The Significant Role of Metadata for Data Marketplaces

Sebastian Lawrenz
Institute for Software and Systems Engineering, Clausthal University of Technology
sebastian.lawrenz@tu-clausthal.de

Priyanka Sharma
Institute for Software and Systems Engineering, Clausthal University of Technology
priyanka.sharma@tu-clausthal.de

Andreas Rausch
Institute for Software and Systems Engineering, Clausthal University of Technology
andreas.rausch@tu-clausthal.de

Abstract

With the shift to a data-driven society, data trading takes on a completely new significance. In the future, data marketplaces will be equivalent to other electronic commerce platforms such as Amazon or eBay. Just like any other online marketplace a data marketplace is a platform that enables convenient buying and selling of products- in this case "data"

Metadata is data about data. Metadata plays a significant role in data trading, as it serves as an orientation for all involved parties in the data marketplace. A seller who wants to sell their data on the marketplace needs metadata to describe the selling offer, and the buyer can use it to search and identify relevant data.

This paper outlines the significance of metadata in data trading on a data marketplace and classifies the levels of metadata. Moreover, in data trading metadata has also a significant role in determining the data quality. In this paper we also discuss the role of metadata in terms of data quality.

Keywords: metadata; data trading; data marketplaces; data quality

1. Motivation and Introduction

In the last decade, data has become more important than ever. An oft-quoted phrase is: "Data are the oil of the 21. Century". This can easily be proven by a looking at the most valuable companies in the world today such as Google, Apple, Microsoft, Facebook and Tencent ("Wertvollste Unternehmen nach Markenwert weltweit 2018 | Ranking," n.d.). All of them have businesses related to data, or their core business model is based on data (see Google, Facebook and Tencent).

Moreover, more and more companies are generating and collecting data, which is nowadays quite easy. But not all of them are able to use this data to its full potential. On the other side, experts for Data Science and Artificial intelligence are growing up and companies are being found, which specialize in this field. But they need data and must buy them. For this new fields of business emerges. These are for example free Data Sharing platforms like Kaggle[1], where users can download and upload data sets for free, or commercial data marketplaces like the iota date marketplace[2].

Data as a commodity sold online, seems similar to any other product sold online. But data is quite different than other products sold online. Although there are many differences, we identify

[1] Cf. https://www.kaggle.com

[2] Cf. https://data.iota.org/

the following as the two biggest differences between other products and data- Product description and product quality.

1. **Product description**- When any other products is sold online except for the description additional details such as pictures etc. can be added. This helps the buyer's a lot in the decision-making process. But for data these additional details cannot be provided hence the description about the data one of the most important factor affecting the buyer's decision.
2. **Product quality**- Most of the products sold online except for digital products such as movies, songs etc. can be returned if the customer is not satisfied with the product. But the same cannot be done with data. As once the data is seen by the buyer it loses its value and the buyer might make copy of it. Thus, return policies for data is not feasible.

To give all the stakeholders of a data marketplace an overview about the content and offers, metadata are essential. Metadata is data, which provides information about other data, or in more simple terms, data about data. Metadata is necessary to identify relevant data in a data marketplace, to help the seller to create an offer, to give the buyer a first overview about the data set and much more.

The main goal of this paper is to show the significance of metadata in data trading on a data marketplace and classify the levels of metadata. The rest of the paper is structured as followed: Section 2 gives an overview about data, metadata and data marketplaces. In Section 3 the significance of metadata for data trading on a data marketplace is presented. Section 4 shows the identified challenges. And finally, Section 5 concludes and gives insights in future work.

2. Background

This section deals with the background and the actual state of the art.

A brief overview about data marketplaces, data and metadata is presented in this section.

2.1. Data Marketplaces

More data marketplaces are emerging, since Data-driven Applications and Artificial Intelligence, becomes more and more popular. Depending on the user requirements, the types of the data being traded can be grouped as follows:

1. **Real time data (RT)**: The buyer needs just the actual data and not the whole history of the data set. An example for this is a navigation application, where the buyer wants to show the actual traffic on the road.
2. **Non-real time data (NRT)**: The buyer does not need the actual data in real time, but he is interested in a whole data set. An example for this need could be a predictive analytics algorithm.

Due to these different requirements, completely different marketplace requirements arise. For the first type, a mechanism to securely subscribe the data in nearly real time is required, whereas for the second type a secure data exchange platform is necessary. Metadata are essential for both kinds of data marketplaces, but in different types and levels. Since the focus of our research is data trading with Non-real time (NRT) data, we will focus on it in the rest of this paper. Furthermore, in this paper we discuss the role of the metadata in detail. More information and challenges in data trading in general can be found in this paper: (Lawrenz, Sharma, & Rausch, 2019)

2.2 Data and Metadata

The term metadata literally means "data about data" (Baca, 2016) and is combination from the Greek word meta, which means beyond or after, and data. Belonged to this, we also could

understand it, as **post** data. Since there is no uniform definition so far, this section describes the difference between data and metadata.

Before exploring the profundity of metadata, first we need a common understanding of data. For this introduction we will use **D**ata **I**nformation **K**nowledge **W**isdom (DIKW) hierarchy also known as "knowledge pyramid" (Rowley, 2007). Even is the original origin of the DIKW Pyramid is still uncertain (Wallace, 2007), an often quoted article for this hierarchy is (Ackoff, 1989). The following definitions for *data*, *information* and *knowledge* are also based on (Ackoff, 1989):

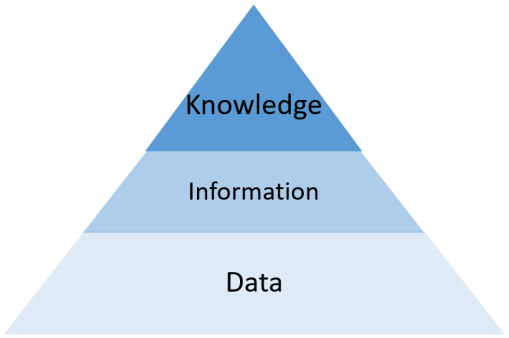

Figure 1: The Information Pyramid (Rowley, 2007)

- **Data**: Data is symbols that represent the properties of objects. They are just a collection of symbols, for example Strings or integer values.
- **Information:** Information is data that is processed to be useful, providing answers to 'who', 'what', 'where', and 'when' questions.
- **Knowledge:** Describes the interpretation of information and explains the use of it. To get knowledge it is very important to answer the 'how' questions about the data.

In some representations, **W**isdom is also introduced as the highest level, and as a link between information, but a consideration of wisdom does not add value in the context of this paper.

Our theory and definition is based on these previously introduced works and places metadata in the context of the DIK(W) pyramid. We present the following definition for metadata:

Definition: Metadata is *Information* about *Data*. It answers the w-Questions about the origin of a Dataset (collection of Data) and is required to generate *Knowledge*.

This means that metadata at least answers the "Where" question I.e. where is the data originating from. Furthermore, many data also answer questions such as "what" is the data about or "when" was it created. As metadata is just not data but some information describing something about the datasets and is required to generate *Knowledge*, which is the base for humans to use data and bring them in a useful way.

3. Metadata for data trading on a data marketplace

The definition of metadata introduced in section 2.2 again underlines the importance of metadata, because generating knowledge is impossible without metadata. Even an algorithm cannot do it, because nobody can design this algorithm and a perfect algorithm is not existing, which is proven by Alan Turing and with the halting problem (Turing, 1937).

In terms of data trading this means, that metadata is essential for creating an offer to sell data on the data marketplace. Furthermore, this means that it is more important to show a part of the metadata to the seller, as a part of the real dataset. In summary, we can state that metadata is the most important part for selling data, because it describes the content of the dataset and helps the buyer to get an overview about it and creating an idea.

Data and its corresponding metadata are generated in various ways and can be used in many different ways. Thus, the lifecycle of data differs in many cases. Figure 2 shows an example of lifecycle for a dataset. Also, the lifecycle and different levels are measured between the level of *interpretation* and level of *processing*. Both levels, interpretation and processing, are crafted. The level of **interpretation** describes the process of making sense out of a collection of data that has

been processed ("Data Interpretation," n.d.). The level of **processing** means "the collection and manipulation of items of data to produce meaningful information" (French, 1996). This is the process from raw data, up to useful data. Both levels are entangled with each other and as deeper look at, the more human intervention is needed.

In this paper we define lifecycle of a dataset as a period from which data is created till it turns into knowledge. This is explained using the example of a weather data set. The lifecycle begins with the phase of the data collection. In a defined time period, the temperature will be measured. Here, also the first part of the metadata starts, for example:

- What will be measured and in which unit? → Temperature in °C
- Where is the measuring station? → Location, for example Clausthal-Zellerfeld
- When it is measured? → for example, from 23.06.2017 until 22.06.2018

After data collection, the data can be preprocessed. This can be for example a filter process to clean the dataset and remove incorrect values. In addition, all this step should be added to the metadata, because this is an important information for end-user.

Finally, when it comes to using data, its usage can be internal, external or both. Here the user begins to link the metadata with the dataset itself to generate new knowledge. This is the highest level of interpretation.

This process can be explained using an example of weather forecasting company who buys some dataset on a data marketplace: A company wants to train an Artificial Intelligence (AI) model for weather forecast. The company has already a set of data [**A**], which are used as training data. They begin to train their algorithm with this dataset, but later they want to extend the forecast area. For this they need more data, to increase the precision and quality of the AI. For this they buy a second dataset [**B**] on a data marketplace. They clean this dataset and filter out some values, which are not relevant for them. Finally, they combine both datasets to a new dataset [**C**].

C = A∪B.

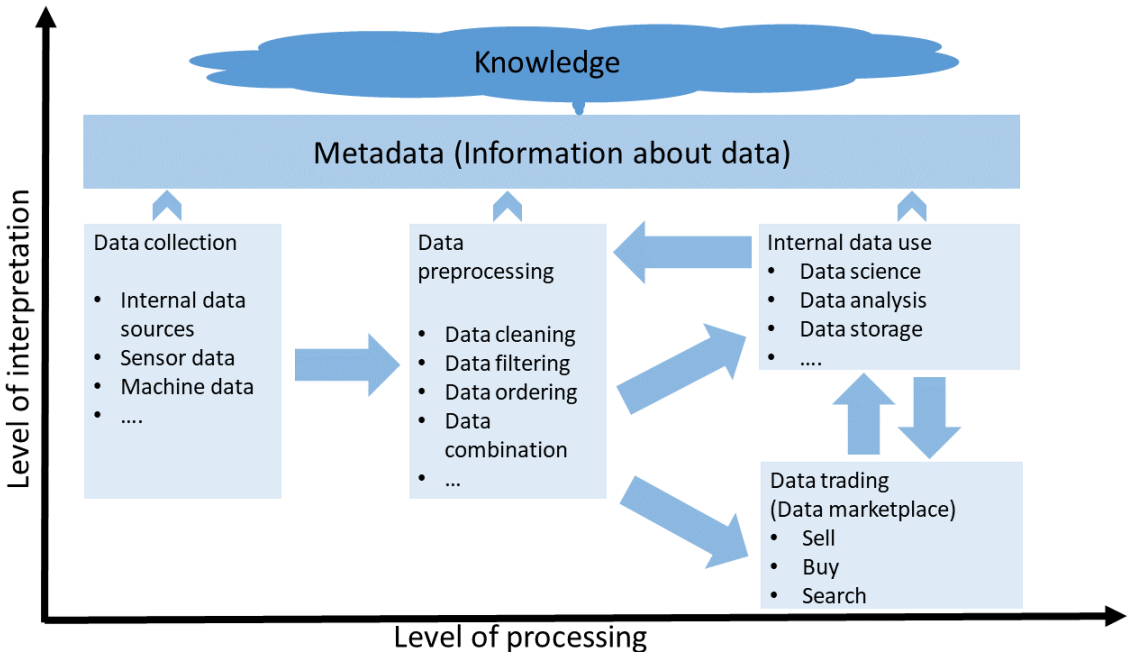

Figure 2: Overview about the data lifecycle

As shown in Figure 2, and as already mentioned, with the deeper Level of interpretation and processing, the automatism to generate Metadata automatically decreases and the expert knowledge from the humans is required increasingly.

The origin of a dataset, for example a timestamp (when) and the data source (where) e.g. a sensor, can be generated automatically easily, while information about the exact content (what) needs a description from a human. Furthermore, some of the steps from the data processing can be added easily and automatically to the metadata, but to choose these steps again is a task for humans. With every new decision and every new processing new pieces of metadata are growing up and they have to be migrated to the existing metadata. In particular, if two previously independent data sets are merged, the corresponding metadata must also be merged. As a result, metadata must always be designed *dynamically*.

The following is a simple example of a metadata record that is abstracted:

$$Data = \{Type, Timestamp, Size\}$$

$$autoInformation = \frac{\partial y}{\partial x} Data$$

$$manualInformation = \{User\ Description\}$$

$$Metadata = \{autoInformation, manualInformation\}$$

Beginning with a dataset, which contains a type e.g float, a timestamp, and the size of the dataset. derivate from this dataset d*ata* some metadata can be already derivate automatically (*autoInformation*), eg. the size of data and the time period (e.g. from 2010 until 2018). In addition, the producer of the data enters some Information about the content (e.g. Weather data from Clausthal-Zellerfeld), which are stored in the *manualInformation*. The content from both describes finally the m*etadata*, which belongs to the dataset *Data*.

3.1 Use of the Metadata in a Data Marketplace

In summary data trading in a data marketplace is not possible without metadata, because it is the most important reference point for a buyer and a seller. Furthermore, it is important for the following points:

- Semantic and filtered search: Metadata build the base for semantic search algorithm and the web 3.0. It is not possible to build an ontology without metadata
- Verification element: After the buyer bought a dataset he/she has never seen before, he/she needs some factors to make sure, if this is really the dataset he has expected. One of these factors are the metadata, whether he matches the record from the dataset with them.
- Data quality

Data quality was deliberately not introduced in the last section, because quality is never objective. Related to this fact, the description of quality cannot be part of the metadata, but a part of the *Knowledge* (the highest level of interpretation).

According to the ISO 9000 quality describes the extent to which something meets the specific requirements. Some orientation values to measure quality are (Mike Sondalini, n.d.):

- Quality is specification driven – does it meet the set performance requirements
- Quality is measured at the start of life – percent passing specification acceptance
- Quality effectiveness is observable by a number of rejects from customers

As seen quality always depends on personal requirements. Data quality is nothing more than the quality of a data set. But as already described in previous research, it is not easy to check data Quality, because a seller cannot provide a whole dataset to a buyer before he/she bought it (Lawrenz et al., 2019). Therefore, a buyer must assess his/her requirements on the basis of the metadata.

4. Identified Challenges

Metadata is one of the most important factors for trading data on a data marketplace. But there are some challenges associated with meta data and data marketplaces. We identify these challenges in this paper. Following are the identified challenges:

- Which kind of meta data are required in a data marketplace?
- How to generate this metadata?
- Which kind of metadata can be generated automatically, and which kind must be entered manually?
- In which form will the metadata be stored (structured, unstructured, semi-structured)?
- Can metadata be used as a verification key for the data exchange?

Specially, because every data is unique, it is much harder to derive the corresponding metadata. All these challenges are still part of our ongoing work and for some of them are already approaches existing, for example the Dublin Core Metadata Standard, to structure metadata. In our ongoing research we also try to solve these challenges in a similar direction. Starting with the identification of relevant metadata, towards the generation and storage of metadata and finally with a possibility to use it for the verification of a data trade.

5. Conclusion

In this paper we outlined the importance of metadata in data trading on a data marketplace. Furthermore, we also introduced the definition of metadata as- "Metadata is *information* about *data*". For data trading metadata is one of the most important factors. With the help of metadata, a seller can describe the dataset and on the other hand it helps the buyer for finding relevant data. Metadata also plays a very important role for determining the quality of datasets. In order to determine data quality which is knowledge, metadata along with personal requirements is required. But unlike other goods sold online there are various challenges associated with data trading on platforms such as a data marketplace, we identified these challenges and presented in this paper. This paper is a part of an on-going research project- Recycling 4.0, our future work will include solutions for the identified challenges in this paper.

Acknowledgements

This paper evolved of the research project "Recycling 4.0" (digitalization as the key to the Advanced Circular Economy using the example of innovative vehicle systems) which is funded by the European Regional Development Fund (EFRE | ZW 6-85017297) and managed by the Project Management Agency NBank.

References

Ackoff, R. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9. https://doi.org/citeulike-article-id:6930744

Baca, M. (2016). *Introduction to Metadata. 3rd ed.* Los Angeles: Getty Publications. Retrieved from http://www.getty.edu/publications/intrometadata.

Data Interpretation. (n.d.). Retrieved May 19, 2019, from https://www.toppr.com/guides/quantitative-aptitude/data-interpretation/

French, C. (1996). *Data processing and information handling*. Thomson. https://doi.org/10.1145/1458043.1458060

Lawrenz, S., Sharma, P., & Rausch, A. (2019). Blockchain Technology as an Approach for Data Marketplaces. In *ICBCT 2019*. Honululu, USA: ACM. https://doi.org/https://doi.org/10.1145/3320154.3320165

Mike Sondalini. (n.d.). What is Quality? What does Quality Mean? How do You Know When You Have Quality? Retrieved May 19, 2019, from https://www.lifetime-reliability.com/cms/free-articles/work-quality-assurance/what-is-quality/

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180. https://doi.org/10.1177/0165551506070706

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Wallace, D. P. (2007). *Knowledge management: Historical and cross-disciplinary themes*. Libraries unlimited.

Wertvollste Unternehmen nach Markenwert weltweit 2018 | Ranking. (n.d.). Retrieved April 25, 2019, from https://de.statista.com/statistik/daten/studie/162524/umfrage/markenwert-der-wertvollsten-unternehmen-weltweit/