applied sciences

MDPI

Article

Metadata Quality for Biobanks

Volodymyr A. Shekhovtsov and Johann Eder *

Department of Informatics Systems, Universität Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt, Austria
* Correspondence: johann.eder@aau.at

Abstract: The mission of biobanks is to provide biological material and data for medical research. Reproducible medical studies of high quality require material and data with established quality. Metadata, defined as data that provides information about other data, represents the content of biobank collections, particularly which data accompanies the stored samples and which quality the available data features. The quality of biobank metadata themselves, however, is currently neither properly defined nor investigated in depth. We list the properties of biobanks that are most important for metadata quality management and emphasize both the role of biobanks as data brokers, which are responsible not for the quality of the data itself but for the quality of its representation, and the importance of supporting the search for biobank collections when the sample data is not accessible. Based on an intensive review of metadata definitions and definitions of quality characteristics, we establish clear definitions of metadata quality attributes and their metrics in a design science approach. In particular, we discuss the quality measures accuracy, completeness, coverage, consistency, timeliness, provenance, reliability, accessibility, and conformance to expectations together with their respective metrics. These definitions are intended as a foundation for establishing metadata quality management systems for biobanks.

Keywords: metadata; data quality; biobank; quality metrics

Citation: Shekhovtsov, V.A.; Eder, J. Metadata Quality for Biobanks. *Appl. Sci.* **2022**, *12*, 9578. https://doi.org/10.3390/app12199578

Academic Editor: Alexander N. Pisarchik

Received: 24 August 2022
Accepted: 21 September 2022
Published: 23 September 2022

1. Introduction

Biobanks are indispensable resources for quality-conscious medical research. Biobanks collect biological material and associated data and provide them for medical studies. In recent years considerable effort improved the management of the quality of the collection, preparation, and storage of biological materials through dedicated quality management and standardized operating procedures. Several proposals and standards were developed to describe the biological material and its quality characteristics (e.g., SPREC [1], BRICS [2], ISO 20387 [3], or MIABIS [4]). Similar efforts are underway to improve the management of these data and additional data associated with biological material like health records, etc.

Here we focus on the quality of metadata, i.e., the quality of the information a biobank maintains and provides about the data it has to offer. This focus on metadata is motivated by several considerations. The mission of biobanks is to support researchers and research projects with material and data of established quality characteristics [5]. This involves supporting researchers to find material and data (sometimes referred to as cases)

fulfilling the requirements for an intended research project. However, data in biobanks (sample data) is very sensitive personal data of the donors of the biological material (e.g., detailed information about the diagnosis, treatment, risks, etc.) and has to be protected with utmost care. Hence it is only possible to give researchers access to this data after careful procedures [6] involving ethical committee approvals, etc. Therefore, researchers cannot simply search in these databases but need information about the collections in a biobank. Only with this information researchers can decide, which of the hundreds of biobanks to contact.

To address this problem, in [5] we proposed to describe biobank collections by means of metadata (data that provides information about other data [7]) describing the content or

quality of the biobank data, and to use this metadata in the search for collections matching the requirements. The main advantage of such a process is that the metadata can be organized to be free of sensitive information (e.g., anonymized [8,9]), so it is safe to make it available for researchers. When a suitable biobank collection is found based on its metadata, the further process and negotiations can be performed to get direct access to sample data.

The usage of the data associated with the biological samples cannot be known in advance as it is collected to support new research projects in the future. Thus, the common definition of data quality as fitness for use is not appropriate for biobanks, and we need to rely on other definitions of the quality of sample data [10]. However, the usage of metadata is known: to support researchers searching for relevant material and data satisfying the requirements for their intended medical studies. Therefore the fitness for use can be applied for biobank metadata. Hence the overarching criterion for defining metadata quality is how well suited these metadata are to support the search for relevant material and data.

This means in particular, that quality characteristics of metadata consider the degree of efficiency for reaching the following metadata goals:

1. supporting the researchers in their search for biobank collections (to find material and data for their research purposes) without direct access to sample data.
2. supporting them in their decisions about the relevance of the specific collection for their research.

With respect to the first goal, in particular, it characterizes how well the metadata can assure that the researchers know about collections that are relevant to their aims (to see what is offered to them), so they can be found when they are needed. With respect to the second goal, it indicates how well the researchers can know if the collection is matching for their requirements.

The following examples illustrate metadata quality with respect to the above metadata goals:

1. Suppose the biobank collections store COVID test result data. In this case, if they all are accompanied by the metadata which stores the probability for this data to be collected with the same test, it allows to search for collections that possess specific values for such a probability (e.g., by issuing queries "find the collections with COVID test result collected with homogeneous tests", "find the collections with COVID test result collected with heterogeneous tests"). This is an example of the good quality of metadata (high completeness of test consistency metadata) with respect to the first metadata goal (supporting the search).
2. If the biobank states that the HIV status data is only defined for 5% of samples in a collection, and this metadata value is present for a collection, the correspondent collection is not suitable for research that requires such status to be available. This is an example of the good quality of metadata (high completeness of HIV status completeness metadata) with respect to the second metadata goal (making the researcher able to decide on collection relevance).

This paper introduces a metadata quality framework based on an extensive review of approaches for defining metadata quality. The framework consists of a set of concrete metadata quality characteristics and metrics, and discusses the possible ways of using such a framework to perform metadata quality assessment on top of some simple aggregations.

The rest of the paper is organized as follows. Section 2 introduces necessary background information. In Section 3, we describe the state of the art in research within the scope of the paper. Section 4 introduces the biobank metadata and its goals. Section 5 provides the definition for the metadata quality in biobanks and describes its structure and possible aggregations. Based on this definition, Section 6 provides a detailed treatment of the nine most commonly used metadata quality characteristics and the corresponding metrics. At the end, the paper provides conclusions and outlines the directions for future research.

2. Background

2.1. Data in Biobanks

Following [11–14], we define biobanks as *collections of biological material (samples) accompanied with data*.

Biobank data [10,15] is a set of collected facts connected with the samples stored in the biobank. Such facts may describe patients, drugs, or diseases and be collected by humans or through equipment (in labs, via X-Ray, with microscopes). Biobank data can be categorized by means of data subjects (general categories to which the data belongs, such as patient, disease, or drug data), originates from data sources (human sources, laboratory sources, etc.), and is stored and represented according to data formats (textual, image, or video data).

Taken together, the data items form the *data item level* of biobank data representation. This level describes the physical samples stored in a biobank.

In [5], we discussed two main problems arising on that level motivating the introduction of a metadata level in biobanks, as proposed in [5].

1. the data on the data item level is not always directly accessible to researchers due to privacy restrictions and other factors; in most cases, only information about collections or biobanks taken as a whole (and not about individual data items) is available for search; such information does not belong to the data item level, as it does not describe individual samples.
2. the data schema is not homogeneous within a specific biobank, e.g., the set of data item attributes can be different for different collections. The domain is characterized by severe heterogeneities of the representation of data items [13].

2.1.1. Data Item Quality in Biobanks

Data item quality in biobanks is discussed in detail in [10]. It can be assessed either by calculating values of data item quality metrics for separate data items or by aggregating such metric values over whole collections or biobanks. The metrics which can be applied for such a purpose quantify *data item quality characteristics*.

In [10], we defined seven data item quality characteristics applicable to the biobank domain: *data item completeness, accuracy, reliability, consistency, timeliness, precision*, and *provenance*. We will address the treatment of the values of data item quality metrics in the metadata context later in this paper.

2.2. Metadata

We follow [16] in defining the metadata as "*the information that provides the context and additional information about the domain data or conditions on the usage of data*".

Let us look at the above definition in more detail. The context of the domain data includes among other elements:

- the information on data sources and data collection methods;
- the descriptions of the external data.

The additional information about the domain data includes, among other elements:

- the semantics necessary to understand the data, e.g., referring to some ontology [17];
- the information on the units of measurement which are used for data interpretation, e.g., meters or grams.
- the metric values characterizing the quality of data, e.g., percentage values of data completeness metrics, or expert-based estimates of data reliability metrics.

The conditions on the usage of data include, among other elements:

- the information on data attributes referring to their meaning, source, etc., e.g., its precision or the collection method;
- the descriptions of the domains which specify the allowed value sets, e.g., the disease code domain defining the set of allowed disease codes;

The metadata differs from data in purpose and usage, but not in format or structure, so it can be considered the data itself: "the data that provides information about other data".

It is possible to distinguish the following categories of metadata:

- metadata describing the data schema; this is the approach most widely used for databases (*database metadata*); it often assumes that the data is homogeneous, so the schema is centralized and stable;
- metadata describing data semantics, e.g., by ontological means (*semantic* or *content metadata*); this is the approach most widely used for digital libraries;
- metadata describing data quality (*quality metadata*), the quality metadata elements may hold the values of metrics that quantify data item quality characteristics.

The most general definition of metadata quality assumes that the metadata always serves a specific purpose (e.g., it can describe the data schema or provide data semantics, support the search for samples, etc.). It is stated as follows: *metadata quality is the degree of success for the metadata in serving its purpose*. Metadata of high quality serves its purpose better than low-quality metadata. This is the kind of quality we will deal with in this paper so we will provide a more detailed metadata quality definition specific to the biobank domain later in the paper.

In the next section, we provide a detailed review of state of the art in the area of metadata in medical research and the quality of such metadata.

3. State of the Art

In this section, we review the existing body of work on metadata quality not specific to medical or biobanking domains. Until now, biobank metadata was not addressed in detail as such, so the existing work in this area is limited to the research on metadata describing different kinds of resources, e.g., library metadata or digital heritage resources metadata. In addition, some generalizations are available without any attempts to operationalize the defined concepts for the biobanking domain.

On the other hand, the specifications and standards specifically targeting the biobanking domain describe forming the data item attributes connected to biobank samples (this is, e.g., the scope for BRISQ [2]), possibly through the measurement process (codified, e.g., by means of SPREC [1]), as a result, the compliance to such standards can be treated as a sample or data item quality characteristic not related to metadata quality. Furthermore, the MIABIS specification [4], which deals with data structures for describing biobank and biobank collection data, does not currently include the notion of data or metadata quality. A proposal for adding data item quality is currently under consideration.

In this section, we limit ourselves to the works which offer the most comprehensive treatment of this topic, postponing dealing with the existing work on concrete quality characteristics and metrics until Section 6, which describes the quality characteristics for the biobank metadata. There, the papers on specific quality characteristics are listed in the corresponding subsections. We do not deal in detail with the papers which provide quality models or further theoretical treatment of the matter; this will be the subject of the follow-up paper.

The quality of the semantic metadata evaluation is dealt with in detail by Lei et al. in [17]. The authors introduce the following two-fold definition of the quality for metadata which is founded in ontological knowledge: fitness for capturing of the data sources and fitness in instantiating the ontologies they subscribe to. The authors define the quality criteria for the semantic metadata based on the degree of accuracy in the representation of its three base elements: the data sources, the real world, and the underlying ontologies.

A framework for metadata quality assessment is introduced by Margaritopoulos et al. in [18]. The assessment considers three metadata characteristics with respect to the resources it describes: correctness, completeness, and relevance. For correctness, it concentrates on semantic correctness, omitting low-level (syntactic) correctness. The authors define the logic rules and the approach for combining such rules to form the assessment

framework. The rules concentrate on assessing the quality of the metadata description of the dependencies (relations) between the resources.

The extensive set of criteria for metadata quality assessment is introduced by Bruce et al. in [19]. This paper defines the continuum for metadata quality consisting of the following:

1. The set of tasks to be accomplished by means of metadata: find entities based on search criteria, identify entities (i.e., confirm that the found entity is the one which was requested), select the entity appropriate for the user's needs, and acquire the selected entity (purchase, etc.);
2. The set of characteristics of metadata quality (referred to as measures, the detailed list of such measures includes completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility);
3. The groups of more concrete criteria (questions to be answered) belonging to these characteristics;
4. The compliance indicators for such measures (e.g., the presence of documentation).

For example, for the quality measure of completeness:

1. the quality criteria reflect the degree to which the element set completely describes the objects;
2. the set of quality indicators is defined to contain application profile and the presence of documentation;
3. the set of improvement suggestions contains only informal advice (better documentation, adding application profiles, cultural change, etc.)

Further advancing the ideas first presented by Stvilia et al. [20], a detailed framework for metadata quality assessment is proposed by the same authors in [21].

A specific approach for measuring quality in metadata repositories was introduced by Gavrilis et al. [22].

The measurement framework introduced by Kiraly [23,24] defines three additional sets of metadata quality indicators (as compared to standard measures): supporting functional requirements, patterns and anti-patterns, and multilingualism.

Goncalves et al. [25] deal with digital libraries, i.e., the ones containing digital objects and their metadata. Some quality characteristics such as relevance or accessibility are specified for the digital objects, whereas accuracy, completeness, and conformance are specified for the metadata records, and completeness and consistency—for the metadata catalogs.

Radulovic et al. [16] limit themselves to the issue of the linked data quality; its quality model includes a set of aspects for applying the model; this set includes the metadata aspect defined as referring "to the information that provides the context and additional information about the domain data or conditions on the usage of data." Some metadata quality characteristics are listed, such as provenance.

An extended set of metrics to assess the quality of metadata in digital repositories was proposed by Ochoa and Duval [26]. The authors concentrated on quality as fitness for search. The list of quality characteristics does not differ from those presented by Bruce et al. in [19], every characteristic is accompanied by a set of metrics that are supposed to be calculated automatically, without human intervention. This set of metrics is applied by Romero-Pelaez et al. [27] to calculate the metadata quality of open courseware.

In the context of the quality of metadata for research data repositories, Koesten et al. [28] described quality requirements for such metadata (also related to relevance and usability), which can be used to establish a list of corresponding quality characteristics. A survey of work related to the metadata quality of such repositories (concentrating primarily on reference/citation databases) is provided by Strecker et al. [29].

J. Park [30] offers a survey of metadata quality assessment techniques. In the proposed definition of metadata quality, support for "discovery, use, provenance, currency, authentication, and administration" is listed as the most important metadata function. After providing a mapping between quality frameworks proposed by Stvilia et al. [20]

and Bruce et al. [19], it distinguishes completeness, consistency, and accuracy as the most commonly used metadata quality characteristics.

A. Tani et al. [31] offer a survey of metadata quality research limited to digital repositories. The authors provide some insights on handling metadata quality in federated environments, where dealing with large volumes of heterogeneous metadata becomes an issue. In particular, it is noted that the metadata considered high-quality locally can become problematic in a federated environment because the criteria applicable in such environments may be different, as well as the purpose of metadata. Furthermore, the assumed knowledge can be lost while going from the local to the federated environment. The authors emphasize the importance of metadata interoperability in such environments. This view is also supported by Stvilia et al. in [20,21].

Wilkinson et al. [32] and GO FAIR Metrics group [33] (in detail) describe the approach to define metrics exemplifying FAIR principles (findability, accessibility, interoperability, and reusability). The proposed set of metrics characterize

1. the availability of publicly available machine-readable metadata and the metadata longevity,
2. the support for identifier management and the possibility for the public registration of the identifier schemes for the metadata,
3. the existence of authorization procedures and secure access protocols,
4. the degree of support for the knowledge representation languages,
5. the availability of the licensing process,
6. the provenance specifications,
7. the evidence of ability to find the digital resource in search results,
8. the linking to other resources and the degree of support of the same FAIR principles for the linked resources,
9. the degree of meeting community standards.

Scheidlin et al. [34] aim at making it possible to create metadata complying with FAIR principles for a specific domain. To support this, they evaluate the existing metadata against the following criteria: richness (similar to the completeness, supporting usefulness), consensus (coherence from the point of view of the source, level of agreement in the community), accessibility and transparency (ease of use for different stakeholders, supporting usefulness and findability via providing different levels of granularity, supporting open standards, etc.), providing linked metadata (supporting compatibility and interoperability), functional implementation (including the security of use via authorization and authentication).

The problem with all the above papers, and, in particular [20–22], is that they target metadata repositories in general, ignoring the specifics of the biobank domain. Up to now, the problem of establishing the framework for collecting and assessing biobank metadata quality was not the target of the state-of-the-art research. This is unfortunate, as the general treatment of quality in metadata repositories is not always applicable to biobank metadata quality, for example, it does not reflect the specifics of biobanks as data brokers, the specific goals of biobank metadata, and the specific structure of its quality. Furthermore, no paper addresses the exact set of metadata quality characteristics and metrics applicable to the biobank domain; some described characteristics may not be applicable to this domain, whereas some biobank-specific characteristics can be missing.

Our paper addresses this void by proposing a comprehensive framework specifically targeting biobank metadata quality, taking into account the specific goals of biobank metadata, its structure, and the structure of its quality. Furthermore, we describe an exact set of quality characteristics and metrics specifically applicable to the biobank domain.

4. Biobank Metadata

In this section, we introduce a specific kind of metadata that describes data in biobanks. This is the kind of metadata for which we will define the notion of quality further in this paper.

4.1. Motivation and Goals for Biobank Metadata

4.1.1. Metadata Motivation

There are two main problems with the biobank data which make introducing biobank metadata necessary (we first discussed these problems in [5]):

1. A significant percentage of data served by a biobank does not originate there (coming from external sources), so it serves as a data broker for this external data. In this capacity, biobanks use and publish the data originating from external sources, providing mediating services and descriptions if the data is incomplete. Such descriptions form a part of the biobank metadata. The main function of the biobank with respect to such external data is to provide appropriate metadata to describe it. The metadata, in this case, is in the realm of responsibility for the biobank.
2. In biobanks, the direct access to the data can be problematic due to privacy and other reasons, so that the data can be not accessible to the researchers by direct search. To solve this problem, it is necessary to establish metadata to facilitate the search. The main function of the biobank with respect to such inaccessible data is to provide an appropriate metadata which
   - can be searched in place of the original data to find collections or biobanks with certain characteristics;
   - can support further decisions on discovered artifacts (e.g., if the discovered samples and data can support an envisioned study).

   In our treatment, we exploit the similarity with the library management domain [5], where library search also checks only the descriptions of books without going into the books themselves.

4.1.2. Metadata Goals

Based on the above, again following [5], it is possible to distinguish the following main goals for the biobank metadata:

1. to support the biobank in describing the data not originating from the biobank itself;
2. to support the researcher in their search for biobank collections without direct access to sample data;
3. to support the researcher in their decisions if the specific biobank collection (possibly a part of the search results) is relevant or not.

The following examples illustrate biobank metadata with respect to the above metadata goals:

1. The biobank obtains the patient data from the external patient registration system and annotates it with semantic information describing the purpose of every patient attribute. This is the example of the semantic metadata which addresses the first metadata goal (supporting the description of the external data).
2. The biobank collections store COVID test result data. In this case, they all can be accompanied by the metadata, which stores the probability for this data to be collected with the same test. This is an example of the test consistency metadata. As such metadata makes it possible to search for the biobank collections possessing specific consistency values; it addresses the second metadata goal (supporting the search).
3. The biobank states that the HIV status data is only defined for 5% of samples in a collection, and this metadata value is present for a collection. This is an example of HIV status completeness metadata. As such metadata enables the researcher to decide if the specific collection is relevant for their research; it addresses the third metadata goal (making the researcher able to decide on collection relevance).

As these goals can be addressed to a different degree, it leads to the concept of metadata quality in biobanks, to be introduced in Section 5.

4.2. Categories for Biobank Metadata

In this section, we show how the metadata categories defined in Section 2.2 can be applied to the biobank metadata.

4.2.1. Database Metadata in Biobanks

This kind of metadata is used to describe the schema of the biobank data. The problem with implementing such metadata in biobanks is related to the fact that the biobank data is heterogeneous, so in most cases, it is not possible to provide a single stable schema for such data, as the schema of such data can vary from collection to collection.

Another problem with this kind of metadata is that, being limited to describing the formal schema of the biobank data, and not its semantics, it is not well suited to addressing the metadata goals, namely the goal of supporting the search for collections when access to the data on the data item level is not allowed, as the knowledge of the data schema is usually not very helpful for such a search.

As a result, this kind of metadata is rarely used to describe biobank collection data, and we will not cover its quality here.

4.2.2. Semantic Metadata in Biobanks

This kind of metadata (also called content metadata) is used to describe the semantics of the biobank data by means of ontologies. With this approach, the specific ontological concepts (e.g., belonging to some reference ontology) are connected to data items to describe their content. Using such metadata in biobanks is more appropriate as it describes the heterogeneous data in a more flexible way.

An example of using a reference ontology to form semantic metadata, covered in more detail further in this paper, is an approach of applying the LOINC standard for such a purpose, where the specific LOINC codes, or the values of the LOINC parts are connected to data items to describe their content.

It is important to note that such semantic metadata can be connected not only to separate data items, but also to whole collections and biobanks, forming collection-level and biobank-level semantic descriptions, i.e., collection-level and biobank-level metadata; we will deal with such aggregated metadata in more detail in the next section.

4.2.3. Quality Metadata in Biobanks

Based on the representation of biobank data as a set of biobank data items, in [10] we proposed to treat the values of data item quality metrics as metadata. Such *biobank quality metadata*, according to [10], consists of the values for the metrics defined for the data item quality characteristics listed in Section 2.1.1 (metrics for data item completeness, data item reliability etc.) calculated for the specific data items.

It is also important to note that in addition, such quality metadata can contain the values of aggregated quality measurements calculated over whole collections or biobanks; we again will deal with such metadata in more detail in the next section.

4.3. Aggregation Levels for Biobank Metadata

Now, we should take into account that metadata values can be connected to the biobank data on different levels of aggregation and can be calculated by applying aggregations over the biobank data on different levels (see the discussion on aggregated semantic and quality metadata in the previous section).

Based on that, it is possible to distinguish the following aggregation levels for biobank metadata: *data item metadata*, *collection-level metadata*, and *biobank-level metadata*.

4.3.1. Data Item Metadata

This kind of metadata describes separate data items corresponding to the specific samples, each data item consisting of a set of values for specific attributes. Examples of such metadata are as follows:

- the descriptions of the sources of the data items;
- the specifications for data collection protocols in scientific studies specific for data items;
- the descriptions of the data items themselves such as time or method of collection, etc.;
- the quality metric values calculated for data items, such as the values for their data precision, reliability, and other metrics;

4.3.2. Collection-Level Metadata

This kind of metadata describes the data related to the whole collections (treated as atomic data entities, which cannot be decomposed further). Examples of such metadata are as follows:

- the descriptions of the sources of the data for the whole collections;
- the specifications for data collection protocols in scientific studies specific for collections;
- the descriptions of the collections themselves, such as their creation or last update time etc.;
- the quality metric values calculated over collections, some based on the quality characteristics for items, such as the average data precision, or maximum reliability, etc., some calculated directly, such as understandability of the collection description.

4.3.3. Biobank-Level Metadata

This kind of metadata describes the data related to the whole biobanks (treated as atomic entities). Examples of such metadata are as follows:

- the descriptions of the sources of the data for the whole biobanks;
- the specifications for data collection protocols in scientific studies specific for biobanks;
- the descriptions of the biobanks themselves such as their affiliation or contact information, etc.
- the quality metric values calculated over biobanks, some based on the quality characteristics for items, such as the average data precision, or maximum reliability, etc., some calculated directly, such as understandability of the biobank description.

5. Metadata Quality in Biobanks

In this section, we discuss the motivation for introducing biobank metadata quality, first outlined in Section 1, in detail, then we provide the definition of the metadata quality for biobanks.

To proceed with the treatment of metadata quality in biobanks, it is necessary to agree upon its *definition*. In achieving this, it is necessary to

1. establish the relationship between metadata quality and metadata goals,
2. define its *structure*, i.e., its specific elements and their relationships, and its possible *aggregation levels*,
3. agree on a set of *metadata quality characteristics and metrics*, and their application on different aggregation levels.

Further in this section, we provide a verbal definition of metadata quality in connection to metadata goals and define the structure of biobank metadata quality, i.e., its specific elements and their relationships, together with its possible aggregation levels. Introducing and discussing metadata quality characteristics and metrics on different levels of aggregation will be the topic of the next section.

5.1. Addressing metadata goals

In Section 4.1.2, we introduced three main goals for the biobank metadata:

1. supporting the description of the external data by the biobank while serving as a data broker;

2. supporting the user in the search for collections while sample data is not available;
3. supporting the user in making decisions on found artifacts.

These goals can be reached to a different degree by different metadata, e.g., some metadata allows for easy search and offers a clear view of the collections which are offered to researchers in a specific biobank, some can hinder the search, making it cumbersome and obscuring the view. Therefore, it seems natural to establish a way of measuring the efficiency of reaching these goals by the specific metadata. This leads us to the notion of *metadata quality in biobanks*, to be introduced in the next section.

5.2. Introducing Metadata Quality in Biobanks

To distinguish between "good" and "bad" metadata with respect to its goals, we propose to think in terms of metadata quality, assuming that the metadata of high quality supports both search and decision making better.

Introduced this way, *biobank metadata quality specifies how well the biobank metadata goals can be addressed*. It means that this quality defines:

1. *the degree of efficiency in describing and assessing the external data not originated in a biobank*; in particular, it characterizes how well the metadata represents the external data.
2. *the degree of efficiency for supporting the researchers in their search for biobank collections* (to find material and data for their research purposes) without direct access to sample data; in particular, it characterizes how well the metadata can assure that the researchers know about collections which are relevant for their aims (to see what is offered to them), so they can be found when they are needed.
3. *the degree of efficiency in supporting the researchers in their decisions about the relevance of the specific biobank collection for their research*; it means that it indicates how well the researchers can know if the collection is important for their research.

Let us look at the process of dealing with external data in the biobanks (i.e., the data produced elsewhere); while dealing with such data, the biobank has to provide a high-quality description of its quality and content; such descriptions are specific cases of the biobank metadata. This differs from dealing with the data produced by and for the biobank, where the biobank is responsible for the data quality itself.

We can illustrate such a process with the metaphor of a store that sells consumer devices (e.g., smartphones): the quality of the smartphone is outside the scope of responsibility for such a store; it is only responsible for the high-quality (complete, consistent, etc.) description of the smartphone, resolving the problems with the quality of such descriptions (dealing with complaints, measuring, etc.).

An example could be as follows. The data item of interest is the *cause of death* (possible value: a stroke), the quality characteristic of interest for this data item is *data item reliability* (with high metric values, e.g., when the cause of death was produced post-mortem by a trained pathologist).

Now we can introduce *collection-level completeness* as a quality characteristic assessing data item quality descriptions (in our case, the values of data item reliability metrics) on a collection level. We define it as the percentage of samples in the collection for which the metric value for reliability of the data item "cause of death" is available. Here, it is possible to see the different levels for quality definitions.

1. The cause of death is on the data item level; it is an attribute of a data item.
2. The reliability of the cause of death is *the metric value for the quality characteristic assessing the data item level*; it is important for the production of the data forming a part of the quality description for the data item, i.e., the corresponding metadata. Therefore, the reliability here is a data item quality characteristic (*data item reliability*).
3. The completeness of the reliability for the cause of death is the quality characteristic assessing the quality description, i.e., its metric values assess the quality of the metadata describing the produced data, this time for the whole collection. *This quality*

characteristic assesses the metadata level, it is a metadata quality characteristic (*metadata completeness*).

The above description is relevant to *quality of quality metadata* (i.e., the quality of data item quality descriptions or quality assessment results). Another kind of metadata quality is the *quality of content or semantic metadata*. For example, it can be the quality of the ontological concept describing the data item, or the quality of the connection between the reference ontology and the collection.

Assessing the quality of content metadata requires dealing with such topics as quality of ontologies, quality of ontology matching, etc.; we postpone discussing this issue to the subsequent paper, in the rest of this paper, we will mostly deal with the quality of quality metadata. Note that some of the metadata quality characteristics considered in this paper (such as metadata completeness or coverage) can apply to the content metadata as well.

While dealing with search and decisions on found artifacts, continuing our connection to the library management domain, we can state that the biobank search is similar to the library search, where the latter checks only the descriptions of books without going into the books themselves. The realms of responsibility in supporting the search are also similar for the biobank and the library. Both are mostly responsible for the quality of the descriptions of the artifacts whose quality is not under their control (books in a library or external/inaccessible data in a biobank), and not in the quality of these artifacts themselves, and these descriptions are supposed to aid the search. Based on all that, the quality of descriptions, i.e., the metadata quality, can be interpreted as fitness for search.

Extending the examples from Section 4.1.2, the following examples further illustrate metadata quality with respect to the metadata goals:

1. For the case, when the biobank obtains the patient data from the external patient registration system and annotates it with semantic information describing the purpose of every patient attribute, suppose such metadata is defined *for most of the attributes*. This is an example of the good quality of metadata (*high completeness of description values*) with respect to the first metadata goal (supporting the description of the external data).
2. For the case when biobank collections store COVID test result data accompanied by the metadata, which stores the probability for this data to be collected with the same test *for most of the samples*. It allows to search for collections that possess specific values for such a probability (e.g., by issuing queries "find the collections with COVID test result collected with homogeneous tests", "find the collections with COVID test result collected with heterogeneous tests"). This is an example of the good quality of metadata (*high completeness of test consistency metadata*) with respect to the first metadata goal (supporting the search).
3. For the case when the biobank states that the HIV status data is only defined for 5% of samples in a collection, and this metadata value is present for a collection, the correspondent collection is not suitable for research that requires such status to be available. This is an example of the good quality of metadata (*high completeness of HIV status completeness metadata on the collection level*) with respect to the second metadata goal (making the researcher able to decide on collection relevance).

5.3. The Structure of Metadata Quality in Biobanks

We propose quite a simple structure for organizing metadata quality in biobanks, mainly following established standards such as ISO/IEC 25012 (Data Quality model) [35] and ISO/IEC 25024 (Measurement of data quality) [36]. It can be defined as follows:

1. The metadata quality as a whole consists of a set of *metadata quality characteristics*. Quality characteristics help in achieving proper understanding of metadata quality, reflecting different aspects of this quality. Examples of metadata quality characteristics can be metadata completeness, consistency, reliability, or provenance. Such characteristics can be used together to obtain the integrated metadata quality.

2. Metadata quality characteristics are supplemented with *metadata quality metrics*, such metrics quantify the degree of achievement for the metadata goal related to the particular quality characteristic. Applying such a metric to measure the quality of a specific metadata element launches an instance of the metadata quality measurement process, which produces the specific metadata quality measurement value connected back to the measured metadata element. An example of such a metric can be the length of the interval of time between establishing a data item and the corresponding metadata element (a metric for metadata timeliness).
3. Metadata quality characteristics affect each other positively or negatively; some of the characteristics boost the impact of certain other characteristics, some—suppress such an impact. The direction and possible quantitative assessment of the affect is a property of the relation between a certain pair of quality characteristics.

The metrics can be applied to metadata values defined on different levels of aggregation (e.g., the metadata quality metric calculated over collection-level metadata). Also, they themselves can be defined on different levels of aggregation (e.g., the collection-level metadata quality metric, i.e., calculated for the whole collection). This will be discussed in more detail below.

5.4. Metadata Quality on Different Levels of Aggregation

Metadata quality metrics can be calculated for metadata elements defined for specific data items; such metrics are *item-level metadata quality metrics*. An example of an item-level metadata quality metric is the accuracy of the reliability of the cause of death calculated just for one data item attribute value. In addition to that, it is also possible to calculate *aggregated metadata quality metrics*.

As stated above, we define two levels of quality assessment values while dealing with metadata quality in biobanks:

1. data item quality characteristics and metrics [10] assess the quality of the values of data item attributes, forming *data item quality values* or *quality metadata*;
2. metadata quality characteristics and metrics assess data item quality values, forming *quality metadata quality values* or *quality of quality metadata* (as stated above, we restrict ourselves with this kind of metadata quality, omitting the treatment of the quality of content metadata, further in this paper we will refer to *metadata quality values* always assuming quality metadata quality values)

Both data item quality values and metadata quality values can be formed on different levels of aggregation: for *data items*, *collections* and *biobanks*. This can lead to complex aggregations while calculating metadata quality values over quality metadata.

5.4.1. Aggregated Data Item Quality Values as Quality Metadata

We start with possible aggregations while calculating data item quality metric values which form quality metadata. There exist three aggregation approaches to calculating such quality metadata (data item quality metric values):

1. The first and the simplest approach is the calculation of quality metadata as scalar metric values separately for each data item attribute value without aggregation. This can be done with formulas performing calculations directly on specific values. An example could be *the accuracy or reliability for a body mass index (BMI) attribute value defined for a specific sample*.
2. The second approach is the calculation of such metadata values directly on higher levels of aggregation. This can be done through formulas calculating quality values for collections or biobanks directly based on corresponding sets of data item attribute values. An example could be *the completeness of the BMI attribute values for a collection which requires the whole set of BMI values defined for all the samples for its calculation*.
3. The last approach is the aggregation of other quality metric values (e.g., calculated for separate attribute values or samples) to obtain values on higher levels of aggregation

(e.g., for collections or biobanks). This can be done through formulas aggregating quality metric values to form other quality metric values. An example could be *the average accuracy of a BMI attribute for a collection calculated based on a set of the BMI accuracy values for the samples*.

Detailed treatment of the calculation of such metric values with examples involving different aggregations is provided in [10].

5.4.2. Aggregated Metadata Quality Values as Quality of Quality Metadata

Similar aggregations may also be used while calculating metadata quality metric values, which form quality of quality metadata.

1. The first approach is calculating the quality of quality metadata as scalar metric values separately for each metadata value without aggregation. An example could be *the accuracy for the reliability of the BMI value defined for a specific sample*.
2. The second approach is the calculation of the quality of quality metadata values directly on higher levels of aggregation. This can be done through formulas calculating such values for collections or biobanks directly based on quality metadata values. An example could be *the completeness of the BMI reliability values for a collection*.
3. The last approach is the aggregation of the quality of quality metric values (e.g., calculated for separate attribute values or samples) to obtain values on higher levels of aggregation (e.g., for collections or biobanks). This can be done through formulas aggregating quality of quality metric values to form other quality of quality metric values. An example could be *the average reliability of accuracy of a BMI attribute for a collection calculated based on a set of the BMI reliability of accuracy values for the samples*.

5.4.3. Two Levels of Aggregation

As a result, we distinguish the following kinds of metadata quality metrics with respect to two levels of aggregation:

1. non-aggregated metadata quality metrics on non-aggregated quality metadata (NN-metrics). An example of such a metric is a sample-level accuracy of cause of death reliability.
2. non-aggregated metadata quality metrics on aggregated quality metadata (NA-metrics). An example of such a metric is a collection-level accuracy of average collection-level reliability of the cause of death.
3. aggregated metadata quality metrics on non-aggregated quality metadata (AN-metrics). An example of such a metric is average collection-level completeness of cause of death reliability.
4. aggregated metadata quality metrics on aggregated quality metadata (AA-metrics). An example of such a metric is a biobank-level average completeness of collection-level completeness of the cause of death.

An exact aggregation formula depends on the kind of metrics being aggregated; we deal with that while describing individual metrics.

This paper will mainly focus on defining quality characteristics and metrics for biobank metadata dealing with relatively trivial cases of aggregations. We define:

1. *sample-level aggregate metadata metrics* calculated over metadata elements related to data items defined for the particular biobank sample;
2. *collection-level aggregate metadata metrics* calculated over metadata elements related to data items defined for the set of samples of the particular biobank collection

The complete treatment of possible aggregations will be the subject of the follow-up paper.

6. Metadata Quality Characteristics and Metrics

In this section, we describe specific metadata quality characteristics and corresponding metrics in detail. For this, we selected the following nine characteristics: *metadata accuracy*, *completeness*, *coverage*, *consistency*, *timeliness*, *provenance*, *reliability*, *conformance to expectations*, and *accessibility*.

6.1. Intrinsic and Relative Metadata Quality Characteristics and Metrics

In [10], following Stvilia et al. [20], we distinguished between intrinsic and relative data item quality metrics. Similarly, this paper distinguishes between *intrinsic and relative metadata quality metrics*:

1. Metadata quality assessment based on the *intrinsic metadata quality metric* must rely only on the existing biobank metadata values, not on human opinions, standard data, or metadata schema. Such metrics are objective.
2. Metadata quality assessment based on the *relative metadata quality metric* may rely on human opinions, standard data, or the metadata schema. Such metrics are subjective.

It is also possible to define intrinsic and relative metadata quality characteristics: for such a characteristic, all possible metrics are, respectively, intrinsic or relative. In addition, mixed metadata quality characteristics may possess intrinsic and relative metrics at the same time.

6.2. Metadata Accuracy

This characteristic reflects the need for the metadata to be accurate. It contributes to addressing the metadata goals as follows:

1. high-accuracy metadata allows for a better description of the external data, being more likely to provide the correct description of the data, adequately representing it for the prospective users;
2. high-accuracy metadata supports the search better as the data item values corresponding to the found artifacts after searching within accurate metadata values are more likely to correspond to the provided search criteria;
3. high-accuracy metadata supports decisions based on the found artifacts better as these decisions are more accurately based on the data item values corresponding to the artifacts.

6.2.1. Related Work on Metadata Accuracy

The following authors address metadata accuracy in their papers:

1. Bruce et al. [19] defined metadata accuracy as accuracy in values, correspondence to factual information;
2. Margaritopoulos et al. [18] defined metadata accuracy as semantic correctness in describing relations;
3. Lei et al. [17] defined metadata accuracy with respect to data sources and with respect to the real world (the accuracy of labeling, the accuracy of classification), also they defined non-spuriousness as contributing to accuracy;
4. Gavrilis et al. [22] also defined accuracy with respect to the real world;
5. Goncalves et al. [25] calculated accuracy values for the metadata records using a specific formula;
6. Radulovic et al. [16] defined *semantic accuracy* (in representing the real-world facts) and *syntactic accuracy* (correspondence of metadata values to a specific domain);
7. Ochoa and Duval [26] defined accuracy as a distance between the information to be taken from the metadata description and from the data object itself; they proposed to calculate the accuracy metrics (defined only for text-based metadata) based on the semantic distance between data attribute and metadata attribute texts;
8. Neumaier et al. [37] defined accuracy as a quality characteristic for open data portals understood as correctness in representing the resource; Quarati et al. [38] proposed

similar characteristic to access the accuracy of metadata for open government data. Furthermore, these researchers proposed to define conformance as a syntactic accuracy in correspondence to the domains.

In the survey of opinions by metadata practitioners regarding the relative importance of metadata quality characteristics conducted by Gentry et al. [39], accuracy was put in second place after consistency.

6.2.2. Calculating Metadata Accuracy Metrics

We propose to limit metadata accuracy metrics by calculating only syntactic accuracy, i.e., *the degree of correspondence of the metadata values to their domain*. Assuming that the metadata values belong to the specific domain, it is calculated as follows:

$$MSAcc_{mc} = \frac{n(I_m) - \sum_{i \in I_m} f_m^{dt}(i)}{n(I_m)}, m \in M_c \quad (1)$$

where M_c is a set of all domains for the metadata values describing data item attributes instantiated by the samples in a biobank collection c, $n(I_m)$ is a cardinality of the set I_m which includes all metadata values belonging to the domain M, $f_m^{dt}(i)$ is a negative threshold function over a metadata value $i \in I_m$ related to its domain $m \in M$, which is equal to 0, if i belongs to m, 1 otherwise. An example of the domain constraint served well by such a metric is the non-negativity constraint for the data item reliability [10] value domain (*data item reliability accuracy metric*).

Suppose there are 200 reliability values available for a collection c_1 (describing different data item attributes), 14 of them being negative (violating the domain constraint). Then the reliability accuracy for this collection is calculated as follows: $MSAcc_{(reliability)c_1} = (200 - 14)/200 = 0.93$.

To calculate the metadata accuracy metric based on all domains defined for a collection, it is necessary, as for the data item accuracy metrics [10], to apply the aggregate function (e.g., average or median) to all $MSAcc_m, m \in M_c$, where M_c is a set of all domains defined for the metadata values describing data item attribute values instantiated by the samples belonging to a biobank collection c. An example could be the *collection-level domain-average-based syntactic metadata accuracy*, which is calculated by applying the following formula:

$$MSAcc_c = \frac{\sum_{m \in M_C} MSAcc_{mc}}{n(M_c)} \quad (2)$$

Suppose the collection-level reliability accuracy for the collection c_1 is equal to 0.6, whereas the completeness accuracy is equal to 0.85, and the conformance to expectations accuracy is equal to 0.8. Then $MSAcc_{c_1} = (0.6 + 0.85 + 0.8)/3 = 0.75$.

6.3. Metadata Completeness

This characteristic reflects the need to supplement all data items with the corresponding quality metadata and is assessed as the percentage of non-empty metadata values. It contributes to addressing the metadata goals as follows:

1. more complete metadata allows for a better description of the external data, as it provides such a description for a higher percentage of data values;
2. more complete metadata supports the search within metadata values better as more metadata values correspond to artifacts, so these artifacts are more likely to be found;
3. more complete metadata supports decisions based on the found artifacts better as these decisions are supported with more data values.

6.3.1. Related Work on Metadata Completeness

A significant number of authors address metadata completeness in their papers:

1. Some authors defined such completeness explicitly based on the presence of the metadata values:
   - Kiraly et al. [40] defined completeness this way for the case of digital heritage repository,
   - Margaritopoulos et al. ([41,42]) defined metadata completeness this way at the field and representation levels exemplified by the detailed set of metrics,
   - Sicilia et al. [43] defined completeness of the metadata records in learning object repositories;
2. Bruce et al. [19] defined metadata completeness by using the criteria such as "being as complete as economically possible", and "completeness in covering the whole population";
3. Margaritopoulos et al. in the further paper [18] defined metadata completeness mostly as completeness in describing the relations between resources;
4. Lei et al. [17] defined metadata completeness as annotation completeness with respect to the data sources and the real world;
5. Goncalves et al. [25] defined completeness
   - for the metadata records—as a degree to which the values of the properties are present in a description, as compared to the standard (e.g., to the Dublin core specification),
   - for the metadata catalog—based on the number of objects with complete metadata records.
6. Phillips et al. [44] defined completeness through implementing the minimally viable record;
7. Nichols et al. [45] defined metadata completeness as completeness representing the quality as a whole visually.
8. Ochoa and Duval [26] calculated completeness taking into account the relative importance of the metadata fields in a given context.
9. Neumaier et al. [37] and Neumaier et al. defined existence as a quality characteristic for open data portals understood as completeness applied to certain attributes; Quarati et al. [38] proposed a similar characteristic to access the quality of metadata for open government data.

Some researchers treated completeness metrics in more detail:

1. Margaritopoulos et al. ([41,42]) described a set of metrics to measure metadata completeness. These metrics are fine-grained, i.e., they go down to the field level. Among these metrics:
   - completeness of a metadata field (differently calculated for simple, and aggregate—single-valued and multi-valued fields by traversing the metadata schema hierarchy)
   - completeness of a metadata record (weighted average of field completeness)
   - completeness at the representation level (considered as supplementary to the field completeness, reflects the additional presentation features influencing the quality, such as completeness of the set of possible audio or video illustrations)
2. Kiraly [40] described the metrics for metadata completeness for digital heritage resources on three levels—individual records, subsets (collections), the whole datasets:
   - simple completeness and completeness of sub-dimensions—ratio of filled fields;
   - existence of fields—the ratio of available fields as compared to the source elements in a record;
   - uniqueness of the descriptive fields (title, alternative title, description);
   - multilinguality.

   This work also investigates the existence of record patterns—i.e., the groups of fields that form the "typical record".

3. Phillips et al. [44] identified a set of standard attributes for the metadata record (constituting a minimally viable record) and defined a quality metric for the degree of such attributes being present.
4. Diaz de la Paz et al. [46] proposed another approach of applying weights to calculate the completeness of metadata elements.
5. Lorenzini et al. [47] split the metadata tags into four categories (compulsory, recommended, domain-specific, optional; with a possible assignment of priorities) and calculated category-specific completeness values.

In the survey of opinions by metadata practitioners regarding the relative importance of metadata quality characteristics conducted by Gentry et al. [39] completeness was put in the sixth place, after consistency, accuracy, timeliness, accessibility, and provenance.

6.3.2. Calculating Metadata Completeness Metrics

We define metadata completeness as *a degree of completeness with respect to the metadata connected to the data values*.

In this section, we limit ourselves to calculating completeness for scalar (atomic) metadata: the values for such metadata are scalar values that cannot be decomposed further. The quality metadata is of this kind, as the quality values are atomic. An example of such metadata is the quality metadata holding values of scalar quality metrics, e.g., data item completeness metadata.

For such metadata, a binary value presence function can be used to indicate its presence for an instance of the specific metadata holder, i.e., the sample, collection, or biobank as a whole.

The difference between completeness and, e.g., reliability or provenance is that completeness does not make sense while connected to a single data item in a collection; it is reduced to a simple presence function.

Sample-Level Metadata Completeness

We define sample-level metadata completeness as *a degree of presence of sample-level metadata values for a sample.* It means that we assume that the sample itself is a metadata holder.

We propose to calculate this metadata metric as a ratio of a total number of instantiated sample-level metadata values to a total number of all metadata values declared for a sample (e.g., on the collection level).

This completeness metric is calculated as follows:

$$MCompl_s^{samp} = \frac{n(A^M_{c(s)}) - \sum\limits_{a^M \in A^M_{c(s)}} f_s^M(a^M)}{n(A^M_{c(s)})}, s \in S_c \quad (3)$$

where $A^M_{c(s)}$ is a set of metadata attributes declared for a collection $c(s)$ owning the sample s, $f_s^M(a^M)$ is a negative metadata value instantiation function which can be equal to 0 if the value for a metadata attribute $a^M \in A^M_{c(s)}$ is instantiated for s, or 1 otherwise.

Suppose there exists a collection c_1 which metadata configuration (specification) includes data item completeness, data item accuracy, data item timeliness, and data item precision, and all values except the data item completeness value are missing for the specific sample s_1. Then $MCompl_{s_1}^{samp} = (4-3)/4 = 0.25$.

Collection-Level Metadata Completeness

Metadata completeness metrics that characterize the biobank collection can be calculated as:

1. *Sample-based collection-level metadata completeness* by aggregating metadata completeness values calculated for all or a subset of the samples in a collection;

2. *Data attribute-scoped collection-level metadata completeness* over the values for a specific set of data attributes instantiated for all or a subset of the samples belonging to a biobank collection; such set can contain a single data attribute, a subset of or all data attributes defined for a biobank collection.
3. *Metadata attribute-scoped collection-level metadata completeness* over the values for a specific set of metadata attributes instantiated for all or a subset of the samples in a collection; such set can contain a single metadata attribute, a subset of or all metadata attributes defined for a biobank collection.

Sample-Based Collection-Level Metadata Completeness

Such metadata completeness metric is calculated by applying the aggregate function (e.g., average, minimum/maximum, or median) to a set of metadata completeness values calculated for all or the subset of the samples in a given collection. For example, *sample-average-based collection-level metadata completeness* is calculated by applying the following formula:

$$MCompl_c^{avgS} = \frac{\sum\limits_{s \in S_c} MC_{sc}^{samp}}{n(S_c)}, \tag{4}$$

where S_c is a set of all samples in a collection c, MC_{sc}^{samp} is a value of a sample metadata completeness calculated for a sample s belonging to the biobank collection c.

Suppose there is a collection c_1 which contains 100 samples, where 70 samples have sample-level metadata completeness of 0.6, and 30 samples have sample-level metadata completeness of 0.9 (to simplify calculations, we restricted the set of possible sample-level metadata completeness values to two members). Then $MCompl_{c_1}^{avgS} = (0.6 \cdot 70 + 0.9 \cdot 30)/100 = 0.69$.

Data Attribute-Scoped Collection-Level Metadata Completeness

We define such completeness as *a degree of metadata value presence for a specific data item attribute in a specific collection*. For example, such completeness could characterize the presence of all metadata values (for all defined data item quality characteristics) for the patient age data item attribute for a specific collection (patient age reliability, patient age accuracy, etc., taken together). It is calculated by applying the following formula:

$$MCompl_{ac}^{attr} = \frac{n(S_c) \cdot n(A_c^M) - \sum\limits_{s \in S_c} \sum\limits_{a^M \in A_c^M} f_{sa}^M(a^M)}{n(S_c) \cdot n(A_c^M)}, a \in A_c \tag{5}$$

where $f_{sa}^M(a^M)$ is a negative metadata value presence function which can be either 0 if the value for a metadata attribute $a^M \in A_{c(s)}^M$ calculated for a data item attribute a is present for s, or 1 otherwise.

Suppose there is a collection c_1 which contains 100 samples, where each data item attribute value instantiated by a sample is described by four metadata attributes, and there are 260 metadata values missing out of those describing the patient age data item attribute. Then $MCompl_{(age)c_1}^{attr} = (100 \cdot 4 - 260)/(100 \cdot 4) = 0.35$.

Metadata Attribute-Scoped Collection-Level Metadata Completeness

We define such completeness as *a degree of metadata value presence for a specific metadata attribute in a specific collection*. For example, such completeness could characterize the presence of data item accuracy or reliability values for a specific collection (describing the values of different data item attributes). It can be calculated as follows:

$$MCompl_{a^M c}^{attr^M} = \frac{n(S_c) \cdot n(A_c) - \sum\limits_{s \in S_c} \sum\limits_{a \in A_c} f_{sa}^M(a^M)}{n(S_c) \cdot n(A_c)}, a^M \in A_c^M \tag{6}$$

Suppose 300 data item accuracy values are missing for a collection c_1, which contains 100 samples, each instantiating 5 data item attributes. Then $MCompl^{attr^M}_{(accuracy)c_1} = (100 \cdot 5 - 300)/(100 \cdot 5) = 0.4$.

6.4. Metadata Coverage

Related to completeness is metadata coverage, which can be defined in the following ways:

1. as an absolute number of non-empty metadata values on a certain level of aggregation, e.g., for a collection (the larger, the better);
2. as an absolute number or a relative ratio of attributes holding non-empty metadata values on a certain level of aggregation, e.g., for a collection;
3. as an absolute number or a relative ratio of diverse values for metadata attributes present on a certain level of aggregation; e.g., it can define which number or which percent of allowed metadata values for a certain attribute is present for a collection;
4. as a synonym to metadata completeness.

It contributes to addressing the metadata goals as follows:

1. metadata with wider (higher) coverage allows for a better description of the external data, supplemented with more diverse information;
2. metadata with wider coverage supports the search within metadata values better as more diverse metadata values correspond to artifacts, so these artifacts are more likely to be found by specifying diverse search criteria;
3. metadata with wider coverage supports decisions based on the found artifacts better as these decisions are supported with richer data values, allowing for better judgment.

6.4.1. Related Work on Metadata Coverage

Metadata coverage is also addressed in the literature, though the number of papers is significantly smaller as compared to the number of papers related to metadata completeness:

1. Liolios et al. [48] defined the Metadata Coverage Index (MCI), which is a standardized metric for quantifying database metadata richness; this index was also used by Bellini and Nesi [49] in their assessment tool for open access cultural heritage repositories.
2. OLAC Metadata Metrics specification [50] included metrics to assess metadata coverage of open language archives.
3. Koesten et al. [28] listed metadata coverage as a contributing factor to metadata relevance;
4. In the biobank domain, Klie et al. [51] dealt with increasing metadata coverage of biobank entries by means of deep-learning entity recognition.

6.4.2. Calculating Metadata Coverage Metrics

As we provided several definitions of metadata coverage above, we can describe several approaches to calculating coverage metric values.

Metadata Coverage as the Absolute Number of Non-Empty Values

The first approach is to define metadata coverage as an absolute number of non-empty metadata values in a biobank on a certain aggregation level, e.g., connected to some collection or to a specific attribute. This way, the calculation of coverage metric values uses formulas similar to those used for completeness calculation, with the difference being that such metrics are absolute and not relative.

For example, we define data attribute-scoped collection-level metadata coverage as *a total number of non-empty metadata values for a specific data item attribute in a specific collection*. Similarly to the corresponding completeness metric, such coverage metric could characterize the total number of all present metadata values (for all defined data item quality characteristics) for the patient age data item attribute for a specific collection (patient age

reliability, patient age accuracy etc., taken together). It can be calculated as follows (see the explanation of the formula elements in the previous section):

$$MCov_{ac}^{attr} = \sum_{s \in S_c} \sum_{a^M \in A_c^M} h_{sa}^M(a^M), a \in A_c \tag{7}$$

where $h_{sa}^M(a^M)$ is a metadata value presence function which can be either 1 if the value for a metadata attribute $a^M \in A_{c(s)}^M$ calculated for a data item attribute a is present for s, or 0 otherwise.

Suppose there is a collection c_1 which contains 100 samples, out of which there are 70 samples with defined two metadata attribute values describing the patient age data item attribute values, and 30 samples with defined three metadata values describing the values instantiating the same data item attribute. Then $MCov_{(age)c_1}^{attr} = 70 \cdot 2 + 30 \cdot 3 = 220$.

Other metadata coverage metrics belonging to this category can be calculated similarly based on the formulas for metadata completeness.

Metadata Coverage as the Number of Attributes Holding Non-Empty Values

Taking into account the definition of metadata coverage as the number of metadata attributes holding values on the specific aggregation level (e.g., for a collection), another approach to calculating coverage is to calculate the ratio of the number of metadata attributes for which the values are present for a collection c to the total number of metadata attributes defined for c. This calculation can be made using the following formula:

$$MCov_c^{a^M num} = \frac{\sum_{a^M \in A_c^M} h_c^M(a^M)}{n(A_c^M)} \tag{8}$$

where $h_c^M(a^M)$ is a presence function that is evaluated to 1 if the metadata attribute a^M contains the non-empty value for any of the samples belonging to a collection c, 0 otherwise.

Suppose the data item attribute values instantiated by the samples in a collection c_1 are described by data item consistency, completeness, reliability, and provenance. Additionally, suppose that no data item completeness values are instantiated to describe data item attribute values for all the samples in c_1, while values for the other three metadata attributes are instantiated to describe at least one data item attribute value for some sample in c_1. Then $MCov_{c_1}^{a^M num} = 3/4 = 0.75$.

Metadata Coverage as the Number of Diverse Values for a Metadata Attribute

Taking into account the definition of metadata coverage as the number of diverse values for an attribute on the specific aggregation level (e.g., for a collection), yet another approach to calculating coverage is to calculate *the number of diverse metadata values for a collection metadata attribute*. It is not always applicable for assessing the quality of quality metadata as the number of diverse values, e.g., for data item reliability, is not very informative.

This calculation can be made using the following formula:

$$MCov_{a^M c}^{adiv} = ndiv_c(a^M), a^M \in A_c^M \tag{9}$$

where $ndiv_c(a^M)$ is a function which returns the number of diverse values for a metadata attribute a^M defined for a collection c. An example could be the number of diverse data item accuracy values for all samples in a collection.

Such coverage can also be calculated as a relative metric as follows:

$$MCov_{a^M c}^{adivr} = \frac{ndiv_c(a^M)}{nall_c(a^M)}, a^M \in A_c^M \tag{10}$$

where $nall_c(a^M)$ is a function that returns a total number of possible metadata values defined for the domain of a^M. This metric makes sense only for the cases when the number of possible values for the metadata attribute is finite.

Suppose the data item attribute values instantiated by the samples in a collection c_1 are described by the data item consistency metric, which takes three diverse values for all samples in c_1, whereas the total number of possible values for the data item consistency domain is ten. Then $MCov^{adivr}_{(consistency)c_1} = 3/10 = 0.3$.

6.5. Metadata Consistency

This characteristic reflects the need for the metadata to be consistent, e.g., corresponding to the same standard. An additional characteristic related to metadata consistency is *metadata coherence* which can be defined as a degree of uniformity of the metadata values.

Metadata consistency contributes to addressing the metadata goals as follows:

1. highly consistent metadata allows for a better description of the external data, as it provides such a description more uniformly, free of conflicts and ambiguity;
2. highly consistent metadata supports the search within metadata values better as more metadata values are likely to match the consistent search criteria, there is less possibility that the specific criteria will match only some consistent subset of the data;
3. highly consistent metadata supports decisions based on the found artifacts better as these decisions are supported with unambiguous and conflict-free data.

6.5.1. Related Work on Metadata Consistency

The following authors address metadata consistency in their papers:

1. Bruce et al. [19] defined metadata consistency (1) with respect to the current collection or to the different versions the current collection (is all the data consistent?); (2) with respect to the data from other collections or the publicly available (community) data;
2. Gavrilis et al. [22] defined the degree of consistency for the metadata values;
3. Lei et al. [17] defined consistency as (1) lack of duplication w.r.t data sources and w.r.t real world (2) non-ambiguity w.r.t data sources (3) consistency in compliance to ontologies;
4. Goncalves et al. [25] defined consistency (1) for the metadata records—as conformance to the schema (i.e., the records are supposed to have types defined by the schema) (2) for the metadata catalog—as the degree of having the same metadata specifications assigned to multiple digital objects;
5. Radulovic et al. [16] defined metadata consistency mostly in linked data context, also as metadata compliance.
6. Ochoa and Duval [26] defined consistency as the degree of matching the standard definition for the metadata element measured by counting the number of violations of the standard, whereas coherence is measured as a semantic distance between different fields belonging to the same metadata element, e.g., between title and description of the data item.
7. Griffiths et al. [52] identified inconsistency as the main problem in sharing the metadata describing microbiological datasets related to food safety and proposes using standards and common ontologies as a means of harmonization of such metadata.
8. In the medical domain, Zaveri et al. [53] dealt with quality assessment of biomedical metadata, on the example of Gene Expression Omnibus (GEO) database, which defines metadata as key-value pairs. The authors identified the most frequent quality problems for both keys and values being ambiguity and lack of consistency.

In the survey of opinions by metadata practitioners regarding the relative importance of metadata quality characteristics conducted by Gentry et al. [39], consistency was distinguished as the most important characteristic.

6.5.2. Calculating Metadata Consistency Metrics

In defining metadata consistency metrics, we follow the approach introduced in [10] to measure data item consistency and limit ourselves to calculating collection-level metadata consistency metrics. We define collection-level metadata consistency as *a reverse degree of variability with respect to the metadata standard within a collection*. This reflects that the collection metadata is more consistent if it conforms to fewer standards. Further in this section, we describe some collection-level consistency metrics.

Metadata Attribute-Scoped Collection-Level Metadata Consistency Degree Based on the Number of Standards

This metric is calculated based on the number of standards all metadata values instantiating a specific metadata attribute belonging to a collection conform to:

$$MConsT_{a^M c}^{totA^M} = 1 - \frac{n(T_{a^M})}{\max\limits_{a^{M'} \in A_c^M} n(T_{a^{M'}})}, a^M \in A_c^M, \tag{11}$$

where T_{a^M} is a set of standards all metadata values instantiating the metadata attribute a^M conform to, A_c^M is a set of all metadata attributes declared for a collection c.

Suppose the data attribute values for the samples in a collection c_1 are described by the data item consistency metadata attribute conforming to 2 standards, data item reliability metadata attribute conforming to 5 standards, and data item completeness metadata attribute conforming to 1 standard. Then, $MConsT_{(consistency)c_1}^{totA^M} = 1 - 2/\max(2,5,1) = 0.6$.

Metadata Attribute-Scoped Collection-Level Metadata Consistency Degree Based on the Most Frequently Applied Standard

This metric is calculated as a ratio of the number of values instantiating a specific metadata attribute belonging to a biobank collection conforming to the most frequently applied standard to the total number of values instantiating this metadata attribute:

$$MConsT_{a^M c}^{freqA^M} = \frac{\max\limits_{t \in T_{a^M}} n(V_{ta^M}^M)}{n(V_{a^M}^M)}, a^M \in A_c^M \tag{12}$$

where T_{a^M} is a set of standards all metadata values instantiating the metadata attribute a^M conform to, $V_{a^M}^M$ is a full set of metadata values instantiating the metadata attribute a^M for the biobank collection, $V_{ta^M}^M$ is a subset of this set consisting of the metadata values conforming to the standard t.

Suppose there are 100 data item completeness values describing data item attribute values in a collection c_1, out of which 70 data item completeness values conform to the data item completeness standard S1, and 30 completeness values conform to the data item completeness standard S2. Then, $MConsT_{(completeness)c_1}^{freqA^M} = \max(30,70)/100 = 0.7$.

Metadata Attribute-Based Collection-Level Average Metadata Consistency Degree

This metric is calculated as an average of the metadata attribute-scoped metadata consistency values for all metadata attributes declared for a collection:

$$MConsT_c^{avgtotA^M} = \frac{\sum\limits_{a^M \in A_c^M} MConsT_{a^M}^{totA^M}}{n(A_c^M)}, \tag{13}$$

or

$$MConsT_c^{avgfreqA^M} = \frac{\sum\limits_{a^M \in A_c^M} MConsT_{a^M}^{freqA^M}}{n(A_c^M)}, \tag{14}$$

Suppose the data attribute values for the samples in a collection c_1 are described by the data item consistency metadata attribute conforming to 3 standards, data item reliability metadata attribute conforming to 5 standards, data item completeness metadata attribute conforming to 2 standards, and data item precision metadata attribute conforming to 1 standard. Then, $MConsT_{c_1}^{avgtotA^M} = ((1-3/\max(3,5,2,1)) + (1-5/\max(3,5,2,1)) + (1-2/\max(3,5,2,1)) + (1-1/\max(3,5,2,1)))/4 = (0.4+0+0.6+0.8)/4 = 0.45$.

Data Item Attribute-Based Metadata Consistency Metrics

These metrics can be calculated, e.g., by aggregating consistency values for all metadata attributes assessing the quality of the values instantiating the specific data item attribute; we will not cover such metrics here due to the lack of space.

6.6. Metadata Timeliness

This characteristic addresses the need for the metadata to reflect the real state of the data items. It can be defined as the length of the time interval between the change in a data item and the time when the metadata reflects that change.

Metadata timeliness contributes to addressing the metadata goals as follows:

1. highly timely metadata allows for a better description of the external data, as such a description better corresponds to the reality and has a lower chance of becoming obsolete;
2. highly timely metadata supports the search within metadata values better as the search is more likely to return relevant and current collections;
3. highly timely metadata supports decisions based on the found artifacts better as these decisions are supported with current data.

Related to timeliness is *metadata volatility* which characterizes instability of the metadata elements, i.e., the frequency of metadata changes. It, as a rule, reflects the changes in quality of the underlying data items. We will not address it in this paper due to lack of space.

6.6.1. Related Work on Metadata Timeliness

The following researchers address metadata timeliness in their papers:

1. Bruce et al. [19] defined timeliness instantiated as
   - *currency* which is the length of the period when the object is changed, but the metadata does not reflect it;
   - *lag*, which is the length of the period when the object is made available, but the metadata is not yet ready.

   The authors state that in dealing with timeliness, it is possible to take into account the cultural differences between collection maintainers and developers (IT people vs. library people, maybe this is also applicable for the IT people vs. biobank people relation). For example, library people see data as stable, so they tend to avoid changes if not absolutely necessary, whereas IT people are more flexible with respect to such changes. Neither of the approaches is seen as preferable.
2. Lei et al. [17] defined timeliness as time-accuracy;
3. Radulovic et al. [16] defined timeliness in a standard way as a part of the set of contextual dimensions.
4. Ochoa and Duval [26] defined timeliness as the ability of the metadata instance to keep its quality over time. They proposed to measure it based on the number and the direction of changes in the values of its quality metrics over time.

In the survey of opinions by metadata practitioners regarding the relative importance of metadata quality characteristics conducted by Gentry et al. [39] timeliness was put in the third place, after consistence and accuracy.

6.6.2. Calculating Metadata Timeliness Metrics

Both data attribute values, and their metadata attribute values are accompanied by creation times. Therefore, it makes sense to calculate metadata timeliness based on the distance between these two types of values, e.g. as *a reverse distance of time between creating the data attribute value and its supplementing metadata values*.

An example of low metadata timeliness is the case when collecting the information about the standard the data item conforms to is done two years after storing the data item: this is usually unacceptable as the standard information can become forgotten or lost in a course of time.

Sample-Level Metadata Timeliness

Sample-level metadata timeliness metric can be calculated, e.g., as *the average metadata attribute-scoped metadata timeliness for a specific sample* by aggregating timeliness distances for all possible "data item attribute value—metadata value" pairs, with every pair consisting of a data item attribute value connected to the specific sample, and the metadata value instantiating the specific metadata attribute, connected to this data item attribute. In this case, the creation time must be available for both elements of the pair.

$$MTimel_{sa^M}^{avgM} = \frac{\sum\limits_{a \in A_{c(s)}} (1 - d_s^M(a, a^M) / \max\limits_{a \in A_{c(s)}} d_s^M(a, a^M))}{n(A_{c(s)})}, a^M \in A^M \tag{15}$$

where $A_{c(s)}$—a set of all data attributes declared for the biobank collection $c(s)$ containing the sample s, for which the information about the creation time is available, $d_s^M(a, a^M)$ is a distance between the creation time of the value instantiating the data item attribute a and the creation time of the value instantiating the metadata attribute a^M, both for the sample s, where the metadata attribute a^M is connected to the data item attribute a.

Suppose the collection data item attributes include patient age, disease code, and BMI, each described by the data item reliability metadata attribute value, and the interval of time between storing the attribute value and the data item reliability value for the specific sample s_1 is 2125 seconds for patient age, 1000 seconds for disease code, and 2500 seconds for BMI. Then $MTimel_{s_1(reliability)}^{avgM} = ((1 - 2125/2500) + (1 - 1000/2500) + (1 - 2500/2500))/3 = (0.15 + 0.6 + 0)/3 = 0.25$.

Finding the Data Item Attribute with Minimal Metadata Timeliness for a Specific Sample

A sample-level metadata timeliness "bottleneck" for a specific metadata attribute can be discovered by finding *the data item attribute with minimal timeliness with respect to this metadata attribute*. For the case when a sample-level timeliness for a specific metadata attribute is calculated by means of the formula (15), the minimal value for this timeliness is equal to 0 (when $d_s^M(a, a^M) = \max\limits_{a \in A_{C(s)}} d_s^M(a, a^M)$), so the corresponding data item attribute can be defined as follows:

$$a_{sa^M}^{minMTimel} = \{a | d_s^M(a, a^M) = \max_{a \in A_{c(s)}} d_s^M(a, a^M)\}, a^M \in A^M \tag{16}$$

For the previous example $a_{s_1(reliability)}^{minMTimel} = \{a | d_{s_1}^M(a, reliability) = \max(2125, 1000, 2500)\}$ which is evaluated to BMI.

It is also possible to find the data item attribute and metadata attribute, resulting in minimal timeliness for the whole sample.

Collection-Level Metadata Timeliness

Aggregate metadata timeliness can be also calculated on the collection level. In this section, we describe the calculation of several collection-level timeliness metrics.

Data attribute-scoped average collection-level metadata timeliness aggregates all metadata timeliness values calculated for a specific data item attribute (first element of the pair) for all samples in a collection. It can be calculated as follows:

$$MTimel_{ac}^{attr} = \frac{\sum_{s \in S_c} \sum_{a^M \in A_c^M} (1 - d_s^M(a, a^M) / \max_{a \in A_c} d_s^M(a, a^M))}{n(S_c) \cdot n(A_c^M)}, a \in A_c \quad (17)$$

Suppose the interval of time between storing the BMI data item attribute value and the metadata attribute value is 1500 seconds for 350 metadata values describing BMI data item attribute values in the collection c_1 and 2500 seconds for all other metadata values describing BMI data item attribute values defined for the same collection. Additionally, suppose that the collection c_1 contains 250 samples, each instantiating data item attributes described by two metadata attributes. Then $MTimel_{(BMI)c_1}^{attr} = (350 \cdot (1 - 1500/2500) + 150 \cdot (1 - 2500/2500))/(250 \cdot 2) = 0.28$.

Metadata attribute-scoped average collection-level metadata timeliness aggregates all metadata timeliness values calculated for a specific metadata attribute (second element of the pair) for all samples in a collection. It can be calculated as follows:

$$MTimel_{a^M c}^{attr^M} = \frac{\sum_{s \in S_c} \sum_{a \in A_c} (1 - d_s^M(a, a^M) / \max_{a \in A_c} d_s^M(a, a^M))}{n(S_c) \cdot n(A_c)}, a^M \in A_c^M \quad (18)$$

Suppose the interval of time between storing the data item attribute value and the data item completeness metadata attribute value is 1000 seconds for 250 data item completeness metadata values describing data item values in the collection c_1 and 2500 seconds for all other data item completeness metadata values defined for the same collection. Additionally, suppose that the collection c_1 contains 250 samples, each instantiating four data item attributes. Then $MTimel_{(completeness)c_1}^{attr^M} = (250 \cdot (1 - 1000/2500) + 750 \cdot (1 - 2500/2500))/(250 \cdot 4) = 0.15$.

It is also possible to define sample-based collection-level metadata timeliness by aggregating sample-level timeliness values with the aggregation formula similar to (4).

Finding the Data Item Attribute Value with Minimal Metadata Timeliness in a Collection

It is also possible to discover the collection-level metadata timeliness "bottleneck," i.e., the data item attribute value with minimal metadata timeliness in a collection. It is done by finding a "data item attribute value—metadata attribute value" pair characterized by minimal timeliness among all such pairs defined for all samples in a collection.

It is done by extending the Formula (16) to search within the set of all "data item attribute value—metadata attribute value" pairs within a collection; we omit the formula and the example here due to the lack of space.

6.7. Metadata Provenance

Similarly to data item provenance [10], metadata provenance is the degree of linking between the metadata sources, collection methods, and standards on the one side and the metadata values on the other side. For the metadata of low provenance (contributing to low metadata quality), it is problematic to see its origin and the standard it conforms to.

We propose the following definition of this characteristic: *the metadata provenance is the completeness with respect to metadata standard, collection method, or source*. Here we follow [10] which proposed similar definition for data item provenance. In this section, we will only consider *metadata standard provenance*, the approach for dealing with metadata source or collection method provenance is similar.

Metadata provenance contributes to addressing the metadata goals as follows:

1. metadata with good provenance allows for the description of the external data which can be trusted more (i.e., which is of better quality);

2. metadata with good provenance supports the search within metadata values better as the search is more likely to return trusted information;
3. metadata with good provenance supports decisions based on the found artifacts better as these decisions are supported with data that can be trusted.

Related quality characteristic is *metadata trustfulness* which is often based on metadata provenance; we will not deal with it in detail due to the lack of space.

6.7.1. Related Work on Metadata Provenance

Several researchers provide definitions for metadata provenance in their papers:

1. Bruce et al. [19] defined provenance as the quality of the source of origin, quality of methodology of origin, quality of transformations;
2. Gavrilis et al. [22] defined provenance as auditability (i.e., the ability to track the metadata record back to the original);
3. Radulovic et al. [16] defined provenance as the metadata quality aspect defined as a part of the quality model, also supporting the trustfulness.
4. Ochoa and Duval [26] defined provenance as the degree of trust in the source of the metadata item, to be measured as an average quality of all instances produced by this source.

In the survey of opinions by metadata practitioners regarding the relative importance of metadata quality characteristics conducted by Gentry et al. [39], provenance was put in the fifth place, after consistency, accuracy, timeliness, and accessibility.

The trustfulness is defined in [16] as the characteristic supported by the provenance; this can probably be combined with conformance to expectations, also as the trustworthiness of a resource or an information provider; defined as a part of the set of contextual dimensions.

6.7.2. Calculating Metadata Provenance Metrics

We define metadata standard provenance as a degree of completeness with respect to the information about the metadata standard. As a result, metadata standard provenance metrics are calculated similarly to the metadata completeness metrics. The question it answers is as follows: do we know the standard the metadata values conform to?

Sample-Level Metadata Provenance

This metric is defined as *a degree of instantiation of the metadata standard information for a biobank sample.* We propose to calculate it as a ratio of a number of metadata standards which are instantiated for the metadata attributes of a sample to an overall number of metadata attributes declared for a biobank collection owning the sample:

$$MProvT_s^{sampM} = \frac{n(A^M_{C(s)}) - \sum\limits_{a^M \in A^M_{C(s)}} f_s^T(a^M)}{n(A_{C(s)})} \tag{19}$$

where $A^M_{C(s)}$ is a set of all metadata attributes defined for a biobank collection $C(s)$ owning the sample s, $n(A^M_{C(s)})$ is its cardinality, $f_s^T(a^M)$ is a negative standard instantiation function over metadata attribute $a^M \in A^M_{C(s)}$ which is equal to 0 if the information on the standard is instantiated for a^M in s, or 1 otherwise.

Suppose the data item attribute values instantiated by the sample s_1 are described by ten metadata attributes, among which the standard information is not instantiated for four metadata attributes, then $MProvT_{s_1}^{sampM} = (10-4)/10 = 0.6$.

Collection-Level Metadata Provenance

Metadata provenance metrics which characterize the whole biobank collection can be calculated as:

1. *Sample-based collection-level metadata provenance* by aggregating sample-level metadata provenance values for all or a subset of the samples in a collection;
2. *Metadata attribute-scoped collection-level metadata provenance* by aggregating the metadata provenance values for a specific set of metadata attributes describing the data item attribute values instantiated for all or a subset of the samples in a collection; such set can contain a single attribute, a subset of or all metadata attributes defined for a biobank collection.

Sample-Based Collection-Level Metadata Provenance

We propose to calculate such a metric by applying the aggregate function to a set of sample-level metadata provenance values calculated for all the samples in a given collection. For example, *sample-average-based collection-level metadata provenance* is calculated by applying the following formula:

$$MProvT_c^{avgSM} = \frac{\sum\limits_{s \in S_c} MProvT_{sc}^{sampM}}{n(S_c)}, \tag{20}$$

where S_c is a set of all samples belonging to a collection c, $Prov_{sc}^{samp}$ is a metadata provenance value calculated for a sample s belonging to c.

Suppose there is a collection c_1 which contains 100 samples, where 80 samples have the sample-level metadata provenance of 0.75, and 20 samples have the sample-level metadata provenance of 0.95. Then $MProvT_{c_1}^{avgSM} = (0.75 \cdot 80 + 0.95 \cdot 20)/100 = 0.79$.

Metadata Attribute-Scoped Collection-Level Metadata Provenance

As before, here we limit ourselves to the case when the metadata attribute set contains a single metadata attribute. For this case, this metric can be calculated as *a degree of standard information instantiation for a specific metadata attribute in a specific biobank collection* by applying the following formula:

$$MProv_{a^M c}^{attr^M} = \frac{n(S_c) \cdot n(A_c) - \sum\limits_{s \in S_c} \sum\limits_{a \in A_c} f_{sa}^T(a^M)}{n(S_c) \cdot n(A_c)}, a^M \in A_c^M \tag{21}$$

where $f_{sa}^T(a^M)$ is a negative metadata standard instantiation function over a sample $s \in S_c$ and a data attribute $a \in A_c$, which is equal to 0 if the information about the metadata standard is instantiated for the metadata attribute a^M describing the value of the data attribute a instantiated by the sample s, or 1 otherwise.

Suppose the standard information is missing for 250 data item reliability metadata values describing data item values in a collection c_1, which contains 200 samples, each accompanied by four data item attributes. Then $MProv_{(reliability)c_1}^{attr^M} = (200 \cdot 4 - 250)/(200 \cdot 4) = 0.6875$.

It is also possible to define data item attribute-scoped collection-level metadata provenance with the aggregation formula similar to (5); we omit its treatment here due to the lack of space.

6.8. Metadata Reliability

Reliability of the metadata can be derived from the reliability of its source or the collection method, or the standard it relies upon. For example, the reliability of the metadata provided by the biobank itself can be higher than the metadata from external sources. Low reliability means that the metadata cannot be trusted, so its quality is low.

Metadata reliability contributes to addressing the metadata goals in a way similar to metadata provenance:

1. highly reliable metadata allows for a better description of the external data, as such a description can be trusted more;
2. highly reliable metadata supports the search within metadata values better as the search is more likely to return information that can be trusted;
3. highly reliable metadata supports decisions based on the found artifacts better as these decisions are supported with data that can be trusted.

6.8.1. Related Work on Metadata Reliability

Very limited number of researchers provide definitions for metadata reliability in their papers:

1. Ceravolo et al. [54] proposed adding a peer-to-peer trust layer to the metadata generators to increase metadata reliability;
2. Kapidakis [55] generalized the results of the above research by defining metadata reliability through the reliability of metadata providers, introducing the notion of metadata source reliability;
3. Hausner and Sommerland [56] defined metadata reliability through persistence, conformance to standards, and quality assurance; they stated that increasing such reliability is a primary goal for metadata quality management in digital libraries.

6.8.2. Calculating Metadata Reliability Metrics

Following [10], we define an indirect metadata quality metric as the metric based not on the measurement of the metadata value itself, but on the quality of an element (or elements) of its context. The quality of the metadata value has to be calculated based on that context quality. In this paper, we only deal with *indirect metadata reliability metrics* based on the reliability metrics defined for metadata sources (following Kapidakis [55]), collection methods, or related standards.

As the context of the metadata value includes *the metadata source*, we further limit ourselves to the metrics based on the *reliability of metadata sources*. For example, the data item completeness metadata for the BMI data item attribute for a collection will be more reliable if it comes from the biobank itself, and not from the non-trusted external source.

Here we will not deal with calculating the reliability of the metadata sources in detail; we just assume that these *source reliability values* are available. A detailed treatment of metadata source reliability metrics will be the topic of the follow-up paper.

After taking hold of the metadata source reliability values, the next step is to connect them to the biobank metadata elements. It can be done in one of the following ways:

1. The description of the biobank collection includes the declarations of the metadata sources. The scope of these sources is limited to all metadata attributes connected to the data item attributes declared for this collection. They are instantiated for some or all the values of the metadata attributes describing the values of all data item attributes instantiated by the samples belonging to this collection.
2. The description of the data item attribute declared for a biobank collection includes the declaration of the metadata sources. The scope of such sources is limited to the metadata attributes connected to this data item attribute. They are instantiated for the values of all metadata attributes describing the values of this data item attribute instantiated by the samples belonging to this collection.
3. The description of the metadata attribute declared for a biobank collection includes the declaration of the metadata sources. The scope of such sources is limited to this metadata attribute. They are instantiated for the values of this metadata attribute describing the values of all data item attributes instantiated by the samples belonging to this collection.

Sample-Level Metadata Reliability

A sample-level aggregate metadata reliability metric can be calculated as *the average degree of source reliability for a specific sample*. It is calculated by aggregating the source reliability values connected to all metadata attribute values belonging to a sample for which the source information is available:

$$MRel_s^{avg^M} = \frac{\sum\limits_{a^M \in A^M_{C(s)}} Rel(r_{a^M}(s))}{n(A^M_{C(s)})}, \tag{22}$$

where $A^M_{C(s)}$—a set of all metadata attributes declared for the collection $C(s)$ containing the sample s, for which the metadata source information is available, $Rel(r_{a^M}(s))$ is a reliability value of the metadata source $r_{a^M}(s)$ which provides the value for the metadata attribute a^M belonging to a sample s.

To simplify calculations, here and below, we assume that all values of a specific metadata attribute describing data item attributes for a specific sample are produced by the same metadata source (with the same reliability), and that this source can be different for different samples. For example, all data item consistency values describing data item attributes (patient age, disease code, etc.) for a given sample are produced by the same source; still, their source can be different for other samples.

Suppose the sample s_1 instantiates data item attribute values described by the data item completeness metadata attribute with the source reliability of 0.7, data item provenance metadata attribute with the source reliability of 0.8, and the data item precision metadata attribute with the source reliability of 0.9. Then $MRel_{s_1}^{avg^M} = (0.7 + 0.8 + 0.9)/3 = 0.8$.

Collection-Level Metadata Reliability

Below we list some of the possible collection-level aggregate reliability metrics.

Sample-based average degree of metadata source reliability for a collection is defined as an average of the sample-level metadata reliability values for all samples in a collection:

$$MRel_c^{avgS^M} = \frac{\sum\limits_{s \in S_c} MRel_s^{avg^M}}{n(S_c)}, \tag{23}$$

where S_c is a set of all samples in c.

Suppose there is a collection c_1 which contains 100 samples, where 40 samples have sample-level average metadata reliability of 0.55, and 60 samples have sample-level average metadata reliability of 0.85. Then $MRel_{c_1}^{avgS^M} = (0.55 \cdot 40 + 0.85 \cdot 60)/100 = 0.73$.

Metadata attribute-scoped average degree of metadata source reliability for a collection is calculated as an average of the metadata source reliability values of the metadata sources related to a specific metadata attribute describing values of all data item attributes instantiated by the samples in a collection:

$$MRel_{a^M c}^{avgattrM} = \frac{\sum\limits_{s \in S_c} Rel(r_{a^M}(s))}{n(S_c)}, a^M \in A_c^M \tag{24}$$

Suppose there is a collection c_1 which contains 100 samples, where the data item attributes of 15 samples are described by the data item accuracy metadata attribute with the source reliability of 0.35, and the data item attributes of 85 samples are described by the data item accuracy with the source reliability of 0.75. Then $MRel_{(accuracy)c_1}^{avgattrM} = (0.35 \cdot 15 + 0.75 \cdot 85)/100 = 0.69$.

The absolute minimum for metadata source reliability for a collection defines the value of the narrowest "bottleneck" for a metadata source reliability that exists in the collection:

$$MRel_c^{\min SM} = \min_{s \in S_c} \min_{a^M \in A^M_{C(s)}} Rel(r_{a^M}(s)) \quad (25)$$

Suppose there is a collection c_1 where the minimal metadata source reliability for one sample s_1 is the reliability of the data item accuracy source equal to 0.4, and the minimal metadata source reliability for the rest of the samples is the reliability of the data item completeness source equal to 0.9. Then $MRel_{c_1}^{\min SM} = \min(0.4, 0.9) = 0.4$.

It is also possible to find a metadata attribute with minimal source reliability in the collection by finding $MRel_c^{\min SM}$ and checking which metadata attribute it corresponds to. In the previous example, this metadata attribute is evaluated to data item accuracy.

6.9. Metadata Conformance to Expectations

We define metadata conformance to expectations in the broader sense as the ability for the metadata to support achieving its user-related goals, namely to support the search and the ability to make decisions based on found artifacts. In this section, we will treat it in the narrower sense as *the ability for the metadata to support finding the requested data-item level artifacts* (e.g., collections or biobanks). This should not be confused with the ability for the metadata *itself* to be found or be accessible; this is characterized by the metadata accessibility and will be addressed in the next section.

Following Ochoa and Duval [26], we state that metadata conformance to expectations, as defined in this section, supports the FAIR principle of metadata findability (defined by GO FAIR metrics group [33] as the ability for the metadata to support finding the data it describes), as the expectations for metadata correspond to metadata goals, and one of the goals is the support for search.

Among other FAIR principles, we treat accessibility as being supported by accessibility metrics in Section 6.10, the rest (interoperability and reusability) will be addressed in the follow-up paper dedicated to FAIR principles supported by quality characteristics for biobank data items and metadata.

Metadata conformance to expectations contributes to addressing the metadata goals in the following way:

1. metadata which conforms to expectations better allows for a better description of the external data, as such a description can contain information which makes described artifacts easier to find;
2. metadata, which conforms to expectations better (allowing for finding data item-level artifacts easier), supports the search based on metadata values better as the search is more likely to return requested artifacts;
3. metadata which conforms to expectations better supports decisions based on the found artifacts better, as these decisions are supported with artifacts that correspond to specified criteria in a more meaningful way.

6.9.1. Related Work on Metadata Conformance to Expectations

Several researchers provide definitions for metadata conformance to expectations or findability in their papers:

1. Bruce et al. [19] defined conformance to expectations as fulfilling promises from the metadata owner to the user, providing all that is expected to find;
2. Margaritopoulos et al. [18] defined conformance to expectations as relevance (can also be treated as a separate characteristic);
3. Gavrilis et al. [22] defined conformance to expectations as appropriateness (for the intended use);
4. Radulovic et al. [16] defined conformance to expectations as relevancy (a part of the set of contextual dimensions).

5. Jaffe et al. [57] motivated the necessity of the conformance to the expectations quality characteristic by the need to address the ethical aspect of metadata quality. They argue for introducing the ethical dimension of metadata evaluation and that "truly good metadata is metadata that goes beyond adherence to technical standards", and that it is preferable to evaluate metadata against the conformance to the expectations of the specific community.
6. The principle of metadata findability introduced by GO FAIR metric group [33] was further exemplified by Wilkinson et al. [58] by the metric measuring the existence of the globally unique and persistent ID for the data included in the metadata, registering in a searchable resource.
7. Ochoa and Duval [26] defined conformance to expectations as the degree to which the metadata instances facilitate the ability for the user to find the resource, i.e., as the synonym to metadata findability; they proposed to measure it as, e.g., the degree of uniqueness of the words in the description.
8. The uniqueness quality metrics proposed by Ochoa and Duval [26] can be calculated using graph-based techniques proposed by Phillips et al. [59]. These techniques assess metadata quality by means of defining metadata record graphs based on available metadata elements. Such graphs interpret metadata elements as nodes and connections between such elements (e.g., based on common source, common value etc.) as edges. They allow to calculate metrics based on relationships between metadata elements:
    - the density of the graph, i.e., the ratio of the number of actual edges to the number of possible edges;
    - the distribution of the degree of the graph, i.e., the number of edges that intersect with a given node;
    - the Gini coefficient based on such distribution, i.e., the degree of uniformity in the distribution.

    Taken together, these metrics allow calculating the entropy for each set of metadata elements understood as a measure of the uniqueness of the corresponding values.

6.9.2. Calculating Metadata Conformance to Expectations Metrics

Following Ochoa and Duval [26], we define conformance to expectations metrics (contributing to findability) to be based on *the relative amount of unique information contained in the metadata element*.

For example, if all metadata elements for average BMI reliability of the available biobank collections contain the value of 1 (i.e., indicating that BMI data for these collections is 100% reliable), adding a new metadata element with the value of 1 cannot help in distinguishing the corresponding biobank collection from the rest, whereas a new element holding the value of 0.5 (indicating that BMI data for this collection is 50% reliable) helps to find the corresponding collection much easier, as this value is unique in the set of metadata elements available for search.

To assess such uniqueness, again following Ochoa and Duval [26], we use the concept of *entropy*, defined, based on Shannon [60], as the reverse measure (negative logarithm) of the probability of encountering a specific value in a set, i.e., the degree of "non-standardness" of this value for this set.

Calculating Metadata Attribute Value Uniqueness

We start with defining the metric for metadata attribute value uniqueness. Suppose we have the metadata attribute which holds values belonging to a finite set (e.g., the data item reliability, which is assessed by humans on the interval from 0 to 1 with the step of 0.1, defining the set of 11 values: {0, 0.1, ..., 0.9, 1}).

Now we can assess the uniqueness of the metadata values instantiating that metadata attribute to describe the values of different data item attributes instantiated by different samples by calculating their entropy. It can be done using the following formula:

$$MUniq_c(a^M, a, s) = 1 - \frac{\log(Num^{a^M}_{A_c S_c}(v_{a^M}(a,s)))}{\log(n(A_c) \cdot n(S_c))}, s \in S_c, a \in A_c, a^M \in A^M \quad (26)$$

where $v_{a^M}(a,s)$ is a value of the metadata attribute a^M describing the value of the data item attribute a instantiated for the sample s, A_c is a set of all data item attributes defined for a biobank collection c, S_c is a set of all samples in c, $Num^{a^M}_{A_c S_c}(v_{a^M}(a,s))$ is the total number of values equal to $v_{a^M}(a,s)$ in the set of values for the metadata attribute a^M describing the values of all data item attributes in A_c instantiated for all samples in S_c.

If $Num^{a^M}_{A_c S_c}(v_{a^M}(a,s)) = 0$ (which means that $v_{a^M}(a,s)$ is not present in the set of values for a^M), the formula (26) is evaluated to 1, if $Num^{a^M}_{A_c S_c}(v_{a^M}(a,s)) = n(S_c) \cdot n(A_c)$ (all data item attributes in A_c instantiated by all the samples in S_c are described by the same $v_{a^M}(a,s)$ metadata value for a^M), this formula is evaluated to 0.

The above metric allows for assessing the conformance to expectations for the given metadata attribute value describing a data item attribute value instantiated for a specific sample.

Suppose a collection c_1 contains 100 samples, each instantiating patient age, disease code, and BMI data item attributes described, in turn, by data item reliability, and consistency metadata attributes. Suppose also that the value of 0.5 for the data item consistency metadata attribute describing BMI data item attribute for the sample s_1 is equal to the data item consistency values describing in total 210 data item attribute values. Then, $MUniq_{c_1}(consistency, BMI, s1) = 1 - \log(210)/\log(3 \cdot 100)$.

Calculating Aggregated Metadata Attribute Value Uniqueness for a Collection

It is also possible to calculate the aggregated uniqueness of metadata attribute values for a collection. For this, we can, for example, calculate *the average sample-based metadata value uniqueness* as follows:

$$MUniq^{avg}_c(a^M) = \frac{\sum_{s \in S_c} \sum_{a \in A_c} MUniq_c(a^M, a, s)}{n(S_c) \cdot n(A_c)}, a^M \in A^M_c, c \in C \quad (27)$$

This metric assesses the average uniqueness of values for the specific metadata attribute for the collection.

Suppose a collection c_1 contains 100 samples, each instantiating patient age, disease code, and BMI data item attributes described, in turn, by data item reliability, and consistency metadata attributes. Suppose also that the value of 0.5 for the data item consistency metadata attribute describes 140 data item attribute values, and the value of 0.6875 describes the rest of the values. Then, $MUniq^{avg}_{c_1}(consistency) = (0.5 \cdot 140 + 0.6875 \cdot 160)/(3 \cdot 100) = 0.6$.

It is also possible to aggregate such values for all metadata attributes in a collection.

6.10. Metadata Accessibility

Metadata accessibility reflects the need for the metadata *itself* to be accessible, i.e., being available when it is necessary, or being able to be found when requested. In concentrating on the ability for the metadata itself to be found, it differs from conformance to expectations as our definition of that characteristic in Section 6.9 reflects the ability of metadata to support finding the data items or collections it describes, not finding the metadata itself.

Metadata accessibility contributes to addressing the metadata goals in the following way:

1. highly accessible metadata allows for a better description of the external data, as such a description can be better available when necessary, or be easier to find;
2. highly accessible metadata (which is easier to find) supports the search within metadata values better as the search is more likely to return requested metadata elements;
3. highly accessible metadata supports decisions based on the found artifacts better as these decisions are supported with data that is easier to access.

The metadata accessibility characteristic, as defined in this section, directly supports the FAIR principle of metadata accessibility (see the discussion of supporting other FAIR principles in Section 6.9).

The related characteristic is *metadata understandability*, e.g., as addressed directly by Radulovic et al. [16]; we will not consider it here in detail, as it is not very useful as the means of assessing quality metadata (quality values are mostly numeric, whereas understandability metrics are applicable to the text data).

6.10.1. Related Work on Metadata Accessibility

Several researchers provide definitions for metadata accessibility in their papers:

1. Bruce et al. [19]: classified the barriers of accessibility as (1) physical (metadata is physically separated from the data or is unreadable etc.); (2) economical (access is not affordable); (3) intellectual (metadata is difficult to understand for the people with other backgrounds);
2. Radulovic et al. [16] defined accessibility as (1) security, i.e., the extent of protecting the data (or the metadata) from alteration and misuse; (2) licensing, i.e., granting permission for using the data; (3) availability—the degree of the data ready for use; (4) performance of the underlying system;
3. Wilkinson et al. [58] defined accessibility as the availability of the standardized access protocol, accessibility of the metadata without data, authentication, and authorization.
4. Ochoa and Duval [26] defined accessibility as a degree to which the metadata instance can be found; they proposed to measure it by calculating the number of links leading to this instance and, for text-based metadata, by calculating readability indexes over metadata texts. Calculating such accessibility metrics can be performed using graph-based techniques by Phillips et al. [59] already discussed in Section 6.9 together with conformance to expectations metrics.

In the survey of opinions by metadata practitioners regarding the relative importance of metadata quality characteristics conducted by Gentry et al. [39], accessibility was put in the fourth place, after consistency, accuracy, and timeliness.

6.10.2. Calculating Metadata Accessibility Metrics

Here, we restrict ourselves with the simple accessibility metrics reflecting that some of the metadata instances can be inaccessible due to privacy restrictions (e.g., when they accidentally contain data that was not anonymized enough).

Based on that, we can define collection-level metadata accessibility, e.g., as described below.

Data Attribute-Scoped Collection-Level Metadata Accessibility

We define such accessibility as *a degree of metadata value accessibility for a specific data item attribute in a specific collection*. For example, it could characterize the accessibility of all metadata values (for all defined data item quality characteristics) for the BMI data item attribute for a specific collection (BMI reliability, BMI accuracy, etc., taken together). It can be calculated as follows:

$$MAcc_{ac}^{attr} = \frac{n(S_c) \cdot n(A_c^M) - \sum\limits_{s \in S_c} \sum\limits_{a^M \in A_c^M} b_{sa}^M(a^M)}{n(S_c) \cdot n(A_c^M)}, a \in A_c \quad (28)$$

where $b^M_{sa}(a^M)$ is a negative metadata value accessibility function which can be either 0 if the value for a metadata attribute $a^M \in A^M_{c(s)}$ calculated for a data item attribute a is accessible for s, or 1 otherwise.

Suppose there are 130 inaccessible metadata values describing BMI data item attribute values in a collection c_1, which contains 250 samples instantiating data item attribute values that are described by the values of two metadata attributes. Then $MAcc^{attr}_{(BMI)c_1} = (250 \cdot 2 - 130)/(250 \cdot 2) = 0.74$.

Metadata Attribute-Scoped Collection-Level Metadata Accessibility

We define such accessibility as *a degree of metadata value accessibility for a specific metadata attribute in a specific collection*. For example, it could characterize the accessibility of metadata accuracy or reliability values for a specific collection. It can be calculated as follows:

$$MAcc^{attr^M}_{a^M_c} = \frac{n(S_c) \cdot n(A_c) - \sum_{s \in S_c} \sum_{a \in A_c} b^M_{sa}(a^M)}{n(S_c) \cdot n(A_c)}, a^M \in A^M_c \quad (29)$$

Suppose there are 230 inaccessible data item provenance values describing data item attribute values in a collection c_1, which contains 250 samples, each accompanied by four data item attributes. Then $MAcc^{attr^M}_{(provenance)c_1} = (250 \cdot 4 - 230)/(250 \cdot 4) = 0.77$.

7. Conclusions

Metadata Quality is an essential asset for improving the efficiency and effectiveness of the search for relevant material and data for planned medical studies. When we started designing and developing a data and metadata quality management system for Austrian biobanks, we soon discovered the need for precise definitions and descriptions of metadata quality. This paper is intended as a reference for these developments serving a shared understanding of which metadata quality characteristics are relevant in the biobanking domain, what these quality characteristics precisely mean, and how they can be measured and communicated.

We provide detailed descriptions and clear definitions for the metadata quality characteristics in biobanks. We started by discussing the motivation for dealing with metadata quality and introduced biobank metadata and its purposes and goals. This was followed by providing the verbal definition of metadata quality as the degree of addressing the metadata goals. The definition of the structure of metadata quality in biobanks includes quality characteristics and metrics. The introduction of possible aggregations can be used to calculate the corresponding metrics on two levels: data item and metadata level. Finally, we describe nine of the most important biobank metadata quality characteristics (accuracy, completeness, coverage, consistency, timeliness, provenance, reliability, accessibility, and conformance to expectations) in detail and define their corresponding metrics.

This carefully defined foundation of metadata characteristics for biobanks is intended to support both medical researchers using biobanks and biobank administrators in understanding the possible ways of defining and assessing the metadata quality in the biobank domain to make the search for relevant material and data for medical studies more efficient and effective: we expect that using metadata characteristics researchers can identify promising collections faster and avoid to interact with many biobanks which are not able to offer relevant data in the quality of the intended project.

Author Contributions: Conceptualization, J.E. and V.A.S. ; methodology, J.E. and V.A.S.; review state-of-the-art: V.A.S., formalization: V.A.S. and J.E., writing—original draft preparation, V.A.S.; writing—review and editing: V.A.S. and J.E.; visualization, V.A.S.; supervision: J.E. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Austrian Bundesministerium für Bildung, Wissenschaft und Forschung within the project BBMRI.AT (GZ 10.470/0010-V/3c/2018).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lehmann, S.; Guadagni, F.; Moore, H.; Ashton, G.; Barnes, M.; Benson, E.; Clements, J.; Koppandi, I.; Coppola, D.; Demiroglu, S.Y.; et al. Standard preanalytical coding for biospecimens: Review and implementation of the Sample PREanalytical Code (SPREC). *Biopreserv. Biobank.* **2012**, *10*, 366–374.
2. Moore, H.M.; Kelly, A.B.; Jewell, S.D.; McShane, L.M.; Clark, D.P.; Greenspan, R.; Hayes, D.F.; Hainaut, P.; Kim, P.; Mansfield, E.; et al. Biospecimen reporting for improved study quality (BRISQ). *J. Proteome Res.* **2011**, *10*, 3429–3438.
3. De Blasio, P.; Biunno, I. New Challenges for Biobanks: Accreditation to the New ISO 20387: 2018 Standard Specific for Biobanks. *BioTech* **2021**, *10*, 13.
4. Merino-Martinez, R.; Norlin, L.; van Enckevort, D.; Anton, G.; Schuffenhauer, S.; Silander, K.; Mook, L.; Holub, P.; Bild, R.; Swertz, M.; et al. Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 Core). *Biopreserv. Biobank.* **2016**, *14*, 298–306.
5. Eder, J.; Shekhovtsov, V.A. Data quality for federated medical data lakes. *Int. J. Web Inf. Syst.* **2021**, *17*, 407–426.
6. Eder, J.; Gottweis, H.; Zatloukal, K. IT solutions for privacy protection in biobanking. *Public Health Genom.* **2012**, *15*, 254–262.
7. Riley, J. *Understanding Metadata*; National Information Standards Organization: Washington, DC, USA, 2017; Volume 23.
8. Ciglic, M.; Eder, J.; Koncilia, C. Anonymization of data sets with null values. *Trans. Large-Scale Data- Knowl.-Centered Syst.* **2016**, *XXIV*, 193–220.
9. Stark, K.; Eder, J.; Zatloukal, K. Priority-based k-anonymity accomplished by weighted generalisation structures. In Proceedings of the International Conference on Data Warehousing and Knowledge Discovery, Krakow, Poland, 4–8 September 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 394–404.
10. Shekhovtsov, V.A.; Eder, J. Data Item Quality for Biobanks. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems L*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 77–115.
11. Müller, H.; Dagher, G.; Loibner, M.; Stumptner, C.; Kungl, P.; Zatloukal, K. Biobanks for life sciences and personalized medicine: Importance of standardization, biosafety, biosecurity, and data management. *Curr. Opin. Biotechnol.* **2020**, *65*, 45–51.
12. Quinlan, P.R.; Gardner, S.; Groves, M.; Emes, R.; Garibaldi, J. A data-centric strategy for modern biobanking. In *Biobanking in the 21st Century*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 165–169.
13. Eder, J.; Dabringer, C.; Schicho, M.; Stark, K. Information systems for federated biobanks. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems I*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 156–190.
14. Karimi-Busheri, F.; Rasouli-Nia, A. Integration, networking, and global biobanking in the age of new biology. In *Biobanking in the 21st Century*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 1–9.
15. ASQ Quality Glossary. https://asq.org/quality-resources/quality-glossary. Accessed 02-08-2022
16. Radulovic, F.; Mihindukulasooriya, N.; García-Castro, R.; Gómez-Pérez, A. A comprehensive quality model for Linked Data. *Semant. Web* **2018**, *9*, 3–24.
17. Lei, Y.; Uren, V.; Motta, E. A framework for evaluating semantic metadata. In Proceedings of the 4th International Conference on Knowledge Capture, Whistler, BC, Canada, 28–31 October 2007; pp. 135–142.
18. Margaritopoulos, T.; Margaritopoulos, M.; Mavridis, I.; Manitsaris, A. A Conceptual Framework for Metadata Quality Assessment. In Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications, Berlin, Germany, 22–26 September 2008.
19. Bruce, T.R.; Hillmann, D.I. The continuum of metadata quality: Defining, expressing, exploiting. In *Metadata in Practice*; ALA Editions: Chicago, USA, 2004; pp. 257–271.
20. Stvilia, B.; Gasser, L.; Twidale, M.B.; Shreeves, S.L.; Cole, T.W. Metadata quality for federated collections. In Proceedings of the Ninth International Conference on Information Quality (ICIQ-04), Cambridge, MA, USA, 5–7 November 2004; pp. 111–125.
21. Stvilia, B.; Gasser, L.; Twidale, M.B.; Smith, L.C. A framework for information quality assessment. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 1720–1733. https://doi.org/10.1002/asi.20652.
22. Gavrilis, D.; Makri, D.N.; Papachristopoulos, L.; Angelis, S.; Kravvaritis, K.; Papatheodorou, C.; Constantopoulos, P. Measuring quality in metadata repositories. In Proceedings of the International Conference on Theory and Practice of Digital Libraries, Poznań, Poland, 14–18 September 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 56–67.
23. Király, P. Towards an extensible measurement of metadata quality. In Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, Göttingen, Germany, 1–2 June 2017; pp. 111–115.
24. Király, P. Measuring Metadata Quality. Ph.D. Thesis, Georg-August-Universität Göttingen, Göttingen, Germany, 2019.
25. Gonçalves, M.A.; Moreira, B.L.; Fox, E.A.; Watson, L.T. "What is a good digital library?"—A quality model for digital libraries. *Inf. Process. Manag.* **2007**, *43*, 1416–1437. https://doi.org/10.1016/j.ipm.2006.11.010.

26. Ochoa, X.; Duval, E. Automatic evaluation of metadata quality in digital repositories. *Int. J. Digit. Libr.* **2009**, *10*, 67–91. https://doi.org/10.1007/s00799-009-0054-4.
27. Romero-Pelaez, A.; Segarra-Faggioni, V.; Alarcon, P.P. Exploring the provenance and accuracy as metadata quality metrics in assessment resources of OCW repositories. In Proceedings of the 10th International Conference on Education Technology and Computers, Tokyo, Japan, 26–28 October 2018; pp. 292–296.
28. Koesten, L.M.; Kacprzak, E.; Tennison, J.F.; Simperl, E. The Trials and Tribulations of Working with Structured Data: A Study on Information Seeking Behaviour. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1277–1289.
29. Strecker, D. Quantitative Assessment of Metadata Collections of Research Data Repositories. Ph.D. Thesis, Humboldt-Universität zu Berlin, Berlin, Germany, 2021.
30. Park, J.R. Metadata quality in digital repositories: A survey of the current state of the art. *Cat. Classif. Q.* **2009**, *47*, 213–228.
31. Tani, A.; Candela, L.; Castelli, D. Dealing with metadata quality: The legacy of digital library efforts. *Inf. Process. Manag.* **2013**, *49*, 1194–1205. https://doi.org/10.1016/j.ipm.2013.05.003.
32. Wilkinson, et al., M.D.; Sansone, S.A.; Schultes, E.; Doorn, P.; Bonino da Silva Santos, L.O.; Dumontier, M. A design framework and exemplar metrics for FAIRness. *Sci. Data* **2018**, *5*, 180118.
33. GO FAIR Metrics Group. FAIR Metrics. http://fairmetrics.org, Accessed 04-08-2022.
34. Scheidlin, C.; Celino, M.; Demir, M.H.; Dennis, R. FAIR Metadata Standards for Low Carbon Energy Research—A Review of Practices and How to Advance. *Energies* **2020**, *14*, 6692.
35. ISO/IEC 25012:2008 Software Engineering – Software Product Quality Requirements and Evaluation (SQuaRE) – Data Quality Model. International Organization for Standardization: Geneva, Switzerland, 2008.
36. ISO/IEC 25024:2015 Software Engineering – Software Product Quality Requirements and Evaluation (SQuaRE) – Measurement of Data Quality. International Organization for Standardization: Geneva, Switzerland, 2015.
37. Neumaier, S.; Umbrich, J.; Polleres, A. Automated quality assessment of metadata across open data portals. *J. Data Inf. Qual. (JDIQ)* **2016**, *8*, 1–29.
38. Quarati, A. Open Government Data: Usage trends and metadata quality. *J. Inf. Sci.* **2021**, 01655515211027775. https://doi.org/10.1177/01655515211027775.
39. Gentry, S.; Hale, M.L.; Payant, A.; Tarver, H.; White, R.; Wittmann, R. Survey of Benchmarks in Metadata Quality: Initial Findings. UNT Digital Library, University of North Texas: Denton, USA, 2020. https://digital.library.unt.edu/ark:/67531/metadc1637685, Accessed 04-08-2022
40. Király, P.; Büchler, M. Measuring completeness as metadata quality metric in Europeana. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 25–30 June 2018; pp. 2711–2720.
41. Margaritopoulos, M.; Margaritopoulos, T.; Mavridis, I.; Manitsaris, A. Quantifying and Measuring Metadata Completeness. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 724–737. https://doi.org/10.1002/asi.21706.
42. Margaritopoulos, T.; Margaritopoulos, M.; Mavridis, I.; Manitsaris, A. A Fine-Grained Metric System for the Completeness of Metadata. In Proceedings of the Conference Paper in Communications in Computer and Information Science, Jeju Island, Korea, 10–12 December 2009; Sartori, F.; Sicilia, M.A., Manouselis, N., Eds.; Springer: Milan, Italy, 2009; pp. 83–94. https://doi.org/10.1007/978-3-642-04590-5_8.
43. Sicilia, M.A.; Garcia, E.; Pages, C.; Martinez, J.J.; Gutierrez, J.M. Complete metadata records in learning object repositories: Some evidence and requirements. *Int. J. Learn. Technol.* **2005**, *1*, 411–424.
44. Phillips, M. Metadata Quality, Completeness, and Minimally Viable Records, 2015. https://vphill.com/journal/post/4075 Accessed 05-08-2022
45. Nichols, D.M.; McKay, D.; Twidale, M.B. A lightweight metadata quality tool. In JCDL'08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries; ACM Press: New York, NY, USA, 2008; pp. 385–388. https://doi.org/10.1145/1378889.1378957.
46. Díaz de la Paz, L.; Riestra Collado, F.N.; García Mendoza, J.L.; González González, L.M.; Leiva Mederos, A.A.; Taboada Crispi, A. Weights Estimation in the Completeness Measurement of Bibliographic Metadata. *Comput. Y Sist.* **2021**, *25*, 47–65.
47. Lorenzini, M.; Rospocher, M.; Tonelli, S. On assessing metadata completeness in digital cultural heritage repositories. *Digit. Scholarsh. Humanit.* **2021**, *36*, ii182–ii188.
48. Liolios, K.; Schriml, L.; Hirschman, L.; Pagani, I.; Nosrat, B.; Sterk, P.; White, O.; Rocca-Serra, P.; Sansone, S.A.; Taylor, C.; et al. The Metadata Coverage Index (MCI): A standardized metric for quantifying database metadata richness. *Stand. Genom. Sci.* **2012**, *6*, 444–453.
49. Bellini, E.; Nesi, P. Metadata Quality Assessment Tool for Open Access Cultural Heritage Institutional Repositories. In *Proceedings of the Information Technologies for Performing Arts, Media Access, and Entertainment*; Lecture Notes in Computer Science; Springer: Heidelberg, Germany, 2013; Volume 7990, pp. 90–103. https://doi.org/10.1007/978-3-642-40050-6_9.
50. Open Language Archives Community (OLAC) Metadata Metrics, 2009.
51. Klie, A.; Tsui, B.Y.; Mollah, S.; Skola, D.; Dow, M.; Hsu, C.N.; Carter, H. Increasing metadata coverage of SRA BioSample entries using deep learning-based named entity recognition. *Database* **2021**, *2021*, baab021.
52. Griffiths, E.; Dooley, D.; Graham, M.; Van Domselaar, G.; Brinkman, F.S.; Hsiao, W.W. Context is everything: Harmonization of critical food microbiology descriptors and metadata for improved food safety and surveillance. *Front. Microbiol.* **2017**, *8*, 1068.

53. Zaveri, A.; Hu, W.; Dumontier, M. MetaCrowd: Crowdsourcing biomedical metadata quality assessment. *Hum. Comput.* **2019**, *6*, 98–112.
54. Ceravolo, P.; Damiani, E.; Viviani, M. Adding a peer-to-peer trust layer to metadata generators. In Proceedings of the OTM Confederated International Conferences "On the Move to Meaningful Internet Systems", Agia Napa, Cyprus, 31 October—4 November 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 809–815.
55. Kapidakis, S. Exploring metadata providers reliability and update behavior. In Proceedings of the International Conference on Theory and Practice of Digital Libraries, Hannover, Germany, 5–9 September 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 417–425.
56. Häusner, E.M.; Sommerland, Y. Assessment of metadata quality of the Swedish National Bibliography through mapping user awareness. *Cat. Classif. Q.* **2018**, *56*, 96–109.
57. Jaffe, R. Rethinking Metadata's Value and How It Is Evaluated. *Tech. Serv. Q.* **2020**, *37*, 432–443.
58. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
59. Phillips, M.E.; Zavalina, O.L.; Tarver, H. Exploring the utility of metadata record graphs and network analysis for metadata quality evaluation and augmentation. *Int. J. Metadata Semant. Ontol.* **2020**, *14*, 112–123.
60. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.