Influence of Information Quality via Implemented German RCD Standard in Research Information Systems

Otmane Azeroual, Joachim Schöpfel, Dragan Ivanović

Influence of Information Quality via Implemented German RCD Standard in Research Information Systems

Otmane Azeroual[1]*[0000-0002-5225-389X], Joachim Schöpfel[2] [0000-0002-4000-807X] and Dragan Ivanovic[3] [0000-0002-9942-5521]

[1] German Center for Higher Education Research and Science Studies (DZHW), Berlin, Germany
[2] GERiiCO-Labor, University of Lille, Villeneuve-d'Ascq, France
[3] University of Novi Sad, Novi Sad, Serbia
* Correspondence: azeroual@dzhw.eu

Abstract: With the steady increase in the number of data sources to be stored and processed by Higher Education and research institutions, it has become necessary to develop Research Information Systems, which will store this research information over the long term and make it accessible for further use, such as reporting and evaluation processes, institutional decision-making and the presentation of research performance. In order to retain control while integrating research information from heterogeneous internal and external data sources and disparate interfaces into RIS and to maximize the benefits of the research information, ensuring data quality in RIS is critical. To facilitate a common understanding of the research information collected and to harmonize data collection processes, various standardization initiatives have emerged in recent decades. These standards support the use of research information in RIS and enable compatibility and interoperability between different information systems. This paper examines the process of securing data quality in RIS and the impact of research information standards on data quality in RIS. We focus on the recently developed German Research Core Dataset standard as a case of application.

Keywords: research information systems (RIS); research information; research information integration; heterogeneous information sources; research core dataset (RCD); standardization; interoperability; data quality management concept.

1 Introduction

The collection and exchange of research information (e.g., information about research staff, organizations, publications, project funding, patents, partners, etc.) are integral parts and success factors in the research information management process. More and more, Higher Education and research institutes collect, integrate and store their research information in Research Information Systems (RIS) for evaluation, reporting and presentation of research results. The term RIS describes a central database or federated information system that aggregates information about research activities, results and impact.

The development of RIS includes the emergence of different research information standardization initiatives at both international and national levels [4]. The most widely used data model for RIS is the euroCRIS Common European Research Information Format (CERIF) [14]. The CERIF Data Model is an international standard for the management and exchange of research information and describes relevant object types from all areas of research and development [13]. Another international initiative, the Canada-based non-profit Consortia Advancing Standards in Research Information (CASRAI), develops and maintains a standard vocabulary for research information focusing on semantics [4][15].

In Germany, the Council of Science and Humanities (WR) initiated in 2013 a standard for the German science system, the so-called Research Core Dataset (RCD)[1] in order to harmonize reporting across a wide variety of institutions [9]. The RCD specifies definitions and aggregation rules for research information on staff, early-career researchers, third-party funding and budget, patents and spin-offs, publications, and research infrastructures [5]. No particular technical system is required for its implementation; the RCD is system-agnostic. Because of the variety of RIS in German institutions, the RCD does not contain specifications for the underlying base data. Implementation of the standard is, however, aided by the provision of a technical data model in XML-format based on CERIF, which describes both base and aggregate data formats and their respective relationships.

The WR recommends the RCD as a voluntary standard for German Higher Education and research institutions [10]. However, its relevance is not limited to Germany or German speaking countries because of the international character of research and related infrastructures. A new standard for German research information management has a direct and immediate impact on RIS in other, international, European or foreign research institutions and networks. Also, the way how the large German research community process its essential data and information may serve as a model for other national communities.

Our paper investigates the potential impact of the RCD standard on RIS data quality and, moreover, on workload, usage and interoperability with other systems. Following the introduction, the second section briefly describes the peculiarities of the current RIS compared to other information systems. The two following sections three and four deal with data quality issues related to RIS implementation and provide a general concept for RIS data quality managing, together with a workflow. Based on these elements, section five explores the potential usefulness of the RCD standard for the management of RIS data quality. The conclusion summarizes our findings and provides some perspectives for further investigation.

2 Specifics of RIS compared to other information systems

Information Systems (IS) has long been concerned with the research and development of socio-technical systems for the capture, storage, dissemination, retrieval and management of information and knowledge [12]. IS are built to obtain information to aid strategic decision-making. Without them, reporting processes in Higher Education and research institutes often take considerable effort: Research information is extracted manually from Excel spreadsheets, accounting systems (such as SAP, PAISY, etc.), publication repositories (ArXiv, SSOAR, etc.), bibliometric databases (e.g., PubMed, Scopus, Web of Science, etc.), identifiers (ORCID, DOI, etc.), and other systems used. This data often goes through the hands of various employees.

Next to institutional requirements of more efficient research information management, researchers themselves desire precise descriptions of their research profiles, taking into account special criteria of the evaluation, the activity of the respective discipline or they want to find and use exact subject-specific research information of other scientists. In addition, they would like to reduce the burden of documenting projects, creating CVs and publications, etc., as well as duplicate data collection for different reporting purposes and at different levels of organization should be avoided for them to gain more time for the research process.

All these requirements can only be met by the RIS (such as Pure, Elements, Converis, FactScience, HISinOne, etc.). RIS provide the technical infrastructure for centralized data collection of research information, which can then be integrated and modified according to institutional needs. These functionalities are of particular importance for institutions needing to comply with various, elaborate reporting requirements, such as the Research Excellence Framework (REF) in the UK. RIS capabilities also include the automatic creation of standardized reports, facilitating institutional decision-making aided by dashboard functionality. Besides administrative relief, RIS can also provide additional value for the researchers themselves. Their profiles, outputs and activities can be

[1] In German: "Kerndatensatz Forschung" (KDSF) https://kerndatensatz-forschung.de/

publically presented in a research portal, thus furthering visibility and possibilities for cooperation within and between organizations. In addition, some RIS allow for the automatic generation of personal CVs and/or publication lists.

3 Concrete problem statement – Introduction of RIS in academic institutions

The handling of research information in RIS and its numerous applications make the issue of data quality more imminent. Many Higher Education and research institutions have difficulties in guaranteeing the quality of their data, for example, at each institution, personal data, information about their scientific activities, projects and publications are entered and recorded. Low data quality states that analyzes and evaluations are faulty or difficult to interpret. Another point is the additional work required due to low data quality. These quality issues (such as spelling errors, duplicates, missing values, incorrect data, incorrect formatting, wrong data representations and contradictions, etc.) may arise when collecting (especially when entering manually), integrating and storing research information in different RIS [3]. For RIS applications, this point is very important, as all calculations and analyzes are based on the research information and even small errors can accumulate and negatively impact the results. The problems of low data quality cause costs and employee dissatisfaction for the academic institutions.

With the growing amount of data and the increasing number of heterogeneous source systems in Higher Education and research institutes, maintaining data quality is crucial for adequate research information use. Therefore, data quality management is an important issue and of increasing relevance for universities and non-university research institutions. The new concepts, techniques and methods of data quality management especially require a high degree of data quality to deliver correct and effective results. Detailed studies on the techniques, methods for improving and increasing the data quality in RIS can be found in the papers ([2], [3]).

To formulate the problem of our paper well, these questions must be answered: firstly, how to restore decision-makers' confidence in the research information that is manipulated and collected in a RIS. These are bulky, completely heterogeneous, distributed and of varying quality. In order to facilitate integration, the standardization of data models and interfaces can reduce the autonomy of the systems and thus enforce homogeneity. Here is the second question to answer: how can the RCD standard analyze the research information and improve its data quality? Understanding the research information is an important task in their integration into RIS. The goal is to better extract, interpret and reuse the research information.

4 Concept for managing data quality in RIS

Constantly growing data volumes and their increasing complexity make effective and automated data management seem more significant. In particular, Higher Education and research institutions employing RIS, which are equipped to manage research information of high complexity, must benefit greatly from systematic monitoring and improvement of data quality. The gain of collection and integration of research information and its sources still poses challenges for these organizations. Thus, there is a need for a concept of data quality assurance for systematically covering and eliminating data quality problems in the context of RIS so that research management stakeholders can obtain reliable results for effective decision-making.

Data quality is defined here as “fitness for use” or “fit for purpose” and refers to the degree of fulfillment of the totality of requirements for the data needed for a particular purpose [16]. A variety of definitions of this term can be found e.g. in [1], [7], [17], [23] and [24]. For the management of data quality in RIS, the question arises as to which concept of quality management the scientific institutions can employ in order to ensure a high level of quality in RIS. We identify five phases which must be run continuously to improve the sustainable quality of research information in RIS. The phases are depicted in Figure 1.

There are many Total Data Quality Management (TDQM) frameworks in the literature that advocate continuous improvement in data quality by following the Define, Measure, Analyze and Improve cycles [6], [16], [19] and [20]. The phases examined in the workflow presented here are particularly suited for research information (such as publication data, project data, patent data, etc.) during the integration into RIS.

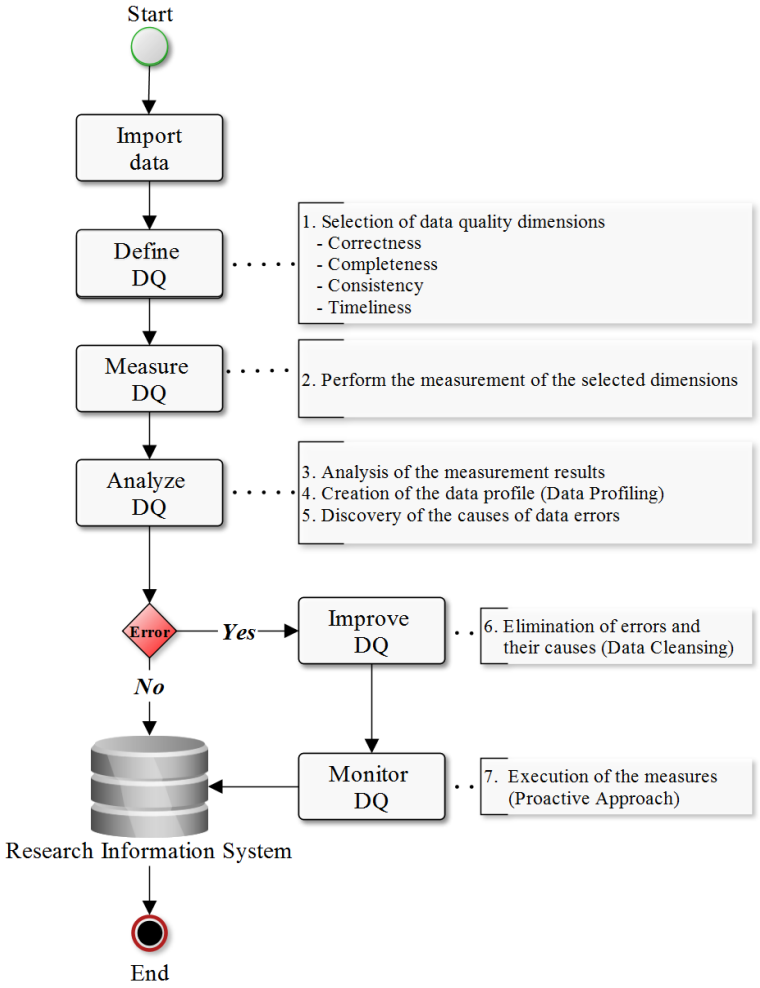

Fig. 1. Iterative workflow for ensuring research information quality.

At the beginning of the workflow, research information from internal and external data sources (e.g., publication databases, project databases, etc.) is collected by the management or technical staff of an organization, and subsequently these data collections are subjected to inspection and their quality is checked. In the first phase (definition of data quality), relevant data quality dimensions must be selected and corresponding requirements defined for them. Requirements can be set by a RIS administrator or manager. Examples of such requirements are presented in Table 1.

Table 1. Data quality dimensions and their requirements in RIS.

Data quality dimensions	**Requirements**
Correctness	Research information must be correct in content and conform to the specified format
Completeness	Research information must be completely extracted from the heterogeneous data sources

Consistency	Research information need not be contradictory
Timeliness (Currency)	Research information must be current (for example, change of name after marriage, change of address, etc.)

During the measurement phase, it is examined to what extent the research information meets the specified requirements. This is quantified and measured with the aid of the four objective data quality dimensions examined and their simple metrics. Data quality can be objectively measured and assessed on the basis of quality dimensions (e.g., completeness, correctness, consistency, timeliness, etc.), as well as subjectively judged by users (e.g., understandability, interpretability, concise representation, relevancy, etc.) [8], [21]. Detailed explanations on this topic and the research question "How can the data quality be measured in RIS?" can be found in the work [2]. After the data quality measurements have been carried out in the RIS, their measurement results must be checked during the analysis phase and the causes of poor data quality must be determined. The causes can be manifold, among other things the cause of erroneous measurements or by the RIS users themselves may have arisen. Therefore, with the help of a data profiling tool (such as Quadient® DataCleaner); it is usually possible to uncover causes of data quality problems in the RIS. Data profiling methods (e.g., attribute analysis, functional dependency and reference analysis) are performed to detect errors. Errors and their causes discovered in the improvement phase should be sustainably eliminated. For this, the methods of Data Cleansing (Parsing, Correcting & Standardizing, Matching, Consolidating and Enhancement) must be applied. If a certain data quality has been achieved, it should be preserved as long as possible. During the control phase, the data is routinely checked before being stored in the RIS, because research information is constantly changing. This requires the implementation of appropriate quality assurance measures such as proactive measures. In order to avoid recurrence of errors in the future, the aim should be to find the causes of these errors and to take proactive measures. Only through the permanent data control is the institution able to provide information about its data quality status at any time, as well as enhancing the confidence in the existing data.

After all phases have been completed, there is always a new iteration, comprising also the re-definition of requirements if necessary. Punctual data cleansing has only a short-term effect. The resulting improvements are quickly lost, especially with frequently modifying research information. Therefore, data quality should not be considered a one-time action. In order to effectively improve data quality in RIS, these investigative methods and techniques need to be applied to research information throughout their lifecycle to guarantee a defined level of quality.

After implementing and testing our workflow with a Quadient® DataCleaner tool using a sample of Web of Science publication data (2,600 datasets) before their integration into RIS, we were able to resolve different data quality errors and abnormalities relating to:

- Incorrect author names (errors and/or name changes)
- Incorrect capture of special characters in author names
- Incorrect and incomplete collection of institutional information of the authors
- Author name disambiguation
- Multiple attributions of identical DOI to different titles and vice versa
- Incorrect order of institutional information
- Faulty multiple detections (e.g., institutions are recorded as several different ones)
- Erroneous separation of institution names (e.g. split if a name contains "and")

5 Application of RCD in RIS

The term standard in research and practice is understood to mean a standardization of data and may differ according to the type of information conveyed by it: quality and compatibility standards. The use of standards is required to ensure compatibility and interoperability between different systems and ultimately to enable their integration. In almost all countries the scientific landscape is very heterogeneous. In addition to a large number of universities, there are a large number of research

institutions. Even if they have very different goals and internal processes vary widely, they all have the same thing to do to compete for funding. This usually happens in the form of rankings or assessments, which presupposes the existence of transparent research reporting. For this reason, the implementation of the standard like RCD in Germany should improve the research coverage of each institution.

RCD aims to standardize the basics for data-driven reports on scientific activities. The RCD specification describes data grouping by content and formats for implementation in RIS. RIS architecture consists of three layers: integration layer, data retention layer, and presentation layer. RCD standard is flexible in its modeling. RCD Standard can be implemented while integrating research information into RIS or exporting data. We focus on integrating RCD into the RIS because universities and research institutions integrate many different heterogeneous data sources into RIS. In fact, collecting and integrating data is a key task in the decision-making process. In addition, the RCD should be able to model and query these multiple data sources.

An application of the RCD standard in RIS is shown in Figure 2. According to RCD standard is intended to eliminate the ambiguity in the collection of research information, thereby improving the quality of these.

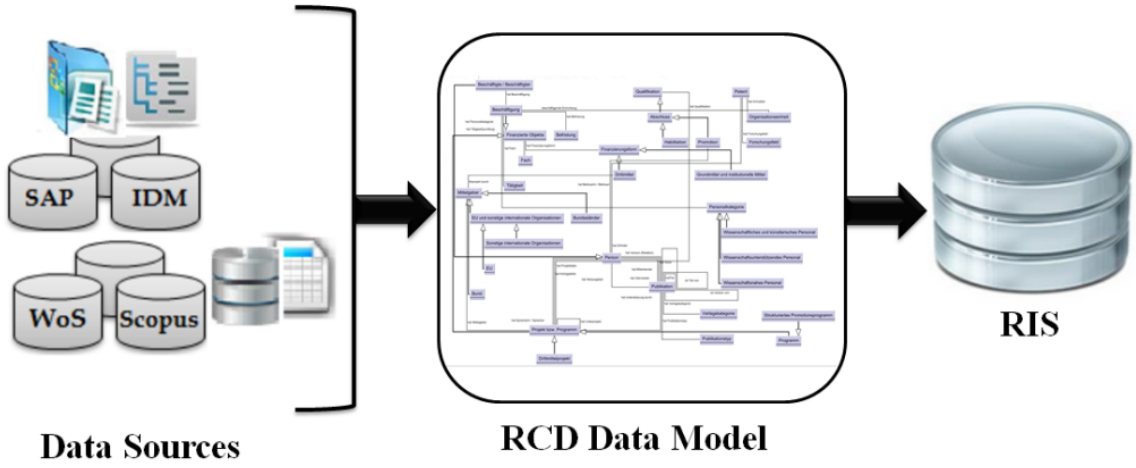

Fig. 2. Practicing RCD Standard in RIS.

Internal and external source data is often contaminated by schema conflicts (such as name and structural conflicts and conflicts by data representation, etc.) and data conflicts (such as conflicting data contents and different definitions of attributes, etc.). According to RCD standard, the ambiguity or the finding of the existing impurities in the collection of research information to eliminate and thereby improve the quality of these. The schema and data conflicts are remedied by necessary integration steps. Using practical examples, different conflict situations and adequate solutions will be considered below using the RCD standard.

The following presents and explains the schema conflicts and data conflicts listed.

- Name conflicts: Name conflicts differ in synonyms and homonyms. Synonyms are semantically equivalent objects that use different relationships for attributes and relations. Homonyms arise when semantically different objects are named the same.
- Structural conflicts: In structural conflicts identity conflicts and missing attributes occur. Identity conflicts are when two relatively identical database tables or contents are identified using different primary keys. Missing attributes must be created in the target database, but care should be taken to allow null values. If there are attributes in the tables of the source databases that only occur exactly in one of the source databases, these attributes must be maintained in the target database.
- Data representation conflicts: It can happen that semantically equivalent attributes with different data types are mapped in the database source. The target database must

have a data type that can represent both. The data sources are migrated to the target database using a conversion function. If different value ranges are specified in the source database for the same attributes, the union must be selected in the target database. If null values are allowed in any of the source databases for an attribute, null values must also be allowed in the target database for that attribute. Default values cannot be taken from the source databases if the same default was not used for all semantically equivalent attributes.

- With regard to the data conflicts, when research information is stored in the source databases according to different conventions, the source databases may contain redundant data. Such can be largely corrected by machine. If the conflicting data content is the result of misspellings or different replication of existing research information, then in most cases manual intervention is required.

Figure 3 explicates these issues and solves them using RCD Standard during their integration into RIS.

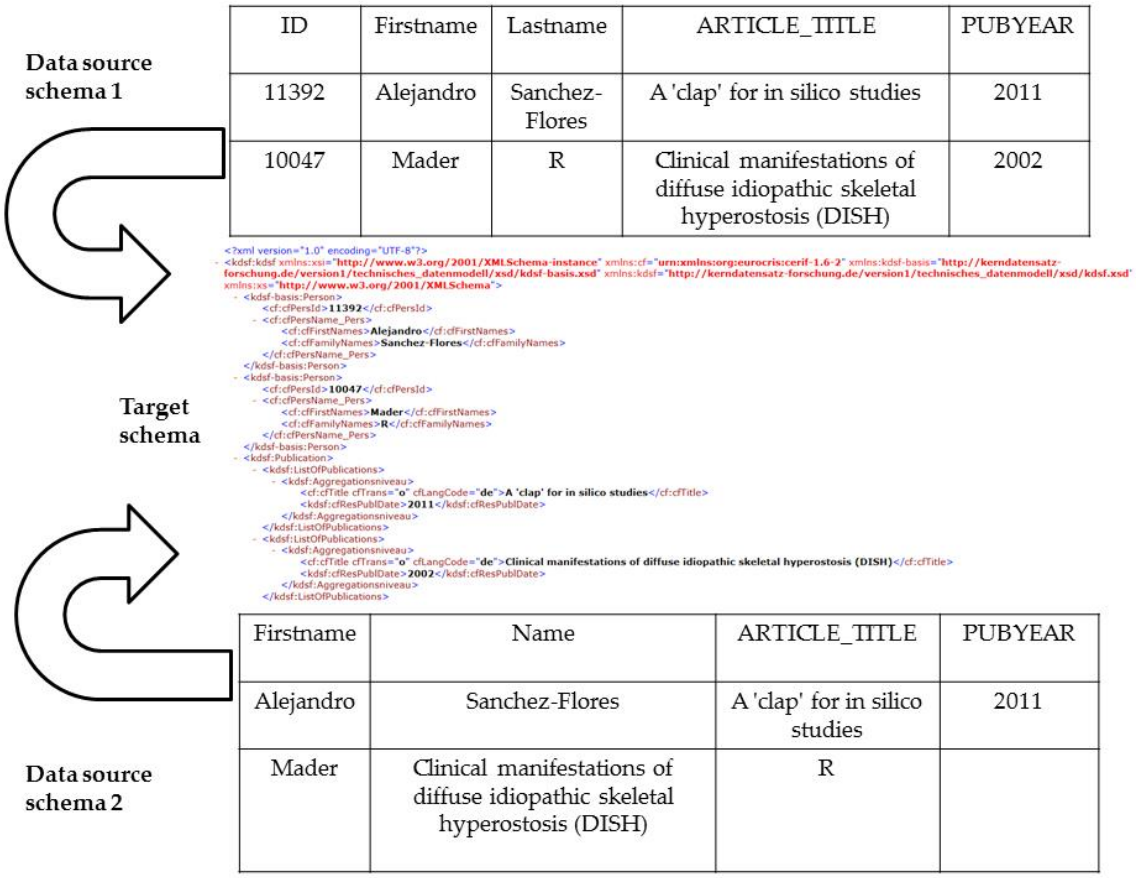

Fig. 3. Solution to the problems of schema and data conflicts.

Standardization of research information would pave the way for sharing comparable data between research institutions. Research information must be maintained in a RIS suitable for both the collection and the preparation and provision of the data. The RCD standard provides the possibility for a simplified comparative presentation. RCD standard exists as XML and OWL (Web Ontology Language) as a data exchange format. In addition, the RCD team also provides an Entity Relationship Model (ERM) that can be used to understand the areas of RCD. However, as the core dataset's goal is to simplify the exchange of data and explicitly not build a centralized database system with research information from various research organizations, the core dataset project does not develop a schema for a database system. On the basis of the RCD-XML schema, transformation rules (e.g. with XQuery or XSLT) can be defined, which convert existing data models (e.g. CERIF) into the data model of the core data record.

RCD Standard creates several advantages for research institutions. This can be listed as follows:

- supports the interoperability of data of different formats
- increases the flexibility of different systems
- facilitates the understanding of information structures
- simplifies data exchange, communication and networking
- minimizes misinterpretations and resulting in additional expenses
- supports the harmonization of business processes within and between research institutions
- increases data quality and thus significantly strengthens transparency and comparability between research institutions
- reduces the administrative effort
- enables the development of systems that work with these structures

6 Conclusion

Research information systems provide information about research activities and their results. This information is essential for insight knowledge of research activities, for the evaluation of research performance and for well-informed decision-making. But as we stated above, the utility and the perceived usefulness of a RIS is related to the quality of ingested and processed data. Data quality problems in research information can lead to wrong decisions and in some cases even jeopardize the existence of the institutions. The workflow presented in our paper offers institutions the opportunity to assess and improve the quality of research information prior to integration in RIS.

In this context, the implementation of the Research Core Dataset standard can contribute to better data quality and better data processing. The RCD specification describes data grouping by content and formats with a flexible model. The standard can be implemented as part of the RIS integration layer (input) as well as part of the presentation layer (export). The RCD contributes as well to the uniform definition and structuring of research information as for ensuring the data quality and comparability of the data. In our paper, we focus on integrating RCD into the RIS because universities and research institutions integrate many different heterogeneous data sources into RIS. Collecting and integrating data is a key task in the decision-making process; increasing the data quality ahead of further data processing is a top priority for RIS data management.

As we mentioned at the beginning, this paper is part of a larger research project and should be considered in the context of other related topics, like user acceptance, text and data mining approaches and other, more traditional data quality management tools. Regarding RCD, we would suggest at least three further research perspectives:

1. Data export: We put the focus on the input, ahead of the RIS data processing, as this appears to be a priority task for RIS data management. However, additional attention should be paid to the potential usefulness of the RCD standard downstream, as part of the RIS presentation layer and the data export.
2. Comparative studies: We highlighted the potential interest of the RCD for non-German speaking countries and institutions. Other countries have developed similar terminological standards, and it may be helpful for further development, harmonization and interoperability to identify other national standards and to compare them with the German RCD.
3. Use cases: This paper provides general information about the RCD and its potential advantages for data quality management in RIS. Academic and research institutions start to make use of the RCD, and it would be interesting, in one year or two, to conduct a survey on the experience with and the impact of the implementation of the RCD in different institutional and technological environments.

Further investigation should also consider new approaches to data quality issues in the larger context of big data [18]. In particular, a recent study on data collaboratives [22] may be helpful for a better understanding and assessment of RIS data quality insofar it provides a taxonomy of some main specific coordination problems, such as matching potential data providers and data users,

maintaining control over the data and its unforeseen uses or matching a particular problem with the attributes of the data, and describes potential mechanisms to deal with these problems in a given organizational environment, including process standardization, transfer of knowledge and negotiation.

References

1. Apel, D.; Behme, W.; Eberlein, R.; Merighi, C.: Datenqualität erfolgreich steuern. Praxislösungen für Business Intelligence-Projekte, 3., überarbeitete und erweiterte Auflage, dpunkt. verlag, (2015).
2. Azeroual, O.; Saake, G.; Wastl, J.: Data measurement in research information systems: metrics for the evaluation of data quality. Scientometrics, vol 115(3), pp. 1271–1290, April (2018).
3. Azeroual, O.; Schöpfel, J.: Quality issues of CRIS data: An exploratory investigation with universities from twelve countries. Publications, vol 7(1), pp.1–18, February (2019).
4. Biesenbender, S.; Hornbostel, S.: The Research Core Dataset for the German science system: challenges, processes and principles of a contested standardization project. Scientometrics, vol 106, pp. 837–847, February (2016).
5. Biesenbender, S.; Hornbostel, S.: The Research Core Dataset for the German science system: developing standards for an integrated management of research information. Scientometrics, vol 108, pp. 401–412, July (2016).
6. Deming, W. E.: Out of the Crisis. MIT Press, Cambridge, MA: MIT Press, (1982).
7. English, L. P.: Improving data warehouse and business information quality: Methods for reducing costs and increasing profits. New York, NY, USA: John Wiley & Sons, Inc., (1999).
8. 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).
9. German Council of Science and Humanities: Recommendations on a research core dataset (Drs. 2855-13), Berlin, Germany, January (2013).
10. German Council of Science and Humanities: Recommendations on a research core dataset (Drs. 5066-16), Berlin, Germany, February (2016).
11. Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M.: Daten- und Informationsqualität. Auf dem Weg zur Information Excellence. 3., erweiterte Auflage. Springer Vieweg, Wiesbaden, (2015).
12. Hovorka, D.; Larsen, K.; Monarchi, D.: Conceptual convergences: Positioning information systems among the business disciplines. In Proceedings of the 17th European Conference on Information Systems (ECIS 2009), Verona, Italy, January (2009).
13. Ivanovic, D.; Surla, D.; Racković, M.: A CERIF data model extension for evaluation and quantitative expression of scientific research results. Scientometrics, vol 86(1), pp. 155–172, January (2011).
14. Jörg, B. CERIF: The common European research information format model. Data Science Journal, vol 9(1), pp. 24–31, July (2010).
15. Jörg, B.; Höllrigl, T.; Baker, D.: Harmonising and formalising research administration profiles CASRAI/CERIF. In: CRIS2014: 12th International Conference on Current Research Information Systems, Rome, May 13–15, (2014).
16. Juran, J.; Goferey, A. B.: Juran's Quality Handbook. 5th ed. New York: McGraw-Hill, (1999).
17. Krcmar, K.: Informationsmanagement. Springer, Gabler, (2015).
18. Lytras, M. D.; Visvizi, A.: Big data research for social science and social impact. Sustainability, vol 12(1), 180, (2020).

19. Madnick, S.; Wang, R. Y.: Introduction to total data quality management (TDQM) research program. TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management, (1992).
20. Madnick, S.E.; Wang, R.Y.; Lee, Y.W.; Zhu, H.: Overview and framework for data and information quality research. ACM Journal of Information and Data Quality, vol 1(1), pp. 1–22, June (2009).
21. Naumann F.; Rolker C.: Assessment methods for information quality criteria. In Proceedings of the 15th International Conference on Information Quality, Cambridge, MA: MIT Press, October (2000).
22. Susha, I.; Janssen, M.; Verhulst, S.: Data collaboratives as "bazaars"? A review of coordination problems and mechanisms to match demand for data with supply. Transforming Government: People, Process and Policy, vol 11(1), pp. 157–172, March (2017).
23. Wang, R.Y.; Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), pp. 5–33, December (2015).
24. Würthele, V.G.: Datenqualitätsmetrik für Informationsprozesse. Norderstedt, (2003).