects cannot be standardized, the algorithm shall name them differently in the metadata according to a uniform scheme'. The implementation is also quite simple, since this again is a text recognition problem. However, the naming scheme must already exist for the results to be usable.

The fifth requirement is 'If similar definitions cannot be standardized, but data objects are similar, the algorithm shall connect the metadata of data objects'. This is possible because linking definitions is a classification problem. However, the complexity of the problem substantially increases the effort required to train the algorithm. The results may also need to be improved by experts.

The sixth requirement, 'If the existing definitions of master data objects do not match newly determined definitions, the algorithm shall replace them in the metadata', can be implemented well with SVM as this is mainly a comparative text recognition problem.

The seventh requirement is 'If two definitions of master data objects exist at the same hierarchical level and if they have the same fields, the algorithm shall indicate that they may be synonyms and establish a link in the referring metadata'. SVM are well suited to support this task, since it is a classification and text recognition issue. As with previous

requirements, the training effort is probably significant and the results may need to be improved by experts.
The eighth requirement (‘The algorithm shall enter definitions of master data objects into a repository') does not require machine learning. Therefore, we have not considered this requirement in our study.

4 Conclusion

4.1 Summary

We have identified 37 requirements for augmenting metadata management in the context of master data management and grouped these requirements in seven categories. Experts evaluated our findings and prioritized the requirements. On this basis, we can now answer RQ 1: Key requirements for augmenting metadata management are documentation of business rules, identification of metadata and visualization of master data.
We have also identified 15 types of machine learning algorithms and grouped them in four categories. We explored which algorithms are fit for meeting which of the requirements identified in the first step of our research. We answer RQ 2 as follows: Supervised and unsupervised learning algorithms and, with certain restrictions, neural networks have the greatest potential to support metadata management effectively. Reinforcement learning, however, does not seem to be suited to augment metadata management.
A more detailed assessment showed that SVM might best support the identification of metadata. Unsupervised machine learning algorithms are appropriate to augment visualization of metadata. These algorithms help to identify and classify elements to be visualized and their relationships. Unfortunately, no algorithm is well suited to support documentation of business rules.

4.2 Limitations

We need to acknowledge some limitations to our research. First, our database for eliciting requirements for algorithms to support metadata management was rather small. While there is a large body of literature on metadata management, sources that address managing metadata referring to master data are rare. The number of experts we interviewed was also limited. A more extensive number of relevant publications and more interviews could lead to a more comprehensive list of requirements. Second, our research has only been able to provide an overview of the topic. We took a comprehensive look at metadata management and assessed 15 algorithms against 37 requirements. It is likely that a more detailed analysis of specific subtasks of metadata management would lead to requirements that are more specific. Focusing on one or a few algorithms would allow assessing the contribution of these algorithms more precisely. Third, we have worked and argued on a conceptual level only. However, we have not practically tested the extent to which concrete algorithms are suitable for meeting specific requirements.

4.3 Implications

Our results can help practitioners who want to support managing metadata through artificial intelligence to pre-select machine learning algorithms. Investigations that are more concrete must be carried out in specific application areas. We believe that our work offers a useful basis for future research. As mentioned above, an expansion of the database with more publications and more interviewees, a more detailed analysis of the contribution of selected algorithms to concrete subtasks of metadata management, and a real-world testing of algorithms in practical applications are promising research opportunities. It would also be interesting to explore in which cases the benefits of using machine learning algorithms exceed the costs involved. We have limited our analysis to assessing the contribution of machine learning algorithms for augmenting metadata management. Research that is more extensive could consider the entire area of master data management.

5 Bibliography

[BBV17] Bahmani, Z.; Bertossi, L.; Vasiloglou, N.: ERBlox: Combining Matching Dependencies with Machine Learning for Entity Resolution. In International Journal of Approximate Reasoning, 2017, 83; pp. 118–141.

[BNP99] Burnett, K.; Ng, K. B.; Park, S.: A comparison of the two traditions of metadata development. In J. Am. Soc. Info. Sci., 1999, 50; pp. 1209–1217.

[CD11] Cruzes, D. S.; Dybå, T.: Recommended Steps for Thematic Synthesis in Software Engineering. In (Ed. IEEE Computer Society): ESEM 2011 Proceedings, 2011; pp. 275–284.

[Fe06] Fettke, P.: State-of-the-Art des State-of-the-Art. In WIRTSCHAFTSINFORMATIK, 2006, 48; pp. 257–266.

[Ga20] Ganesan, B. et al.: Link Prediction using Graph Neural Networks for Master Data Management, 2020.

[GAM20] Guérin, E.; Aydin, O.; Mahdavi-Amiri, A.: Artificial Intelligence. In (Eds. Guo, H.; Goodchild, M. F.; Annoni, A.): Manual of Digital Earth. Springer, Singapore, 2020; pp. 357–385.

[Ha09] Haykin, S. S.: Neural networks and learning machines. Pearson, New York, 2009.

[Hi21] Eds. Hildebrand, K. et al.: Daten- und Informationsqualität. Auf dem Weg zur Information Excellence. Springer Vieweg, Wiesbaden, 2021.

[HOÖ11] Hüner, K. M.; Otto, B.; Österle, H.: Collaborative management of business metadata. In International Journal of Information Management, 2011, 31; pp. 366–373.

[IS18] ISO/IEC/IEEE: Systems and Software Engineering: Life Cycle Processes Requirements Engineering (ISO/IEC/IEEE 29148), 2018.

[JW20] Judah, S.; White, A.: Hype Cycle for Data and Analytics Governance and Master Data Management, 2020, 05.04.2021.

[Lo08] Loshin, D.: Master Data Management. Elsevier, Amsterdam, 2008.

[Ma15] Mayring, P.: Qualitative Inhaltsanalyse. Grundlagen und Techniken. Beltz Verlag, Basel, 2015.

[Mu10] Murthy, K. et al.: Content-Aware Master Data Management: Proceedings of the 16th COMAD, Nagpur, India, 2010; pp. 206–210.

[Mu12] Murphy, K. P.: Machine learning. A probabilistic perspective. MIT Press, Cambridge, 2012.

[NN06] Nock, R.; Nielsen, F.: On weighting clustering. In IEEE transactions on pattern analysis and machine intelligence, 2006, 28; pp. 1223–1235.

[OH09] Otto, B.; Hüner, K. M.: Funktionsarchitektur für unternehmensweites Stammdatenmanagement. Universität St. Gallen, St. Gallen, 2009.

[OÖ16] Otto, B.; Österle, H.: Corporate Data Quality. Springer, Berlin, 2016.

[Pe12] Peffers, K. et al.: Design Science Research Evaluation. In (Eds. Peffers, K.; Rothenberger, M.; Kuechler, B.): Design science research in information systems. Springer, Berlin, 2012; pp. 398–410.

[Ru16] Russell, S. J. et al.: Artificial intelligence. A modern approach. Pearson, Boston, 2016.

[SS20] Sindhu Meena, K.; Suriya, S.: A Survey on Supervised and Unsupervised Learning Techniques. In (Eds. Kumar, L. A.; Jayashree, L. S.; Manimegalai, R.): AISGSC. Springer, Cham, 2020; pp. 627–644.

[STS16] Singh, A.; Thakur, N.; Sharma, A.: A review of supervised machine learning algorithms: 2016 3rd INDIACom, 2016; pp. 1310–1315.

[To99] Tozer, G. V.: Metadata management for information control and business success. Artech House, Boston, 1999.

[We20a] Ed. Weber, F.: Künstliche Intelligenz für Business Analytics. Springer Fachmedien, Wiesbaden, 2020.

[We20b] Ed. Wennker, P.: Künstliche Intelligenz in der Praxis. Springer Fachmedien Wiesbaden, Wiesbaden, 2020.

[WW02] Webster, J.; Watson, R. T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. In MIS Quarterly, 2002, 26; pp. xiii–xxiii.