Accountable Data Analytics Start with Accountable Data: The LiQuID Metadata Model

Sarah Oppold[1] and Melanie Herschel[2,1]

[1] University of Stuttgart, Germany
[2] National University of Singapore, Singapore
{Sarah.Oppold, Melanie.Herschel}@ipvs.uni-stuttgart.de

Abstract. Insights based on data are omnipresent. However, in particular in modern data analytics applications, information about the underlying data often remain obscure, hindering accountable data analytics. Recent efforts have been put into better describing such data based on metadata, similarly to what has been done in various scientific disciplines for transparent and reproducible research. Based on a detailed study of various metadata standards and proposals, we observe that existing metadata models do not yet sufficiently cover information that is relevant for data accountability. To fill this gap, this paper proposes LiQuID, a novel metadata model to make datasets accountable throughout their life cycle. It is more general than existing metadata models, which can be mapped to LiQuID. We validate LiQuID for the purpose of dataset accountability based on a real-world workload we created.

Keywords: Metadata Model · Accountability

1 Introduction

Data underly various insights and decisions today, e.g., for dating recommendations, marketing decisions, scientific findings, or responses to pandemics such as COVID-19. The result of analyzing these data potentially influences various aspects of people's lives. Unfortunately, the development of data analysis pipelines, that rely on data, is prone to errors. Even though developers of such pipelines may have the best intentions, mistakes are likely to occur [10]. To understand and account for decisions or insights drawn from data, an important aspect is to account for the underlying data itself, which includes being transparent about the creation, handling, purpose, and meaning of the data.

Not being aware of properties or intended purpose of data and (possibly inadvertently) mishandling and misinterpreting the data as a consequence can have significant repercussions. One example is the introduction of discrimination into decision support systems, as could be observed with the recidivism prediction system COMPAS [2]. Another example arises in the COVID-19 pandemic, where lots of data have been shared in a word-wide effort to gain insights. However, as sites like Our World in Data [3] point out, caution has to be applied when reading reported numbers, as it is often unclear what they mean. For instance, are

[3] https://ourworldindata.org/coronavirus

number of tests counted as swabs or individuals tested? What is the source of symptoms reported with cases? Was publishing the data rightful? The datasets would clearly benefit from accompanying descriptive data, i.e., *metadata*, to answer such questions.

More generally, governments, ethical review boards, scientists, engineers, policy makers, and many more stakeholders need to assess and scrutinize data, e.g., to determine the appropriateness of the data for their purpose or to ensure that data are used correctly, ethically, and lawfully. Information pertinent to this assessment is not included in the data itself, it needs to be provided alongside the data as metadata. This information makes datasets more transparent, and can serve as evidence to verify compliance or appropriateness of data with respect to rules or requirements [13, 16]. Thereby we obtain *accountable datasets*, which we understand as follows in this paper: Accountable datasets are datasets about which there is sufficient information to justify and explain the actions on these datasets to a forum of persons, in addition to descriptive information and information on the people responsible for it. In this paper, we propose to convey the necessary information in the form of metadata. This information enables *dataset accountability*, where all persons responsible for a dataset, i.e., all persons who have been involved in the life cycle of the dataset, must justify and explain their actions on the dataset with respect to a set of rules, e.g., laws, contracts, or moral rules, to a forum of persons in authority. Our notion of dataset accountability goes beyond information accountability [16], which focuses on the appropriate use of data, leaving out all other steps in the life cycle of a dataset such as its creation or maintenance. It also complements algorithmic accountability [17], which is about the justification of entire algorithmic systems.

Clearly, the metadata for accountable datasets are very diverse and broad. They cover all phases of the life cycle of a dataset (including data collection, processing, maintenance, and usage) and address different aspects (e.g., meaning, purpose, responsible parties, or ethical considerations). While it is possible to obtain some of the necessary information when "releasing" the data for further use, some pieces of information such as design decisions or responsibilities require collection along the dataset generation process or even prior to its start. Planning in advance what information should be gathered and incorporating this into the design process is therefore beneficial for holistically accountable datasets.

This paper presents a metadata model for accountable dataset that gives a clear structure on what information is possibly relevant and provides guidance on what questions to consider when handling datasets. It is systematically designed along two dimensions: the first dimension models the different phases of the data life cycle, while the second dimension models essential questions (how, what, why, etc.) that can be asked about each phase. The information answering each question in each life cycle step is structured following five key fields or attributes. Overall, the metadata model, which we call LiQuID[4] is defined such that it can accompany any dataset, e.g., from initial data sources to datasets resulting from complex processing.

[4] Name refers to the modeled Life cycle steps, Questions, and Information about Data

There are plenty of existing, highly domain-specific metadata models that can be considered candidates for supporting dataset accountability, as they have been established to make items of interest and corresponding metadata, e.g., findable, accessible, interoperable, reusable, and repeatable [1, 4, 11, 14], also known as FAIR principles [18]. Focusing on datasets as particular items of interest, [3, 8, 9] can be considered emerging approaches towards metadata models for accountable datasets. However, these have been defined in a rather ad-hoc fashion. We show in this paper that LiQuID generalizes the aforementioned existing metadata models. A detailed study of how existing models fit into our metadata model demonstrates both the appropriateness of LiQuID and the gaps in existing models in terms of dataset accountability.

To be of practical use, it is important that a metadata model for accountable datasets covers the metadata necessary for typical questions that arise when datasets are evaluated or verified. Therefore, we determine a real-world query workload on accountable data, based on analyzing audit literature, the GDPR, and conducting an expert survey. We observe that LiQuID is the only model we are aware of that can answer all queries of the workload, validating LiQuID's completeness. We further see that the workload requires a substantial fraction (75%) of the fields modeled by LiQuID, indicating its conciseness. No other metadata model can fully handle the workload, and 10% of fields required by the workload are not present in any considered existing metadata model.

In summary, we make the following contributions: (Section 2) a novel metadata model for accountable datasets, called LiQuID; (Section 3) a detailed analysis of existing metadata models with respect to dataset accountability that demonstrates both the appropriateness of our model and the gaps in existing models for accountable data; and (Section 4) a real-world data accountability query workload which we use to validate the completeness of LiQuID.

2 LiQuID: a metadata model for accountable data sets

This section presents LiQuID, for which we set the following requirements:

1. **Holistic view**: The metadata model covers the whole life cycle of a dataset.
2. **Systematic structure**: A systematic structure offers a clear guidance on what information is potentially relevant.
3. **Accountability**: Following our notion of dataset accountability, the metadata model should (i) include information on responsible entities (e.g., creators, dataset managers) who can be held responsible for the handling of the data, as well as (ii) leave room for explanations and justifications in anticipation of an accountability discussion.
4. **Extension**: The metadata model builds on existing and time-tested approaches, maintaining and supporting features that have proven to be important (e.g., type descriptions, ontologies, FAIR principles [18]).

After describing the general hierarchical metadata model in Section 2.1, we provide selected details in Section 2.2. Section 2.3 discusses how LiQuID meets the requirements mentioned above.

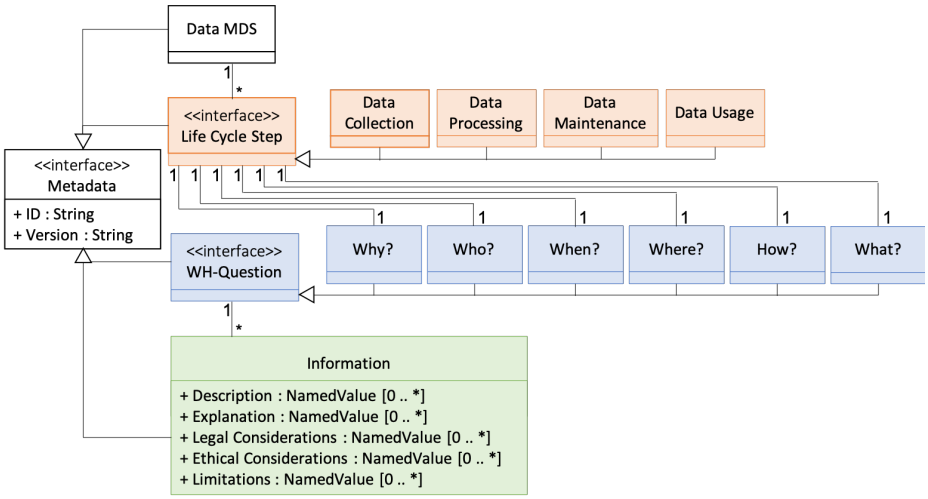

Fig. 1. Overview of LiQuID, showing the different levels of the hierarchical model, namely the *life cycle level*, *question level*, and: *information level*.

2.1 Metadata model overview

Figure 1 summarizes our metadata model for accountable data. It is a hierarchical model comprising three levels: the top level models the different data life cycle steps, we therefore call it *lifecycle level*. For each such step, the second level, named *question level*, structures metadata according to questions relevant for accountable data. The *information level* at the leaf level models the actual information per life cycle step and question. Note that by extending a general *metadata interface*, each element of the metadata model is uniquely identified and may have multiple versions. Note, that a full XSD is also available on our project website. [5]

Life cycle level. In the life cycle level, we consider four essential steps in the life cycle of a dataset. The first step is *data collection* that relates to the creation, gathering, or capture of the data. Data collection typically involves manual entry or gathering, automatic capture of data produced through various processes, or the acquisition of third party data. The *data processing* step covers all data manipulations that have altered or transformed the data. Frequently applied data manipulations during preprocessing include data standardization, data cleaning, or aggregation. Under *data maintenance*, we understand the handling of the dataset once it has been released for further use. This includes a wide range of data management operations, e.g., updates, additions, deletions of (some) data, its archival, or destruction. Finally, *data usage* takes into account past, present and anticipated activities supported by or applied on the dataset, e.g., input to machine learning algorithms or distribution to other parties. Although some-

[5] https://www.ipvs.uni-stuttgart.de/departments/de/research/projects/fat_dss/

times seen as separate data life cycle steps, note that we consider information on data storage and distribution as part of the information on data usage, as they are essential information when the need to account for proper use of data arises and are thus jointly queried with the data usage information.

Question level. The second level structures information that LiQuID covers for every step of the life cycle by commonly used WH-Questions: *Why?*, *Who?*, *When?*, *Where?*, *How?*, *What?*. This categorization follows the general human rationale of asking for information about an entity of interest. While this may appear simplistic, we believe this simplicity allows to easily understand and use LiQuID. We show later that this structure actually covers all the information contained in other metadata models - and more, providing evidence that this intuitive model is nevertheless effective in covering the required information.

Information level. While the first two levels essentially serve to contextualize the information to be provided for accountable data, the information level organizes the information needed for each life cycle step and question in five fields. The first field is a *description* that answers a WH-question for a data life cycle step. In order to invite explanations and justifications, which are essential for accountability, the information level additionally models fields for *explanation*, *legal considerations*, *ethical considerations* as well as *limitations* of the answer.

2.2 Information details

To get a better understanding of what information our metadata model covers, Table 2 summarizes what is understood as relevant information for three combinations of a life cycle step S and a question Q, denoted as $S.Q$. An exhaustive description for all combinations is available on our project website. As the examples in Figure 2 show, we associate a list of questions with each field of each $S.Q$ combination. These are intended to help populate the metadata sheets.

In the subsequent discussion, we focus on the questions to consider when filling out the information relating to *collection.who*. We make up a simplistic example to illustrate the potential content of each field. The example considers a hospital that collects case numbers for a particular disease.

- **Description:** Considering the question *Who?* during data collection, the description includes information on who (people, organizations) was involved in the data collection process. It further encompasses any information relevant to their identification, their role in the data collection process, information necessary to assess their qualifications to fulfil this role, and any details about these people or organizations that may impact the data. In our example, we would report the hospital and the head of the service responsible for collecting accurate numbers.
- **Explanation:** As part of the explanation, a justification on why these particular people were involved in the data collection process can be provided. Continuing our example, we explain that this hospital is collecting these numbers as they are the only medical facility to treat the disease in a larger area. The responsible person is justified by her job description.

	collection.who	processing.how	usage.where
Description	Who (people, organizations) was involved in data collection? Provide all information relevant to their identification, their role in data collection, all information necessary to assess their qualifications to fulfill this role, and all characteristics which could have an influence on the data set.	What was the methodology/ procedure for data processing? Which methods and tools were used in each step and what was the (technical) environment?	Where is the data set published/ available? Where (place, geographically) can the published data set be used?
Explanation	Why were these particular entities involved in data collection?	Why was the data processed using this particular methodology/ procedure, methods, tools and (technical) environment?	Why is the published data set made available at this place? Why can the published data set be used at this place?
Legal consid.	Why was it lawful that these people participated in data collection?	Why was it lawful to process the data using this methodology/ procedure, methods, tools, and(technical) environment?	Why is it lawful to publish the data set at this place? Why is it lawful to use the published data set at this place?
Ethical consid.	Why was it ethically justifiable that these people were involved in the data collection process?	Why was it ethically justifiable to process the data using this procedure, methods, tools, and(technical) environment?	Why is it ethically justifiable to publish the data set at this place? Why is it ethically justifiable to use the data set at this place?
Limitations	What general limitations in the data set could result from the selection of entities involved in the data set collection and their characteristics or qualifications? What limitations for the overall objective (Why?) could result from the choice of people? What efforts have been made to mitigate the identified problems and limitations? Why are there no limitations?	What general limitations in the data set could result from this choice of methodology/ procedure, methods, tools, and (technical) environment of the data processing? What limitations for the overall objective (Why?) could result from this choice? What efforts have been made to mitigate the identified problems and limitations? Why are there no limitations?	What general limitations for data set usage could result from the selection of place where the data set is made available? Where should the published data set not be used?

Fig. 2. Examples of information covered for different $S.Q$ combinations.

- **Legal Considerations:** To ensure that the persons involved in data collection had the right to do so, legal considerations recording why it was lawful that these people were involved in the data collection process are included in the metadata sheet. In our example, this includes an acknowledgement that the hospital is legally allowed to collect these data, e.g., based on disease control regulations.
- **Ethical Considerations:** We also consider ethical questions, asking why it was ethically justifiable that these people were involved in the data collection? For instance, if the hospital receives funding depending on the number of cases, has a conflict of interest been ruled out?
- **Limitations:** Finally, the metadata model offers the possibility to clarify (i) what limitations in the data set could result from the selection of persons involved in the data collection (based on their their characteristics or qualifications available in the description); (ii) what limitations for the overall objective (Why?) could result from the choice of people; (iii) what efforts have been made to mitigate the identified limitations; or (iv) why there are no limitations. In our example, a limitation is that the data may lag behind the actual situation given internal processes at the hospital. But mechanisms have been put in place to not lag behind by more than 24 hours.

2.3 Discussion

Having introduced both requirements for our metadata model and the model itself, we now review how LiQuID meets the requirements.

First, the metadata model should offer a holistic view on a dataset, covering its whole life cycle. This is achieved by the life cycle level of LiQuID that considers the essential steps of a dataset's life cycle.

A systematic structure that provides guidance on what information to consider is given by the overall hierarchical structure of LiQuID. For each data life cycle step, it asks WH-questions, which are the self-evident human rationale for assessment. Each question can be answered in a structured way, dictated by the information level.

Let us now review how LiQuID supports our accountability requirement. On the one hand, accountability is supported by asking for responsible entities which should be described in the *Who?* question of each life cycle step. On the other hand, the anticipated accountability discussion has to be modeled without actually being able to know the questions. But since the questions can be expected to be critical inquiries of the decisions made in the different life cycle steps, the metadata model encourages to think about such critical questions and leaves room for responses by providing the information fields for explanation, legal considerations, ethical considerations, and limitations.

Finally, the metadata model should be compatible with existing metadata models by extending these. As we will discuss in detail in Section 3, LiQuID generalizes existing models, which can be mapped into our metadata model. We also assume that details modeled by well-established standards, definitions, and ontologies are "docked" at the information level, i.e., each field modeled at the information level contains further structured elements that are application dependent.

3 Comparative assessment

This section compares our metadata model for accountable data with established, time-tested, and revised metadata models used in various disciplines. Even though their subject of interest and purpose differ from our metadata model for accountable datasets, they implicitly represent accumulated knowledge of what information is deemed important to describe some subject of interest. More specifically, we map nine existing metadata models to LiQuID. To this end, any field specified by an existing model is mapped to the corresponding field(s) in LiQuID. Note that we obtain a complete mapping, in the sense that we could map all information modeled by an existing metadata model to LiQuID.

For our comparative assessment, we choose metadata models with varying specificity and from various domains. These include two general models [5, 15], four standards to describe an item of interest arising in various domains [1, 11, 4, 14], and three emerging metadata models for fair, accountable, and transparent datasets [3, 8, 9].

- **Dublin Core (DC)** [5], one very general metadata model we consider is a conceptual generic model often used as base for other models;
- **W3C PROV (PROV) [15], another very general metadata model which focuses on describing the lineage of some end product;**

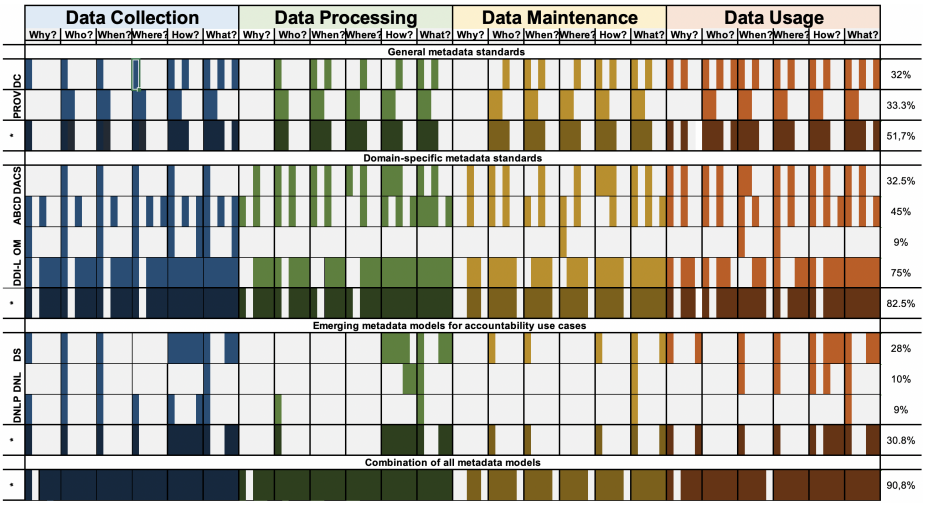

Fig. 3. Mapping between LiQuID (columns) and other existing metadata models.

- **Describing Archives: A Content Standard (DACS)** [14] specifies archiving principles and a metadata model for (aggregations of) archival records on, e.g., books, reports, or movies;
- **Access to Biological Collection Data (ABCD)** [1], a metadata model implementation for biological sample collections;
- **Observations and Measurements (OM)** [11] emerged from the Open Geospatial Consortium and defines a general conceptual schema for observations and measurements as well as sampling details;
- **Data Documentation Initiative Lifecycle (DDI-L)** [4], a metadata implementation describing (groups of) social studies based on questionnaires;
- **Datasheets for Datasets (DS)** [8] document (personal) datasets allowing them to be examined for new machine learning applications within the context of fair machine learning;
- **Data Nutrition Labels (DNL)** [9] provide automatically generated modular labels that describe datasets and are intended to enable accountable AI;
- **Data Statements for NLP (DNLP)** [3] describe spoken or written texts in order to enable fair natural language processing.

To determine to what extent the existing metadata models cover our model, we study the existing models in detail and map the entries they specify to LiQuID. Figure 3 depicts a visualization of this mapping. The columns reflect the leaves of our hierarchical model (i.e., each column corresponds to an information field under a question and life cycle step, information fields being in the order listed in Section 2.2). Rows filled with colors represent the metadata models listed above. Rows with label $*$ filled with dark color aggregate a group of metadata models. A colored cell indicates that there is *at least one* specified entry in the metadata model (row) which corresponds to the respective information field

in this metadata model (column). Different colors are used to distinguish the different life cycle steps to enhance readability. We color the fields generously, which means fields (i) with only little information on the respective combination of life cycle step S, question Q, and detail D, denoted $S.Q.D$ or (ii) not explicitly meant but amenable for the specific $S.Q.D$ have been colored. If a field is left blank, this indicates that there is no entry in the metadata model of the row ("notes" or "additional comments" set aside) that corresponds to $S.Q.D$ identified by the column. At the end of each row, we also provide a coverage percentage, calculated as the number of details (cells) covered by a metadata model, divided by the number of detailed fields in LiQuID.

Interestingly, Figure 3 shows that both general metadata models cover about 30% of LiQuID. Even combined they only cover 51.7%. Both standards contain few fields, some of them too general to be mapped to specific LiQuID fields.

Figure 3 shows that the lowest coverage of 9% is achieved by OM and DNLP. The low coverage of OM can be explained, since the standard describes geological specimen for which an accountability discussion is unlikely. Additionally, these specimen typically do not undergo processing and maintenance life cycle steps.

More interestingly, we observe that while DNL and DS have higher coverage than DNLP, the coverage of these metadata models, which have been proposed with accountability use cases in mind, is generally low. Aggregating them still only covers around 31% of the details considered in LiQuID. This clearly shows that while the proposed metadata models may serve well the specific application they were engineered for (e.g., information for developers of machine learning pipelines [9]), they do not provide a general metadata model for accountable datasets. This is further validated once we consider a query workload over accountable datasets in Section 4.

Focusing on the domain-specific metadata-models, we see that their coverage highly varies between 9% (OM) and 75% (DDI-L). While their individual coverage may only be moderate, we observe that they cover different details to a different degree. Indeed, the standards complement each other, as shown by the aggregated coverage on this group of 82.5%.

This high combined coverage of 90.8% shows that many fields included in LiQuID are already deemed important by existing metadata models. However, the systematic structure of LiQuID also reveals "blind spots", as it includes additional fields where information is still missing from any of the considered metadata models. For instance, we note that while all data life cycle phases are considered, data maintenance is covered less. However, it is reasonable to assume that accountability questions on data maintenance arise, for example when personal data has to be corrected or deleted due to an opt-out of a data subject. Looking at the questions level, the Why? question is the least covered element, which is surprising since the management of data should ideally have a goal. Finally, at the information level, both explanations and ethical considerations are scarcely covered by the considered existing data models.

In summary, we observe that existing standards and emerging data models with accountability use cases in mind, can all be fully mapped to our metadata

model for accountable datasets. The converse does not hold, as LiQuID is not covered 100% by any considered data model. To understand how relevant the information that LiQuID covers is for dataset accountability, we determine an accountability workload and study which information it actually queries.

4 A query workload over accountable datasets

Ideally, a predefined benchmark would be used in order to assess the metadata model objectively. However, we are not aware of any benchmark considering accountability by including a set of questions or queries which are realistic in datset accountability scenarios. We therefore contribute a first such benchmark by creating a workload of queries on accountable datasets and then assess LiQuID with respect to this workload. We first introduce our methodology to create the workload in Section 4.1. Section 4.2 then discusses how LiQuID fits this workload. The full workload is also available on our project website.

4.1 Creating the workload

Sources of real-world accountability questions. To determine a realistic workload of queries on metadata models for accountable datasets, sources are needed that describe existing practices, regulations, and questions that arise in settings requiring accountability with respect to data.

We identify three such sources. Our first source comes from the Federal Trade Commission (FTC) [7] and establishes a list of guidelines or statements relating to accountability as part of an in-depth study on data brokers. Data brokers collect personal data about individuals from different sources and sell these data to companies. In an effort to create more transparency on data brokers, the FTC made data brokers answer questions, which they provide in their report. This shows how regulators conduct real audits and the report includes 101 statements relating to accountability. One sample statement asks to *"Provide a list and description as to the nature and purpose of all the products and services (both online and offline) that the Company offers or sells that use personal data. Include a separate description of each product or service identified[.]"*.

Second, we consider regulations from the General Data Protection Regulation (GDPR) [6], which aims at protecting personal data. It is one of the most restrictive data protection regulations and focuses on data processing and therefore we expect it to be a tough test for metadata models for accountable datasets. Beyond clarifying what data protection regulators will test for, it also takes into account data subjects who have the right to contest who uses data about them. From the GDPR regulations, we derive possible questions that aim at verifying the regulations. As an example, consider the following regulation from GDPR Article 3(2): *"This Regulation applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union, where the processing activities are related to: (a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to*

such data subjects in the Union; or (b) the monitoring of their behaviour as far as their behaviour takes place within the Union.". We associate this to the following accountability questions that may be asked when verifying if and how a regulation applies: *"Are the data subjects of the personal data you process in the European Union? Are the personal data processing activities related to the offering of goods or services to data subjects in the European Union? Are the personal data processing activities related to the monitoring of data subjects behavior that takes place in the European Union?"*.

Lastly, we conducted an expert surveyin order to determine questions asked in dataset assessment deemed relevant by experts. This expert survey extends beyond personal data, which is the focus of both FTC and GDPR. Ten experts from various domains (including librarians, data management experts, doctors, social scientists) participated in the survey. They explained what criteria are important to them when they assess a dataset and what questions they would ask to assess these criteria in the different data life cycle phases.

Overall, from these sources, we obtain 183 textual descriptions of what information is relevant in real-world data accountability scenarios.

From textual descriptions to structured queries. In a next step, we determine a query language that allows us to query the data corresponding to the 183 textual descriptions, assuming the data are represented hierarchically (as in our metadata model). Given the textual descriptions, we observe that the query language needs to support different constructs, in particular, conditions, comparisons, for-loops, and the use of equality constraints. Given these requirements, we opt to use XQuery as query language, as it meets all requirements.

Following the questions and statements from the three sources, we derive XQueries, defined over an XML Schema that follows our metadata model. That is, when writing the queries, we determine from which fields of our model the relevant information can reasonably be retrieved. Note that we do not claim that our queries cover all possibilities. Also, note that our queries are the result of a best-effort approach to resolve ambiguities in questions or statements. To simplify our queries, we assume further elements nested under the elements defined by our metadata model that structure the data. In practice, such elements may be the result of a domain-specific ontology for different accountability use cases, as supported by our extension requirement.

For example, Algorithm 1 shows the XQuery that translates the FTC statement provided above (repeated in the algorithm's header for convenience). The color coding indicates semantic correspondences between the text and the query. First, the query identifies *(who?)* the Company named "myCompany" who acts as "Service provider" by offering or selling products and services. It further checks that the company uses *(what?)* personal data *(why?)* to include in their products or services. When all these conditions are met, we return the description of the identified product or service, assuming it includes a name and a description. We explain the purpose of the product or service that processes personal data.

Algorithm 1: XQuery example derived from the FTC statement *"Provide a list and description as to the nature and purpose of all the products and services (both online and offline) that the Company offers or sells that use personal data. Include a separate description of each product or service identified[.]"*

```
for $x in MDSStore/DataMDS/DataUsage/Who/Information
where $x/Description/Name="myCompany"
and $x/Description/Type="Service provider"
and ($x/../../Why/Information/Description/Type="Product" or
     $x/../../Why/Information/Description/Type="Service")
and $x/../../What/Information/Description/Type="Personal data"
return
<Result>
<Name>{$x/../../Why/Information/Description/Name}</Name>
<Description>{$x/../../Why/Information/Description/Desc}</Description>
<Purpose>{$x/../../Why/Information/Explanation/Purpose}</Purpose>
</Result>
```

4.2 Evaluation of the metadata model wrt. the workload

This section studies how our metadata model supports the real-world accountability workload obtained as described in the previous section.

From the 183 textual statements and questions, we were able to express 97% with our metadata model. The remaining 3% are statements that refer to (i) information on why some measure was *not* taken, which is not modeled, because it did not happen, or (ii) unintentional data manipulations, which would be intentional as soon as they are modeled in the metadata.

Next, we study which fields of our metadata model are covered by the queries of our workload. Figure 4 shows the coverage of queries based on FTC, GDPR, and the expert survey, as well as the coverage when unifying all queries (row marked with $*$ and with black color, ignore flags for now). The visualization is analogous to the visualization in Figure 3. A field is colored when at least one query of the workload refers to it, and coverage is the number of fields referred to by a workload, divided by the 120 fields available in our metadata model.

First, we observe that the coverage of workloads of different sources varies between 43.3% for GDPR and 52.2% for the expert survey. However, the workloads complement each other and, when combined, access 75% of the fields in our metadata model. While the necessity of 30 fields of our metadata model is not demonstrated by the workload, we clearly see that a substantial number of fields not covered by any metadata model devised with accountability scenarios is relevant in our workload (cf. Figure 3).

Among the fields accessed when combining all three workloads, we see that 9 of these fields (flagged fields among the black fields in Figure 4) are among the 11 fields not covered by any other considered metadata model (left white in Figure 3). This validates that our systematic structure and approach in defining

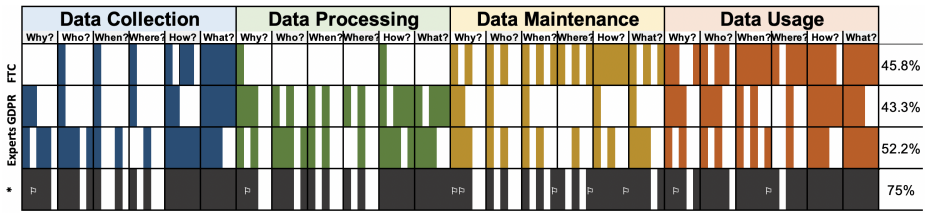

Fig. 4. Workload coverage.

the metadata model has contributed to identifying relevant fields not considered by other metadata models.

Finally, assuming that any field that is either accessed by our real-world accountability workload or has been defined by a previous metadata model is relevant, we see that 94.2% of fields modeled by LiQuID are relevant.

In conclusion, our study of how LiQuID relates to related work and real-world workloads demonstrates that our metadata model successfully covers a wide range of accountability queries and generalizes existing metadata models well, indicating the completeness of the proposed metadata model. At the same time, it is sufficiently concise, as it does not model a significant amount of information for which relevance still needs to be demonstrated.

5 Conclusion and Outlook

To summarize, we presented a novel metadata model for accountable datasets. It hierarchically models information relevant in scenarios requiring dataset accountability, covering different steps of the data life cycle, various questions arising at each step, and structuring the answers based on five attributes. We presented a detailed review of metadata models that can be considered candidates to enable accountable datasets. We observed that our metadata model can fully cover these, while being more general by modeling additional information. That this additional information is indeed relevant for accountable datasets is validated based on a real-world workload of queries arising in dataset accountability scenarios. Overall, our metadata model is the first model we are aware of that is rich enough to answer all queries of the defined workload.

While the insights gained through the research conducted in this paper are encouraging for using the proposed metadata model in practice, there are still quite a few challenges to tackle as part of future research. First, as we experienced ourselves, informing all fields of the metadata model is a tedious and time-consuming task. Therefore, we plan to investigate how to automatically or semi-automatically fill fields. Another avenue of future research is the integration of accountable datasets in a larger environment, such as a system for accountable decision support [12]. For this, the metadata about datasets needs to be linked to metadata collected about other parts of a system to give a holistic view.

References

1. Access to Biological Collection Data task group: Access to Biological Collection Data (ABCD) (2007), http://www.tdwg.org/standards/115
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics **6**, 587–604 (2018)
4. Data Documentation Initiative: DDI lifecycle 3.2 (2014), https://ddialliance.org/Specification/DDI-Lifecycle/3.2/
5. DCMI Usage Board: DCMI metadata terms (2020), https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
6. European Parliament, Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (2016)
7. Federal Trade Commission: Data brokers: A call for transparency and accountability (2014)
8. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumeé III, H., Crawford, K.: Datasheets for datasets. In: Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning. p. 17 (2018)
9. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The dataset nutrition label: A framework to drive higher data quality standards. CoRR p. 21 (2018)
10. Olteanu, A., Castillo, C., Diaz, F., Kiciman, E.: Social data: Biases, methodological pitfalls, and ethical boundaries. SSRN Electronic Journal p. 47 (2016)
11. Open Geospartial Consortium: Observations and Measurements (2013), https://www.ogc.org/standards/om
12. Oppold, S., Herschel, M.: A system framework for personalized and transparent data-driven decisions. In: Advanced Information Systems Engineering. p. 16. Springer International Publishing (2020)
13. Singh, J., Cobbe, J., Norval, C.: Decision provenance: Harnessing data flow for accountable systems. IEEE Access **7**, 6562–6574 (2019)
14. The Society of American Archivists: Describing archives: A content standard (2019), https://saa-ts-dacs.github.io/
15. W3C Working Group: An overview of the PROV family of documents (2013), https://www.w3.org/TR/prov-overview/
16. Weitzner, D.J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., Sussman, G.J.: Information accountability. Communications of the ACM **51**(6), 82–87 (2008)
17. Wieringa, M.: What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In: Fairness, Accountability, and Transparency. p. 1–18 (2020)
18. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3**(1) (2016)