ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2025, Vol. 86, No. 4, pp. 343–357.

Data Quality Management in Problem-Solving Using Research Infrastructures over Heterogeneous Data Sources

N. A. Skvortsov

Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russia
e-mail: nskv@mail.ru

Received November 29, 2024
Revised January 10, 2025
Accepted January 14, 2025

Abstract—Problem-solving based on available research data, especially in the context of open science and research infrastructures, should ensure the possibility of their multiple reuse. Data quality metrics are important characteristics that affect not only the accuracy of research results but also the assessment of data suitability, the feasibility of solving specific research problems, the choice of methods for working with data, object matching, data compatibility, and other aspects of reuse. This requires an assessment of various data quality dimensions at different levels of aggregation, from entire datasets to individual values. This study presents an approach to integrated data quality management based on data specifications, as well as data and metadata quality requirements. Various data quality assessment dimensions, including accuracy, completeness, and provenance, are discussed. The developed approach is applied to problem-solving using multiple data sources in stellar astronomy.

Keywords: data quality, data reuse, formal specifications, non-functional requirements

DOI: 10.31857/S0005117925040055

1. INTRODUCTION

Research inevitably encounters the need to assess the quality of the available data used in problem-solving [1]. At different stages of scientific research, there is a need to assess the quality of data. When searching and selecting suitable datasets, it is important to evaluate their applicability based on information about the quality of the data in these datasets. When creating samples that include only suitable data, as well as when data cleaning and improving to prepare them for research, it is necessary to evaluate the quality of data related to specific objects or characteristics. When using research methods, it is important to take into account the quality of the source data both to assess the quality of the results obtained and to assess the quality of the methods themselves. Thus, data quality metrics are an important and even necessary element in solving various research problems. The lack of information about the quality of the data or ignoring it can significantly reduce the quality of the results, even to their falseness, or make it impossible to conduct a study.

The volume of research data is steadily growing in many disciplines. Given the heterogeneity of research data sources and their diversity, problems arise with the compatibility of data obtained by different methods, created for different purposes, and with different quality requirements.

In some sources, information about the quality of the data presented may be contained in the description of datasets, accompanying documentation, or be known from external research. In other sources, the data is accompanied by quality assessments included in the structure of the

sets themselves. In some cases, there is no information about the quality of the data, but it can be assessed using statistical methods by analyzing the data itself or its samples. Data quality assessment methods can also vary significantly. Depending on the tasks being performed, different quality dimensions and different criteria for their evaluation may be important. In some cases, it is necessary to evaluate the quality of solutions in different ways based on information about the quality of the source data. Thus, heterogeneity is observed both in the metadata itself, concerning data quality, and in their application.

To support research and manage relevant data, research infrastructures are being created that accumulate data, provide services and metadata, ensuring their reuse. Due to their importance, approaches to data quality management in research infrastructures require thorough research. When working with large amounts of data and many heterogeneous sources, it is necessary to strive for automated quality management. Therefore, metadata regarding data quality must be clearly defined, accessible, and understandable to both humans and machines.

This article sets out the problem of developing the presentation and application of various data quality dimensions in research infrastructures. The following section provides an analysis of the state of research in this area. Section 3 provides a classification of approaches to data quality management. Metadata specifications and principles of their use in research infrastructures are described. Section 4 provides an example of data quality management for problem-solving in stellar astronomy using data from large multicolor photometric surveys of the sky. General conclusions are drawn about the development of research infrastructures to increase the effectiveness of research.

2. THE STATE OF RESEARCH IN DATA QUALITY MANAGEMENT

Quality problems are so sensitive in the national and global economy that approaches, methods, and standards have long been worked out to solve them. In computer science, data quality issues, as part of the general problem of the quality of real-world objects, play an equally important role. Databases and information systems are considered independent objects [1], which, like other objects in the real world, may have an acceptable or unacceptable state, which can be assessed by data quality metrics. At the same time, the data in them reflect the state of real-world objects, and it is necessary to evaluate both the correspondence of the state of information objects to the characteristics of real-world objects and the possibility of identifying real-world objects from the data. Ignoring data quality issues can lead to serious consequences not only in the information sphere but also negatively affect other areas of human activity.

Data on the qualitative and quantitative characteristics of real-world objects are used to assess their quality. Thus, the quality of objects (products, materials, and others) is always related to the quality of data about them. Moreover, in the modern research paradigm, which is based on the extraction of new knowledge from research data, collecting a large volume of data of varying quality about the studied objects from various sources requires assessment and consideration of their quality in research, which ultimately determines the development of science.

There may be requirements for data used in human activities. The quality of data (for example, characteristics such as accuracy and reliability) about real-world objects can determine the applicability and usage characteristics of both the data itself and the objects of research. Poor data quality, even if the objects under study themselves are of high quality, can make it difficult to solve problems and interfere with the proper use of both data and research objects.

For a long time, multi-criteria models that include sets of quality dimensions have been the prevailing approach in assessing data quality. Even in early studies, such as [2], the importance and relevance of various data quality dimensions were discussed. The main dimensions are accuracy, completeness, integrity, and relevance, as well as veracity, unambiguity, reliability, and volume of data. At the same time, the set of dimensions used depends on the objectives defined for a specific

research domain. Some studies emphasize the relationship between data quality and the quality of products and real-world objects, which confirms the effectiveness of such approaches to data quality assessment. From this point of view, the value of data is determined by its consumer, which draws attention to such quality dimensions as the significance and correctness of the data [3].

Over the course of research, an approach has been established in which the quality model is defined as a space of quality dimensions. In [4], the quality assessment methodology is defined by a model that includes a specific set of metrics, methods for evaluating them, and possibly methods for summarizing them. Depending on the tasks being performed, different methods for evaluating quality metrics can be used in each specific model, such as algorithms, rules, heuristics, or machine learning models that provide solutions to problems related to certain aspects of data quality.

For example, data completeness can be evaluated as the completeness of tuples (filling in all attributes), attributes (estimating the number of missing values among attribute values in tuples), or as coverage of all existing objects of a given type. The dimension of accuracy can be defined as the error of values, syntactic accuracy (the coincidence of names used for one object), or semantic accuracy (consistency of facts). The data volume can be determined by the amount of memory used or the expected number of tuples returned, and so on. In other words, the methods and implementations of quality assessment may depend both on the characteristics of datasets or samples and on specific limitations of the domain to which the data relate. The semantics of a particular metric, despite the same names, can vary significantly in different tasks. Accordingly, data quality requirements can be formulated in research problems both in terms of datasets or data warehouses and the applied problem domain.

Similar approaches are defined in current data quality standards. There are international and corresponding Russian national standards related to data quality management, in particular, ISO/TS 8000-1:2011 (GOST R 56214-2014 in Russia [5]) and a set of related standards. These standards comply with established principles of data quality modeling and introduce quality dimensions such as accuracy, completeness, and provenance (or source) of data without limiting specific implementations of quality assessment methods based on these criteria. Also, the standards define conceptual schemes for quality management and principles for developing criteria for evaluating these metrics, including principles of unambiguous syntactic and semantic coding. The quality of data within these standards indicates the extent to which the data meets the requirements of consumers and meets the established criteria. The principles of presenting requirements and improving data quality are established, allowing consumers to request data of appropriate quality and accurately determine whether the data obtained meets the established standards.

The ISO 19157:2013 [6] standard (GOST R 57773-2017 in Russia) is also interesting. It establishes the practice of quality management when working with spatial data sources with different resolutions and granularities. This standard allows you to select the most suitable data in terms of quality to solve problems at the required scale.

Problems with the availability of suitable research data arise in many data-intensive domains. In this regard, the field of research infrastructures is actively developing. They combine research data, services, and tools, allowing them to be used repeatedly.

One of the ideological foundations of their development has become the guiding principles of FAIR-data [7], which propose a direction for ensuring the findability (F), accessibility (A), interoperability (I) and, as a result, the reusability (R) of research data. When developing research infrastructures and how they work, data quality issues are often reduced to approaches for evaluating data warehouses in terms of compliance with FAIR principles.

Major interdisciplinary initiatives of the research community aimed at providing and managing data, as well as analyzing it, address issues of data quality, mainly from the perspective of FAIR

principles, shifting the focus from data quality management to assessing the quality of the data management infrastructure for the possibility of improving data quality.

The RDA initiative has developed the FAIR Data Maturity Model [8], which defines a set of indicators, their priorities, and methods for assessing compliance with the principles of FAIR. It is used as a general approach to evaluate different methodologies. In particular, the FAIRsFAIR [9] project was created to promote the principles of FAIR for data generated by researchers. Popular explanations, examples, and solutions have been developed that promote data interoperability and reuse, as well as tools, standards, and practices for data management in various scientific disciplines. Based on the FAIR Data Maturity Model and FAIRsFAIR developments, the F-UJI [10] tool has been implemented to assess the degree of compliance with the principles of FAIR research datasets and provide recommendations. Thus, such developments create an alternative approach to assessing data quality. Evaluating compliance with FAIR principles provides a tangible assessment of the quality of data management, rather than the data itself, basically. On the other hand, this may devalue the principles of FAIR itself as a strategic direction in the development of data management. After all, these principles remain promising until the declared possibility of autonomous data management by a machine (machine-actionability) is sufficiently impossible, that is, the transition from processing predefined sets of metadata and instructions by a machine to ensuring the correct interpretation of data and metadata that the machine has not previously worked with.

The ELIXIR initiative specializes in biomedical and biological information resources. Here, data quality issues are also more often discussed precisely in the context of compliance with FAIR principles, in particular, in the recommendations of the FAIR CookBook [11]. However, no specific recommendations have been developed on data quality management.

The ELIXIR [12, 13] research infrastructure (in particular, the Data and Interoperability platforms within it) supports the storage of large amounts of data, advanced metadata management, data integrity control, duplicate and data detection mechanisms, and other tools used to manage data quality. At the same time, there are no specialized tools for data quality specification apart from assessing compliance with the FAIR principles.

Within the framework of the ESIP initiative dedicated to data management in geo-sciences, research related to data quality management was conducted. The Data Quality Working Group (DQWG) and Information Quality Cluster (IQC) have developed recommendations for the application of best practices and standards in this domain, principles of data life-cycle management to ensure their integrity, quality and reuse, and have contributed to the implementation of these recommendations among data providers, software developers and research groups [14–16]. These solutions are primarily related to the established practice of working with spatial data, such as standardizing attributes and metadata, using quality flags in data, and compliance with specialized standards.

Also, within the framework of ESIP, a matrix for assessing the maturity of research data has been developed, combining the basic principles of data curation during long-term storage and utilization. Part of these principles is data quality management [17].

International standards for data quality management also continue to evolve. The set of spatial data quality management standards is in the process of updating [18].

The most common products related to data quality management are data cleaning tools as part of integration systems. The tasks solved in such products are detecting errors and duplicates, typecasting and standardizing the representation of values, solving the problem of missing data, and others. However, these products can also provide data quality assessment tools and algorithms for calculating quality metrics. Such products include, in particular, Talend Data Quality [19], which is part of a relational database integration system. This tool uses SQL-based business rules to

represent data requirements and monitor data quality based on dimensions such as completeness, accuracy, and consistency of data, with the ability to detect specific problem areas. IBM InfoSphere Quality Stage [20], part of IBM's product line for supporting data management processes, provides continuous monitoring of quality events in data streams, error correction, data cleanup, and metadata enrichment.

Semantic Data Quality Management (SDQM) [21] is based on the analysis of ISO 9000 standards and research of data quality models with sets of dimensions, as well as semantic Web technologies [22], including RDF technologies [23]. Semantic approaches based on RDF, a language with extensible semantics for describing resources, allow accompanying resources identified in the global information space with metadata, the semantics of which is defined by dictionaries or ontologies in various namespaces, setting resource requirements in terms of these dictionaries, and finding relevant resources. Moreover, the RDF model has become the most commonly used for defining metadata schemas. To define the quality model, the data quality management dictionary [24] is introduced, the data provenance model for web resources [25] and other dictionaries are reused.

Big Data Quality Management Framework (BDQMF) [26] is designed to manage the quality of big data at all stages of its life-cycle. BDQMF covers the entire data quality management process, from requirements definition and preprocessing to data analysis and quality monitoring. The main approaches include creating a data quality profile with quality targets, evaluating and improving quality at each stage, validating and optimizing quality rules to improve data accuracy, and monitoring and visualizing data after processing.

From the point of view of FAIR principles, data quality issues should be considered in the context of ensuring data reuse (R) [27]. The main principle (R1) in this direction indicates the need to describe data richly with accurate and up-to-date information (attributes). Such information refers to the resources and metadata needed to evaluate the reusability of data, rather than just describing its semantics. This information may include details about the provenance of the data, the license terms of its use, and other non-functional properties, which nevertheless play an important role in the selection of data by researchers to solve their research problems [28]. Metadata based on data quality metrics should also be included in this category.

The data quality standards developed by the W3C consortium include the PROV [29] standard, which establishes a model for describing the provenance of data. This standard defines a lot of non-functional metadata that is captured during any data manipulation. They include information about the authorship, the method of obtaining, transformations of the data, information about how some data were used as the sources when creating the data in question, and much more. However, in practice, the available data is most often accompanied by minimal information, limited by information about the authorship, affiliation, and time of creation of datasets, and rarely uses all the rich features provided by this standard.

In addition, the W3C consortium has developed an approach to data quality specification DQV (Data Quality Vocabulary) [30] based on RDF technologies. The definitions in this model allow you to describe a vocabulary of quality metrics for datasets, including sets of dimensions and methods for evaluating quality metrics, such as estimates of the volume, completeness, accuracy, and other characteristics of the contents of data catalogs. Specifications in the DCAT (Data Catalog Vocabulary) [31] format are used to communicate with the described data catalogs.

The DCAT specification allows you to describe catalogs, datasets, more general concepts of resources, their representations in various formats and sources, individual records, as well as data services that provide access to data through software interfaces for querying and obtaining the necessary parts of data. Thus, data quality specifications can be associated with entire catalogs and datasets, as well as with individual fragments of them. DCAT recommendations are evolving. Version 2 adds some data identification capabilities and backward references to the DQV model.

The recent version 3 adds support for versions and series of datasets. The mutual references with the DQV model provide enhanced data quality specification capabilities in the development of DCAT.

In general, approaches to building data quality models can be considered well-established and flexible enough for further development within the framework of existing concepts. However, despite this, such approaches are rarely used to describe data in the form of well-defined models. Metadata related to data quality is most often present in documentation in an arbitrary form or included directly in datasets but is not isolated as separate quality metadata and is poorly structured. This is especially noticeable, for example, in the case of errors in measuring the characteristics of objects. Often, the values of characteristics and their errors are presented in catalogs as equivalent attributes of an object.

3. QUALITY METADATA AND A DATA QUALITY MANAGEMENT APPROACH

Data quality specifications in the DQV [30] format offer an RDF schema that allows you to specify a structure for describing data quality. This scheme, among other things, includes:

- Sets of quality dimensions (Dimension): characteristics that are used to evaluate various aspects of data quality;
- Metric categories (Category): classification of metrics according to certain criteria, which facilitates their organization and use;
- Values of quality metrics (QualityMeasuration): specific values that reflect the quality level according to the specified metrics;
- Methods for evaluating quality metrics (Metric): methods for calculating or generating quality metrics, they describe exactly how the quality of the data was assessed, as well as the types of data that are used to represent the values of quality metrics;
- Datasets and their Representations (Distribution): linking quality metadata to specific datasets or copies of them in specific formats;
- Data Services (DataService): linking quality metadata to services that provide access to data.

A quality metadata model for a quality specification can be built based on specifications similar to those offered in DQV. However, such a basic model requires significant expansion to take into account the specification of quality metadata at various levels of data aggregation, the definition of actions with data depending on their quality, the establishment of quality requirements for the data being searched or created, and other important features.

In order to form the necessary types of specifications included in the quality model, the quality metadata was classified according to various criteria. Linking the elements of this classification allows you to define the specifications of quality metrics, establish their relationship with the described data, determine actions with data based on their quality, and formulate data requirements that need to be taken into account.

1. Classification by quality dimensions:
   - volume;
   - completeness;
   - accuracy;
   - provenance;
   - relevance;
   - reliability;
   - and others.
2. Classification by type of quality metric values:
   - boolean flag or bit vector;
   - real value;

- a set of categorical values of quality grades;
- a set of categorical values distinguishing types of violations or compliance with quality requirements;
- the same types with empty values (unknown quality).

3. Classification according to the method of value specification:
- constant;
- formulas or rules;
- tabular.

4. Classification by metadata source:
- metadata of the dataset;
- metadata in the data schema;
- external metadata repositories;
- metadata in publications;
- evaluation based on an experiment;
- statistical evaluation of data;
- evaluation on a representative sample of data;
- volatile metadata (requiring reassessment).

5. Classification by data aggregation method:
- dataset;
- relation;
- data sample (slice, query conditions);
- related semantics of entities or processes;
- tuple;
- specific entity or process;
- attribute;
- related attribute semantics;
- attribute value.

6. Classification of actions of data quality improvement:
- deleting data according to the level of aggregation (dataset, sample, tuple, entity, attribute, value);
- assigning quality metadata (metric, weight);
- selection of quality data;
- data enrichment from external sources;
- averaging methods based on data quality;
- averaging methods without taking into account data quality;
- ignoring data quality (including all data without regard to quality).

7. Classification of data quality requirements:
- restrictions on the composition and quality of the source data;
- requirements for improving the quality of the source data;
- declared required quality assessments of the results;
- characterization of the dependence of the quality of the results on the quality and characteristics of the source data;
- the need to evaluate the quality of the results or the inability to evaluate them.

In cases where the necessary metadata at the dataset level is missing or requires adjustments on subsets of data (for example, parameter slices), they can be evaluated statistically on all data or representative samples of them, and then existing metadata can be supplemented in the same format.

Quality metadata at the level of tuples or attribute values is usually included directly in the data presented in catalogs or datasets as additional relationship attributes.

Quality metadata related to a specific level of data aggregation is distributed to nested levels of aggregation. Thus, any data tuple or value used can be associated with quality metadata collected from different levels of data representation, including metadata of the dataset as a whole, metadata of attributes, values, and others. When defining requirements for quality metadata, restrictions can be used on the metadata of entire datasets, as well as on the metadata of specific data values or their provenance.

Quality weight assessment approaches based on quality metrics are often used to make decisions about further actions. The weight can be expressed as numerical values, as well as special labels, such as the absence of an assessment (nan) or indication of data removal (del). If there are no quality attribute values, the quality weight is set to nan, which does not affect the total quality weight of the tuple. If at least one weight value for a tuple is marked as del, it means that the tuple should be excluded from the analysis, and other weight values are not taken into account. The tuple quality weights defined at the sample or dataset level affect the overall quality weight of the tuple. The presence of multiple weights for a tuple is generalized by calculating their average value. The same principles apply to the quality weights of individual attribute values. Thus, the quality weights of tuples and attributes are also extended to the weights of each of the attribute values.

4. AN EXAMPLE OF PROBLEM-SOLVING USING QUALITY SPECIFICATIONS

As an example of the application of the data quality management model, a quality model was developed, and the problem of improving the quality of data in the domain of stellar astronomy was solved based on a variety of large multi-color photometric sky surveys. To study different types of stars, their spectra reconstructed from photometric data in different emission bands, and their observed parameters and estimated values of astrophysical parameters, the problem was set to cross-match and merge data on the same objects from different photometric catalogs. An example of the analysis is presented on the data of the catalogs SDSS [32], UKIDSS [33], GALEX [34, 35].

Catalog analysis shows a variety of factors that affect the quality of data and affect the quality of object matching:

- Catalogs have different sky coverage due to the location of the observatories that conducted the observations.
- Each catalog contains data obtained through various sets of photometric filters available in telescopes and providing observations in different bands of the emission spectrum of objects.
- Telescopes have different optical resolutions and, accordingly, have different positioning (calibration) accuracy and object discrimination. It is expected that the accuracy of distinguishing objects in observations in the Galactic plane may be lower since the number of objects in the field of view is greater.
- Telescopes have different limits of the minimum and maximum magnitudes of objects: brightness values below the minimum are recorded within the margin of error, and values above the maximum are saturated and do not reflect the actual values.
- Catalogs include various sets of observation quality flags, as well as flags for classifying objects, detecting artifacts, and other information about observations.
- Observations carried out in different epochs are reflected in the position of objects due to their movement and in the intensity of the brightness as a result of its variability.
- Data distortions can also be related to various observing conditions: the time of year and day, the object's position relative to the zenith, and weather conditions.
- The location of an object at the edges of the field of view affects the noise level compared to observations in the center of the field and may also cause glare associated with the equipment.

The problem of cross-matching catalogs and sky surveys remains relevant in astronomy and is usually solved for a pair of catalogs with different matching quality requirements. In accordance with the requirements of specific tasks, data is preprocessed, or observations with low quality are deleted. However, the procedure may vary depending on the specifics of each case.

Based on the analysis of the problem, the following tasks were set:

1. Catalog coverage assessment: It is necessary to determine whether the catalogs sufficiently cover sites in certain sky viewing directions. It is important that one field of view contains data from several surveys with different sets of filters, which will cover most of the pass bands in the emission spectrum of objects.

2. Cross-matching method: It is necessary to develop a method for solving the problem of cross-matching surveys of different qualities (resolutions), including matching multiple observations of objects both within the same catalog and between different catalogs.

3. Estimation of the matching radii: It is necessary to solve the problem of estimating the radii of matching objects between catalogs depending on the direction of observation relative to the plane of the Galaxy.

4. Data quality assessment: It is required to evaluate the data quality on observational objects based on the values of catalog metadata and flags present in catalog structures.

5. Matching list generation: It is necessary to develop an approach to the formation of object matching lists and merging data to obtain tuples describing the magnitudes of objects in different emission bands.

6. Presentation of the results of the merge: It is required to present the results of the merge in the form of a catalog and associated metadata.

4.1. Data Completeness Issues

Task 1, related to the problem of completeness of data in surveys, appears in two aspects.

First, this refers to the data coverage of most directions in the sky by the catalogs used. This coverage significantly affects the solution of problems such as estimating the positional accuracy of catalogs and the ability to estimate absorption in the interstellar medium in various directions in the sky, which also depends on the angle of observation relative to the Galactic plane. Data completeness values in the specified directions are evaluated statistically based on catalog data. In this context, it is not the percentage coverage of the sky that is important but the coverage of large areas in different directions. For a specific field of view, a selection of catalogs can be chosen to provide good coverage.

Secondly, the completeness of the data from the used photometric catalogs is estimated from the point of view of covering most bands of the observed spectrum of stellar emission. This characteristic is crucial for estimating the spectrum based on photometry data and depends on the set of filters used in telescopes. It is defined by a set of attributes in the catalog structure.

It is evident that the data from any single large photometric survey do not meet these two requirements. That is why it becomes necessary to integrate and match multiple catalogs so that the data obtained as a result of the catalog merging are suitable for stellar spectrum estimation, further classification, and parametrization.

4.2. Data Accuracy Issues

In task 2, the problem of errors in matching multiple observations of objects in the low-resolution catalog (GALEX) and then averaging the erroneous matching results to a single tuple is solved by choosing the sequence of catalog matching. First, all tuples of this catalog are matched with higher-quality catalogs. Then, if it is necessary to obtain an average tuple for multiple catalog

observations, the results of cross-matching with higher-quality catalogs are averaged. The cross-matching of objects from different catalogs by coordinates makes a joint accuracy calculated as a mean squared error that takes into account both uncertainties. Thus, a lower estimate of the joint accuracy is achieved and, accordingly, the probability of incorrect matching is reduced.

When solving task 3, the radius of object matching is usually taken as a constant: 1 angular second for SDSS and UKIDSS, 3 angular seconds for GALEX. These values are related to the positional accuracy and optical resolution of catalogs (the accuracy of distinguishing objects). Different values of coordinate accuracy can be found in the catalog-related publications, but the description of coordinate quality remains ambiguous. Moreover, it is desirable to estimate the radius of matching objects depending on the direction in the sky. Therefore, the usual approach turns out to be insufficient, and it becomes necessary to determine the optimal matching radius.

The problem of choosing the matching radius is solved by applying a statistical estimate that maximizes the number of unique coordinate matches for a given radius in a specific direction of observation relative to the Galactic plane [36]. Thus, the positional accuracy of catalogs is detected on data or representative samples, and metadata of accuracy specifications are generated for them. These metadata can be presented tabularly depending on the field in a certain direction in the sky and later used for each value of the coordinates value of the field objects as its accuracy.

4.3. Data Reliability Issues

Task 4 is related to evaluating the accuracy of data in different catalogs, which depends on a number of factors:

- Positional accuracy of objects: depends on the resolution of the equipment used to observe the sky.
- Location of objects in the field of view: objects located in the center of the field have higher accuracy than those located at the edge.
- Background overexposure: bright stars located close to the observed objects can cause interference.
- Artifacts: such as planets or satellites passing in the object field of view, as well as lens flares.
- Saturation of bright objects: bright stars at the upper limit of the magnitudes determined by the equipment can distort data on their brightness.
- Undetection of objects: objects at the lower limit of magnitudes may remain unnoticed.
- Accuracy of magnitude values: for each magnitude value, its accuracy in a certain emission band is indicated.

Some catalog attributes are flags that indicate the type of object, additional characteristics of its observations, and their quality. Data preparation takes into account the limitations imposed on catalog attributes that are not related to data quality, in accordance with the semantics of the task.

- The attributes Fr and Nr in GALEX indicate the angular size of the object at half-light. The restriction cuts off tuples with values of these attributes exceeding 0.003 (for stars).
- The class attribute in SDSS provides information about the type of object: star, galaxy, artifact, and others. The restriction cuts off tuples with values other than 6 (star).

Next, it is necessary to evaluate the quality of the tuples as a whole, both the observations and the values of the characteristics of the objects in them. The GALEX catalog uses the following flags for additional information about observations and data quality:

- Attributes Fexf and Nexf are vectors of bit values indicating various types of low-quality observations: objects with neighbors, mixed with other objects, truncated by the observation

boundary, and others. When these flags are not equal to 0, the magnitude value is assigned a quality weight of del.

- Attributes Fafl and Nafl are vectors of bit values indicating various types of artifacts: flares, spots, and others. If the values of these flags are not equal to 0, the magnitude value is also assigned a quality weight of del.
- Attributes nS/G and fS/G are real values that evaluate whether an object is a star or a galaxy. We can say that this is a classification of an object indicating the degree of its reliability. The tuple quality weight is set to 1 for parameter values exceeding 0.5; otherwise, the weight is 0.

The SDSS catalog uses one essential tuple quality flag:

- The attribute Q introduces categorical gradations and evaluates the quality of observation: 1 — low, 2 — acceptable, 3 — high. If the attribute value is 1, the quality weight of the tuple is set to 0; if the value is 2, then the weight is 0.5; if the value is 3, then the weight is 1.

The following quality flags are used in the UKIDSS catalog:

- The attribute cl is a mixed categorical and gradation metric that evaluates whether an object is a star, galaxy, or noise. Attribute values may indicate assignment to a particular class, indicating the degree of reliability. The quality weight value of the del tuple is assigned for all values except -2 and -1 (stars).
- The attribute p* is a real value that partially duplicates and clarifies the value of the cl attribute, estimating the probability that the object is a star. If the attribute value exceeds 0.7, the tuple quality weight value is 1; if the value is below 0.3, the value is 0; otherwise, the value is 0.5.
- Attributes pG and pN estimate a similar probability for galaxies and noise and are not considered in this task. Tuples classified as galaxy or noise using the cl attribute are deleted.

The exact way in which the weights of qualities, tuples, or values are formed based on the values of the flags was set by setting the problem in collaboration with experts in stellar astronomy. If there are no quality attribute values, the quality weight is assigned the value nan. Thus, based on the flags, the quality weights of the tuples and some attribute values in the tuples were determined. The quality weights are generalized to obtain general estimates of the quality of tuples or values in accordance with the principles described in the previous section. The generalized data quality weights are taken into account later when merging data.

4.4. Merging Data

When performing task 5, related to the formation of object matching lists, the principles of working with identified objects from several catalogs, proposed in [37], are applied. Match bundles are used that exclude transitive relationships of matched objects in multiple catalogs. Data is merged based on tuples included in match bundles and their generalized quality weights.

Depending on the tasks being solved on the identified objects, data merging can be performed according to different principles. For example, to solve the problem of the parametrization of objects, only high-quality data are important, and any low-quality observations are removed from consideration. However, if it is necessary to save low-quality objects to preserve the completeness of the data and improve its quality, such data can be included in the analysis.

Coordinates from the most accurate catalog are used for the identification of objects after merging and positioning them (in the example, this is SDSS). For identified multiple observations from the same catalog, collected in a single tuple, the coordinates can be averaged without taking into account the weights.

Flags mainly relate to the quality of observations of the object brightness in the sky. The fusion brightness data can be averaged from observational data in the same emission band.

Most catalogs contain attributes indicating the error (σ) of the magnitude measurement (m). To estimate the error of the merge result when selecting only reliable tuples (with a high quality weight: $r = 1$) a weighted mean can be used. In this case, the weight is inversely proportional to the square of the error, which minimizes the total variance. If more than reliable tuples are used in solving the problem, it is necessary to take into account the estimation of the quality weight (r) of the tuple or the value together with the error of the magnitude value. In this case, when calculating the weighted mean (2), the weight inverse to the square of the error is multiplied by the weight of the quality value (1).

$$w = \frac{r}{\sigma^2}, \tag{1}$$

$$m = \frac{\sum_i m_i w_i}{\sum_i w_i}. \tag{2}$$

The error of the result in this case can be estimated as follows (3):

$$\sigma = \sqrt{\frac{1}{\sum_i w_i}}. \tag{3}$$

4.5. Result Generation

The solution to problem 6 is to form the resulting relationship. It is created by combining matched tuples from different catalogs with a generalized evaluation of attributes for which multiple values have been found. The result should combine the following elements in a common tuple:

- coordinates from the most accurate catalog, which are also used to identify the object;
- weighted mean magnitude estimates for each emission band corresponding to the catalog bands;
- error estimates for each of the magnitude values.

This relationship can be represented as an independent catalog, which should be abundantly supplied with metadata. Such metadata should include a description of the provenance of the data and its quality, which will make it possible to assess the applicability of the data in problem-solving and ensure that it can be reused.

The description of the data provenance of the new catalog may include the following elements:

- at the catalog level: links to catalogs whose data were used to generate the result;
- at the attribute level: links to attributes (coordinates, magnitudes in the corresponding bands) from which attribute values were formed;
- a link to the description of the tuple matching method;
- a link to the description of the selection method (for coordinates) and the averaging of attribute values (for magnitudes).

The description of the catalog data quality may include:

- at the relation level: information about the completeness of the sky coverage and the completeness of the attributes (covered observation spectral bands), depending on the direction in the sky;
- information about evaluating the positional accuracy of the resulting data;
- information about the quality of fused tuples;
- information about the presence of weighted estimates of magnitude errors in the catalog attributes and their relationship to the attributes of the magnitude values themselves.

We thank O.Y. Malkov (INASAN) for setting the problem used to demonstrate the presented approach.

5. CONCLUSION

The approaches to data quality specification in research infrastructures with multiple heterogeneous data sources of different qualities are investigated. Standards for describing data quality are based on multi-criteria models, which usually do not limit the composition and methods of evaluating quality metrics. The development of the principles of data quality specification is proposed to determine how to store and access quality metadata, the level of aggregation of the evaluated data and take into account non-functional data requirements when problem-solving and data processing tasks. These principles are illustrated when solving a problem in stellar astronomy. At different stages of its solution, various quality dimensions are evaluated, including completeness, accuracy, and reliability of data, different sources of quality metadata are used, such as information on the accuracy of physical quantities in catalogs, tuple quality flags, guaranteed quality from catalog documentation, and statistical estimates based on a sample of catalog data. Depending on the obtained estimates of quality dimensions, different data sources are used, data that is not suitable in quality is filtered out, the sequence of data processing in problem-solving algorithms is changed, and metadata of both data sources and intermediate and final results of solving the problem is generated and stored.

REFERENCES

1. Wand, Y. and Wang, R., Anchoring Data Quality Dimensions in Ontological Foundations, *Commun. ACM*, New York: ACM, 1996, vol. 39, no. 11, pp. 86–95.
2. Ballou, D. and Pazer, H., Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems, *Manage. Sci.*, New York: TIMS, 1985, vol. 31, no. 2, pp. 150–162. https://doi.org/10.1287/mnsc.31.2.150
3. Wang, R. and Strong, D., Beyond Accuracy: What Data Quality Means to Data Consumers, *J. Manage. Inf. Syst.*, Abingdon: Taylor & Francis, 1996, vol. 12, no. 4, pp. 5–33. URL: http://www.jstor.org/stable/40398176
4. Batini, C. and Scannapieco, M., *Data Quality: Concepts, Methodologies and Techniques*, Heidelberg: Springer, 2006, 262 p. https://doi.org/10.1007/3-540-33173-5
5. ISO 8000-1:2022 Data quality – Part 1: Overview. Geneva: ISO, 2022. URL: https://www.iso.org/standard/81745.html
6. ISO 19157:2013 Geographic information – Data quality. Geneva: ISO, 2013. URL: https://www.iso.org/standard/32575.html
7. Wilkinson, M., Dumontier, M., Aalbersberg, I., et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, *Sci. Data*, London: Nature Research, 2016, vol. 3, art. 160018. https://doi.org/10.1038/sdata.2016.18
8. FAIR Data Maturity Model. Specification and Guidelines. Version 1.0, RDA FAIR Data Maturity Model Working Group, Geneva: Zenodo, 2020. https://doi.org/10.15497/rda00050
9. FAIRsFAIR, Fostering FAIR Data Practices in Europe. URL: https://www.fairsfair.eu/
10. Devaraju, A., Mokrane, M., Cepinskas, L., et al., From Conceptualization to Implementation: FAIR Assessment of Research Data Objects, *Data Sci. J.*, 2021, vol. 20, no. 1, art. 4. https://doi.org/10.5334/dsj-2021-004.
11. The FAIR Cookbook for FAIR Doers. URL: https://faircookbook.elixir-europe.org/
12. Harrow, J., Drysdale, R., Smith, A., et al., ELIXIR: Providing a Sustainable Infrastructure for Life Science Data at European Scale, *Bioinformatics*, Oxford: Oxford University, 2021, vol. 37, no. 16, pp. 2506–2511. https://doi.org/10.1093/bioinformatics/btab481
13. ELIXIR Platforms. URL: https://elixir-europe.org/platforms

14. Recommendations from the Data Quality Working Group, NASA ES DSWG, 2019.
URL: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/recommendations-from-the-data-quality-working-group

15. Data Quality Working Group's Comprehensive Recommendations for Data Producers and Distributors, NASA ES DSWG, 2019.
URL: https://www.earthdata.nasa.gov/s3fs-public/imported/ESDS-RFC-033.pdf

16. ESIP Information Quality Cluster, Earth Science Information Partners (ESIP).
URL: http://wiki.esipfed.org/index.php/Information_Quality

17. Peng, G., Privette, J., Kearns, E., et al., A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets, *Data Science Journal*, London: Ubiquity, 2015, vol. 13, no. 2, pp. 231–253. https://doi.org/10.2481/dsj.14-049

18. ISO 19157-1:2023, Geographic Information — Data Quality. Part 1: General Requirements, Geneva: ISO, 2023. URL: https://www.iso.org/standard/78900.html

19. Sirotnak, C. and Cook, J., The Total Economic Impact of Talend: Cost Savings and Business Benefits Enabled by Talend Solutions, Cambridge: Forrester, 2023.
URL: https://www.talend.com/lp/the-total-economic-impact-of-talend/

20. Chien, M. and Medd, J., Magic Quadrant for Augmented Data Quality Solutions, Stamford: Gartner, 2024. URL: https://www.gartner.com/en/documents/5257863

21. Fürber, C., *Data Quality Management with Semantic Technologies*, Thesis, Wiesbaden: Springer Gabler, 2016. https://doi.org/10.1007/978-3-658-12225-6

22. Berners-Lee, T., Hendler, J., and Lassila, O., The Semantic Web, *Scientific American*, New York: Springer Nature (US), 2001, vol. 284, no. 5, pp. 34–43. URL: https://www.jstor.org/stable/26059207

23. Cyganiak, R., Wood, D., and Lanthaler, M. (eds.), RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation, Wakefield: W3C, 2014. URL: http://www.w3.org/TR/rdf11-primer/

24. Fürber, C. and Hepp, M., Towards a Vocabulary for Data Quality Management in Semantic Web Architectures, *Proc. 1st Int. Workshop on Linked Web Data Management (LWDM2011)*, New York: ACM, 2011, pp. 1–8. https://doi.org/10.1145/1966901.1966903

25. Hartig, O. and Zhao, J., Provenance Vocabulary Core Ontology Specification, San Diego: SourceForge, 2012. URL: https://trdf.sourceforge.net/provenance/ns.html

26. Taleb, I., Serhani, M., Bouhaddioui, C., et al., Big Data Quality Framework: A Holistic Approach to Continuous Quality Management, *Journal of Big Data*, Heidelberg: SpringerOpen, 2021, vol. 8, article no. 76. https://doi.org/10.1186/s40537-021-00468-0

27. Gallo, R., Data Quality with FAIR Principles, an Introduction, The Hyve, 2024.
URL: https://www.thehyve.nl/articles/data-quality-with-fair-principles

28. Skvortsov, N., The Principles of Data Reuse in Research Infrastructures, *Proc. Int. Conf. Common Digital Space of Scientific Knowledge: Problems and Solutions (CDSSK 2020)*, Aachen: CEUR WS, 2021, vol. 2990, pp. 62–74. URL: https://ceur-ws.org/Vol-2990/rpaper6.pdf

29. PROV-Overview: An Overview of the PROV Family of Documents, W3C Working Group Note, Wakefield: W3C, 2013. URL: http://www.w3.org/TR/prov-overview/

30. Data on the Web Best Practices: Data Quality Vocabulary, W3C Working Group Note, Wakefield: W3C, 2016. URL: https://www.w3.org/TR/vocab-dqv/

31. Albertoni, R. and Isaac, A. (eds.), Data Catalog Vocabulary (DCAT), Version 3, W3C Recommendation, Wakefield: W3C, 2024. URL: https://www.w3.org/TR/vocab-dcat/

32. Alam, S., Albareti, F., Prieto, C., et al., The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III, *Astrophysical Journal Supplement Series*, Bristol: IOP Publishing, 2015, vol. 219, no. 1, p. 12. https://doi.org/10.1088/0067-0049/219/1/12

33. Lawrence, A., Warren, S., Almaini, O., et al., The UKIRT Infrared Deep Sky Survey (UKIDSS), *Monthly Notices of the Royal Astronomical Society*, Oxford: Oxford University, 2007, vol. 379, no. 4, pp. 1599–1617. https://doi.org/10.1111/j.1365-2966.2007.12040.x

34. Bianchi, L., Herald, J., Efremova, B., et al., GALEX Catalogs of UV Sources: Statistical Properties and Sample Science Applications: Hot White Dwarfs in the Milky Way, *Astrophysics and Space Science*, Heidelberg: Springer, 2011, vol. 335, no. 1, pp. 161–169. https://doi.org/10.1007/s10509-010-0581-x

35. Bianchi, L., Shiao, B., and Thilker, D., Revised Catalog of GALEX Ultraviolet Sources. I. The All-Sky Survey: GUVcat_AIS, *Astrophysical Journal Supplement Series*, Bristol: IOP Publishing, 2017, vol. 230, no. 2, p. 24. https://doi.org/10.3847/1538-4365/aa7053

36. Malkov, O., Dluzhnevskaya, O., Karpov, S., et al., Cross Catalogue Matching with Virtual Observatory and Parameterization of Stars, *Open Astronomy*, Warsaw: De Gruyter Open, 2012, vol. 21, no. 3, pp. 319–330. https://doi.org/10.1515/astro-2017-0390

37. Gray, J., Szalay, A., Budavari, T., et al., Cross-Matching Multiple Spatial Observations and Dealing with Missing Data, Microsoft Technical Report, MSR-TR-2006-175, Redmond: Microsoft Research, 2006. https://doi.org/10.48550/arXiv.cs/0701172

This paper was recommended for publication by A.A. Galyaev, a member of the Editorial Board