Quality of Research Information in RIS Databases: A Multidimensional Approach

Otmane Azeroual[1,2,3](✉), Gunter Saake[2], Mohammad Abuosba[3], and Joachim Schöpfel[4]

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

[2] Otto-von-Guericke-University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany

[3] University of Applied Sciences (HTW) Berlin, Wilhelminenhofstraße 75 A, 12459 Berlin, Germany

[4] GERiiCO-Labor, University of Lille, 59650 Villeneuve-d'Ascq, France

Abstract. For the permanent establishment and use of a RIS in universities and academic institutions, it is absolutely necessary to ensure the quality of the research information, so that the stakeholders of the science system can make an adequate and reliable basis for decision-making. However, to assess and improve data quality in RIS, it must be possible to measure them and effectively distinguish between valid and invalid research information. Because research information is very diverse and occurs in a variety of formats and contexts, it is often difficult to define what data quality is. In the context of this present paper, the data quality of RIS or rather their influence on user acceptance will be examined as well as objective quality dimensions (correctness, completeness, consistency and timeliness) to identify possible data quality deficits in RIS. Based on a quantitative survey of RIS users, a reliable and valid framework for the four relevant quality dimensions will be developed in the context of RIS to allow for the enhancement of research information driven decision support.

Keywords: Research information systems (RIS) · Research information · Utility · System acceptance · Data quality dimensions · Data quality measurement · Data quality improvement · Reliability · Validity · Structural equation modeling

1 Introduction

For the operation of a research information system (RIS) as a central source of information in academic institutions, the quality of the research information and the reliability of derived statements is of central importance. RIS is a database and tool of research administration that specifically supports the management and provision of research information and its activities (such as affiliation of persons to institutions, publications, research projects, patents, etc.). The peculiarity of RIS is to understand, manage, evaluate and further develop the portfolio of scientific research activities of academic institutions.

W. Abramowicz and R. Corchuelo (Eds.): BIS 2019, LNBIP 353, pp. 337–349, 2019.
https://doi.org/10.1007/978-3-030-20485-3_26

In addition, RIS provides them with a sound basis for decision-making and reporting, in which the research information from different heterogeneous sources (e.g. human resources, financial budgets, libraries, etc.) are brought together. The reason for this is not at least the intention to merge the collected research information into a homogeneous amount, to bring it into logical context and to be able to evaluate and present research-relevant decisions consequently. Since research information serves the interests of various data users (e.g., academic institutions, funding bodies, companies, etc.), the reports should be generated from a high-quality RIS. If this research information is incorrect, incomplete or inconsistent, this may have significant implications for institutions.

To have valid and valuable results, it is indispensable to define quality dimensions for data management, measuring, achieving, maintaining and ensuring the highest quality of research information, in addition to the application of methods and techniques (e.g., data profiling and data cleansing) [1, 2].

Data quality dimensions help to structure the research information in RIS and make the success measurable for the decision maker [3]. They provide a way to measure and manage data quality and information [13]. As discussed in the various studies [7, 9, 10, 12], there are a diversity of data quality problems in definition and measurement that are essential to ensuring high data quality [19]. Without quality control, data quality will progressively decrease [5].

The paper firstly examines the quality of RIS and its impact on user acceptance, and then proposes a framework as a structural equation model (SEM) to support quality measurement in RIS. With this model, it is possible to find out to what extent the investigated data quality dimensions have an influence on the improvement of the research information in RIS.

Research on this topic so far, by euroCRIS or the German DINI AG FIS, often stressed the general importance of data quality. Our paper tries to add more detailed insight, based on the four quality dimensions (correctness, completeness, consistency and timeliness) and their relationship to the process of improvement in the RIS. To estimate the reliability and validity of the data quality dimensions for the improvement process in the RIS, results of a quantitative online survey by the "QuestionPro" software (between February 2018 and September 2018) with universities and academic institutions from Germany and other European countries are presented. More information about the survey is provided in [4].

The paper tries to answer the following questions:

- Which aspects are important for describing the data quality in RIS?
- Which data quality dimensions are important for RIS to check and measure research information?
- What data quality problems will be exchanged during collection, integration and storage of research information in RIS?
- How to detect data quality problems in RIS?
- At which point of data processing does a data quality check by the RIS take place?
- Which methods and techniques are used to improve and increase data quality in RIS?
- How high is the data quality in RIS?

Factor analysis and Cronbach alpha test are used to assess the consistency, reliability and validity of the results [11, 14, 16].

The paper has four sections: (1) the introduction to the topic and methodology; (2) the concept of data quality in the context of RIS and the user acceptance based on data quality in RIS; (3) presentation of results; (4) a framework for measuring and improving the quality of research information in RIS. Finally, the paper ends with a conclusion of the most important results and an outlook.

2 State-of-the-Art Data Quality

The increase in research information and its sources presents universities and academic institutions with difficult challenges, furthermore data quality is becoming more important. The term data quality is defined in various ways both in the literature and by experts. Wand and Wang [17] conclude that data quality issues occur with inconsistencies between the view of the information system and the view of the real world. The occurring deviations can be determined based on data quality dimensions such as completeness, correctness and consistency. English [8] differentiates between the quality of the data definition, the architecture, the data values and the data presentation. Wang and Strong [18] evaluate in an empirical study of general data quality dimensions across four categories. Data quality was therefore determined contextually or based on the data values (inner data quality). Furthermore, the data quality must satisfy what the user and system demands.

From these different approaches to the topic of data quality can be defined in the context of RIS as *fitness for use* and describes the suitability of the data objects for users in a particular context [18], they must be correct, complete, consistent and current.

Four data quality dimensions, as defined by Wand and Wang [17] will be explained, which are considered relevant in the context of data quality in RIS.

- *Correctness:* The research information is consistent in content and form with the data definition. Correct research information contains the contentwise correct information in the predefined formats of the attributes.
- *Completeness:* On the one hand, the criterion refers to the completeness of the research information in the transmission of data between the different systems. A record is complete if no data has been lost during the transformation from System A to System B. On the other hand, a record is complete if all necessary values are included.
- *Consistency:* The consistency of the research information refers to the correctness of the stored data in the sense of a consistent and complete representation of the relevant aspects of reality.
- *Timeliness:* The research information is current if it reflects the actual property of the described object in a timely manner. The research information is not outdated.

In addition to data security, ease of use and other variables, data quality is one of the main conditions of user acceptance of RIS. This is primarily about trust - trust in the system, in its provider and in its administration. A system that does not reliably identify or correct data problems, or that itself is a source of data quality defects, can (and will) not be trusted. Perceived quality problems affect the subjective performance expectations of the system. Data quality problems or poor data quality can have different causes. In order to improve data quality, the cause must be known. Because only if the cause is remedied, a lasting improvement of the data quality can be achieved. However, in the case of the RIS, poor data quality is all the more problematic in terms of strategic and sometimes highly sensitive information and decision-making aids, such as personal or financial data. The perceived data quality has a direct impact on the expected benefit and thus, indirectly, on the intended and real use of the system. User acceptance is not only a matter of ergonomics and system quality, but also of organization, communication and legal protection. In this sense, data quality is a necessary but not sufficient condition for user acceptance. But one can also ask the question differently: What incentives does the system and its organizational environment provide for the scientists involved? What "facilitating conditions" are created by science management to support acceptance by scientific staff?

3 Results

This section presents the research results of the quantitative study. The survey was addressed to 240 German universities and research institutions and 30 European universities. A total of 51 German universities and research institutions and 17 European universities responded. According to the survey, the responding institutions implemented their RIS 2 to 8 years ago.

The survey's main objective was to assess the management of data quality, i.e. processes to define, measure, and improve data quality. The first question identifies main aspects of the concept of "data quality". According to the respondents, most important is the quality of data provision (overall RIS), as well as the quality of the data content and the quality of the definitions (see Fig. 1).

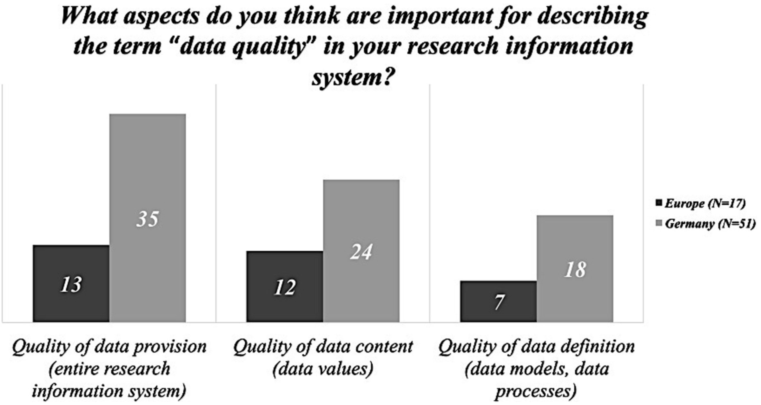

Fig. 1. Aspects describing the concept of data quality in the context of RIS (*N* = *68*).

Data quality can differ between the data definition and architectural quality, the quality of the data values as well as the quality of the data presentation and groups these into [8]:

- The quality of the data standards (guidelines that support a consistent, accurate, clear and understandable data definition).
- The quality of the data definitions (semantic aspects and business rules).
- The quality of the information system architecture (general design of data models and databases in terms of reuse, stability and flexibility).

Former research defined data quality with some universal dimensions. The survey tries to assess which of these dimensions are of particular importance to institutions for examining and measuring data quality in RIS.

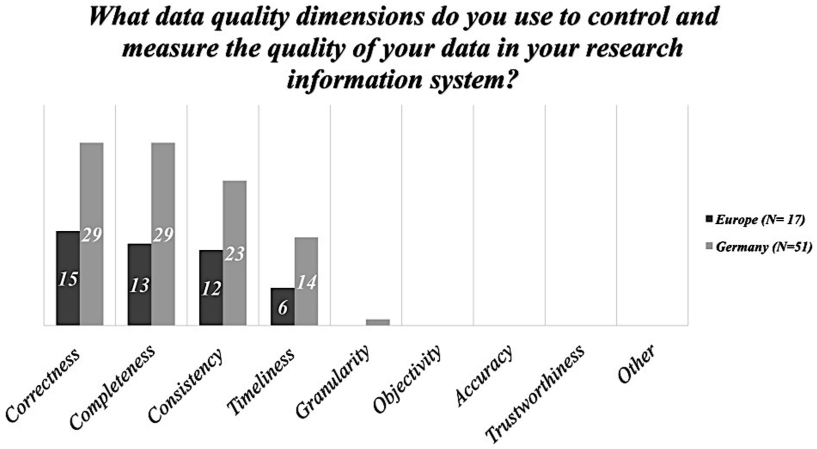

Fig. 2. Data quality dimensions for RIS (*N* = *68*).

The survey reveals that the respondents evaluate the correctness, completeness, consistency and timeliness of the research information as most important (see Fig. 2). To monitor data quality, these dimensions should be objectively measurable, automatically collected which requires querying the data sources to have values for processing. For larger data sources, good sampling and extrapolation techniques should be used. Automatic assessments should be conducted as often as possible, and simple procedures should be used to not burden RIS unnecessarily. Therewith, e.g. the correctness, completeness and consistency of the research information is verified or at least well assessed. Research information that meets 80% (good) to 100% (very good) of the correctness, completeness, and consistency of data represents a precise reflection of real-world system states to information system states and can be used to justify about data quality [17]. Because such reasoning can be made to improve data quality [17]. In addition, the respective degree of fulfillment of the requirements can be determined by the data user. Figure 3 presents a model for classifying data quality dimensions in the context of RIS.

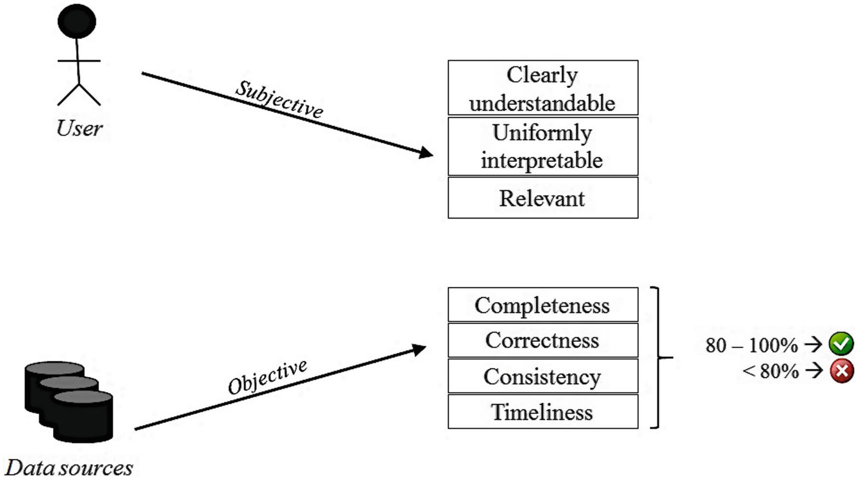

Fig. 3. Classification of data quality dimensions in RIS.

During the collection and storage phase of important research information from various internal and external data sources of the institutions in RIS, a large variety of data problems arise which must be processed by the RIS. From the point of view of universities and academic institutions, Fig. 4 shows possible data quality problems of data quality in RIS.

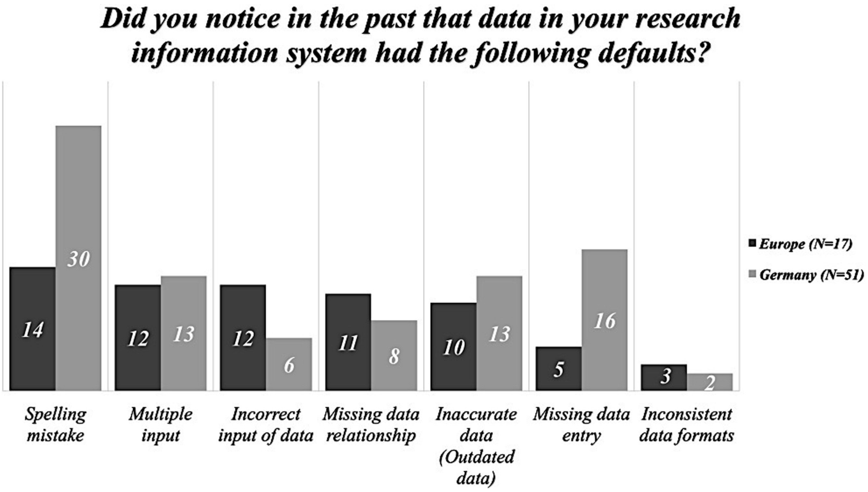

Fig. 4. Data quality problems in RIS (*N* = *68*).

Poor data quality leads to wrong decisions, employee dissatisfaction and rising costs. In order to be able to recognize errors at an early stage and treat them efficiently, the following questions must be answered in institutions:

- Will the data quality in RIS get worse or better?

- Which source system causes the most/least data quality problems in RIS?
- Can patterns or trends be recognized by the data quality check in RIS?

Data quality problems are continuously detected in institutions by plausibility checks. The analysis of quality problems in RIS are illustrated in Fig. 5.

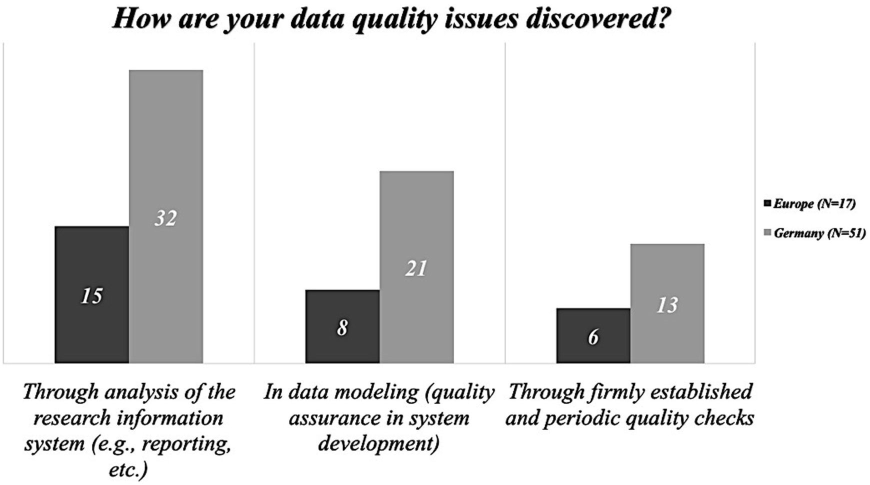

Fig. 5. Data quality checks in RIS (*N* = *68*).

The quality checks performed by RIS on institutions take place most during the data processing in the data storage in the RIS as well as the import of internal and external data sources and data presentation (see Fig. 6).

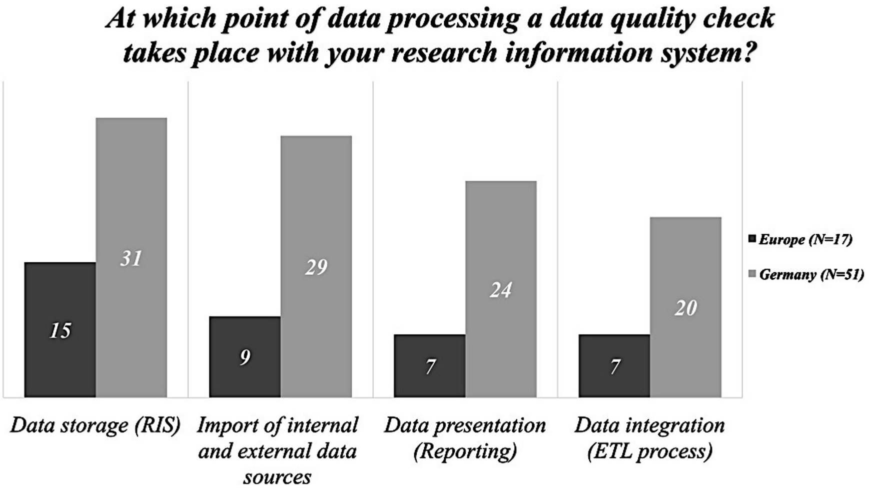

Fig. 6. Data processing with the quality checks by RIS (*N* = *68*).

Which techniques, methods and measures are used to improve the RIS data quality? The majority of the respondents use data cleansing methods, while pro-active approaches and data profiling rank second. Re-active approaches and ETL processes seem rather rare. Figure 7 shows the results.

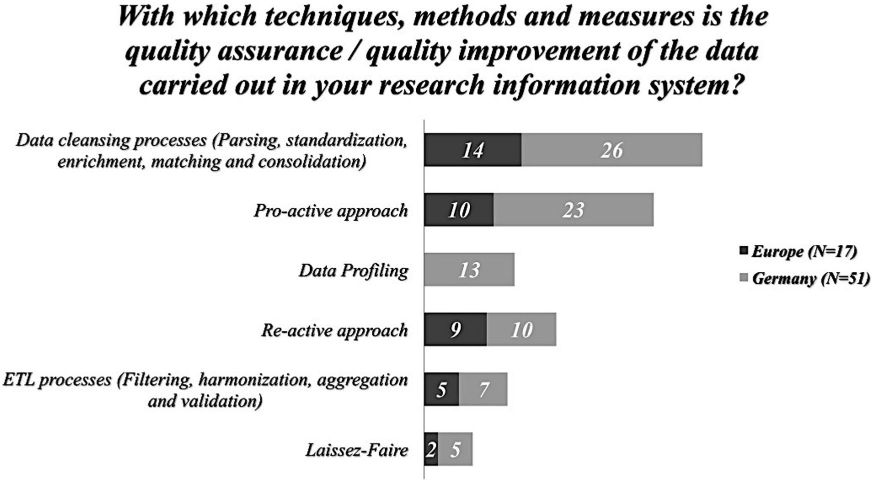

Fig. 7. Techniques, methods and measures to improve data quality in RIS (*N* = *68*).

Many respondents attach great importance to data quality in RIS (see Fig. 8). High quality contributes to the fact that working with the RIS is perceived by the users as pleasant and easy. High and reliable data quality creates trust in RIS. Users not only work more efficiently and more powerfully, but also more securely, which in turn increases user acceptance. High data quality adds value and provides benefits to universities and research institutions which will further increase user acceptance.

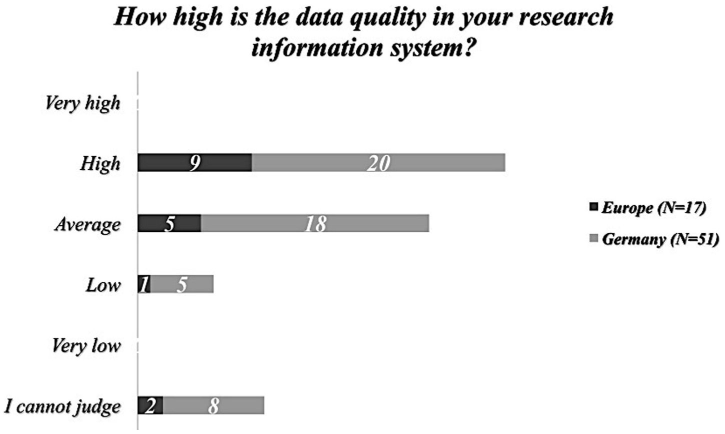

Fig. 8. Degree of data quality of RIS in German and European universities and academic institutions (*N* = *68*).

4 Supporting Framework for Research Information Quality Dimensions

To make a statement about the dependency relationship of the data quality dimensions for the improvement process in the RIS, it is necessary to consider important dimensions to each other. For such a consideration and estimation of reliability and validity based on the results of the quantitative survey, a flexible framework will be used as a structural equation model (SEM) [14]. The data quality in RIS is measured by the variables correctness, completeness, consistency and timeliness. The framework is shown in Fig. 9.

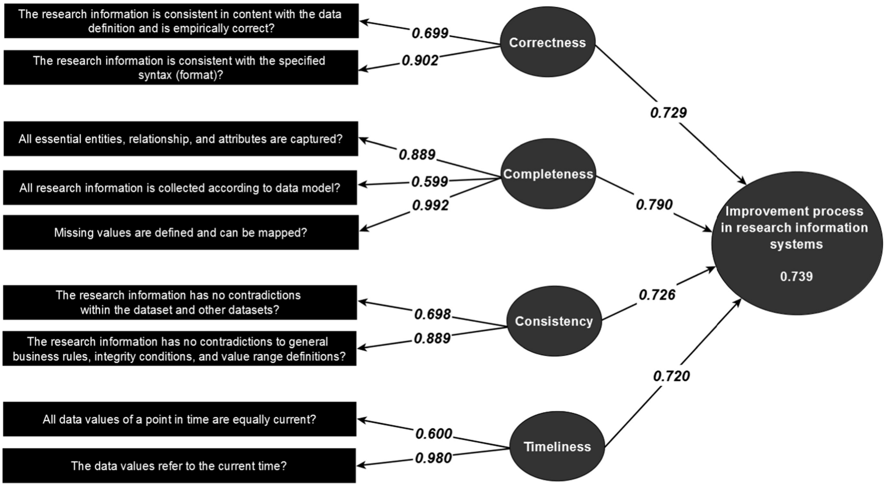
The research information is consistent in content with the data definition and is empirically correct?
The research information is consistent with the specified syntax (format)?
0.699
0.902
Correctness
0.729
All essential entities, relationship, and attributes are captured?
All research information is collected according to data model?
Missing values are defined and can be mapped?
0.889
0.599
0.992
Completeness
0.790
Improvement process in research information systems
0.739
The research information has no contradictions within the dataset and other datasets?
The research information has no contradictions to general business rules, integrity conditions, and value range definitions?
0.698
0.889
Consistency
0.726
0.720
All data values of a point in time are equally current?
The data values refer to the current time?
0.600
0.980
Timeliness

Fig. 9. Framework as a structural equation model for data quality dimensions in RIS.

The framework allows the measurement of the observable variables of data quality dimensions. They represent the latent variables of the constructor. Each latent variable is operationalized by directly observable or ascertainable variables. In this framework, a reflective model is used because the latent variables affect the respective indicators. The number next to the arrow describes the relationship between the latent variable and the corresponding indicator. This number is to be interpreted as a factor load and indicates how strong the reliability and validity is to the latent variable. The developed framework can support institutions at every step from design, execution, analysis, and improvement to assessing and overcoming the data quality issues of a RIS.

To evaluate the reliability and validity of the scales for data quality dimensions and design factors for the improvement process in the RIS, a statistical analysis with R on the survey data was applied to assess Cronbach's alpha analysis and principal component analysis (PCA). Ensuring that respondents can accurately answer questions about data quality dimensions has been limited to the universities and academic institutions that have been using RIS for a long time. For the survey, a Likert scale was used, with four possible answers from "very important" to "unimportant".

Cronbach’s alpha determines the reliability of the dimensions and its value is between 0 and 1, with values less than 0.5 considered “unacceptable” and higher than 0.65 “good” [6]. Table 1 calculates and displays the reliability of the coefficient for each individual dimension.

Table 1. Result of Cronbach’s alpha for data quality dimensions.

Dimensions	Number of items	Alpha
Correctness	2	0.729
Completeness	3	0.790
Consistency	2	0.726
Timeliness	2	0.720
Overall Cronbach’s alpha	0.739	

The results for the dimensions consistency and timeliness were respectively 0.72, for the dimension correctness it was around 0.73 and for the dimension completeness it was 0.79. The total Cronbach’s alpha value reaches a value of 0.74 and is therefore considered as a good value. Therefore, the Cronbach’s alpha reliability coefficients values show that the instrument is reliable for calculation and all dimensions have a relative consistency for each construct.

To further investigate the relationships of data quality dimensions or factors, the determination of content and construct validity was made using PCA. This is a “method of data reduction or factor extraction based on the correlation matrix of the dimensions involved” [15]. The aim with this method is to create the measurement of the dimensions with the orthogonal rotation technique “varimax rotation”, which minimizes the number of connections and simplifies the interpretation of the dimensions [11]. To calculate the factors, a coefficient greater than 0.5 was chosen to make the factor matrix more reliable, with the eigenvalue (variance) greater than 1 and Kaiser-Meyer-Olkin (KMO) greater than 0.5 for measuring the adequacy of the sample [15]. The correlation of the factor values is called loadings and these explain the relationship between dimensions and factor. Using the factor loading, one can see which dimensions are highly correlated with which factor and which dimensions can be assigned to that factor [15]. Table 2 below shows the results and calculation of PCA for the validity of the data quality dimensions.

Table 2. Result of PCA for data quality dimensions.

Dimensions	Number of items	Factor loading	Eigenvalues	% of variance
Correctness	CorrQ1	0.699	1.73	19.34
	CorrQ2	0.902		
Completeness	CompQ1	0.889	5.01	51.84
	CompQ2	0.599		
	CompQ3	0.992		
Consistency	ConsQ1	0.698	1.33	15.75
	ConsQ2	0.889		
Timeliness	TimeQ1	0.600	1.01	13.07
	TimeQ2	0.980		

The KMO-value for items of the dimension correctness was 1.20, for the completeness 1.64, for the consistency 0.86 and for the timeliness the KMO-value was 0.80. Thus, the KMO values were above 0.5 for all four dimensions, indicating that the sample size was adequate and that there were enough items for each factor. All factors of the tested dimensions had a factor load of more than 0.5, which means that all items can be loaded with the same factor. For the first two factors, correctness and completeness, the extracted variance was 19.34% and for others 51.84%. The eigenvalue for both factors is thus greater than 1. However, for consistency and timeliness, the eigenvalue is also greater than 1, with an extracted variance of 15.75% and 13.07%. Thus, for all factors, the items are compared to related dimensions and can be grouped into one factor.

As an overall assessment of the framework, it can be summarized that the data quality dimensions positively influence the improvement process in the RIS. The analysis results prove the good reliability and validity of the nine data quality items in RIS. The result of the PCA shows that the items were highly valid in the construct and demonstrate good statistical properties for testing the developed framework. For academic institutions that have problems integrating different information systems or external data sources, it is advisable to consider these four reliable and valid dimensions to optimize data processing processes and ensure data quality in the RIS.

5 Conclusion

An institution needs research information to monitor and evaluate its research activities and make strategic decisions about different application and usage scenarios. For a holistic view of the research activities and their results, the introduction of a RIS is therefore essential. It is equally essential that such a system provide the required information in a secured quality. In order to make the best possible decisions in academic institutions, they must be based on research information that has to meet high requirements. The right research information must exist and be available in the right place at the right time.

Decisions made on the basis of bad or inadequate information due to poor data quality may not be optimal. Poor data quality in RIS poses a challenge in terms of time and cost to institutions, which should not be underestimated. Especially when integrating research information from heterogeneous systems into the RIS, it may happen that the data formats of the fields in the source system do not match. It is possible that the source data is in the wrong format or in the wrong range of values. To overcome this challenge, the source systems must be measured, adjusted and controlled so that these constellations can no longer occur.

The term data quality in the context of RIS refers primarily to the first aspect, especially the correctness, completeness and consistency of data and the information derived from it. Implicitly, the second aspect is where the timeliness of the information is to be considered, because outdated information is generally no longer correct in dynamic environments, such as in academic institution. Through the analysis of the survey results and the developed framework, the most important dimensions for the improvement process in RIS could be identified, which are crucial for the measurement of data quality in the RIS. The measurement of these four dimensions can be done with each RIS (see [3] for further details on the data measurement in RIS). The concept presented in the paper offers an appropriate way to measure and improve the processing of data quality in RIS.

The quantitative analysis of this paper has shown that data quality is a critical success factor in user acceptance. To ensure the sustainable use of such a system, it requires the greatest possible user acceptance on the part of the science management, the system administrators and the scientists themselves. User acceptance is based on trust in data quality, which requires continuous quality management. Data quality should therefore be treated as a high priority business process, not only to guarantee and enhance the added value of the information produced, but also to ensure the confidence or user acceptance in universities and academic institutions with a RIS, which in turn the responsible use of such systems is indispensable and at the same time, in the sense of positive feedback, can contribute to the quality of all data. Further work is needed for a better understanding of the relationships between data quality, satisfaction, acceptance and perceived usefulness of RIS, with a larger sample including in-house developments and second generation systems.

References

1. Azeroual, O., Saake, G., Abuosba, M.: Data quality measures and data cleansing for research information systems. J. Digital Inf. Manage. **16**(1), 12–21 (2018)
2. Azeroual, O., Saake, G., Schallehn, E.: Analyzing data quality issues in research information systems via data profiling. Int. J. Inf. Manage. **41**, 50–56 (2018)
3. Azeroual, O., Saake, G., Wastl, J.: Data measurement in research information systems: metrics for the evaluation of data quality. Scientometrics **115**(3), 1271–1290 (2018)
4. Azeroual, O., Schöpfel, J.: Quality issues of CRIS data: an exploratory investigation with universities from twelve countries. Publications **7**(1), 1–18 (2019)
5. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 1–52 (2009)
6. Bovee, M., Srivastava, R.P., Mak, B.: A conceptual framework and belief-function approach to assessing overall information quality. Int. J. Intell. Syst. **18**(1), 51–74 (2003)
7. Engemann, K.: Measuring data quality for ongoing improvement: a data quality assessment framework. Benchmarking Int. J. **21**(3), 481–482 (2014)
8. English, L.P.: Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. Wiley, New York (1999)
9. Ge, M., Helfert, M.: A review of information quality research - develop a research agenda. In: Proceedings of the 12th International Conference on Information Quality, MIT, Cambridge, MA, USA, November 9–11, January 2007 (2007)
10. Heinrich, B., Kaiser, M., Heinrich, B.: How to measure data quality? A metric-based approach. In: Twenty Eighth International Conference on Information Systems, Montreal, pp. 101–122, December 2007 (2007)
11. Jolliffe, L.T., Cadima, J.: Principal component analysis: a review and recent developments. Phil. Trans. A Math. Phys. Eng. Soc. **374**(2065), 20150202 (2016)
12. Madnick, S.E., Wang, R.Y., Lee, Y.W., Zhu, H.: Overview and Framework for data and information quality research. J. Data Inf. Qual. (JDIQ) **1**(1), 1–22 (2009)

13. McGilvray, D.: Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Morgan Kaufmann, Boston (2008)
14. Miller, M.B.: Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modeling. Struct. Equ. Model. Multi. J. **2**(3), 255–273 (1995)
15. Panahy, P.H.S., Sidi, F., Affendey, L.S., Jabar, M.A.: A framework to construct data quality dimensions relationships. Indian J. Sci. Technol. **6**(5), 4422–4431 (2013)
16. Schmitt, N.: Uses and abuses of coefficient alpha. Psychol. Assess. **8**, 350–353 (1996)
17. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM **39**(11), 86–95 (1996)
18. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. **12**(4), 5–33 (1996)
19. Wang, R.Y., Ziad, M., Lee, Y.W.: Data Quality, vol. 23. Springer, New York (2002)