Scientometrics
https://doi.org/10.1007/s11192-018-2735-5

Data measurement in research information systems: metrics for the evaluation of data quality

Otmane Azeroual[1,2] · Gunter Saake[2] · Jürgen Wastl[3]

Received: 7 October 2017

Abstract In recent years, research information systems (RIS) have become an integral part of the university's IT landscape. At the same time, many universities and research institutions are still working on the implementation of such information systems. Research information systems support institutions in the measurement, documentation, evaluation and communication of research activities. Implementing such integrative systems requires that institutions assure the quality of the information on research activities entered into them. Since many information and data sources are interwoven, these different data sources can have a negative impact on data quality in different research information systems. Because the topic is currently of interest to many institutions, the aim of the present paper is firstly to consider how data quality can be investigated in the context of RIS, and then to explain how various dimensions of data quality described in the literature can be measured in research information systems. Finally, a framework as a process flow according to UML activity diagram notation is developed for monitoring and improvement of the quality of these data; this framework can be implemented by technical personnel in universities and research institutions.

✉ Otmane Azeroual
Azeroual@dzhw.eu

Gunter Saake
Saake@iti.cs.uni-magdeburg.de

Jürgen Wastl
Juergen.Wastl@admin.cam.ac.uk

[1] German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, 10117 Berlin, Germany

[2] Department of Computer Science, Institute for Technical and Business Information Systems Database Research Group, Otto-von-Guericke-University Magdeburg, P.O. Box 4120, 39106 Magdeburg, Germany

[3] The Old Schools, University of Cambridge, Trinity Lane, Cambridge CB2 1TT, UK

Published online: 06 April 2018

Keywords Current research information systems (CRIS) · Research information systems (RIS) · Research information · Data quality · Data quality dimensions · Data measurement · Data monitoring · Science system · Standardization

Introduction

The topic of "research databases and research information systems" is by no means new. In recent years, the introduction of research information systems at universities and research institutions has strongly increased in Germany and throughout Europe (DINI AG Research Information Systems 2015). Research information systems can provide universities and research institutions with a current overview of their research activities, collect information on their scientific activities, projects and publications and manage and integrate into their website. For researchers, they offer opportunities to collect, categorize and make use of research information, be that for publication lists, for the development of new projects, to reduce the effort required to produce reports, or in the external presentation of their research and scientific expertise.

Data quality plays an important role in the usability and interpretation of institution-specific data. The quality of data is however also a significant consideration for external data sources. University administrations and researchers have since the early 1990s begun to recognize the importance of quality of data that are electronically stored databases. A few years ago, almost all German universities and research institutes were interested in the topic of quality of data in their RIS—a development that has been since further progressed. The growing quantity of data and the increasing number of source systems are becoming serious problems for institutions. In order to keep control and gain the greatest possible benefit from such information not only infrastructure measures, but also measures for the observance and increase of the data security and data quality are necessary (Apel et al. 2015). Almost all institutions rate high data quality as an essential consideration for their information on research activities. But only a few invest the time and resources to maintain and improve this data quality. In most cases, a poor data basis in individual departments is either reluctantly tolerated, or in the worst case not even perceived (Apel et al. 2015).

This paper presents a holistic view of procedures for guaranteeing data quality in RIS. The handling of large data sources are daily operations for these institutions. Since data errors such as missing values, duplicates, spelling mistakes, incorrect formatting and inconsistencies occur during the collection, transmission and integration of research information in different systems and can spread over different areas, it is necessary to recognize these errors early and to treat them efficiently (Azeroual and Abusoba 2017). If users were unable to access the information they needed to make decisions, the value of the data they used and their confidence in the RIS would decrease. It is already sufficient if a small error renders the data unusable throughout the institution. Here, the completeness, correctness, timeliness and consistency of the data play a decisive role. It is important to understand that there cannot be data quality—and thus no data quality management—without measurements of data quality. Beyond that, the aim of this paper is to investigate the data quality of the research information given in RIS and to measure the quality dimension, based on the literature, and to develop a data quality framework with the objective of monitoring and improving the data quality of RIS.

The paper is constructed in the following manner. In "Research Information System (RIS)" section, the term RIS is briefly explained and defined. Moreover, a brief overview of the objectives and the added value of RIS will also be given. In "Data Quality and Data

Quality Dimensions" section data quality and its dimensions will be described based on the literature. In "Measurement of Data Quality of the RIS" section the measuring points are introduced with the help of a RIS architecture and it is indicated how data quality can be measured in RIS. In this case, the focus lies upon the measurement of the selected data quality dimensions using the example of data on publications held in RIS. The paper ends with a summary of the most important results and final remarks. The framework as a process flow according to UML activity diagram notation is developed for the purposes of monitoring and improvement of quality of these data is made available for all institutions which may wish to use it.

Research information system (RIS)

A RIS is a **central database** that can be used to collect, manage and provide information on research activities and research results. The information presented here are metadata about research activities, e.g. projects, third-party funds, patents, cooperation partners, prizes and publications and are referred to as research information (Azeroual et al. 2018a, b).

For the purpose of this article, the authors distinguish between the following groups of interests and their information need with regard to RIS. The following Fig. 1 illustrates these groups of interests.

Starting from the variety of systems which fall under the term "research database", three more specific types of research information systems may be differentiated (DINI AG Research Information Systems 2015):

1. Simple detection systems (such as university bibliographies or research portals)
2. Research profile services (such as Linked Open Data applications)

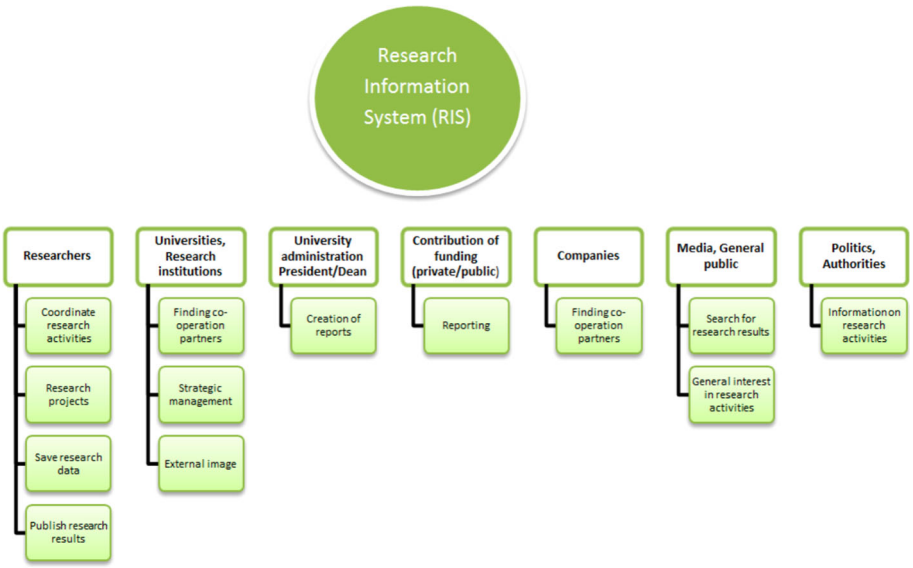

Fig. 1 Interest groups of research information

3. Integrated research information systems with multiple output and analysis functionalities

Simple detection systems capture several entities (like for example, projects, organization, people, publications and promotion etc.), are optimized only on single use scenarios and are thereby not linked up with each other. Traditional research portals, university bibliographies, patent databases and repositories for publications and research primary data are examples of this kind of system (DINI AG Research Information Systems 2015).

The term "**research profile services**" describes information systems, e.g. linked open data concepts for the exchange and merging of data available from public and institutional data sources (DINI AG Research Information Systems 2015).

Integrated research information systems are combined database and reporting systems that enable an institution to document, evaluate and further develop its research activities (DINI AG Research Information Systems 2015). The focus is on building a quality-assured corpus of metadata that brings together and maintains many of the external and internal information. In addition, the value-added services for the external communication of the institution can be excluded through research portals and web services for institutional websites (DINI AG Research Information Systems 2015). Integrated research information systems have the following characteristic (DINI AG Research Information Systems 2015):

(a) In a data model, the entities as well as their relationships to one another are described.
(b) Information from different internal and external data sources is brought together and semantically enriched.
(c) The uses of IT-solutions helps with the concept distributing of data maintenance and quality assurance across content, hierarchy and organizational boundaries.
(d) Research information can be provided for different use scenarios, report procedures and decision-making processes. The systems allow multiple issue- and analysis functions, and permit multiple uses of the collected data.

A research information system links all research information from a university or research institution and presents these in a compact form. The main objective of a RIS is to present the research activities of the overall university and its researchers to show the central vision with clarity, which makes work easier for the researchers. Data is entered once into the research information system and can then be used several times, e.g. on websites, for project applications or reporting processes. Double data retention and with it additional work for users is avoided. As a further objective, research information systems are established as a central instrument for the uniform and continuous communication and documentation of the various research activities and results. Beyond legal obligations on the documentation of research activities, the external representation of the universities and research institution will be improved by the illustration of current data (Azeroual et al. 2018a, b).

The better discoverability of information helps all: researchers in search of cooperation partners, companies in the awarding of research contracts, the public through transparency and general information about their university and research institution. In Fig. 2 the added value of research information systems for different user groups are set out and summarized.

Since there are different understandings in the literature on research information systems, the technical architectures for research information systems are presented below. For

Fig. 2 Added value of research information systems

this paper, the following elements of research information system architecture are distinguished:

1. **Data Access Layer**
2. **Application System Layer**
3. **Presentation Layer**

The **data access layer** contains the internal and external data sources. This level includes, for example, databases from the administration or publication repositories of libraries, identifiers such as e.g. ORCID or bibliographic data from the Web of Science or Scopus. Models for the standardized provision of research data in RIS are the Research Core Dataset (RCD) and the Common European Research Information Format (CERIF) data model.

A moving of source data into the RIS takes place via classical Extract, Transform, and Load (ETL) process.

The **application system layer** contains the research information system and its applications that merge, manage, and analyze the data held in the underlying level.

In the **presentation layer**, representations of the analysis results are prepared and presented for specific user groups, in the form of reports using business intelligence tools. In addition to various reporting possibilities, here also portals and websites of institutions can be supplied with data.

The following figure (see Fig. 3) gives an overview of the individual processes and shows which components belong to which process step.

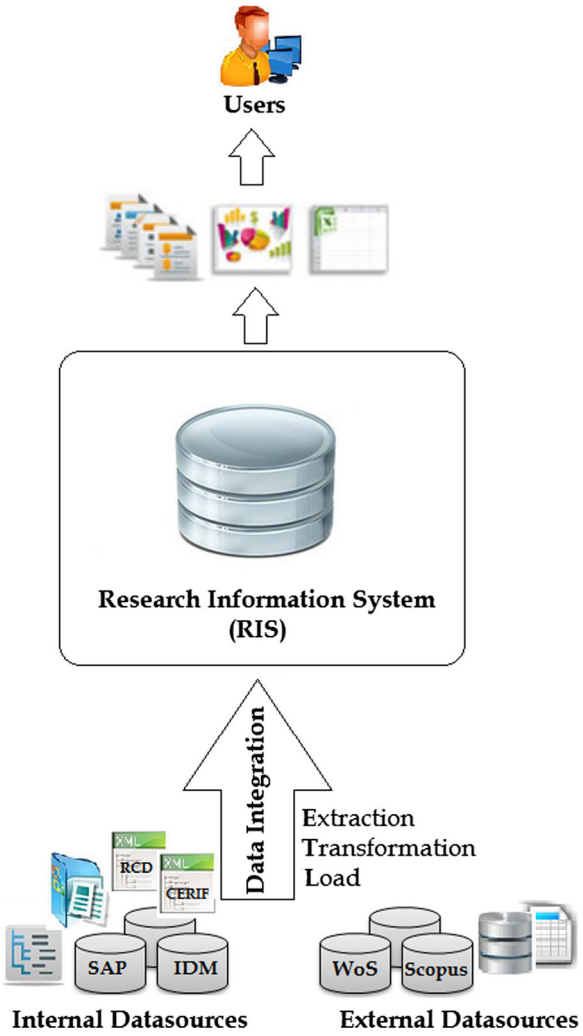

Fig. 3 Architecture of research information system (RIS)

Data quality and data quality dimensions

The subject of data quality is very strongly subjective and can be best described in the words of Larry English in this context. He says, to assess the best method of data quality, is to look at what data quality generally means and to define for itself what quality means for the data (English 1999). The questions that to answer it therefore for the definition of data qualities are:

- How do we define requirements for data quality in a particular environment/context?
- Who makes these data quality requirements?
- How do we make these requirements measurable and how do we control them?

In the literature, the terms ‘data quality’ and ‘information quality’ are often used synonymously (Gebauer and Windheuser 2015). In general it can be said that the quality of information or data is the degree to which features of this data product meet the requirements (Hinrichs 2002). Data quality thus means the “degree of usability of information for the respective application” (Krcmar 2015).

The following definition is used in this paper for data quality: The quality of data is a “multi-dimensional measure of the suitability of data to fulfill the purpose bound in its acquisition/generation. This suitability may change over time as needs change” (Würthele 2003). This underlines the subjective requirements for data quality in respective institutions and illustrates a possible dynamic data quality process. The definition makes it clear that “the quality of data depends on the time of the consideration and on the level of claims placed at the time on the data” (Apel et al. 2015).

If we are talking about data quality, then we talk about a certain reliability of the data in the institution context at a certain point in time. This reliability should be measured and evaluated on the basis of defined standards in order to determine the quality of the data. These defined standards can change over time, so they should not be anchored. To ensure this, an objective consensus must first be created to assess the reliability of data.

With the help of the data quality pyramid shown in Fig. 4 three steps of successful operationalization can be identified (Gebauer and Windheuser 2015).

Data quality can be understood as a superset of all data quality dimensions that form the second level of the pyramid. Data quality metrics are necessary to evaluate data quality dimensions. They form the quality measure with which a quantitative statement is possible. These quality metrics form the operational basis for determining data quality (Gebauer and Windheuser 2015). A metric is understood to mean methods and systems which, as a result, provide quantifiable values and key figures.

In order to make the quality of data measurable, certain characteristics (quality dimensions) are required, which must be assigned to the data (Apel et al. 2015). Mostly, the reliability of data resulting from a process of collection or delivery can only be ascertained after an adjustment and adaptation in the data storage system. This means that processes, which are part of the preprocessing, should keep to the requirements placed on the reliability of the delivered data in order to enable effective data cleansing. Besides, the institution context must always be considered.

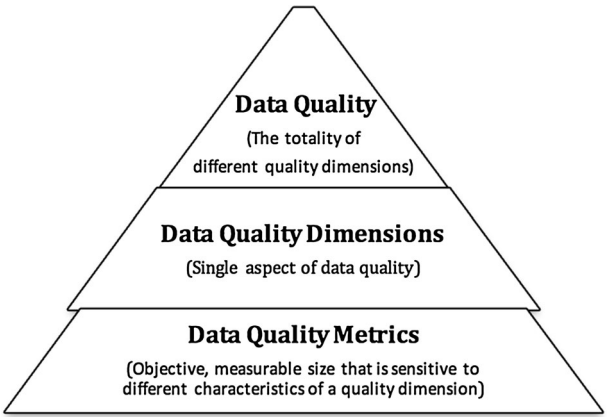

Fig. 4 Data quality pyramid (Reproduced with permission from Gebauer and Windheuser 2015)

According to Wang and Strong (1996), the data quality requirements can be divided into the following four categories (Table 1).

In each of these categories, defining dimensions can contribute to high data quality. In the literature, several dimensions are identified and classified. In the following a list of the most discussed dimensions is performed based on Krcmar (2015) and Wang and Strong (1996) for the quality of data (Table 2).

In this paper, only the four data quality dimensions will be considered in the context of RIS. Incidentally these four dimensions were selected because on the one hand they are discussed extensively in scientific publications and, on the other hand, they play an important role in practice (Hildebrand et al. 2015).

Measurement of data quality of the RIS

It is not possible to carry out meaningful measurement of data quality without defining the relevant dimensions clearly beforehand (Martin 2005). Besides, institutions must decide which data quality dimensions are important and how these should be measured, because many dimensions are multivariate. A general quantitative definition of the metric of a data quality dimension appears as follows (Lee et al. 2006):

$$\text{Rating score} = 1 - \frac{\text{Number of unwanted results}}{\text{Number of all results}}$$

In section four, all data quality dimensions were considered; this section now presents metrics for the most important dimensions (completeness, timeliness, correctness and consistency) that are encountered and explains how these dimensions can be measured within an RIS. As a final point, a framework will be developed for oversight and improvement of data quality.

The measuring points of the data quality should stem from the RIS architecture. With this in mind areas of measurement are indicated which are suitable for a measurement and oversight of the data quality (Apel et al. 2015).

As can be seen in Fig. 5, possibilities for measurement may be found in the following areas of an RIS:

- Internal and External Datasources (or Source Systems)
- Central Research Information System (RIS)
- RIS-Frontend

The RIS collects information about the research activities and research results of the organization and the scientists affiliated to it by automated synchronization of the data set with different external data sources. For an automated import of data from existing systems, a linking of internal and external application systems can be implemented (Herwig

Table 1 Requirements categories of data quality (Wang and Strong 1996)

Accuracy	The extent to which data values are in conformance with the actual or true values
Relevancy	The extent to which data are applicable (pertinent) to the task of the data user
Representation	The extent to which data are presented in an intelligible and clear manner
Accessibility	The extent to which data are available or obtainable

Table 2 Dimensions of the data quality (Krcmar 2015; Wang and Strong 1996)

Completeness	The extent to which data are of sufficient breadth, depth, and scope for the task at hand
Correctness/free of error	The extent to which data is correct and reliable
Timeliness	The extent to which the age of the data is appropriate for the task at hand
Consistency	The extent to which data are always presented in the same format and are compatible with previous data

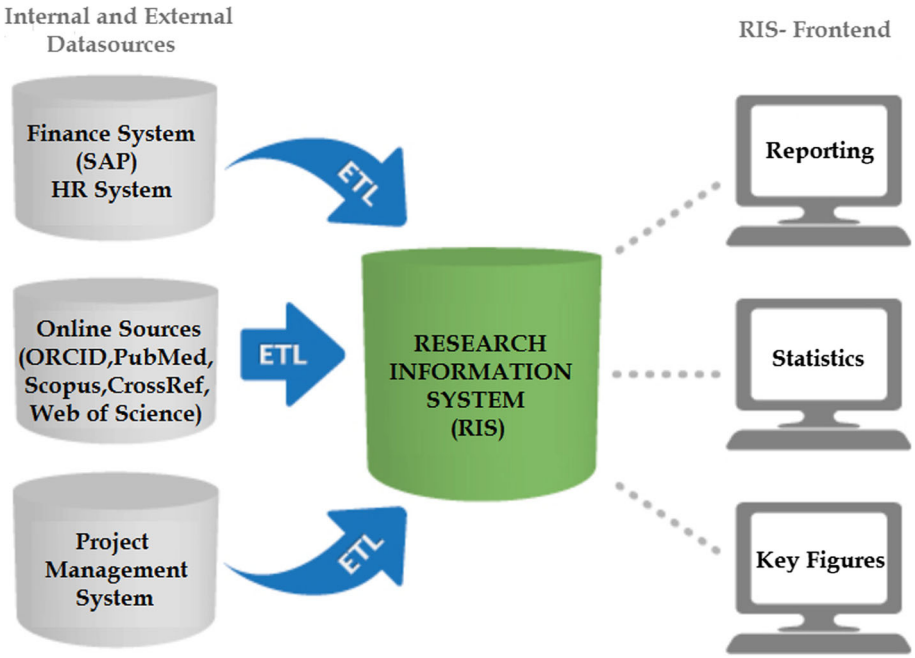

Fig. 5 Measurement capabilities within the RIS architecture

and Schlattmann 2016). The application systems to be included in the capture by research information ideally include the campus management system, as well as the identity management system for internal systems and the public publication and project databases (Herwig and Schlattmann 2016). Web of Science, Scopus or PubMed, are examples for the case of publication data. The financial systems of the administration of grant funding might provide data on externally funded projects, and the personnel administration system information about scientific staff.

In the area of source systems, a measurement is very beneficial, as an improvement of the data quality in this area would ensure that data for all systems and areas are correct for all desired future uses.

The technical implementation aside, checks on data collection make very good sense in the area of source systems. Moreover, masks or style sheets could be created, such as date inputs (YYYYMMDD) or prepared selections by use of drop-down lists (Apel et al. 2015).

The data of the whole research institutions are given to a central RIS. For this reason measurements, which concern several areas of a research institution, can be compared with each other. Therefore, the opportunity would be provided here to check different source systems for their consistency (Apel et al. 2015). In the last step of the RIS process the data presentation (RIS-Frontend) is shown. By means of portals, reports and other front final uses, the information coming from the systems will be visualized. By different use of components, the processed information and analyses are made available to the user. In the area of FIS-Frontend, measurement or test procedures may be carried out by persons. In the simplest case, investigations of operating figures are conducted in areas. In addition, operating numbers of the RIS can be compared with the source systems. In this case, a specialized technical person is necessary (Apel et al. 2015).

Measurement of completeness

Here the special importance of null values and empty places has to be considered. The dimension of completeness can be subdivided into five different types (Scannapieco et al. 2005):

1. **Value completeness** All attribute values of a tuple defined as relevant are available.
2. **Tuple completeness** All attribute values of a tuple are available.
3. **Attribute completeness** All values of an attribute (a column) are available.
4. **Relation completeness** All values in the entire relation are available.
5. **Tuple relation completeness** All tuples are present in the relation.

In this paper, value and tuple relation completeness are relevant. The assessment of value completeness can be carried out in a research information system. To examine the completeness, attribute values which are marked as important but which are nonetheless not available must be identified. This can be done with a control of the attribute values, whilst checking the values for null values.

To determine whether all tuples exist in a relation, the source systems need to be consulted. The source system is used as a reference system to carry out the comparison with the content of the RIS. Discrepancies indicate that objects in the source system do not correspond with those in the RIS, or are absent from it.

The opposite case, that is, data exist in the RIS which do not exist in the source systems, plays an important role in the checking of data quality. Although completeness would be ensured at the measuring point of the RIS, this would not necessarily indicate a high data quality. The appearance of such events is due to the fact that a manual fire-fighting process has taken place in the source system, but was not understood by the RIS. The cause for such a mistake can be occurring by missing communication through rules forbidding a deletion of data in research information systems.

Completeness can be measurement as follows (Lee et al. 2006):

$$Q_{\text{completeness}} = 1 - \frac{\text{Number of incomplete units}}{\text{Number of checked units}} \tag{1}$$

If NULL is used in RIS for the calculation of the completeness, it must be considered that NULL can have three different meanings (Cordts 2013):

1. A value does not exist
2. A value exists, but is unknown
3. It is not known whether a value exists

In the first meaning NULL does not count as incompleteness, in the second it does, and in the third meaning, both would be possible (Batini and Scannapieco 2006).

Depending on the granularity of completeness one can distinguish between record-completeness, column-completeness and table-completeness (Cordts 2013). Record-completeness refers to the number of NULL values of all the columns of a record, column-completeness to that of a single column in relation to all the values of that column (Cordts 2013). Table-completeness determines the number of all NULL values in a table sets them in comparison to all values in the table (Cordts 2013).

According to the following (see Table 3) an example of the calculation of completeness is made clear in a list of publications:

Measurement of timeliness

Timeliness is assessed according to how current a data value is. As an active examination would be too cost-intensive or would not be possible at all in some cases, the timeliness metric is based on an estimate (Heinrich et al. 2007). With the estimation occurs a probability, which estimates how current the examined data values are. For the timeliness metric some parameters are defined again (Heinrich and Klier 2009):

- "A" is a data attribute (e.g. "professional status")
- "w" is a suitable data value (e.g. "a student")
- "Age(w,A)" is the age of the data value. This can be calculated from the time of measurement and the data acquisition time
- "Decline(A)" is an empirically ascertained value which describes the decay rate of the data value and of the data attribute

The metric of timeliness as a data quality dimension can be measured as follows (Heinrich et al. 2007):

$$Q_{\text{timeliness}}(w,A) := e^{(-\text{decline}(A)\cdot\text{age}(w,A))} \tag{2}$$

$Q_{\text{timeliness}}(w,A)$ expresses a probability of the actuality of the data attribute (A) of the data value (w). In the above-mentioned formula, the assumption is made that the decay (Decline(A)), period of validity, is distributed exponentially. This assumption is made because it can be empirically shown that this fits best to a life span distribution of this data attribute (Heinrich et al. 2007). With unchanging data, like birthplace or date of birth, it is Decline(A) = 0. This applied on formula 2: Metrics for the DQ dimension timeliness set to one ("1") (Heinrich and Klier 2009).

$$Q_{\text{timeliness}}(w,A) := e^{(-\text{decline}(A)\cdot\text{age}(w,A))} = e^{(-0\cdot\text{age}(w,A))} = e^{(0)} = 1 \tag{3}$$

The same is valid for data values where $\text{age}(w,A) = 0$. This can occur when the time of observation is equal to the time of acquisition. It follows that (Heinrich and Klier 2009):

$$Q_{\text{timeliness}}(w,A) := e^{(-\text{decline}(A)\cdot\text{age}(w,A))} = e^{(-\text{decline}(A)\cdot 0)} = e^{(0)} = 1 \tag{4}$$

The following example illustrates the calculation of the data quality dimension timeliness by reference to the master data of a person:

In the first step, the decisive attributes of the decay rates are picked out. These values can be taken either from available statistics or, if not available, on the basis of subjective appraisals estimated by experts (Heinrich et al. 2007). The Federal Statistical Office

Table 3 Example of the metric completeness calculation in a publication list

AuthorID	Firstname	Lastname	DOI	ARTICLE_TITLE	PUBTYPE	PUBYEAR	Tuple completeness
352908	Sven	Svenson	https://doi.org/10.1061/9780784412350.0029	Compressibility characteristics of a soft soil stabilized by deep mixing-Volumetric creep deformations	P	2007	1
353035	Jacob	NULL	NULL	Arsenic in the contemporary antitumour chemotherapy	NULL	2015	0.25
353463	Jamila	Morjane	NULL	The quality-a factor for competitiveness of organizations	P	NULL	0.5
353585	Olivia	NULL	https://doi.org/10.1016/j.enganabound.2011.08.010	CMRH method as iterative solver for boundary element acoustic systems	J	2017	0.75
353759	Jean	Jackson	https://doi.org/10.1061/9780784412138.fm	Ground reaction force analysis in normal gait using footwear with various heel heights on different surfaces	P	2014	1
Columns completeness	1	0.5	0.5	1	0.75	0.75	1 − (6/35) = 0.83

provides information about the duration of the validity of names (meaning last name) and addresses.

- **Last Name** 2% of the persons change the last name
- **First Name** 0% of the persons change the first name
- **Address** 10% of the persons change their address
- **E-Mail** 40% of the persons change their email address

In the next step, weightings (gi) for each attribute will be applied; these are statements about the importance of the attributes for the owner of the information. The value attributed to each area ranges from null ("insignificant") to one ("important"). Here the attributes with 1.0 are weighted.

The timeliness of the single attributes is to be seen in following (Cf. Table 4).

To be able to deliver complete information about timeliness, the values of the attribute timeliness and the weightings must now be used. This step must be carried out in aggregate (Gebauer and Windheuser 2015):

$$Q_{\text{timeliness}}\left(T, A_{1,\ldots,}A_4\right) := \frac{1.00 * 1.00 + 0.99 * 1.00 + 0.82 * 1.00 + 0.82 * 1.00}{1.00 + 1.00 + 1.00 + 1.00} \approx 0.9075 \tag{5}$$

The result says that the data relative to the master data of a person are topical to a degree of **90.75%**.

The requirements for timeliness are considered not to be fulfilled if the filling or loading process has not been completed. The data would thus not be faithful to the present and current. The same is valid by non-completion of the loading process, be that because it is abandoned or the time frame for it exceeded. The measurability of the timeliness cannot be carried out directly about the contained data. Metadata are necessary for it. Metadata contains information about the status, duration, time, and both the incidence appearance of errors and reasons for them. With the help of these points the requirement for the timeliness can be measured in the indicator system.

Measurement of correctness

Correctness, or freedom from error, determines to what extent the data is correct, i.e. to which degree two values v and v' agree. Besides, v' corresponds to the value in the real world. The correctness can be measured as follows (Lee et al. 2006):

Table 4 Example of the metric timeliness calculation (Heinrich et al. 2007)

Attributes (Ai)	Firstname	Lastname	Address	E-Mail address
gi	1.00	1.00	1.00	1.00
age(T,Ai,Ai) [year]	0.50	0.50	2.00	0.50
decline(Ai) [1/year]	0.00	0.02	0.10	0.40
Qtimeliness(T.Ai,Ai) = exp(− decline(Ai)*age(T.Ai,Ai))	1.00	0.99	0.82	0.82

$$Q_{\text{correctness}} = 1 - \frac{\text{Number of incorrect data units}}{\text{Total number of data units}} \quad (6)$$

This metric must be used to determine how the granularity of a data unit is specified (database, table, column), resulting in an error (Scannapieco et al. 2005). Correctness can refer to either a column, a table, or even a whole database. Thus, e.g. Table or database correctness can be determined, while the relation is calculated between the correct values and all values. In the case of a complete table, all the correct cell values are determined and counted relative to the number of all cells in the table (Batini and Scannapieco 2006).

In the literature the concepts of syntactic and semantic correctness also appear in connection with data quality (Scannapieco et al. 2005). Helfert (2002) requires as an indicator of **syntactic correctness** that "the data match the specified syntax (format)". This means, for example, the wrong manner of writing of an attribute value is identified. A breach of syntactic correctness occurs when, for example, instead of the correct attribute value ("Article") the attribute value ("Aricle") exists. **Semantic correctness** is not satisfied when, though an attribute value is syntactically correct, a wrong value is assigned. This is the case when for an article, for example, the wrong author is stored. These definitions are adopted in this article because with the RIS the default presumption is that that a record is correct if it agrees (taking into account the transformations in data loading) with the data of the source systems.

If all attribute values have been loaded exactly according to the transformation rules defined in the business rules, a data record can be evaluated as correct. In the case of one-to-one transformations, values in the source system as well as in the research database must be identical. Correctness is given, if with the transformations, with which values can be calculated and then can be only filed afterwards in the RIS, are right.

Based on this definition an assessment of correctness can take place in the RIS. Similar to the verification of timeliness the correctness of data is controlled. Indeed, in this case, the correctness of the content is checked. More precisely a tuple will be considered. This is checked in the source system and in the RIS for correspondence and thereby also for correctness.

If calculated data records in an RIS are to be examined, the steps must be traced back in a test of some sort and be reproduced in order to be able to suggest the data of the source systems. Such an investigation is very taxing and laborious on account of its complexity. This can lead to the non-fulfillment of the criterion. The same applies to calculations which store their original data in different source systems and are not available simultaneously.

The control of historical data in the RIS is problematic, too. If no historical data exist in the source system, a comparison with the historical data of the RIS is not practicable. This case requires that correctness of the data be checked whilst they are still actual. If data were checked at the time of loading, it may be assumed that they are also correct in the archived state. A special case of correctness checking occurs when one or more data records are changed during the active loading process from the source system to the research information system. A control would identify, on this occasion in the RIS, this incorrectly as an error.

The following Fig. 6 illustrates an example of incorrect data in a publication list and calculates the degree of correctness of this information (Formula 7).

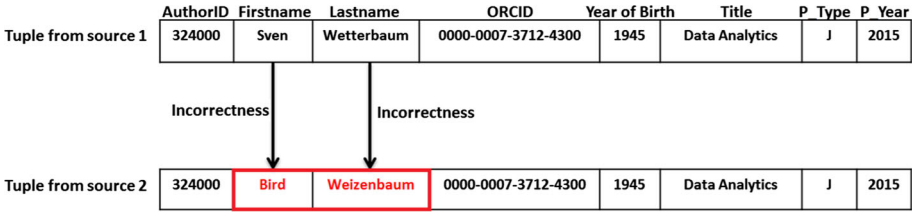

Fig. 6 Example of incorrect data in a publication list

$$Q_{\text{correctness}}(\text{"Wetterbaum"}, \text{"Weizenbaum"}) = 1 - \frac{3}{10} \approx 70.0\%$$
$$Q_{\text{correctness}}(\text{"Sven}, \text{"Bird"}) = 1 - \frac{4}{4} \approx 0.0\% \quad (7)$$

The example makes clear that the metrics of the correctness are determined in the name of the author are quite differently: Under the consideration of the last name "Wetterbaum" and "Weizenbaum", the value of the metrics for correctness amounts to 70.0%, while by the first name "Sven" and "Bird", a value of 0.0% arises, since both sides have no common character.

As an alternative to this example, the Levenshtein distance is used. Levenshtein distance calculates the minimum number of insertions, deletions, substitutions, and match operations to convert a given string to a second string, as well as transform strings of unequal length or to measure the effort on the basis of the minimum number of these operations (Levenshtein 1965).

The following figure shows the measurement of the Levenshtein distance with its operators **I**nsertion, **D**eletion, **S**ubstitution and **M**atch for the example of the author's name "Sven Wetterbaum" (Fig. 7).

As a result of the Levenshtein distance, this example is seven substitution steps that are required in this case.

Measurement of consistency

Consistency (freedom from contradiction) can be looked at from different perspectives. In RIS consistency stands for the observance of the integrity rules which are defined for specific fields, tuple, attributes, or relations. The specified control amount defines logical connections of the checked data of amount, which must be kept (Hildebrand et al. 2015).

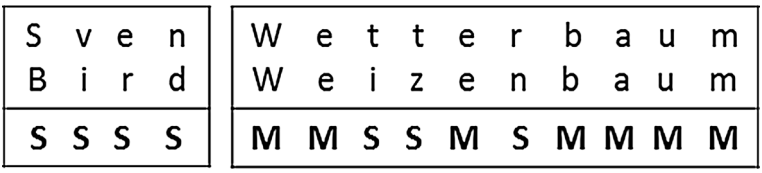

Fig. 7 Calculation of the Levenshtein distance for the example of the author's name in the publication list

Looked at in general, consistency can be described as the contradiction between two or more related data elements (e.g. zip code and location) (Batini and Scannapieco 2006). In addition, consistency in the format is also important for a single data element.

The assessment of contradiction can be carried out by measurement of the data quality within the RIS in order to examine redundantly stored data for contradictions.

Consistency can be measured as follows (Lee et al. 2006):

$$Q_{\text{consistency}} = 1 - \frac{\text{Number of inconsistent units}}{\text{Number of consistency checks performed}} \tag{8}$$

For the sake of clarity, an example of the data quality dimension consistency will be calculated using authors from various publications. First, the publication list is presented by various authors in the following tables, and then the degree of consistency of that information is calculated (Fig. 8).

For the author "Virginia Lopez", the data sources are contradictory both in terms of her surname and her ORCID (Open Researcher and Contributor ID). The background of these discrepancies can be manifold and arise from for example, the fact that the author has not regularly updated her ORCID profile or registered a new ORCID after a name change. Similarly, the operational systems behind the data import may contain outdated or incorrect data.

To do this, the degree of consistency error for the above example is calculated as follows:

$$Q_{\text{consistency}} = 1 - \frac{2}{32} \approx 0.9375 \tag{9}$$

The result states that the data concerning the publication list of authors 93.75% not in contradiction. The extent to which these (consistent) data are correct cannot be deduced from this measured value.

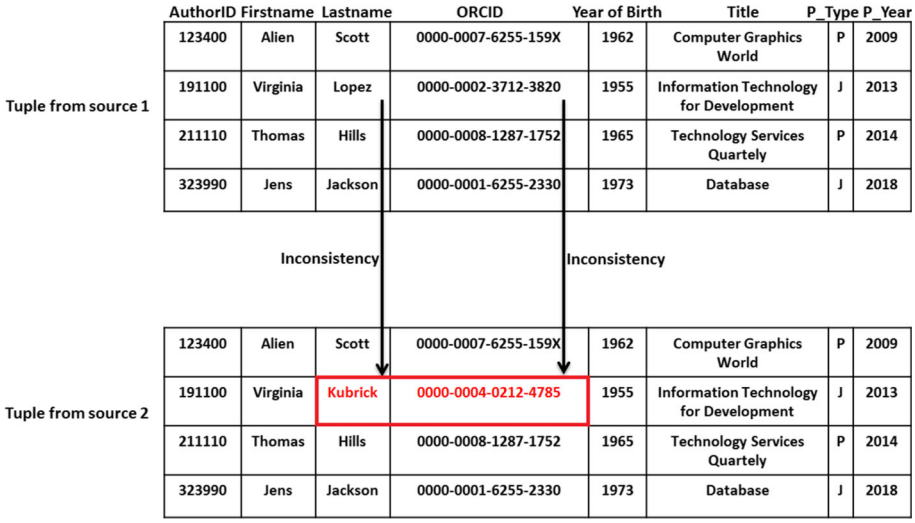

Fig. 8 Example of inconsistent data in a publication list

As an alternative measurement one could use the Levenshtein distance for this example of the ORCID and author's name. Figure 9 shows the calculation of the Levenshtein distance with its operators.

As a result of the Levenshtein distance for this example, there are five substitution steps and two deletion steps in the name, and for their ORCID, seven substitution steps are necessary and the rest are the same.

Discussion

The measuring of data quality of the RIS is an indispensable basis for its lasting improvement. The resultant dimensions are measurable and refer to completeness, correctness, timeliness and consistency of the data. The measurement of these dimensions can occur with every RIS in an institution. After the measurement of the data quality dimensions, it is a matter of providing in the next step for the monitoring and improvement of the data quality in all institutions. Against this background, a framework as a process flow according to UML activity diagram notation is developed in this present paper and should serve as a model to indicate how the quality of the data of the RIS is ensured, supervised and improved.

With the help of the following picture, the mentioned framework is introduced for the monitoring and improvement of the data quality in the RIS (Fig. 10).

At the beginning of the framework external and internal data sources of an establishment are collected by the management or technical staff and afterwards these compilations of data are subjected to control as well as measurement to register the completeness, correctness, timeliness and consistency of the data records. In this case, there are two possibilities: either the data records are identified as faulty, or they turn out to be complete and correct, so that they can be loaded directly in the research information system.

Should faulty data records appear, a correction and updating of the data records must be carried out by the management or the technical staff. Besides, the improved data records are then checked again according to the four dimensions. Only then is it advisable that the improved data are loaded into the research information system.

Finally, the data loaded in the FIS are presented. Using portals, reporting and other front-end applications, the information coming from the system is visualized. Here, the user is provided the prepared information and analyzes in a clear form by various application components.

L o p e z - - K u b r i c k	0000-0008-1287-1752 0000-0004-0212-4785
S S S S S D D	MMMM-MMMS-SMSS-SMSS

Fig. 9 Calculation of the Levenshtein distance for the example of the author's name and ORCID in the publication list

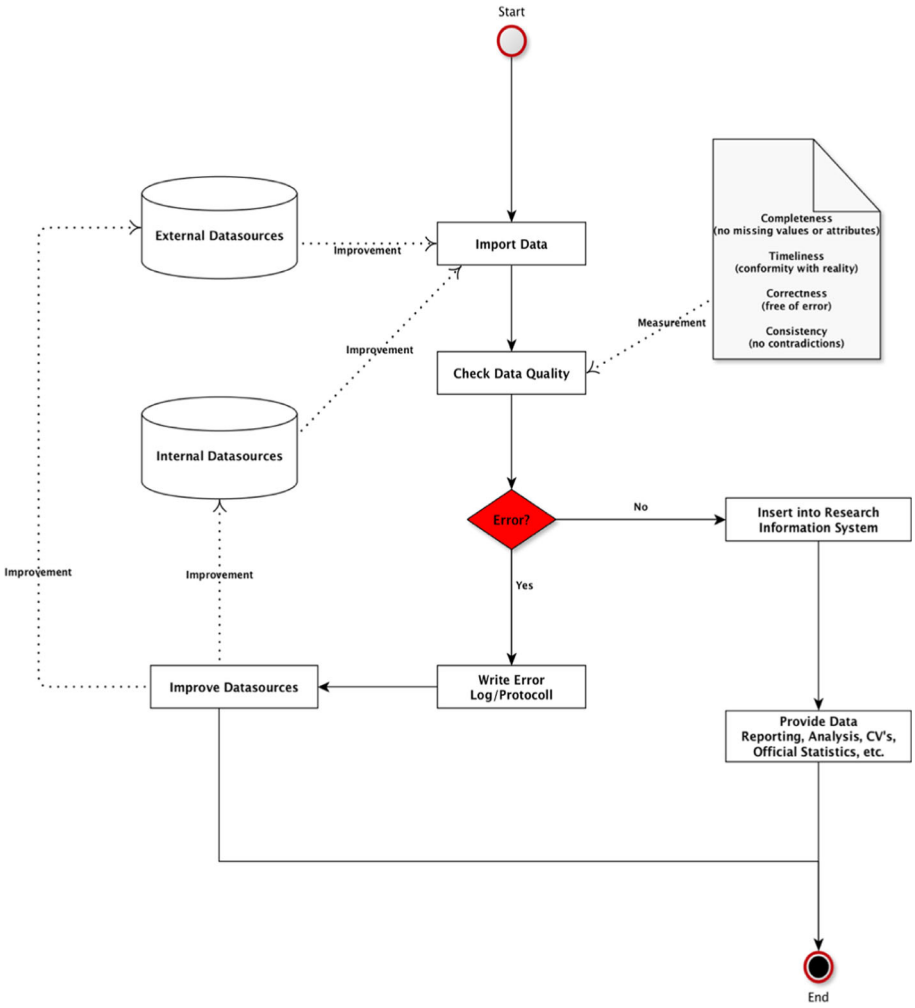

Fig. 10 Framework for monitoring and improving data quality in the RIS

Conclusion and outlook

This paper took up the question of how data quality in RIS can be measured. The aim was to present the selected metrics for the data quality dimensions of completeness, correctness, timeliness and consistency as found in the literature. These enable an objective, effective and largely automated measurement within the research information system. The four selected metrics have proven to be respectably easy to measure. In addition, these represent a particularly representative illustration of the reporting for the RIS users and lead to an improved basis for decision-making.

The results of the measurements of the metrics have shown that measurements are theoretically possible with every RIS, and that a high data quality is available to universities and research institutions, e.g. research information systems. In this respect, the

verification of the data quality must always be carried out with specific reference to its context. The concept which was introduced here provides a suitable approach and a framework to measure data quality in any RIS. The approach shown here and the framework can be used as a basis for the examination and improvement of the data quality in an RIS in institutions making use of such.

As I have explored new techniques and methods such as data cleansing and data profiling to secure and improve data quality issues in the context of RIS, future work will consider other topics such as the extraction-transformation-load (ETL) process and semantic text analysis, which are also applicable to RIS. Finally, expert interviews and quantitative surveys with universities and research institutes are planned to find out how high the quality of data is in their research information system.

References

Apel, D., Behme, W., Eberlein, R., & Merighi, C. (2015). *Successfully control data quality: Practice solutions for business intelligence projects*, 3rd, Revised and Extended Edition. Heidelberg: dpunkt.verlag.

Azeroual, O., & Abuosba, M. (2017). Improving the data quality in the research information systems. *International Journal of Computer Science and Information Security, 15*(11), 82–86.

Azeroual, O., Saake, G., & Abuosba, M. (2018a). Data quality measures and data cleansing for research information systems. *Journal of Digital Information Management, 16*(1), 12–21.

Azeroual, O., Saake, G., & Schallehn, E. (2018b). Analyzing data quality issues in research information systems via data profiling. *International Journal of Information Management, 41*(8), 50–56. https://doi.org/10.1016/j.ijinfomgt.2018.02.007.

Batini, C., & Scannapieco, M. (2006). *Data quality—Concepts methodologies and techniques*. Heidelberg: Springer.

Cordts, S. (2013). *Data Quality in databases*. Hamburg: Maren Nasutta Mana-book-Verlag.

DINI AG Research Information Systems. (2015). *Research information systems at universities and research institutions-position-paper*. https://dini.de/fileadmin/docs/AG_Positionspapier_engl_final.pdf.

English, L. P. (1999). *Improving data warehouse and business information quality: Methods for reducing costs and increasing profits*. New York, NY: Wiley.

Gebauer, M., & Windheuser, U. (2015). *Structured data analysis, profiling and business rules*. Wiesbaden: Springer Fachmedien Wiesbaden.

Heinrich, B., Kaiser, M., & Klier, M. (2007). How to measure data quality? A metric based approach. In *28th international conference on information systems (ICIS)*. Montreal.

Heinrich, B., & Klier, M. (2009). Die Messung der Datenqualität im Controlling – Ein metrikbasierter Ansatz und seine Anwendung im Kundenwertcontrolling. *Controlling & Management: ZfCM ; Zeitschrift für Controlling und Management, 53*(1), S34–42.

Helfert, M. (2002). *Planning and measurement of data quality in data warehouse systems*. Dissertation, University of St. Gallen, Difo-Druck, Bamberg.

Herwig, S., & Schlattmann, S. (2016). *An economics-based location determination of research information systems*. Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn.

Hildebrand, K., Gebauer, M., Hinrichs, H., & Mielke, M. (2015). *Data and information quality. On the way to the information excellence*, 3rd, extended edition. Wiesbaden: Springer Fachmedien Wiesbaden.

Hinrichs, H. (2002). *Data quality management in data warehouse systems*. Dissertation. Oldenburg: Oldenburg University.

Krcmar, H. (2015). *Information management*. Wiesbaden: Springer Gabler.

Lee, Y. M., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to data quality*. Cambridge, MA: MIT Press.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR, 163*(4), 845–848. **(Russisch, Englische Übersetzung in: Soviet Physics Doklady, 10(8): 707–710, 1966)**.

Martin, M. (2005). *Measuring and improving data quality. Part II: Measuring data quality*. NAHSS Outlook. Ausgabe 5.

Scannapieco, M., Missier, P., & Batini, C. (2005). Data quality at a glance. *Datenbank-Spektrum, 5*(14), 6–14.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Würthele, V. (2003). *Data quality metrics for information processes*. Zurich: ETH Zurich.