Empowering Data Consumers to Work with Data: Data Documentation for the Enterprise Context

Clément Labadie[1], Markus Eurich[1], Christine Legner[1]

Faculty of Business and Economics (HEC), University of Lausanne, Switzerland

{clement.labadie, markus.eurich, christine.legner}@unil.ch

Abstract. Enterprises that are engaging in digital transformation need to empower an increasing number of data consumers (sometimes referred to as "data citizens") to work with data. A prerequisite is data documentation – data assets should be inventoried and well-described to facilitate data selection by non-data experts, who need to both find and understand them. This research paper proposes a reference model for data documentation in the enterprise context. It was developed in collaboration with 25 large enterprises, following a Design Science Research process. Compared to existing metadata standards that contain flat lists of metadata attributes, the reference model organizes metadata objects in logical and physical layers and features views dedicated to usage and governance contexts. It thereby improves maintenance and consistency in data documentation, when dealing with hundreds of interdependent data resources, and allows to express inherent relationships between metadata attributes.

Keywords: metadata, data documentation, reference model, design science research.

1 Introduction

"Data consumers don't speak physical and logical. They speak business [1]." This statement echoes the larger, well-documented phenomenon of the digital transformation, prompting enterprises to integrate digital resources as part of their value creation process [2] and rethink their strategy accordingly [3]. With the digital transformation, data has become a critical asset for enterprises [4–6]. In research, this trend has found an echo in the growing body of knowledge on Data-Driven Business Models (DDBMs) [7], [8], which are "designed to create additional business value by extracting, refining and ultimately capitalizing on data [8]".

Enterprises that are engaging in digital transformation need to empower an increasing number of data consumers to work with data. These data consumers

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_m7-labadie

include traditional data experts (e.g., business intelligence specialists, data managers or data architects), but increasingly comprise employees who start employing self-service business intelligence (sometimes referred to as "data citizens") or more advanced data science tools. In order to enable their activities, data that is spread in enterprise systems need to be discoverable for humans and machines alike. These challenges, i.e., extracting data from various sources and refining them to make them ready-for-use have also be conceptualized through the FAIR principles, according to which data should be Findable, Accessible, Interoperable and Reusable [9].

A prerequisite for making data FAIR is data documentation – data assets should be inventoried and well-described to facilitate data use by non-data experts, who need to both find and understand them. Metadata, i.e., data about data [10], [11], is a natural candidate when it comes to documenting data. It is a long-standing topic that features a large number of standards, such as Dublin Core (DC) [12] and the Data Catalog Vocabulary (DCAT) [13]. While these standards emphasize data discovery, they are unable to accurately address the specific requirements of the enterprise context, which brings added layers of complexity, both in terms of systems (e.g., highly distributed and siloed applications) and organization (e.g., governance, roles and responsibilities).

Conversely, our study aims at answering the following research question: how to organize data documentation to support data discovery and data use in the enterprise context? To that end, following a Design Science Research (DSR) process, we developed a metadata model in close collaboration with 25 multi-national enterprises with varying levels of experience in setting up enterprise-wide data documentation, using metadata and data catalog tools. As a reference model, this model is intended to serve as blueprint for enterprises seeking to design their own, company-specific metadata model in support of providing data documentation for the increasing number of data consumers (i.e., both data experts and data citizens). Compared to existing metadata standards that contain flat lists of attributes, the suggested metadata model groups objects in logical and physical layers and features views dedicated to usage and governance contexts.

The remainder of this paper is structured as follows. We start with an analysis of metadata-related research and standards, and by further specifying the research gap. We then detail our research methodology and related steps that went into the model's development. We continue by presenting the reference model for data documentation, its components. We conclude by presenting the model's demonstration and discussing its contribution.

2 Background and Motivation

2.1 Metadata Definition and Categories

To this day, data documentation is most often associated with metadata, that are commonly defined as "data about data" [10], and "aim at facilitating access, management and sharing of large sets of structured and/or unstructured data [14]". Various initiatives have attempted to describe them for specific applications, e.g.,

https://doi.org/10.30844/wi_2020_m7-labadie

earth sciences or multimedia systems [14]. Although these initiatives make context-dependent suggestions for organizing metadata, general sub-categories can be identified [11], [15], [16]:

- **Structural** metadata describes the general data model, e.g., type, attributes of objects and relationships between objects.
- **Administrative** metadata provides information to help manage a resource, e.g., users (with rights) and dates (creation, last update).
- **Terminological** metadata provides an understanding on the data, e.g., definitions, abbreviations, cataloging records and comments from creators and users.
- **Governance** metadata provides an overview of the data landscape from a management point of view, e.g., ownership, roles, responsibilities and level of confidentiality [10].
- **Context** metadata provides information on the environment in which the data exists, e.g., business processes and business purposes (use cases).
- **Use** metadata provides information on how data are consumed, e.g., search logs, usage statistics, processing systems.

The literature highlights that metadata has found a variety of technical applications over the years [17], starting from the bibliography and library domains [16]. Yet, the ones that relate to the enterprise context tend to focus on technical descriptions [17]. However, in an enterprise context, metadata should exceed technical aspects and describe business aspects of data, such as their use in business processes [18].

2.2 Metadata Standards

As mentioned above, metadata is often discussed in specific contexts, leading to a wide variety of metadata standards. In order to acquire a broad overview, we performed a wide-ranging review of existing standards from academic and non-academic sources, resulting in a list of 129 standards. Resources such as the Research Data Alliance's directory of metadata standards [19] were used as a starting point and the list was enriched with standards highlighted in academia [11], [16], [20–23]. To also embrace non-academic sources, we used search string "metadata standard" on Google Search and Google Scholar. The identified metadata standards cover a vast array of topics ranging from archiving (e.g., Open Archival Information System - OAIS), bibliography (e.g., Library of Congress Classification and Subject Headings – LCC, LCSH), cultural material (e.g., Categories for the Description of Works of Art – CDWA), multimedia content (e.g., Synchronized Multimedia Integration Language – SMIL) and geographical data (e.g., ISO 19115). These standards address different aspects of metadata, and they can be classified in the following types [24]:

- **Data structure** standards define metadata element sets and schemas, which are categories (or containers) of metadata. They define a predefined set of attributes meant to describe objects pertaining to the domain of interest. *Example: Library of Congress Classification – LCC.*

https://doi.org/10.30844/wi_2020_m7-labadie

- **Data value** standards are controlled vocabularies, which define terms, names and other values used to populate metadata elements, i.e., they restrict the value domain to fill metadata structures. *Example: Library of Congress d Subject Headings – LCSH.*
- **Data content** standards define cataloging rules and codes, which provide guidelines for the format and syntax of the values used to populate metadata elements. *Example: Synchronized Multimedia Integration Language – SMIL.*
- **Data format / technical interchange** standards define the encoding in machine-readable form of metadata elements, i.e., a manifestation of a particular data standard, encoded and marked up for machine processing. *Example: Extensible Markup Language – XML.*

Out of the 129 identified standards, we excluded all domain-specific standards (e.g., related to bibliography/libraries, archiving/preservation, finance, cultural material, natural sciences, geography, medicine, climate, museum curation, government, education, agriculture, astronomy), which resulted in 18 remaining standards. We then further refined our selection by excluding standards solely specifying data formats or technical interchange and encoding schemes. As a result, 4 standards were retained as relevant for the enterprise context and data catalogs (s. Table 1 below): the Dublin Core Schema (DC) [12], the Data Catalog Vocabulary (DCAT) [13], the Common Warehouse Metamodel (CWM) [25] and the ISO 11179-3 Metadata Registry Metamodel and Basic Attributes (MDR) [26].

Dublin Core presents a flat list of terms (or attributes), comprising, e.g., creator, description, date, identifier, relation and rights. DCAT and MDR go a step further by grouping attributes and specifying relationships between groups in a metamodel, which focuses on the internal logic between the concepts they introduce, but do not integrate external concepts. Finally, CWM provides a metadata interchange standard that enables XML-based exchange of business intelligence and data warehouse metadata between different tools, platforms and repositories.

In terms of metadata categories (s. Section 2.1 above), all comprise structural, administrative and terminological metadata. DCAT and CWM additionally include context metadata, but none of them specifically cover governance and use metadata, which are critical in an enterprise context.

Table 1. Overview of selected metadata standards

Name	*Source*	*Metadata Standard Type*	*Domain*
Dublin Core (DC)	Working group (DCMI)	Structure	General
Data Catalog Vocabulary (DCAT)	Standards organization (W3C)	Structure, Format	Data catalogs
Common Warehouse Metamodel (CWM)	Industry consortium (OMG)	Structure, Format	Business intelligence
Metadata Registry Metamodel and Basic Attributes (MDR)	Standards organization	Structure	General

https://doi.org/10.30844/wi_2020_m7-labadie

	(ISO/IEC)		

2.3 Research Gap

Although available standards may constitute adequate starting points, DC and MDR focus on metadata for describing single data resources or datasets but are not entirely suitable for the enterprise context. In fact, DC and MDR provide fundamental metadata attributes for describing data resources, but they are presented as a flat list. This makes it difficult to maintain them, when dealing with hundreds of data resources, and to express inherent relationships between metadata attributes. In that regard, it is not surprising that DC has been extended in further domain-specific applications[2]. On the other hand, standards such as CWM focus on a specific aspect of metadata management (in CWM's case, business intelligence), but do not cover the overall enterprise context and specifically the business processes that create and consume data. DCAT goes further and introduces metadata objects with relationships but adopts a data publication perspective, as it is meant to link data catalogs available on the web. While all of them provide useful building blocks for metadata management, they lack an integrating, enterprise-driven framework.

3 Research Process

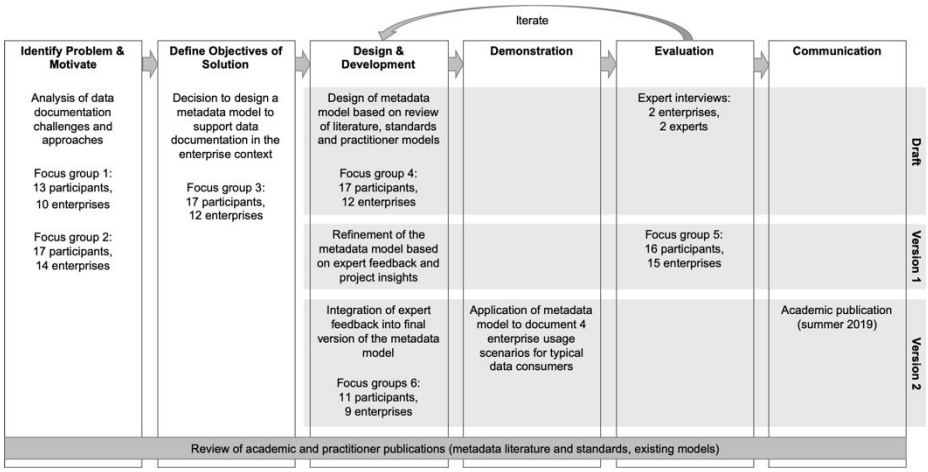

Figure 1. Research process

To provide a reference framework for data documentation in the enterprise context, we developed a metadata model following the iterative DSR process suggested by

[2] For instance, the Visual Resource Association Core Categories (VRA Core) and the Australian Government Locator Service (AGLS) both rely on DC, extending it for cultural material and government domains, respectively.

https://doi.org/10.30844/wi_2020_m7-labadie

Peffers et al. [27]. Figure 1 depicts all research steps and highlights the interaction between the research team and practitioners, which took place in the form of 6 focus group meetings with representatives from 25 large enterprises, as well as 2 expert interviews. Participants in the focus groups had extensive experience with data documentation and were involved in metadata and data catalog initiatives to support data discovery and use in their enterprises. Focus group meetings were part of a collaborative effort between researchers and a stable core team of practitioners over the course of more than 12 months. This collaboration included reviews of metadata standards and analysis of existing practitioner models and led to the development of the metadata model.

During the problem identification phase, we analyzed the issues faced by practitioners in democratizing data within their enterprise and documenting data for various purposes. In focus group 1, we identified and validated typical data consumers, i.e., for which enterprise stakeholders data should be documented. In focus group 2, we outlined data usage of these typical consumers, as well as the type of data documentation to support them. These discussions lead to the conclusion that there is a need for a core understanding of fundamental data objects to be documented, as well as several extensions, describing more specific data objects that may not apply to all business contexts (either company-specific or domain-specific).

In the objective definition phase, the decision to design a metadata model was made (focus group 3). This model is meant to represent the core model for enterprise data documentation, i.e., the recommended minimum approach, that addresses the requirements of different types of data consumers. The core metadata model is meant to serve as a blueprint for enterprise-wide data documentation, for instance in the context of data catalog initiatives.

The subsequent phases consisted of iterative design cycles, each comprising design and evaluation steps, with the last cycle also featuring a demonstration of the finalized model. The first design iteration incorporated insights from our review of metadata standards (specifically Dublin Core and DCAT), as well as existing models from practice at different level of maturity, i.e., a completed model and two models in development, as depicted in Table 2. It resulted in a draft model as alpha artifact that was discussed in a broader group during focus group 4. It was also evaluated through expert interviews from Company B and C which were in the process of developing their approach to data documentation. We also reviewed data documentation models from both enterprises and performed a mapping with our draft version.

The second design iteration led to the first stable version of the metadata model (version 1). We consolidated feedback from focus group 4 and both expert interviews, and incorporated insights from two additional practitioner models (s. Table 2). Version 1 was discussed during focus group 5, and it was decided to improve terminological clarity and put an emphasis on establishing shared metadata object definitions in the next iteration.

The third design iteration led to the final version of the metadata model (version 2). In this step, we reworked the model according to participant feedback and refined the definition of metadata objects, which were extensively discussed during focus group

https://doi.org/10.30844/wi_2020_m7-labadie

6. Once all metadata objects were agreed upon, we demonstrated the metadata model by applying it to document 4 enterprise usage scenarios.

Table 2. Input models for artefact development

Source	*Type*	*Level of maturity*	*Context / industry*	*Version (design iteration)*
DCAT	Standard (W3C)	Established	Facilitating interoperability between data catalogs published on the web	Draft (first iteration)
Dublin Core	Standard (DCMI)	Established	Digital resources	Draft (first iteration)
Company A	Practitioner	Finalized	Pharmaceuticals	Draft (first iteration)
Company B	Practitioner	In development	Packaging	Draft (first iteration)
Company C	Practitioner	In development	IT	Draft (first iteration)
Company D	Practitioner	In development	Manufacturing	Version 1 (second iteration)
Company E	Practitioner	In development	Automation	Version 1 (second iteration)

4 A Metadata Model for Enterprise Data Documentation

4.1 Model Objectives

As shown in section 2, existing approaches to data documentation do not fully address the specific requirements and the complexity of the enterprise-wide context. Out of the relevant standards identified, DC and MDR focus on metadata describing single data resources or datasets. On the other hand, DCAT and CWM do address a multiplicity of data sources but focus on the online distribution of datasets and on data interchange between warehouse and business intelligence platforms, respectively. The enterprise context is characterized by complex data landscapes, i.e., large numbers of interdependent data resources stored in enterprise systems, that underlie strong data governance in order to ensure data quality and comply with internal and external standards and regulations. In the course of digitalization, enterprises need to enable an increasing number and variety of data consumers, with different data needs and degrees of data literacy, to access and use these data resources.

https://doi.org/10.30844/wi_2020_m7-labadie

Conversely, our proposed model is designed as a reference framework for data documentation in an enterprise context, and addresses the following requirements that were identified in the focus group sessions: First, it should support data democratization and provide data documentation for typical data consumers (both experts and non-experts, e.g., data citizens, data analysts, data protection officers, data architects, data stewards, data owners). Second, it should align the different perspectives on data, specifically the business-oriented and the system-oriented perspectives. Third, it was found that for using data in an enterprise context, data consumers need to also understand responsibilities and relevant regulations for data. Lastly, the reference model should identify and structure metadata attributes in a "core model" for data documentation in an enterprise context, i.e. act as a baseline of minimum requirements for all enterprises, regardless of their area of business. The core model could then be extended to include documentation aspects dedicated to specific enterprise functions, such as metadata specific to research and development, for instance.

The suggested model should be able to support data catalog implementation and use cases, which we demonstrate in section 4.4 below. As for implementation, we recommend a use-case driven approach, where enterprises would select and prioritize the metadata objects and attributes that need to be implemented for a specific business purpose.

4.2 Model Structure

The reference model for enterprise data documentation is expressed in form of a metadata model [14], comprising relevant metadata objects that are to be documented as well as their relationships. This allows for distinguishing different perspectives on data and for defining relationships between data and other relevant objects in the enterprise context. Defining data documentation based on a metadata model also eases maintenance and improves consistency of data documentation. The metadata model, which is depicted by Figure 2, is organized in different views, starting with the logical and physical view on the data at the center.

https://doi.org/10.30844/wi_2020_m7-labadie

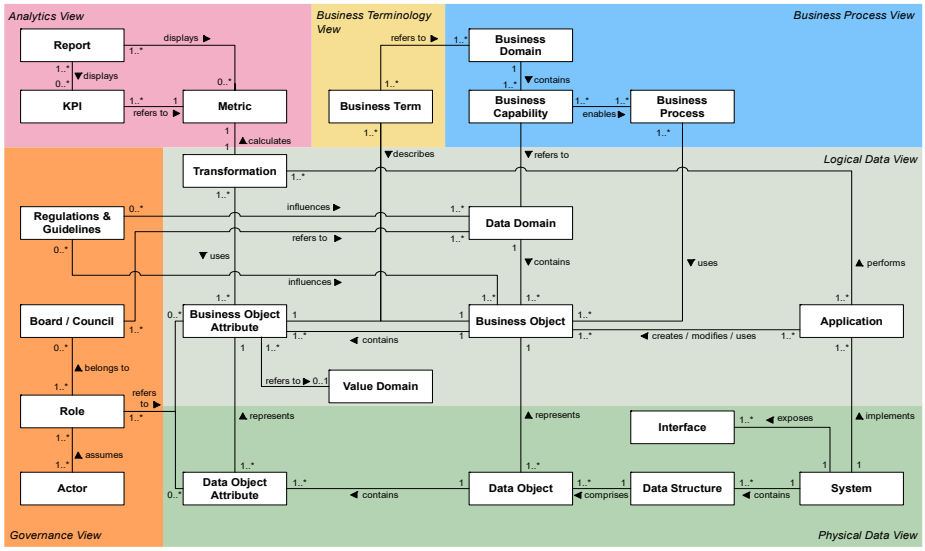

Figure 2. Metadata model overview

The **logical view** represents the conceptual or business view on data, whereas the **physical view** represents the implementation perspective and makes the link with the way data is implemented in systems. The other views are specific to the enterprise context. They comprise the different **usage contexts** that depict where and how data is created and used in the enterprise (i.e., business process, analytics, and the related business terminology). The **governance view** depicts the relevant regulations, guidelines and responsibilities for data. Each view comprises several metadata objects, which we will define in the next section.

4.3 Metadata Objects

The central element[3] of the model is the *business object*, which describes a reoccurring set of information used in multiple business contexts and minimum one data domain. It can be either created, used, or changed in *business processes* [28] and analyzed in reports that aggregate *metrics* and *KPIs*. A business object is specified by *business object attributes* that are characteristics of the *business object* and can contain either free text or a restricted set of values (*value domain*). For instance, in an enterprise context, a *business object* can be a record of a supplier – *attributes* could comprise, e.g., name, street, zip code, country, VAT number, and the country attribute's restricted value domain could consist in a list of ISO country codes. Additionally, *business objects* belong to a *data domain*, which specifies their context – in our example, a supplier is part of the business partner domain. Finally, *business objects* can be created, used or changed by *applications*.

[3] In this section, the term "element" is used in substitute for the term "metadata object".

https://doi.org/10.30844/wi_2020_m7-labadie

Business object and *business object attribute* are both reflected on the physical view, representing their specific implementation in the respective *systems*, with the *data object* and the *data object attribute*, respectively. They are projected into a *data structure*, which unifies *data objects* and *data objects attributes* into a single, distinct format, stored in a *system*. Depending on the database paradigm, the structure can represent, e.g., a table (relational database), a class (object-oriented database) or a key value store (graph database). *Systems* are the physical counterpart of *applications* and expose interfaces, which are meant to transfer data to other *systems*.

Moving onto the usage context, *business objects* are used within *business processes*, either as input or output. *Processes* represent how an enterprise performs its activities, and are enabled by *business capabilities*, which consist in a combination of technological, informational, and organizational resources, and representing what a company does. The business domain represents strategic business areas of an enterprise and reflect is strategic goals. In our example (i.e., supplier as *business object*), procurement is the business domain. It features business capabilities such as a global sourcing capability consisting, e.g., of an ERP system and/or supplier relationship management system (technology), a purchasing team (organization) – it is realized through e.g., strategic sourcing and procure-to-pay *business processes*, which make use of supplier (business partner *data domain*) and product (material *data domain*) *business objects* (information).

Business domains contain several, related *business objects* and *business object attributes*. In addition, *business terms* specify synonyms or alternative expressions for business objects and attributes. Since they are domain-specific, *business terms* are necessarily linked with at least one *business domain*. For instance, the term “debtor” may be used within procurement and financial *business domains* as synonym for "supplier".

On the logical data view, the *transformation* is the gateway to the analytics view. It queries data from one or several *business object attributes* to produce a *metric*, which is a quantifiable measure reflecting the state of the enterprise; in our example, the number of defective parts received from suppliers. *Metrics* are the input for *key performance indicators* (KPIs), which evaluate the success of the enterprise at specific activities, and show the degree of fulfillment of a *metric*, with regards to a stated objective. A *KPI* expresses this number in terms of e.g., percentages, setting a threshold stating that a deficiency rate higher than e.g., 5% is not acceptable. Finally, reports organize and present metrics and / or KPIs in human-readable form, enabling visualization by different dimensions. In our example, a *report* can then display this information by several dimensions, e.g., supplier name, supplier location, date and / or material category.

The governance view features organizational and regulation aspects. On the organizational side, the *actor* assumes certain data-related *roles*. According to their *role* assignment, *actors* refer to designated *business object attributes* and *data object attributes* in the performance of their tasks and responsibilities. For instance, the *actor* Jane Doe bears the *role* of lead strategic buyer – in this context, she is responsible for overseeing negotiations with suppliers, and thus interacts with *attributes* such as product specifications, name and reference price. Several *roles* can be involved in

https://doi.org/10.30844/wi_2020_m7-labadie

boards and councils, which designate working groups bearing advisory and / or decision-making power, with regards to one or more *data domain*. In our example, Jane Doe may be part of a data stewardship council setting guidelines for the procurement *data domain*.

The governance view also contains an element depicting *regulations and guidelines*. It designates any guideline, or set of guidelines, that constrain the structure and / or behavior of an enterprise [29]. It can refer to legal texts (e.g., the General Data Protection Regulation – GDPR), contracts (which may specify binding service level agreements) and standardization documents (e.g., use of standardized customs codes in shipping documents) [29], among others. Requirements can apply to either entire data domains (e.g., GDPR applies to all data domains containing personally identifiable information, such as the business partner data domain) or specific business objects (e.g., product safety regulations apply solely to the materials data domain).

In the following section, we will demonstrate how these metadata object come into play when realizing selected use cases.

4.4 Demonstration

In order to demonstrate the metadata model, we opted for a use-case driven approach. During focus group 5, we gathered usage scenarios from 8 companies representing 8 industries. They depict how typical data consumer groups (e.g., data analysts, data stewards, data owners, data protection officers, data architects and data citizen) may benefit from data documentation that is organized according to the metadata model. The demonstration's purpose is to show how our proposed model may empower data consumers to find and use appropriate data, suited to their needs and issues. Based on practitioner input from the focus groups, we have retained 4 key usage scenarios (s. Table 3): 1) selecting appropriate data for analysis, 2) understanding governance impact of planned data-related changes, 3) investigating data-related business process failure, and 4) assessing data protection requirements.

The first use case is set in the context of business intelligence and analytics, in an enterprise that transforms from traditional manufacturing towards producing connected, portable appliances. A data analyst wishes to understand possible causes of battery drain, and needs to extract battery-related information, as well as data from the appliance's sensors (e.g., location, temperature). Without proper data documentation, the data analyst would have to take guesses as to where (i.e., in which *system*) to find appropriate information (*data object* and *attributes*). Additionally, the way *data objects* and *attributes* are defined and phrased in the system may not be transparent. Using data documentation relying on our metadata model, the analyst could browse definitions from *business terms* in order to isolate *business (and data) objects* containing the information they are looking for, and trace it back to designated *applications and systems*, eventually enabling them to select the relevant data objects and attributes for the analysis or check existing *reports* and the way *metrics* and *KPIs* are calculated.

https://doi.org/10.30844/wi_2020_m7-labadie

In the second use case, the emphasis is placed on the governance context. Here, a data steward plans to make changes to an *attribute*, e.g., by restricting the *value domain* in order to remedy data quality issues. This will be reflected in an update of the internal *guidelines*. Before implementing any change, the data steward needs to understand who will be impacted by planned changes, in order to involve affected stakeholders. This is made possible by the metadata model, as it connects *roles* with the *business object attributes* and *data object attributes* they refer to. Additionally, the data steward could reach out to any *board / council* in charge, by tracing the *attribute* back to its parent *data domain*.

The third use case, centered around investigating business process failures, follows a similar rationale of impact-analysis, this time with regards to the business process context. Here, a process owner witnesses errors in business process, and wishes to investigate its root cause. Starting from the *business process* in question, the metadata model enables him or her to discover *business objects* and related *attributes* used, as well as the processing *application*. The process owner can also find out how business objects and attributes are implemented in different systems, with their respective *data objects* and *attributes*. In order to identify and remedy data defects, process owners could also apply data quality *metrics* or *KPIs*, and identify *roles* and *actors* responsible for defective *objects* or *attributes* to prompt rectifying measures.

The fourth use case focuses on regulatory issues, specifically on data protection, e.g., in the context of GDPR. Here, a data protection officer needs to compile a list of where personally identifiable data is stored, and how it is used, as well as document compliance requirements. The latter can be achieved through specification of the *regulations & guidelines* metadata object, e.g., by documenting business rules and policies. In order to identify affected *business objects* and *business object attributes*, *business terms* (e.g., specifying personal data), *data domains* (e.g., business partner inherently contains personal data) and *business processes* (e.g., customer account management processes necessarily deal with personal information) can be a starting point. Furthermore, thanks to the logical – physical *attribute* inheritance built into the metadata model, a link can be made to the specific *systems* processing the data of interest. In the context of data protection, this is crucial for identifying processing systems located in non-EU countries, to which additional GDPR requirements apply.

Table 3. Summary of usage scenarios

Designation	*Role*	*Purpose*	*Metadata objects*
Selecting data for reports	Data analyst, data citizen	- Find metadata definitions - Locate relevant business objects - Generate reports	Business term, Business object, Business object attribute, Application, Transformation, Report
Understanding impact of data changes	Data steward	- Find stakeholders impacted by planned data change - Identify relationship between business objects	Data domain, Business object attribute, Data object attribute, Value domain, Role, Actor, Board / Council
Investigation	Process	- Identify faulty data	Business process, business

https://doi.org/10.30844/wi_2020_m7-labadie

process failures	owner	- Business process drill-down - Data quality reporting	object, Business object attribute, Data object attribute, Application, Role, Actor, Metric, Key Performance Indicator
Assessing data protection requirements	Data protection officer	- Locate compliance sensitive data - Document policies and business rules - Identify international data transmissions	Regulations & Guidelines, Data domain, Business term, Business process, Business object, Business object attribute, Data object attribute, System,

5 Discussion and Outlook

In the course of the digital transformation, democratizing and generalizing the use of data is on top of the management agenda in many organizations that set out to enhance business processes, inform business decisions and implement data-driven business models. In this context, this research presents enterprises with a reference model meant to act as the foundation of data documentation and support the implementation of tools to empower data consumers to work with data. The suggested metadata model reflects the enterprise context and provides a business-oriented view on data. Compared to existing standards such as DC and MDR, which consist of flat lists of metadata attributes, it provides enterprises with a structure for organizing metadata objects and their relationships. In contrast to CWM, our model has not primarily been developed to standardize the interorganizational exchange of metadata information. Instead, its major focus is on providing relevant context information around data. While the available models may support organizations in finding (i.e., data discovery) data, the suggested metadata model introduces logical and physical views on data and comprises usage and governance contexts. Specifically, it enriches existing, generic models by addressing enterprise-specific contextual aspects, i.e., business processes, business glossary, business intelligence and analytics. It also integrates the governance perspective and suggests metadata objects representing both organizational roles and relevant guidelines and regulations. Our research thereby contributes to providing a holistic perspective on data in the enterprise context and links metadata concepts to enterprise (data) architecture literature.

As previously stated, the model has been assessed and partially tested in selected cases, e.g., by performing a mapping with existing, company-specific enterprise metadata models. Moreover, it has been tested for four concrete usage scenarios. While these steps provide evidence for the validity, consistency and completeness of the model, the next research steps include a broader evaluation of the final model, both artificial (e.g., questionnaire-based) and naturalistic (e.g., implementation of the metadata model in a corporate setting).

https://doi.org/10.30844/wi_2020_m7-labadie

The metadata model presented in this paper constitutes a core model representing fundamental data documentation concepts, and thereby is also meant to provide a foundation for future research. Most importantly, we are interested in understanding how data documentation - and its publication in data catalogs - impacts data citizens' satisfaction, productivity and quality of work. Interesting avenues for future research relate to the design of extensions relying on the core, providing metadata objects and attributes more specifically catered to specific business areas (e.g., research and development, finance). Future research could also address overlaps of the presented metadata model with enterprise architecture repositories. Although we have observed these overlaps within our focus group activities, we have not yet elaborated on them in more detail. However, we have witnessed that in practice, data catalog tools can be populated with information maintained in enterprise architecture management tools, i.e. applications, infrastructure (incl. interfaces) and processes.

In addition, while unstructured data (e.g., video clips, audio files) was not prominent within our focus group discussions, we appreciate the growing need to include this aspect in future iterations of the metadata model.

Acknowledgements

This research was supported by the Competence Center Corporate Data Quality (CC CDQ). The authors would like to thank the experts that contributed to this research.

References

1. Goetz, M., Leganza, G., Hoberman, E., Vale, J.: STIR Your Data For Context. Analyst Report, Forrester Research (2018)
2. Bharadwaj, A., El Sawy, O.A., Pavlou, P.A., Venkatraman, N.: Digital Business Strategy: Toward a Next Generation of Insights. MIS Quart. 37, 471–482 (2013)
3. Matt, C., Hess, T., Benlian, A.: Digital Transformation Strategies. Bus. Inf. Syst. Eng. 57, 339–343 (2015)
4. Pentek, T., Legner, C., Otto, B.: Towards a Reference Model for Data Management in the Digital Economy. In: Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology (DESRIST), Karlsruhe (2017)
5. Mohr, N., Hürtgen, H.: Achieving Business Impact with Data. Analyst Report, Digital McKinsey (2018)
6. Belissent, J., Leganza, G., Vale, J.: Determine Your Data's Worth: Data Plus Use Equals Value. Analyst Report, Forrester Research (2019)
7. Schüritz, R., Seebacher, S., Dorner, R.: Capturing Value from Data: Revenue Models for Data-Driven Services. In: Proceedings of the 50th Hawaii International Conference on System Sciences, pp. 5348–5357, Waikoloa Village (2017)
8. Brownlow, J., Zaki, M., Neely, A., Urmetzer, F.: Data-Driven Business Models: A Blueprint for Innovation. Research Report, Cambridge Service Alliance (2015)
9. Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016)

https://doi.org/10.30844/wi_2020_m7-labadie

10. Roszkiewicz, R.: Enterprise metadata management: How consolidation simplifies control. J. Digit. Asset Manag. 6, 291–297 (2010)
11. Inmon, W.H., O'Neil, B., Fryman, L.: Business Metadata: Capturing Enterprise Knowledge. Morgan Kaufmann, Burlington (2010)
12. Dublin Core Metadata Initiative: DCMI: DCMI Metadata Terms, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (Accessed: 25.08.2019)
13. World Wide Web Consortium (W3C): Data Catalog Vocabulary (DCAT), https://www.w3.org/TR/vocab-dcat/ (Accessed: 25.08.2019)
14. Kerhervé, B., Gerbé, O.: Models for Metadata or Metamodels for Data? In: Proceedings of the 2nd IEEE Metadata Conference, Silver Spring (1997)
15. Marco, D.: Building and Managing the Meta Data Repository - A Full Lifecycle Guide. John Wiley & Sons, Hoboken (2000)
16. Hillmann, D.I., Marker, R., Brady, C.: Metadata Standards and Applications. Ser. Libr. 54, 7–21 (2008)
17. Sen, A.: Metadata Management: Past, Present and Future. Decis. Sup. Syst. 37, 151–173 (2004)
18. Burnett, K., Ng, K.B., Park, S.: A Comparison of the Two Traditions of Metadata Development. J. Am. Soc. Inf. Sci. 50, 1209–1217 (1999)
19. Chen, S., Alderete, K.A., Ball, A.: RDA Metadata Standards Directory, https://rd-alliance.github.io/metadata-directory/standards/ (Accessed: 19.11.2019)
20. Ferguson, S., Hebels, R.: Access to information resources. In: Ferguson, S. and Hebels, R. (eds.) Computers for Librarians (Third Edition), pp. 81–109. Centre for Information Studies, Wagga Wagga (2003)
21. Clobridge, A.: Metadata. In: Clobridge, A. (ed.) Building a Digital Repository Program with Limited Resources, pp. 85–109. Chandos Publishing, Oxford (2010)
22. Vetterli, T., Vaduva, A., Staudt, M.: Metadata Standards for Data Warehousing: Open Information Model vs. Common Warehouse Metadata. SIGMOD Rec. 29, 68–75 (2000)
23. Xiao, B., Zhang, C., Mao, Y., Qian, G.: Review and exploration of metadata management in data warehouse. In: 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), pp. 928–933 (2015)
24. Baca, M.: Introduction to Metadata. The Getty Research Institute, Los Angeles (2016)
25. Poole, J., Chang, D., Tolbert, D., Mellor, D.: Common Warehouse Metamodel. An Introduction to the Standard for Data Warehouse Integration. John Wiley & Sons, New York (2002)
26. International Organization for Standards / International Electrotechnical Commission (ISO/IEC): International Standard ISO/IEC 11179-3. Information Technology - Metadata Registries (MDR) - Part 3: Registry Metamodel and Basic Attributes (2013)
27. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. J. Manag. Inf. Syst. 24, 45–77 (2007)
28. Martin, J.: Computer Data-base Organization. Prentice Hall, Engelwood Cliffs (1977)
29. El Kharbili, M.: Business Process Regulatory Compliance Management Solution Frameworks: A Comparative Evaluation. In: Proceedings of the Eighth Asia-Pacific Conference on Conceptual Modelling, vol. 130, pp. 23–32. Australian Computer Society, Darlinghurst (2012)

https://doi.org/10.30844/wi_2020_m7-labadie