Data quality for federated medical data lakes

Johann Eder and Vladimir A. Shekhovtsov
University of Klagenfurt, Klagenfurt, Austria

Received 19 March 2021
Revised 11 May 2021
Accepted 17 May 2021

Abstract

Purpose – Medical research requires biological material and data collected through biobanks in reliable processes with quality assurance. Medical studies based on data with unknown or questionable quality are useless or even dangerous, as evidenced by recent examples of withdrawn studies. Medical data sets consist of highly sensitive personal data, which has to be protected carefully and is available for research only after the approval of ethics committees. The purpose of this research is to propose an architecture to support researchers to efficiently and effectively identify relevant collections of material and data with documented quality for their research projects while observing strict privacy rules.

Design/methodology/approach – Following a design science approach, this paper develops a conceptual model for capturing and relating metadata of medical data in biobanks to support medical research.

Findings – This study describes the landscape of biobanks as federated medical data lakes such as the collections of samples and their annotations in the European federation of biobanks (Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium, BBMRI-ERIC) and develops a conceptual model capturing schema information with quality annotation. This paper discusses the quality dimensions for data sets for medical research in-depth and proposes representations of both the metadata and data quality documentation with the aim to support researchers to effectively and efficiently identify suitable data sets for medical studies.

Originality/value – This novel conceptual model for metadata for medical data lakes has a unique focus on the high privacy requirements of the data sets contained in medical data lakes and also stands out in the detailed representation of data quality and metadata quality of medical data sets.

Keywords Biobank, Metadata, Data quality, Data lake, Privacy, LOINC, Metadata and ontologies

Paper type Research paper

1. Introduction

Data lakes are architectures for the storage of data for further use (Inmon, 2016; Giebler *et al.*, 2019; Sawadogo and Darmont, 2020). The data lake concept arose with the advent of *big data* as organizations were not able to keep up with the ever-increasing possibilities for collecting and storing data and to integrate all these data in structured data repositories. Data warehouses (Golfarelli and Rizzi, 2018; Vaisman and Zimányi, 2014) require that data, which should be stored in a data warehouse or a data mart, is structured, cleaned, harmonized and integrated, before it is entered into the data warehouse – usually through a carefully designed process of extracting data from the sources, transforming the data into

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).

International Journal of Web Information Systems
Vol. 17 No. 5, 2021
pp. 407-426
Emerald Publishing Limited
1744-0084
DOI 10.1108/IJWIS-03-2021-0026

the structure and formats defined in the data warehouse and loading the data into the data warehouse in defined intervals (ETL process).

Data lakes, in contrast, do not require that the collected data is integrated, pre-processed and harmonized when it is included in the data repository. Transformation and integration of data sets are only performed, when it is needed for a specific purpose when performing big data analytics (statistics, data mining, and machine learning). Until such usage, the data remains in its initial form and format.

Nevertheless, data lakes cannot be mere stores of unrelated data sets, as this would leave the data unsuitable for the intended usage [called data swamps or data dumps (Brackenbury *et al.*, 2018; Hai *et al.*, 2016)]. Storage architectures to effectively store data sets, catalog the data sets to make them available, when needed, is the subject of ongoing research efforts (Nargesian *et al.*, 2019). Most of the approaches follow the principle to characterize data sets with metadata such that the data sets can be found and used when they are needed for exploitation (Sawadogo and Darmont, 2020).

Most of these approaches, however, assume that both the data sets and their schemas are available to support the search for useful data sets. Here, we will focus on situations, where data and schemas are not easily available. Such situations are common in medical research, where data sets are collected by various stakeholders. Novel and unforeseeable research questions trigger innovative analytical processing of available data and require adequate support for finding relevant data sets and prepare them for the intended processing methods. This profile matches the characteristics of data lake architectures to a very high degree. However, the legal and operational constraints for the highly relevant but also highly sensitive data sets imply that they can only be accessed through complex processes (Eder *et al.*, 2012; Lemke *et al.*, 2010).

Medical data is usually produced and collected in institutions for providing health care (e.g. hospitals and clinics) and public and private institutions for medical research (e.g. research institutes, pharmaceutical companies, and medical universities). Each of these institutions or even their subunits (e.g. clinical departments) might organize the data in its own data lake. Access to the content of other data lakes is complex and time-consuming. Nevertheless, the combination and joint, integrated processing of these data is indispensable for progress in biomedical research, and thus for the development of future drugs and therapies. Central storage of such data in a general anonymized form is no viable alternative, as it suffers from the problem of information loss leaving the data typically useless for many scientific studies as anonymization cannot be optimized for the specific priorities of a particular project (Stark *et al.*, 2006).

We call such a distributed system a *federated data lake*, a federation of data lakes with commonalities in the scope, purpose or methods of data collections, which ought to be used together for analytical processing, but which are not directly accessible by the involved or interested parties. Federations typically consist of autonomous participants cooperating for a specific purpose (Sheth and Larson, 1990; Berger and Schrefl, 2008; Eder *et al.*, 2006; Skripcak *et al.*, 2014). This paper is an extension of Eder and Shekhovtsov (2020), where we first discussed the requirements and sketched an architecture for such a system under the term of *data Lakelands*.

The specific contributions of the work presented here are the combination of meta-data management with a special focus on data quality in a federation with very high privacy requirements and the application and specialization of such a system to the requirements of medical data in biobanks.

There is a wealth of related work in the different areas we build upon, integrate and address in this work: the concepts of data lakes and their metadata management (Inmon, 2016; Giebler

et al., 2019; Sawadogo and Darmont, 2020), the discovery of data sets in data lakes (Bogatu *et al.*, 2020), research on data quality (Batini and Scannapieco, 2016; Nahm, 2012) and in particular meta-data quality (Mihaila *et al.*, 2000; Stvilia *et al.*, 2004; Bruce and Hillmann, 2004) and finally the application of semantic techniques in the domain of medical data (Dogac *et al.*, 2006).

Based on all these research efforts, we propose a generic architecture and a conceptual model based on the exchange of ontological metadata and metadata quality characteristics to support the location of data sets, which are potentially useful for specified needs. We specialize this generic model with the ontologies and particularities for medical data sets associated with biobanks and biorepositories (Zatloukal and Hainaut, 2010; Spjuth *et al.*, 2016; Müller *et al.*, 2020) and show how such a federated data lake can be organized, emphasizing the need of metadata quality assessment. Overall, we aim at reducing the efforts of researchers to locate useful data sets for performing their studies.

In this paper, we follow a design science approach (Wieringa, 2014). The goal of this research is to create a conceptual model of metadata of medical data sets maintaining the privacy of the data and sample donors and with a strong focus on the representation of data and metadata quality which can be used by medical researcher to find suitable data and cases for their research more efficient and effective.

2. Biobanks and medical data sets

2.1 Biobanks and biobank networks

Medical research requires biological material and data to perform medical studies, to launch hypotheses generating projects and to test hypotheses about diseases, therapies and drugs.

Biobanks are infrastructures for medical research collecting and storing *samples* (i.e. biological material such as tissue, blood, and cell cultures) organized in *collections* together with data describing these materials. Biobanks are interdisciplinary research platforms providing material and data for (medical) research projects (Asslaber *et al.*, 2007; Eder *et al.*, 2009; Hainaut *et al.*, 2017; Hofer-Picout *et al.*, 2017).

For a particular research project data from different biobanks might be necessary for various reasons. The European research infrastructure Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium (BBMRI-ERIC) (Vuorio, 2017; van Ommen *et al.*, 2015; Eder *et al.*, 2009; Litton, 2018; Holub *et al.*, 2016; Hofer-Picout *et al.*, 2017; Merino-Martinez *et al.*, 2016) provides an infrastructure to connect biobanks in Europe and provide researchers with means to efficiently and effectively search for suitable material and data needed for their studies.

While biobanks traditionally were mainly concerned with the harvesting, processing, conserving and storing of biological materials, the importance of data associated with these biological materials is gaining importance and attention. The annotation of the biological samples with quality-controlled data about the donors considerably increases the usefulness of biological material collected in biobanks. Data in biobanks can come from various sources: from data derived in medical studies, from health-care provisioning or from data gathered together with the material. The data is very heterogeneous in every aspect possible. Medical data comprises a wealth of different types of data: classical records with numeric data (e.g. lab measures), alphanumeric data, data in natural language texts (e.g. discharge letters, pathology findings), images (e.g. MRT scans), to specific data formats (e.g. gene expression profile), etc. In addition, also many different ontologies and taxonomies are used (e.g. ICD or SNOMED for diagnostic codes). A data set can be very homogeneous, if it was collected in the course of a clinical study or it might be very heterogeneous, like data from health care, where tests are only performed when necessary for treating a patient.

The data is derived from different sources and different processes, it is collected and stored for supporting different purposes and by different stakeholders. Adequate data quality management for medical data sets is a quite difficult and complex undertaking (Roski *et al.*, 2014), nevertheless, it is absolutely essential – not only for treating patients but also for medical research. The data might be contained in different information systems and storage architectures. Therefore, data is only integrated for an acceptable and approved purpose. Hence, the situation of the data sets maintained in a biobank can be characterized as a data lake, i.e. for a certain research project usually data has to be identified, collected, harmonized and prepared to make it useful for pursuing a research project. The requirements of the research projects are hugely varying and by the very nature of research projects, these requirements cannot be known, when the data for the biobank is collected.

Data in biobanks is extremely sensitive as it represents the health status and bodily characteristics of individuals. Therefore, the privacy of data has to be maintained and sophisticated processes and protocols ascertain strict confidentiality and pedantically control that material and data are only used for consented and approved purposes. As a consequence, storing biobank data in central repositories easily accessible by many researchers is not admissible in many countries (Shaw, 2014).

As each biobank data collection can be seen as a data lake, we view this infrastructure landscape as a *federation of data lakes*, an ensemble with a high number of data lakes, which are connected by some (communication) channels. Then, the number of these lakes is huge – the BBMRI-ERIC directory already lists more than 650 biobanks with almost 2,500 collections (www.bbmri-eric.eu).

An important challenge for biobanks and biobank federation infrastructures is how a researcher can find collections, which might have appropriate data sets and materials for a planned research project. In particular, which information biobanks should offer and are able to provide for making such a search efficient and effective.

2.2 Biobanks as data brokers

The biobanks frequently serve as *data brokers*. They do not produce the data, they might not even store the data associated with their biological material collections. The associated data might reside in departmental databases, general hospital information systems or external data repositories such as registries of health records or data of social security insurances.

Biobanks are also not users of the data. The data is needed by researchers in research organizations. Thus, biobanks serve as mediators or brokers matching the information needs of researchers with the data availability of the data producers and data providers.

Data in biobanks can come from the following sources:

- *The data produced by the biobank.* This contains mainly data about handling and storage of biological materials, standard operating procedures applied (e.g. ischemia time, storage temperature, etc.)
- *The data produced for the biobank.* This includes in particular data about material collected in cohort studies for population-based biobanks.
- *The data produced by scientific studies.* Such data is added to the biobank because the materials used in these studies are stored there.
- *The data produced by routine health care.* The most important category of this data is the data from electronic health records (EHRs).

- *The data from linked collections*. Such data comes from any data collection containing the data produced elsewhere but used in the biobank (e.g. tumor register).
- *The data collected from donors*. This might range from general electronic health records to questionnaires of lifestyle properties.

As a consequence, biobank operators are not responsible for the quality of all these data from external sources but are responsible for proper documentation of the data quality of these sources (quality metadata).

3. Obtaining data for research projects

A typical process for a researcher to obtain material and data sets for a research project or a clinical study contains the following steps (simplified) (Eder *et al.*, 2012). The researcher specifies the data requirements. These include search attributes describing the conditions the cases, which the researcher intends to study, have to fulfill (e.g. data about patients over 50 years of age, body mass index above 30, blood group "A+", who tested positive for Covid-19). In addition, these requirements also specify, which (other) data has to be available in the data set and in which quality (e.g. prior diseases, lab data and lifestyle information).

With this profile, a researcher then searches for biobanks, resp. collections in biobanks, which likely have the required data. As mentioned before, the researcher cannot access the data directly. So currently, such requests are mostly communicated to a number of biobanks through person-to-person communication channels (letters, emails and phone calls) or through tools provided by BBMRI-ERIC biobank directory, sample locator, and sample/data negotiator or biobank providers. When a suitable collection is found, the researcher formulates a project, which states in detail for which purpose the data and material are requested, why the data is necessary and the expected outcome of the project. The proposal is then submitted to an ethics committee. The data is only accessible to a researcher after the project is approved by an ethics committee.

The difficulty in this process is that a researcher might not just turn to a database and run a series of queries to search for promising data sources. The biobank can only offer metadata describing the data for such a search. The biobank directory of BBMRI-ERIC or of BBMRI.AT are tools to help researchers to identify potentially useful collections and biobanks in a rather coarse-grained way. To improve the situation is the aim of the following considerations.

The crucial question is: which kind of metadata can a biobank provide for facilitating the search for relevant data sets without risking exhibiting the personal data of the donors? Another question is: how can such metadata be published in a harmonized way in face of the heterogeneities outlined above? Which kind of data quality documentation is needed such that researchers can judge whether the data might be appropriate for their purposes?

4. Data quality

4.1 Data quality definition and issues

The working definition of data quality used for this paper is: *data quality serves as a measure of how well the data represents facts of the real world*. The usual definition of data quality as *the degree to which the data meets the expectations of data consumers based on their intended use of the data* (Batini and Scannapieco, 2016) is problematic for scientific databases as this use is unknown in advance. We have to focus, therefore, on the intrinsic data quality characteristics and, on the other hand, on the quality of metadata as the representation of the data quality and content. High-quality metadata meets consumer

expectations and truthfully represents the content and quality of data collections to a greater degree than low-quality metadata, i.e. higher quality supports search for appropriate data sets better.

To elaborate on this definition, it is necessary to agree on a set of data quality characteristics, which help in achieving a proper understanding of data quality, reflecting its different aspects specific to a medical domain (Nahm, 2012). Such characteristics can be supplemented with quality metrics, which allow for their quantification by means of a measurement process resulting in the specific quality values. These characteristics can be combined to obtain the integrated data quality.

The data quality values can be also interpreted as metadata as they contain information describing the data (Radulovic *et al.*, 2018) such as precision and validity period.

Data quality characteristics can be categorized:

- based on the meta-level of the data elements being characterized – as either *data item quality characteristics* assessing the quality of the data item level (Batini and Scannapieco, 2016), i.e. both the data originated in the biobank and the external data or *metadata quality characteristics* (Bruce and Hillmann, 2004; Stvilia *et al.*, 2004) assessing the quality of the metadata level;
- based on the origin of the data being characterized (Stvilia *et al.*, 2004), as either *intrinsic data quality metrics and characteristics* assessed by relying only on the existing biobank data and *relative metrics and characteristics* assessed by relying on subjective judgments or external data;
- based on the aggregation level of the data being characterized, as a *single item*, *sample* and *collection metrics and characteristics*.

4.2 Data quality characteristics for data items, samples and collections

We define first the data quality criteria for the level of data items, resp. measurements, i.e. for the instantiation of an attribute. Then we aggregate these characteristics for the level of a sample and the level of a collection. The important characteristic of our framework is that we do not deal with individual samples in our architecture, so *all kinds of quality except collection-level quality are only used to calculate the aggregated quality values on the collection level.*

We use the term combination for the aggregation of data quality measures, as different formulas have been proposed and can be used to aggregate the individual data quality measures deriving data quality measures for aggregates for different purposes.

Data completeness (Batini and Scannapieco, 2016) reflects the need in collecting all required data. Sufficient completeness (contributing to high quality) usually means that all such data is present. For biobank data, insufficient completeness is detected when some data attributes are missing because they were either not recorded due to unavailability or simply not transmitted. In particular, in data derived from health care, typically only those data are determined, which are necessary and useful for a diagnosis, treatment or therapy.

On the level of a data item (measurement, attribute instance), completeness refers to whether a data item is available or not. For complex data items (e.g. fever chart, a time series, etc.) completeness means to which degree all elements of a data item are recorded. Again, this kind of completeness is only used to calculate the aggregated completeness on the collection level.

On the collection level, we can define sample-based completeness as an aggregation of the completeness values of all its samples. In addition, attribute-based completeness can be defined for each declared attribute in a collection e.g. as the fraction of non-missing values.

Data accuracy (Olson, 2003) reflects the need to represent the real world truthfully. Low accuracy means that the data is vague and not precise enough or plainly incorrect (not corresponding to reality). High-quality data is accurate. We distinguish between *syntactic accuracy metrics*, which measure the degree of correspondence between the data items and the syntactic constraints related to a given domain e.g. the ratio of valid calendar dates in the birthday attribute and *method-based accuracy metrics* which reflect the accuracy of the diagnostic methods used for collecting the data.

As an example of the method-based accuracy, the data for a tentative diagnosis will be less accurate, if this diagnosis is made through a rapid test rather than a thorough test because the rapid tests are typically designed to have a minimum fraction of false negatives, even at the expense of higher rates of false positives. Usually, thorough tests aim at minimizing both false negatives and positives. Method-based metrics are based on the metrics for method accuracy e.g. their sensitivity, specificity, or likelihood ratio (Mandrekar, 2010). Such metrics are connected to samples and collections e.g. by making the collection methods declared in their descriptions and instantiated for the sample attributes. Then the collection metrics are aggregated over samples e.g. as the average degree of method sensitivity for a collection.

Data reliability characterizes the underlying measured concept. For the medical data related to a concept (such as e.g. the depressive mood), reliability (Greiver *et al.*, 2012) can be defined as a degree to which a question triggering the collection of this data, represents a reliable measure of that concept. Another possible definition derives reliability of the data from the reliability of its source (Mavrogiorgou *et al.*, 2019). For example, the reliability of the data about the cause of death is much higher in a coroner's inquest, if it was provided by a trained pathologist after an autopsy, than if it was given by a general practitioner. Also, the reliability of immunity against measles is higher after a titer assessment, than if it was based on patients' memories of childhood diseases. Low reliability means that the data cannot be trusted, so its quality is low. Method-based data reliability reflects the reliability of the diagnostic methods (Kyriacou, 2001) e.g. calculated as their test-retest coefficients or split-half measures; it is connected to samples and collections e.g. as the sample-based average test-retest coefficient for a collection.

Data consistency reflects the need for non-conflicting data. For the medical data, sufficient consistency (Almeida *et al.*, 2019) (contributing to high quality) means that there is no contradiction in the data as the real-world states reflected by data items are not in conflict. It is measured as a reverse degree of variability with respect to the data collection method used within a sample or collection. Consistency on the collection level (sometimes also called uniformity) mainly informs whether the data items were derived with the same methods, with the intention that data derived with different methods might not be easily comparable. Consider, for example, the different tests and test procedures for a COVID-19 infection with their different properties, sensitivities and selectivities. We define data as more consistent if it was collected by a smaller number of methods or most of the items were collected by a small number of methods.

Data precision is defined as the degree of rule resolution (e.g. the number of used categories) for the values of categorical attributes (Lozano *et al.*, 2008) and the number of significant digits – for the values of numeric data attributes. For example, having just three categories for blood pressure is obviously not very precise and contributes to quality negatively. Other examples of precision are the scale for the ischemia time (which can be specified in hours, minutes or seconds) or the precision of blood pressure measurement

equipment. On the collection level, the precision is usually calculated for specific data attributes declared for a collection.

4.3 Metadata quality characteristics

The description of data quality is a part of the metadata of the data sets, i.e. it describes certain properties of the data. Now the values quantifying data quality characteristics are data as well and as such, they possess data quality characteristics. For example, it might be helpful to know, whether the accuracy of the documentation of data precision is high or low, i.e. to which degree we can trust the precision characteristics. Therefore, we now elaborate on metadata quality characteristics (Bruce and Hillmann, 2004; Margaritopoulos *et al.*, 2012; Stvilia *et al.*, 2004).

Metadata accuracy (Bruce and Hillmann, 2004; Stvilia *et al.*, 2004) reflects the need for the metadata values to correspond to its domain. An example of the domain constraint served well by such metric, is the non-negativity constraint for the data accuracy value domain. In a more general sense, metadata accuracy signifies the level of meeting the requirements of metadata management.

Metadata completeness (Bruce and Hillmann, 2004; Király and Büchler, 2018; Margaritopoulos *et al.*, 2012) reflects the need of supplementing all data items with the corresponding quality metadata. We define it as a degree of completeness with respect to the metadata connected to the data values. It can be calculated for a specific sample as a degree of presence of metadata values for its attributes or for a collection – as an average of all metadata completeness values for its samples. As mentioned above, in our architecture we deal directly only with collection completeness. It can be also based on a subset of attributes for a given collection, reflecting the degree of presence of the metadata values connected to these attributes.

Metadata timeliness (Bruce and Hillmann, 2004) reflects the need for the metadata to reflect the real state of the data items. It can be defined as a reverse distance of time between creating the data attribute value and creating its supplementing metadata values. An example of low metadata timeliness is the case when collecting the information about the diagnostic method used to collect the disease data is done in two years after collecting the data item.

Biobanks have the responsibility to assess and describe the data quality management, documentation and assessment of the data sets they offer. Only with reliable descriptions of the data quality research may assess, whether data can be used for their intended studies and whether these data sets can be processed in combination with other data sets.

5. Generic architectures for federated data lakes

5.1 Introduction

Networks of biobanks can be seen as a federated data lake architecture providing an infrastructure for searching for useful data sets. We focus on the question, which information biobanks can publish to support the search for appropriate cases for medical studies.

Biobanks store biological material called samples (e.g. piece of tissue from a biopsy, blood, saliva, etc.). Samples collected together according to some criteria are called a collection (e.g. a collection of colon cancer tissue and a collection of blood samples from marathon runners).

Samples and collections are annotated with *data*. As direct access to the sample data is not possible, as outlined above, researchers have to search within the metadata describing

the sample data instead of the sample data themselves. Data lakes support such searches as they rely on the management of *metadata* to avoid data swamps (Nargesian *et al.*, 2019).

We consider that such search-supporting metadata has to describe mainly two types of information:

(1) *Data content*, i.e. the information stored as the medical data accompanying the biobank samples, but not directly accessible to the public; and

(2) *Data quality*, i.e. the information about the quality of the medical data stored for the biobank.

Content-describing metadata includes the schema of sample data, the concepts of reference ontologies connected to collections. Quality-describing metadata is represented by quality metric values calculated from data quality properties of sample data or collection data. This information, which can be shared without any privacy or security concerns allows to search for relevant collections but not for individual samples. Queries can refer to values of content metadata (e.g. find all collections containing *BMI* data) and data quality values (e.g. find the collections where the ratio of empty values for the BMI attribute is below 50%). Our approach for including values of data attributes in addition to the publicly available data on collections in the BBMRI biobank directory (Holub *et al.*, 2016) in the search is out of scope for this paper.

The result of a query against the proposed meta-structure is a set of biobanks and collections. Researchers then contact the biobanks (e.g. through the BBMRI negotiator or through email) for gaining access to the sample data.

As biobanks and also data lakes usually accumulate data sets, the maintenance typically requires registering meta-data about additional collections or additional content descriptors (e.g. the descriptors for ontological concepts related to COVID-19). Quality descriptors can be updated whenever new relevant data is available. For example, modifying non-empty values for sample attributes does not cause updates of data completeness descriptors, whereas filling previously empty values may trigger such updates.

5.2 Basic set of concepts

Based on the above considerations, we propose to organize the metadata schema storage facilitating the search in federated medical data lakes based on the set of abstract concepts depicted in Figure 1 (we depict entity concepts as boxes and relation concepts as diamonds). Below we provide a detailed description of these concepts.

We start with defining collections. A *Collection* contains a set of *Collection Data Items* and declares a set of *Data Attributes* to be instantiated within the scope of data items.

To address the biobank domain, sample collections are specialized as *biobank sample collections*, whereas collection data items are specialized as *Sample Data Items* i.e. the data items accompanying biobank samples. An example of the biobank collection can be the colorectal cancer cohorT developed as a part of the ADOPT BBMRI project (http://www.

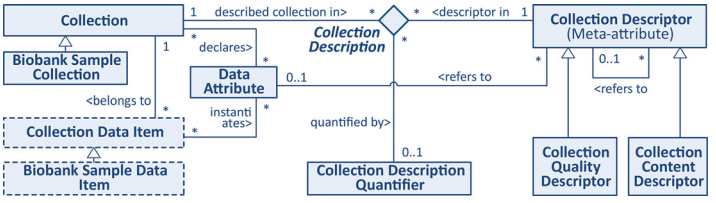

Figure 1.
Basic set of concepts for federated search in medical data lakes

bbmri-eric.eu/scientific-collaboration/adopt-bbmri-eric), which includes a set of sample data items with such attributes as the patient's age, diagnosis code, and BMI.

To facilitate metadata-based search, we propose to describe a collection by the collection meta-schema, which consists of a set of meta-attributes *(Collection Descriptors*). A *Collection Description* relation connects collections and collection descriptors: its instance indicates that the specific instance of the collection descriptor describes the specific collection.

A collection descriptor always describes a collection but can refer to additional concepts. We distinguish the following cases:

- A Collection Descriptor describes a Collection as a whole: it can be based on all values within a Collection or collection-wide properties such as its size. In this case, only the instance of the Collection Description relation is necessary.
- It refers to a Data Attribute declared by a Collection: e.g. being based on values of that attribute. An example can be a percentage of empty values of a given attribute within a Collection. It is represented by an optional "refer to" relation pointing to the involved Data Attribute.
- It refers to another Collection Descriptor declared by a Collection, this allows describing metadata characteristics. It is represented by an optional "refer to" relation pointing to the involved Collection Descriptor.

We define two types of Collection Descriptors to be treated in detail in the following sections:

- *Collection Content Descriptors* describing the content of the collection; these descriptors can e.g. refer to concepts belonging to reference ontologies; and
- *Collection Quality Descriptors* describing the quality of the collection as a whole or the aggregated quality of its items; these descriptors can e.g. hold values of data quality metrics.

It is important to note that, as the collection meta-schema includes only Collection Descriptors, individual Collection Data Items are not accessible within our search architecture, they cannot be returned by queries or described individually, instead, they can only be used to calculate aggregate values for Collection Descriptors. In Figure 1, this is indicated by depicting the Collection Data Item and the Biobank Sample Data Item with a dashed border.

A Collection Description relation can be optionally quantified by a *Collection Description Quantifier*. It can be a quantifying value itself or possess attributes holding quantities calculated for the specific Collection given the specific instance of the collection descriptor such as the number of data items in a given collection which can be described with such an instance (*Data Item Cardinality* attribute), we omit such *quantifier attributes* on a diagram. The queries supported by our search architecture ask for the values of quantifier attributes or for the presence of the given instance of the Collection Description relation.

5.3 Describing collection content

Publishing metadata in form of schemas faces the difficulty that the schemas are very heterogeneous. Even if a uniform schema representation (such as XML Schema) is used, too many differences between different biobank data collections remain and, therefore, searching through a single point of service would be quite difficult. In addition, the schema contains the data model but does not contain sufficient information about the content.

Due to this heterogeneity and the differences in metadata representation, the contents of a data set are best represented by concepts of a reference ontology. Thus, we propose that

each federation of data lakes should declare such an ontology. Then in each data lake, the available data sets provide metadata characterizing their content by referring to concepts of this reference ontology. Each data lake offers then possibilities to search for data sets specifying a set of these concepts. The metadata can also be exchanged between data lakes and search infrastructures for the whole federation can be established. The level of support and the specificity of both queries and results then depend on the level of details provided by the reference ontology.

It is not necessary that in the whole federation only one reference ontology is used. If there are several reference ontologies, then there is a need for providing mappings between the different ontologies (Kalfoglou and Schorlemmer, 2003). As ontologies might evolve over time (e.g. the different versions of the ICD ontology for encoding diagnosis), they require a temporal representation and mappings between different ontology versions to support search over evolving data collections or data collected in different periods (Eder and Koncilia, 2004).

Conceptually, we propose to make a collection content descriptor refer to an *ontological concept* belonging to a specific *reference ontology* (Figure 2). Such concepts are connected by *ontological relations* (such as *is-a* and *part-of*). A collection can own several reference ontologies, we omit details of their mapping here.

In the following section, we propose an approach to use concepts belonging to a particular medical data description standard to define collection content descriptors. Note that we do not consider this approach to be uniquely applicable: alternative approaches (e.g. involving other public ontologies) also can be used for this purpose.

5.3.1 Using LOINC concepts in collection content descriptors. For medical data sets, LOINC [Logical Observation Identifiers Names and Codes (https://loinc.org)] provides a suitable basis for representing the contents of a data collection, so we propose to specialize reference ontology used for describing collection content by means of *LOINC* and the concepts described by this standard – by means of *LOINC concepts*. Together with HL7 (www.hl7.org), LOINC is recommended as a standard for the exchange of health-care documents.

In this section, we propose an approach for using LOINC concepts to define collection content descriptors, which is based on the conceptual schema shown in Figure 3.

LOINC is a standard for clinical nomenclature, which provides a code system for identifying laboratory and clinical observations and test results (examples of such information could be vital signs, level of blood hemoglobin, etc.) It defines a set of *LOINC codes,* which can be used to substitute for the observation and test information in documents, software applications or messages.

The information coded by a single LOINC code (*coded information*) is structured as containing five or six *LOINC parts* (https://loinc.org/download/loinc-users-guide):

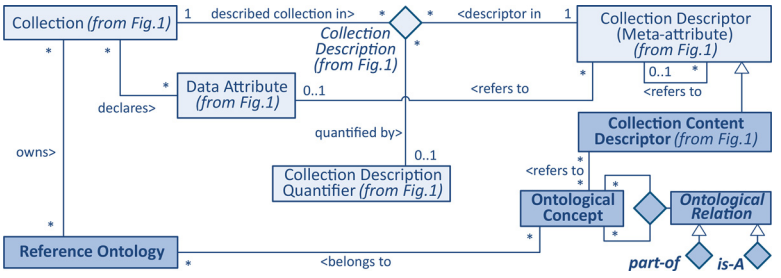

Figure 2.
Concepts for describing collection content

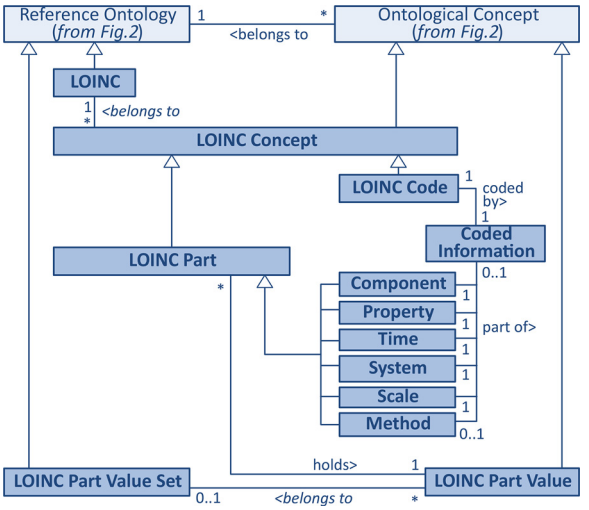

Figure 3.
LOINC concepts applicable for describing collection content

- *Component*: the name of the measured component or analyte (e.g. glucose and propranolol); for diagnostic tests, it can refer to the disease information e.g. its ICD code;
- *Property*: the property which was observed as a result of a test or observation (e.g. substance concentration, mass and volume);
- *Time*: the time aspect of the measurement, e.g. is it periodical (over time) or singular (momentary);
- *System*: the type of a system or sample (e.g. urine, serum or the patient as a whole);
- *Scale*: the scale of measurement (e.g. nominal or range); and
- *Method*: the method of the measurement (e.g. radioimmunoassay and immune blot); this part is optional.

A LOINC part serves as a container holding a *LOINC part value*, for example, a *property* part can contain the value of "CCnc (catalytic concentration)." LOINC part values belong to the domains represented by *LOINC part value sets*, e.g. "CCnc (catalytic concentration)" can belong to a value set of "enzymatic activity properties." We propose to treat such values as ontological concepts, so a LOINC part value set can be seen as a specification of a reference ontology.

An example of the LOINC code for the primary diagnosis ICD code and the corresponding coded information is as follows:

(1) LOINC code: *86255-7*.
(2) Coded information:
(3) *Component:* Primary diagnosis ICD code.
(4) *Property:* Type (known type).
(5) *Time:* Pt (point measure).
(6) *System:* Patient (patient as a whole).
(7) *Scale:* Nom (nominal).

5.3.2 LOINC as collection content descriptors. The specific LOINC code can serve as a content descriptor for a specific collection or its specific data attribute, declaring them as matching the corresponding coded information as a whole. This representation affects the search as follows:

- If the search value is the LOINC code, all collections connected to this code are returned by the search. e.g. the LOINC code "86255-7" can be used as a search value to match all the collections described by this code.
- If the specific LOINC part value is specified as a search value and the collection is described by the specific LOINC code, the value for the given LOINC part of the corresponding coded information must be retrieved and if it matches the search value, the collection is returned by the search. For example, the value "nominal" for a "scale" part can be used as a search value to match all the collections described by LOINC codes representing the coded information where the "scale" part holds this value.

5.3.3 LOINC part values as collection content descriptors. The specific LOINC part value together with the information to which LOINC part it belongs can also serve as a descriptor for the specific collection or its specific data attribute. For example, all the collections where the data was collected by measurement using the nominal scale can be described by the "nominal" part value connected to the "scale" LOINC part.

This representation affects the search as follows:

- If the search value refers to a LOINC part value, the collection is returned if this value matches the value specified as its descriptor. For example, the above "nominal" part value can be used to match all the Collections described by this part value.
- If the search value is the LOINC code, the collection is returned if it is described by all part values forming this code.

5.3.4 Content-related query example. We illustrate the usage of collection content descriptors by means of the following content-related query:

> Find all collections referring to a primary diagnosis.

We suppose that:

- the query UI is able to connect the LOINC codes to the concepts mentioned in the query; and
- the meta-schema describes a set of Collections with LOINC codes; a subset of them (*subset ICD*) is described by the "86255-7" LOINC code indicating that they hold ICD disease codes.

The match is performed as follows:

- Select the LOINC code descriptor corresponding to the information specified in the query. Here, we rely on the ability of the UI to connect LOINC codes to concepts: as the query mentions primary diagnosis, the UI provides the corresponding "86255-7" LOINC code to the system as well, so the corresponding LOINC code descriptor is selected.
- Return all collections described by the LOINC code descriptor selected on Step 1: for this, check the description relations for all collections and select the collections for which the other end of such a relation points to the selected descriptor (returning collections belonging to subset ICD).

5.4 Describing collection quality

Besides the content of data sets, the search requires information about the quality of the data in the data sets, as was discussed in Section 4. We propose to treat collection data and metadata quality in our framework based on the conceptual schema shown in Figure 4. In the following sections, we explain these concepts in detail.

5.4.1 Measuring quality. Quality is measured by applying a *quality metric* to a *measurable element* or a set of measurable elements by means of *quality measurement* to obtain a *quality metric value*. A quality metric defines for its values: a *measurement unit* (e.g. meters or seconds) and a *scale* (a set of acceptable values, optionally with a defined order), which belongs to a *scale type* (ratio, ordinal, nominal, etc.) (Bertoa *et al.*, 2006).

A quality metric is defined for a *data quality characteristic* e.g. one of the characteristics defined in Section 4. For example, a completeness metric such as "the percentage of empty values" is defined for data completeness.

Measurable elements represent different types of quality described in Section 4. They can be one of the following:

- Collection data items, corresponding to the *data item quality*; such elements participate in the calculation of aggregated quality metrics and do not define quality metric values directly.
- Collections or their specific data attributes, corresponding to the *collection quality*; such quality values can be either the result of the aggregation of data item quality values or calculated directly based on the collection properties.
- Collection descriptors, corresponding to the *metadata quality* such as its completeness or accuracy.

For the biobank domain, we specialize the data quality metrics as *biobank data quality metrics* and data quality characteristics as *biobank data quality characteristics*.

5.4.2 Collection quality descriptors. A *collection quality descriptor* does not directly refer to a measured quality (i.e. a quality metric value obtained as a result of a quality measurement). Instead, it defines a set or a range of quality metric values for a specific data quality metric (defined by a "refers to" relation pointing to such a metric). The measured quality metric value for a specific measurable element (e.g. a collection) must fall into this

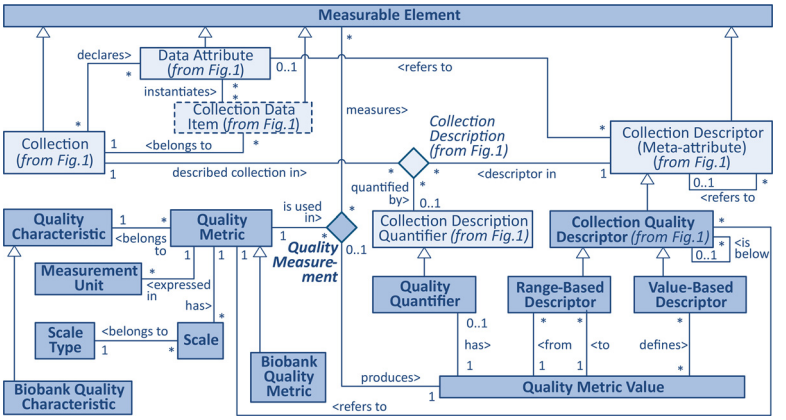

Figure 4.
Concepts for describing collection quality

range or be an element of this set for this descriptor to be connected to the given element by a collection description relation (more about that below).

The collection quality descriptor inherits the ability to be applied to a whole collection or to refer to a data attribute or to another collection descriptor from a collection descriptor. By default, it defines a range or a set of values for a quality metric applied to a whole Collection. If it is necessary to apply a metric to a specific attribute or to a meta-attribute (another collection descriptor), the corresponding attribute or meta-attribute can be connected to a descriptor by means of "refers to" relations. This way e.g. it is possible to define a collection quality descriptor for a "percentage of empty values" metric applied to a "patient age" attribute in a collection.

We distinguish two types of collection quality descriptors:

(1) A *range-based descriptor* defines a range of quality metric values by referring to a pair of such values by means of *from* and *to* relations. The value at the end of a *from* relation defines a lower limit for a range, the value at the end of a *to* relation defines an upper limit. Such descriptors correspond to data quality metrics with the possible values forming a continuous range or when the cardinality of the scale set is high. For example, such a descriptor can define a range between 0 and 20% for a "percentage of empty values" metric applied to a "patient age" attribute.

(2) A *value-based descriptor* defines a set of quality metric values. Such descriptors correspond to data quality metrics with a Scale set of low cardinality (e.g. containing just values of "low," "medium" and "high"). For example, such a descriptor can define a set of values for a "reliability of a collection method" metric containing just a value of "high."

Collection quality descriptors can form a hierarchy, this is represented by the "is below" relation pointing to the upper-level descriptor. For example, the range-based descriptors defining wider ranges can reside on higher levels of the hierarchy, with the descriptors for narrower ranges residing below them. e.g. the descriptors for the ranges of 0–25% and 25–50% can reside below the descriptor for the range of 0–50%. For the value-based descriptors, the sets for the upper-level descriptors can include the sets for the descriptors below them. This way, it can be possible to drill down to a higher level of detail.

If the range for a quality metric value is known (e.g. the ratio scale implies the range of 0–100%), it is possible to create a hierarchy of collection quality descriptors in advance, e.g. in a form of an interval or segment tree, to facilitate search by specifying intervals or specific values. The detailed description of the involved data structures is outside the scope of this paper.

5.4.3 Forming collection description relations. The presence of a collection description relation indicates that the specific collection possesses a quality metric value (measured based on its own properties or, optionally – on additional measurable elements such as its attributes or other collection descriptors) which matches a specific collection quality descriptor:

- for range-based descriptors, the collection description relation is present, if the measured quality metric value for a collection falls into the range defined for this descriptor; and
- for value-based descriptors, this relation is present, if the measured quality metric value for a collection belongs to the set of values defined for this descriptor.

For example, suppose that Collection A has 70% of empty values for a patient's age and Collection B has 30% of such values. If the meta-schema defines a range-based

collection descriptor for a "percentage of empty values" quality Metric referring to a "patient age" data attribute and instantiates it for two ranges: 0–50% ("Descriptor < 50") and 50–100% including the 50% value ("Descriptor >= 50"), then the Collection A will be connected to the "Descriptor >= 50" and Collection B – to the "Descriptor < 50."

If we introduce the upper-level descriptor e.g. "all" for the range of 0–100%, both collections have to be additionally connected to that descriptor. It is also possible to keep connections only for bottom-level descriptors and treat the sets of connections for the upper-level descriptors as unions of lower-level sets.

5.4.4 Quality quantifiers. The above concepts allow checking if the quality metric value for a collection falls within a predefined range. To support queries referring to exact quality metric values for collections, the collection description quantifier is specialized as a *quality quantifier,* which holds a specific quality metric value measured over a set of measurable elements, e.g. over a whole collection, over its data attributes or other collection descriptors.

In our above example, the collection description connecting Collection B and "Descriptor < 50" will refer to a quality quantifier holding the value of 30, whereas the quality quantifier referred to by a collection description connecting Collection A and "Descriptor >= 50" will hold the value of 70.

5.4.5 Quality-related query example. We illustrate the usage of collection quality descriptors by means of a sample query, which asks only for quality-related information.

First, we suppose that the query interface allows the user to select from the predefined intervals for quality values, where the interval boundaries are taken from the existing collection quality descriptors (e.g. the user can choose from "the patient age is empty in less than 50% cases" and "the patient age is empty in not less than 50% cases" for our above set of descriptors).

Now the query can be formulated as follows: "find all collections where the percentage of empty values for the patient age is less than 50%."

To answer this query, the system first relies on the fact, that the range in the above query was taken from the UI referring to the existing "Descriptor < 50," so it takes this descriptor as the starting point. After that, it selects all instances of the collection description relation pointing to that descriptor and returns the corresponding collections. In our case, only Collection B is returned.

If the range in a query does not match any descriptor, it is possible to start from an upper-level descriptor with a range covering the specified range and then filter the connected collections based on their quality quantifier values.

Suppose the query is formulated as follows: "find all collections where the percentage of empty values for the patient age is between 20% and 80%." In this case, no lower-level descriptor matches the specified range. The system starts with the upper-level descriptor with a range covering 0–80% i.e. from the "all" descriptor covering the full range of values. After that, the system takes all the quality description relations pointing to the "all" descriptor and filters them based on their quality quantifier values to select only those which fit into the 20–80% range. As a result, both Collections A and B are returned.

6. Discussion and application

The proposed meta-schema with a strong focus on data quality and metadata quality is a big step in supporting the search for relevant data sets and collections in federated data lakes with high privacy requirements. For our application areas of biobanks and medical data sets, this meta-schema can be seen as an extension to the biobank catalog (Hofer-Picout

et al., 2017) which gives additional information about the data sets annotating the sample collections. With this additional information available the search infrastructure effectively filters collections and biobanks which do not collect and offer the requested data in the necessary data quality. This filtering step reduces the search space considerably.

The quality information can, in particular, also be used for an objective ranking of candidate data sets. We also observed that disease groups and typically available content descriptors together with data quality descriptors already provide quite reasonable filters for collections to reduce search spaces. For example, cohort studies of population-based biobanks typically have no LOINC code 22637–3 (pathology report final diagnosis narrative) or feature such data only with very high levels of empty values. On the other hand, cases of cancer therapy typically offer all this data content. The significance of such correlations in general and their usefulness is currently explored. Therefore, the proposed meta-schema is an important building block to scale up to other search infrastructures such as *sample negotiator* or *sample locator* to the ever-increasing demands of rapidly growing numbers of biobanks, collections and available data sets.

In this paper we did not discuss the problems related to the data instances, (e.g. to search for cases with particular diseases, body characteristics, etc.) because a thorough discussion of the approaches to support such searches without imperiling the privacy of sample donors exceeds the scope of this paper. Search in the proposed meta-structure can easily be combined with the available data in the biobank directory. Precisely answering queries involving additional data instances would open attack possibilities for adversaries applying trackers seeking to wrongfully extract information from the data sets. Nevertheless, it is possible to implement truly anonymous multi-dimensional indexes based on truly anonymous data sets derived from the original data sets to further reduce the search space for promising cases and collections while guaranteeing all necessary privacy requirements (Eder *et al.*, 2012).

7. Conclusions

Supporting effective and efficient search for data sets for medical studies is the premier aim of the presented research. The search is complicated as these data sets are not easily available, are heterogeneous and are maintained anomalously by many different institutions. We described the problem as the more generic problem of a federation of data lakes and developed a generic metamodel, which allows registering the contents, i.e. the schemas of the available data sets without risking violating privacy requirements. The particular focus of this work is the integration of data quality into the meta schema to be able to consider data quality requirements of medical studies to improve the search for suitable data sets. Management of metadata and data quality, thus, aims to contribute to improving medical research in both quality and efficiency.

References

Almeida, J., Santos, M., Polónia, D. and Rocha, N.P. (2019), "Analysis of the data consistency of medical imaging information systems: an exploratory study", *Procedia Computer Science*, Vol. 164, pp. 508-515.

Asslaber, M., Abuja, P., Stark, K., Eder, J., Gottweis, H., Trauner, M., Samonigg, H., Mischinger, H.J., Schippinger, W., Berghold, A. and Denk, H. (2007), "The genome Austria tissue bank (GATIB)", *Pathobiology*, Vol. 74 No. 4, pp. 251-258.

Batini, C. and Scannapieco, M. (2016), *Data and Information Quality: Dimensions, Principles and Techniques*, Springer.

Berger, S. and Schrefl, M. (2008), "From federated databases to a federated data warehouse system", *Proceedings of the 41st Annual HI International Conference on System Sciences (HICSS 2008)*, IEEE, pp. 394-394.

Bertoa, M.F., Vallecillo, A. and García, F. (2006), *An Ontology for Software Measurement, Ontologies for Software Engineering and Software Technology*, Springer, pp. 175-196.

Bogatu, A., Fernandes, A.A., Paton, N.W. and Konstantinou, N. (2020), "Dataset discovery in data Lakes", *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, IEEE, pp. 709-720.

Brackenbury, W., Liu, R., Mondal, M., Elmore, A.J., Ur, B., Chard, K. and Franklin, M.J. (2018), "Draining the data swamp: a similarity-based approach", *HILDA 2018*, pp. 1-7.

Bruce, T.R. and Hillmann, D.I. (2004), *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*, Metadata in Practice, American Library Association, pp. 238-256.

Dogac, A., Laleci, G.B., Kirbas, S., Kabak, Y., Sinir, S.S., Yildiz, A. and Gurcan, Y. (2006), "Artemis: deploying semantically enriched web services in the healthcare domain", *Information Systems*, Vol. 31 Nos 4/5, pp. 321-339.

Eder, J. and Koncilia, C. (2004), "Modelling changes in ontologies", *OTM 2004*, Springer, pp. 662-673.

Eder, J. and Shekhovtsov, V.A. (2020), "Data quality for medical data lakelands", in Dang, T.K., Küng, J., Takizawa, M. and Chung, T.M. (Eds), *FDSE 2020, Vol. 12466 of LNCS*, Springer, pp. 28-43.

Eder, J., Lehmann, M. and Tahamtan, A. (2006), "Choreographies as federations of choreographies and orchestrations", *International Conference on Conceptual Modeling*, Springer, pp. 183-192.

Eder, J., Gottweis, H. and Zatloukal, K. (2012), "IT solutions for privacy protection in biobanking", *Public Health Genomics*, Vol. 15 No. 5, pp. 254-262.

Eder, J., Dabringer, C., Schicho, M. and Stark, K. (2009), *Information Systems for Federated Biobanks, Transactions on Large-Scale Data- and Knowledge-Centered Systems I*, Springer, pp. 156-190.

Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2019), "Leveraging the data lake: Current state and challenges", *DaWaK 2019*, Springer, pp. 179-188.

Golfarelli, M. and Rizzi, S. (2018), "From star schemas to big data: 20+ years of data warehouse research", *A Comprehensive Guide through the Italian Database Research over the Last 25 Years*, Springer, pp. 93-107.

Greiver, M., Barnsley, J., Glazier, R.H., Harvey, B.J. and Moineddin, R. (2012), "Measuring data reliability for preventive services in electronic medical records", *BMC Health Services Research*, Vol. 12 No. 1, p. 116.

Hai, R., Geisler, S. and Quix, C. (2016), "Constance: an intelligent data lake system", *SIGMOD/PODS 2016*, pp. 2097-2100.

Hainaut, P., Vaught, J., Zatloukal, K. and Pasterk, M. (2017), *Biobanking of Human Biospecimens: principles and Practice*, Springer.

Hofer-Picout, P., Pichler, H., Eder, J., Neururer, S.B., Müller, H., Reihs, R., Holub, P., Insam, T. and Goebel, G. (2017), "Conception and implementation of an Austrian biobank directory integration framework", *Biopreservation and Biobanking*, Vol. 15 No. 4, pp. 332-340.

Holub, P., Swertz, M., Reihs, R., van Enckevort, D., Müller, H. and Litton, J.E. (2016), "BBMRI-ERIC directory: 515 biobanks with over 60 million biological samples", *Biopreservation and Biobanking*, Vol. 14 No. 6, pp. 559-562.

Inmon, B. (2016), *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*, Technics publications.

Kalfoglou, Y. and Schorlemmer, M. (2003), "Ontology mapping: the state of the art", *The Knowledge Engineering Review*, Vol. 18 No. 1, pp. 1-31.

Király, P. and Büchler, M. (2018), "Measuring completeness as metadata quality metric in Europeana", *Big Data 2018*, IEEE, pp. 2711-2720.

Kyriacou, D.N. (2001), "Reliability and validity of diagnostic tests", *Academic Emergency Medicine*, Vol. 8 No. 4, pp. 404-405.

Lemke, A.A., Wolf, W.A., Hebert-Beirne, J. and Smith, M.E. (2010), "Public and biobank participant attitudes toward genetic research participation and data sharing", *Public Health Genomics*, Vol. 13 No. 6, pp. 368-377.

Litton, J.E. (2018), "BBMRI-ERIC", *Bioreservation and Biobanking*, Vol. 16 No. 3.

Lozano, L.M., García-Cueto, E. and Muñiz, J. (2008), "Effect of the number of response categories on the reliability and validity of rating scales", *Methodology*, Vol. 4 No. 2, pp. 73-79.

Mandrekar, J.N. (2010), "Simple statistical measures for diagnostic accuracy assessment", *Journal of Thoracic Oncology*, Vol. 5 No. 6, pp. 763-764.

Margaritopoulos, M., Margaritopoulos, T., Mavridis, I. and Manitsaris, A. (2012), "Quantifying and measuring metadata completeness", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 4, pp. 724-737.

Mavrogiorgou, A., Kiourtis, A. and Kyriazis, D. (2019), "Delivering reliability of data sources in IoT healthcare ecosystems", *FRUCT 2019*, IEEE, pp. 211-219.

Merino-Martinez, R., Norlin, L., van Enckevort, D., Anton, G., Schuffenhauer, S., Silander, K., Mook, L., Holub, P., Bild, R., Swertz, M. and Litton, J.E. (2016), "Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 core)", *Biopreservation and Biobanking*, Vol. 14 No. 4, pp. 298-306.

Mihaila, G.A., Raschid, L. and Vidal, M.E. (2000), "Using quality of data metadata for source selection and ranking", *WebDB (informal proceedings)*, pp. 93-98.

Müller, H., Dagher, G., Loibner, M., Stumptner, C., Kungl, P. and Zatloukal, K. (2020), "Biobanks for life sciences and personalized medicine: importance of standardization, biosafety, biosecurity, and data management", *Current Opinion in Biotechnology*, Vol. 65, pp. 45-51.

Nahm, M. (2012), *Data Quality in Clinical Research, Clinical Research Informatics*, Springer, pp. 175-201.

Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q. and Arocena, P.C. (2019), "Data lake management: challenges and opportunities", *Proceedings of the VLDB Endowment*, Vol. 12 No. 12, pp. 1986-1989.

Olson, J.E. (2003), *Data Quality: The Accuracy Dimension*, Morgan Kaufmann.

Radulovic, F., Mihindukulasooriya, N., García-Castro, R. and Gómez-Pérez, A. (2018), "A comprehensive quality model for linked data", *Semantic Web*, Vol. 9 No. 1, pp. 3-24.

Roski, J., Bo-Linn, G.W. and Andrews, T.A. (2014), "Creating value in health care through big data: opportunities and policy implications", *Health Affairs*, Vol. 33 No. 7, pp. 1115-1122.

Sawadogo, P. and Darmont, J. (2020), "On data lake architectures and metadata management", *Journal of Intelligent Information Systems*, Vol. 56 No. 1, pp. 1-24.

Shaw, D. (2014), "Care. Data, consent, and confidentiality", *The Lancet*, Vol. 383 No. 9924, p. 1205.

Sheth, A.P. and Larson, J.A. (1990), "Federated database systems for managing distributed, heterogeneous, and autonomous databases", *ACM Computing Surveys*, Vol. 22 No. 3, pp. 183-236.

Skripcak, T., Belka, C., Bosch, W., Brink, C., Brunner, T., Budach, V., Büttner, D., Debus, J., Dekker, A., Grau, C. and Gulliford, S. (2014), "Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets", *Radiotherapy and Oncology*, Vol. 113 No. 3, pp. 303-309.

Spjuth, O., Krestyaninova, M., Hastings, J., Shen, H.Y., Heikkinen, J., Waldenberger, M., Langhammer, A., Ladenvall, C., Esko, T., Persson, M.Å. and Heggland, J. (2016), "Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research", *European Journal of Human Genetics*, Vol. 24 No. 4, pp. 521-528.

Stark, K., Eder, J. and Zatloukal, K. (2006), "Priority-based k-anonymity accomplished by weighted generalisation structures", *DaWaK 2006*, Springer, pp. 394-404.

Stvilia, B., Gasser, L., Twidale, M.B., Shreeves, S.L. and Cole, T.W. (2004), "Metadata quality for federated collections", *ICIQ 2004*, pp. 111-125.

Vaisman, A. and Zimányi, E. (2014), *Data Warehouse Systems*, Springer.

van Ommen, G.J.B., Törnwall, O., Bréchot, C., Dagher, G., Galli, J., Hveem, K., Landegren, U., Luchinat, C., Metspalu, A., Nilsson, C. and Solesvik, O.V. (2015), "BBMRI-ERIC as a resource for pharmaceutical and life science industries: the development of biobank-based expert centres", *European Journal of Human Genetics*, Vol. 23 No. 7, pp. 893-900.

Vuorio, E. (2017), "Networking biobanks throughout Europe: the development of BBMRI-ERIC", *Biobanking of Human Biospecimens*, Springer, pp. 137-153.

Wieringa, R.J. (2014), *Design Science Methodology for Information Systems and Software Engineering*, Springer.

Zatloukal, K. and Hainaut, P. (2010), "Human tissue biobanks as instruments for drug discovery and development: impact on personalized medicine", *Biomarkers in Medicine*, Vol. 4 No. 6, pp. 895-903.

Corresponding author
Vladimir A. Shekhovtsov can be contacted at: volodymyr.shekhovtsov@aau.at