Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries

Otmane Azeroual, Joachim Schöpfel

publications

MDPI

Article

Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries

Otmane Azeroual [1,2,3,*] and Joachim Schöpfel [4]

[1] German Center for Higher Education Research and Science Studies (DZHW), 10117 Berlin, Germany
[2] Institute for Technical and Business Information Systems—Database Research Group, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany
[3] Department of Computer Science and Engineering, University of Applied Science—HTW Berlin, 12459 Berlin, Germany
[4] GERiiCO Laboratory, University of Lille, 59653 Villeneuve-d'Ascq, France; Joachim.schopfel@univ-lille.fr
* Correspondence: Azeroual@dzhw.eu

Received: 26 November 2018; Accepted: 19 February 2019; Published: 22 February 2019

Abstract: Collecting, integrating, storing and analyzing data in a database system is nothing new in itself. To introduce a current research information system (CRIS) means that scientific institutions must provide the required information on their research activities and research results at a high quality. A one-time cleanup is not sufficient; data must be continuously curated and maintained. Some data errors (such as missing values, spelling errors, inaccurate data, incorrect formatting, inconsistencies, etc.) can be traced across different data sources and are difficult to find. Small mistakes can make data unusable, and corrupted data can have serious consequences. The sooner quality issues are identified and remedied, the better. For this reason, new techniques and methods of data cleansing and data monitoring are required to ensure data quality and its measurability in the long term. This paper examines data quality issues in current research information systems and introduces new techniques and methods of data cleansing and data monitoring with which organizations can guarantee the quality of their data.

Keywords: current research information systems (CRIS); research information systems (RIS); research information management systems (RIMS); research information; data quality issues; data quality perception; quality management task; quality enhancement methods

1. Introduction

Through the use of current research information systems (CRIS)[1], scientific institutions can provide a current overview of their research activities, collect, process and manage information about their scientific activities, projects and publications, as well as integrate them into their web presence. As an integrated system, the CRIS interweaves the individual sources of data and workflows between science and administration, and is thus not only an information provider, but also a working tool. For scientists, a CRIS can be helpful for the management of research activities and outputs, such as publications, research projects, lectures, prices, etc. Also, CRIS should avoid unnecessary work and reduce the effort involved in preparing reports or in presenting research performance and scientific expertise.

[1] The nomenclature for research information systems is more or less unstandardized, including RIMS (Research Information Management System), RIS (Research Information System), RNS (Research Networking System), RPS (Research Profiling System) or FAR (Faculty Activity Reporting). In this paper, the preferred term is CRIS (Current Research Information System), because it is widely used in European countries, i.e., the main context of the survey.

Data quality plays an important role not only in terms of interpretation and usability of the data, but also when evaluating external data sources; data quality is crucial for the perceived utility and the acceptance of implemented systems. Growing volumes of data and source systems are increasingly becoming a serious problem for institutions. Particularly during the collection, transmission and integration of research information in the CRIS, different data errors can arise that can have a variety of negative effects on the data quality. Therefore, a one-time clean-up of the data sources is not enough; data must be maintained continuously.

To better understand the importance of data quality, this paper provides an overview of former studies on data quality and quality issues in CRIS and presents the results of an exploratory investigation with a small panel of CRIS experts from universities in different European countries. Subsequently, new techniques and methods of data cleansing and data monitoring are discussed in order to demonstrate how these data errors can be identified, remedied and improved.

2. Literature Overview

Ref [1] was the first to address data quality as an issue for CRIS. In light of Stempfhuber's paper, the notion of quality should be made more explicit to support the connection of distributed sources of research information at the level of the European infrastructure cloud. Also, he insisted on the need for a formal way to describe quality at the level of the individual data item and at the level of a CRIS as a whole, i.e., a model that "generalizes on the rich set of experiences available in the CRIS community (good or even best practices) in a way that it is directly applicable in a variety of contexts and reduces there the risk of failure and inefficiency". He suggested a strong approach to the term "purpose", a form of "total quality management" for the development of CRIS software, iterative processes with loopback cycles, a user-centered approach of personalized features, and the inclusion of quality attributes in the CERIF data model, to characterize the quality level of external data sources.

Since Stempfhuber's review of CRIS quality, more papers and case studies have been published on quality-related aspects of research information systems. Yet, his expectation of a "formal model" has not been fulfilled. Our purpose is to review the relevant literature and to add empirical evidence for further development of a formal model.

2.1. The Concept of Data Quality

In the field of databases and large information systems, data quality has been defined as "data that are fit for use by data consumers", i.e., those who use data [2]. This means that data quality is not an absolute value but a dynamic concept [3], correlated with the usage of the system and thus, with user satisfaction and system acceptance.

Studies on data quality most often assess multiple dimensions of the concept, with attributes such as accuracy, timeliness, precision, reliability, currency, completeness, relevancy, accessibility and interpretability. Emphasizing the consumer viewpoint, some attributes appear as preliminary or necessary conditions, e.g.,

- the data must be accessible to the user,
- the user must be able to interpret the data, and
- the data must be relevant to the user.

Based on empirical evidence and theoretical studies, data quality can be conceptualized as a framework with four major dimensions: intrinsic data quality (e.g., believability, accuracy, and reputation), contextual data quality (value-added, accuracy, timeliness, etc.), representational data quality (interpretability, ease of understanding, etc.) and accessibility data quality, including access security [2].

Data quality is crucial for business and security [4]. However, as English points out in his white paper on misconceptions about information quality, it is more than accuracy, data cleansing and fitness for purpose, insofar as the same data can be used for different purposes, by different data

consumers, in different contexts and at different times. It is "fitness for all purposes in the enterprise processes that require it", and quality information is data that "consistently meets knowledge workers and end-consumers expectations" [5]. Data quality should therefore be an object of constant care and attention ("quality management"), in order to improve the system performance and the user satisfaction and acceptance.

More recently, ref [6] developed a methodology for the standardization of data and services, for large computational research data infrastructures. The "fitness for (all) purpose(s)" is again in the heart of the model, as the data quality is assessed through the control of the compliance with recognized community standards, and as the quality assurance includes all common research data platforms, services and tools. Moreover, [6] highlight the importance of having a consistent data structure and metadata, two crucial aspects for the evaluation of research data in CRIS [7].

How such a "product-governed quality assurance" model can be applied to CRIS has been described by [8]. The declared goal is to achieve the highest possible data quality through reporting of errors and omissions in the data (feedback, "closed-loop process"). Each information system "addresses a specific group of users with specific interests (...) able to 'react' to any given product". In other words, a quality assurance process based on the data users' feedback would contribute to increasing the quality of data in line with the frequency with which the data are used.

2.2. Current Research Information Systems

A current research information system (CRIS) is a specialized database or federated information system to collect, manage and provide information on research activities and results, such as projects, third-party funds, patents, cooperation partners, prices and publications [9–11]. The building blocks of a CRIS architecture can be described as a three-stage structure [10] (see Figure 1).

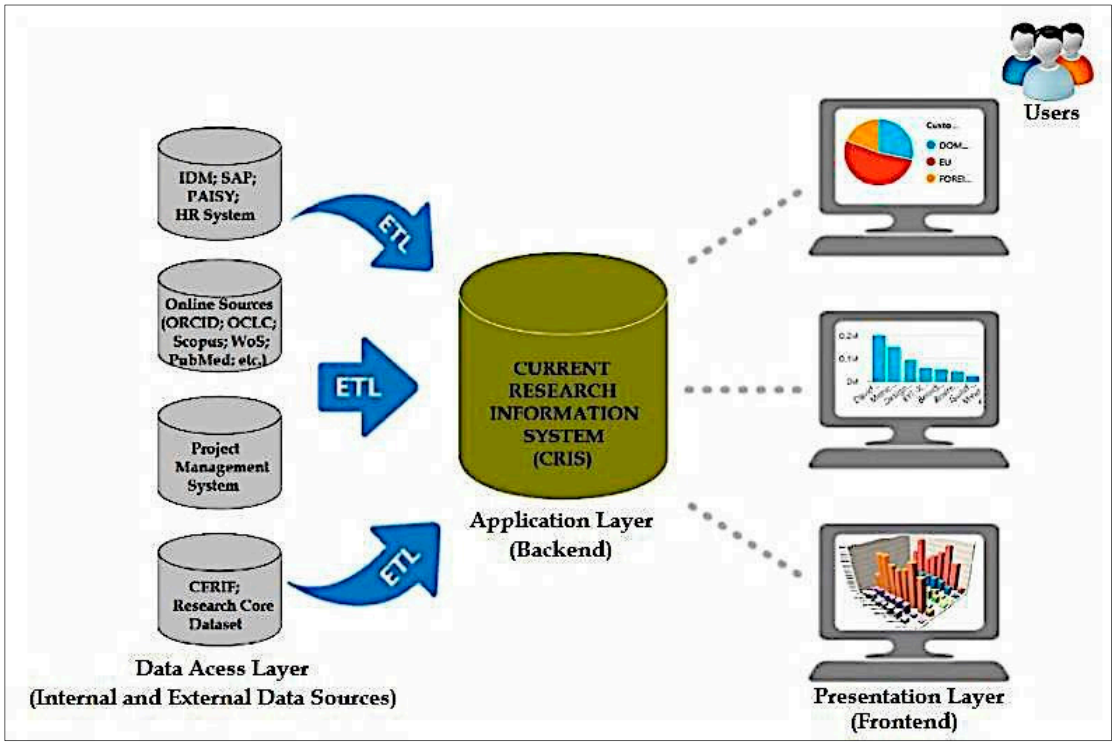

Figure 1. Architecture of CRIS.

The data access layer contains the internal and external data sources, e.g., operational databases (human resources, finance, project management . . .), open repositories, identifiers (ORCID, DOI, etc.), bibliographic data from the Web of Science, Scopus or PubMed, etc. This layer includes data

models for the standardized collection, provision and exchange of research information, such as the Research Core Dataset (RCD) and the Common European Research Information Format (CERIF). The integration of these data sources into the CRIS takes place via classical Extract, Transform and Load (ETL) processes. The application layer (backend) contains the CRIS and its applications, which merge, manage and analyze the data held at the underlying level. The presentation layer (frontend) shows the target group-specific preparation and presentation of the analysis results for the user, which are made available in the form of reports using business intelligence tools, via portals, websites, etc. (for more details see the papers from [10–13]).

2.3. Data Quality Issues in CRIS

Good-quality CRIS data are a "crucial foundation of any successful monitoring and evaluation strategy" [14]. The data must be trustworthy and "fit for purpose" [1]. Without a minimum level of reliability and accuracy of information on persons, organizations, projects and results, a CRIS will be virtually useless for research management and science policy. Ref [15] underscores another challenge, in the context of the Finnish academic CRIS: "The quality of the data has great importance, as 13% of all state funding for the universities (more than 200 million euros a year) is distributed based on the number and quality of the publications". Data quality thus has a tangible financial impact, for the scientific authority as well as for the individual researcher.

A major problem for CRIS quality assurance is the variety and complexity of relevant information (persons, institutions, projects, publications, patents, facilities, etc.) and the large diversity of its sources and providers (internal and external databases, directories and registries of research projects, programs, results, calls, events, budgets, human resources, etc.), all with their own formats and standards which can vary over time. In the words of [16], the main problem to be solved is as follows: "users must be supplied with heterogeneous data from different sources, modalities and content analysis processes via a visual user interface without inconsistencies in content analysis, for example, seriously impairing the quality of the search results".

Thus, completeness of data, i.e., the presence of data about all entities which are subject of a given CRIS [17], is one of the main parameters of data quality, along with correctness, consistency and timeliness.

Obtaining data relevant for management requirements is the first step in the CRIS process and is one of the core tasks. First of all, the relevant research information of a university is collected from the many different data stocks of the individual IT systems via standardized interfaces based on standardized data models and able to deal with different formats. The efficiency and effectiveness of the data sources and the information they contain are of central importance. It is based on internal and external data from various operational systems, which are distributed across the entire research information systems of the universities, such as: system applications products (SAP), human resources system (HRS), financial accounting system (FAS), identity management system (IDM), institutional repositories, etc.

In addition to the internal sources, external sources, such as other publication platforms, commercial databases, research data repositories, etc., are also used for data collection. This results in a large variety of data types that must be further processed by the CRIS. The quality of the source data has a direct influence on the quality of the CRIS. At each university, research information is entered and recorded into the CRIS, so the processing and management of these data must be of good quality so that users can get good results [9,11].

The problem is even more complicated when it comes to national or international research information systems. Although specific scientific, political and institutional environments may be different, the challenges of data quality remain more or less similar, as different studies have shown from the Czech Republic [18], Finland [15], Norway [19], Sweden [20], Poland [21] and the United Kingdom [22]. Ref [14] provide some insight on quality management in the CRIS of the European Research Council, on a supranational level.

The different studies mention different ways for the control, monitoring and improvement of data quality, e.g.,

- a highly structured data model (cf. CERIF),
- the comprehensiveness of the data model
- definition and standardization of a core dataset,
- definition and standardization of terminology and dictionaries,
- mapping of data from different sources,
- a batch-oriented data collection system with controls before data are entered into CRIS which helps to improve, in particular, data completeness,
- generation of quality assurance reports on different levels,
- usage of unique identifiers for persons and organizations,
- semantic enrichment of data by automatic indexing and classification of objects,
- an iterative process for quality enhancement,
- availability of data through an open web interface so that researchers can correct errors involving their own data,
- allowing dynamic links to be added from local web pages to the CRIS, thus encouraging researchers to register more errors and to assist in improving the quality of the data,
- crowdsourcing to improve linkages between publications and CRIS records,
- text mining to help efforts for the attribution of articles to authors or their linking to a project and to enhance the completeness of data,

All data providing systems have their own, independent data models, describing publications, projects, institutions, people etc., and their curation and maintenance need domain-specific knowledge which must be mobilized for the processing and curation of those data that are exported and ingested into the CRIS.

Standardization is one major solution to reduce the impact of heterogeneity and inconsistency. Traditional methods of standardization "appear indispensable and substantially improve quality and cost-effectiveness in subsets" but, as [16] notes, "they can still only be partially implemented, with increasing costs, within the framework of global provider structures and changed general conditions" and should be considered and revised from the aspect of the remaining heterogeneity. Therefore, standardization must be accompanied by automatic and intelligent treatment of data inconsistency.

The need for globally unique identifiers is a specific case of the standardization issue. "In the research world, the global identifier gap has been recognized being critical for quality improvements in information systems and at the same time to enable large-scale information sharing or reuse" [23]. Among the relevant unique identifiers are ORCID, ResearcherID, handle, DOI and URI (see also [24]) and there are many national or other identifiers for institutions and authors. However, as [23] add, even if the idea of globally unique identifiers is exciting, "governance and security issues must not be neglected", such as those issues which may affect the willingness of scientists and organizations "to become a number". Other approaches to reduce the heterogeneity of data and increase the systems' interoperability are standardization of metadata formats[2] and terminology[3].

Another approach to further standardization has been presented in the German context, i.e., the definition of a core dataset [25]. This approach may be helpful in ensuring a level of "minimum interoperability" with clearly defined data requirements, but will reduce the range of purposes for which the data are fit and thus, impact the data quality.

The specific situation of migrating to a new CRIS has been addressed by [26], who describe how the process "unearthed massive data quality issues" and admit that "even after a year and a half of

---

[2] Mainly based on the Dublin Core and bibliographic standards.
[3] See, for instance, the work of CASRAI and euroCRIS.

cleaning, much work remains to be done and some parts of the dataset may never be fully improved". Their conclusion is that careful assessment of the problems, decentralizing a certain amount of the actual cleaning, commodifying data quality improvement operations and the automation of some of the tasks involved can reduce the costs and improve the efficiency of data quality management.

The preparation of a research information system with specific quality procedures may also encourage the development or improvement of external data pools, with bibliographic data, author registers (institutional and personal data, etc.), research data, etc., or the development of efficient data identification and acquisition tools such as web mining tools aimed at discovering knowledge about conferences. This can include additional training for the operational staff of internal data sources, in order to improve data input and system management. Regarding research data, related metadata may contain useful information about data quality (procedures, measurements used to collect data, etc.). Ref [27] describe how the integration of a CRIS into software for higher education administration may contribute to data quality (consistency, completeness) and availability.

Last but not of least importance, the human and community factor should not be neglected as "the high level of expertise the makers of CRISs are able to contribute during the development of these systems" [1] is essential for the implementation of quality assurance and control procedures. The recent EUNIS and euroCRIS joint survey on research information systems and institutional repositories reveals that at least in Europe the libraries and research and innovation departments are the two main institutional services responsible for CRIS data quality [24].

Specific functionalities can be helpful to assist researchers in controlling and correcting their own data. Ref [28] presents a crowdsourcing model for research information systems to validate or correct automatic citation links to publications.

Based on his literature overview, ref [1] described four main areas of CRIS quality, i.e., information quality (or data quality), data integration (networking), quality as a process and personalization. For our own study, we define data quality as the suitability of the data for use in certain required usage objectives, these must be error-free, complete, correct, relevant and consistent [12]. Requirements can be set by different stakeholders, in the CRIS context, e.g., especially by different CRIS user groups, but also by the CRIS administrator.

3. Methodology

To obtain additional empirical evidence on data quality management in research information systems, we conducted a survey with academic institutions which are implementing or already running a CRIS. The intent of the study was not a large-scale survey with a representative sample, but to conduct a quantitative, exploratory investigation with a small international sample of recognized experts, i.e., a preliminary study to prepare a larger and representative study with a large sample of nearly 250 German universities and research organizations.

CRIS experts affiliated to universities were identified from papers and case studies presented at euroCRIS meetings and conferences during the last five years[4]. Thus, contact was made by email with CRIS experts from 30 universities in 14 countries, inviting them to answer a short, anonymous online-based questionnaire with nine questions. After up to three reminders over a period of two months, we obtained 17 completed questionnaires from 12 mainly European countries (see Table 1).

[4] euroCRIS, founded in 2002, is an international not-for-profit association, that brings together experts on research information in general and research information systems (CRIS) in particular, see https://www.eurocris.org/.

Table 1. Survey sample (*N* = 17).

Country	Number of Respondents
Netherland	3
United Kingdom	3
Finland	2
Austria	1
Belgium	1
Norway	1
Russia	1
Serbia	1
Singapore	1
Slovakia	1
Sweden	1
Switzerland	1

The academic institutions that responded to our survey run different research information systems, including Pure (Elsevier), Converis (Clarivate Analytics), METIS (Radboud University Nijmegen), Elements (Symplectic), CRIStin (Norwegian Directorate for ICT and Joint Services in Higher Education and Research) and SoleCRIS (University of Tampere). Some of which are commercial software, others are public and institutional (local) solutions, including open source solutions, such as DSpaceCRIS or VIVO.

4. Results

When examining the concept of data quality in CRIS, the following results are found from the survey with the institutions and CRIS experts. First of all, a distinction must be made between design quality and execution quality. Design quality concretizes in the specification of standards, data definitions and documentation. With regard to execution quality, the focus is on the data contents and the data values. Another important aspect of the data quality is the quality of the data provision, which particularly takes into account the software components of the overall system. The significance of the individual aspects is summarized in Figure 2.

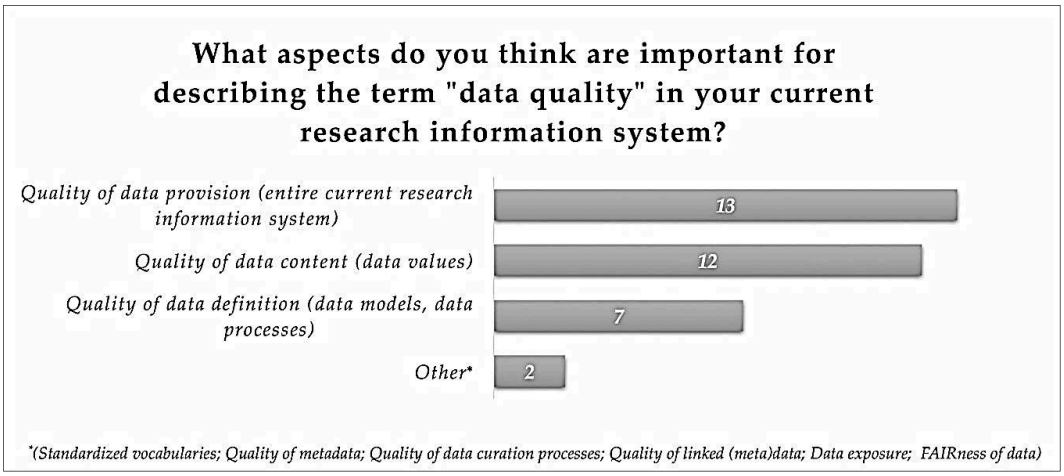

Figure 2. Aspects describing the concept of data quality in CRIS (*N* = 17).

The majority of the respondents considered the quality of data provision and the quality of data content (values) to be most important; more important than data models, processes or other aspects.

Which are the real problems the experts observe in their everyday work with their CRIS? Thirteen issues were mentioned, ranging from the most frequent problems, like spelling mistakes, multiple and

incorrect input or missing data relationships, to less frequent problems like inconsistent data formats, different data types, or wrong rules (see Figure 3).

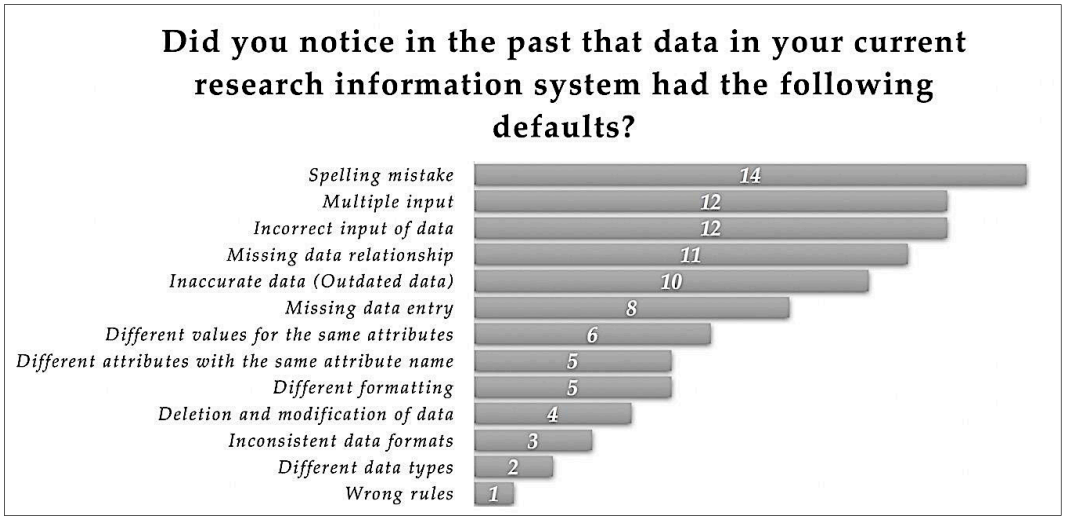

Figure 3. Data quality issues in CRIS (N = 17).

The origins of these problems are multiple, and they need a differentiated approach. How do the respondents of this investigation deal with data quality issues? How do they discover the problems? Figure 4 provides some elements about quality checks.

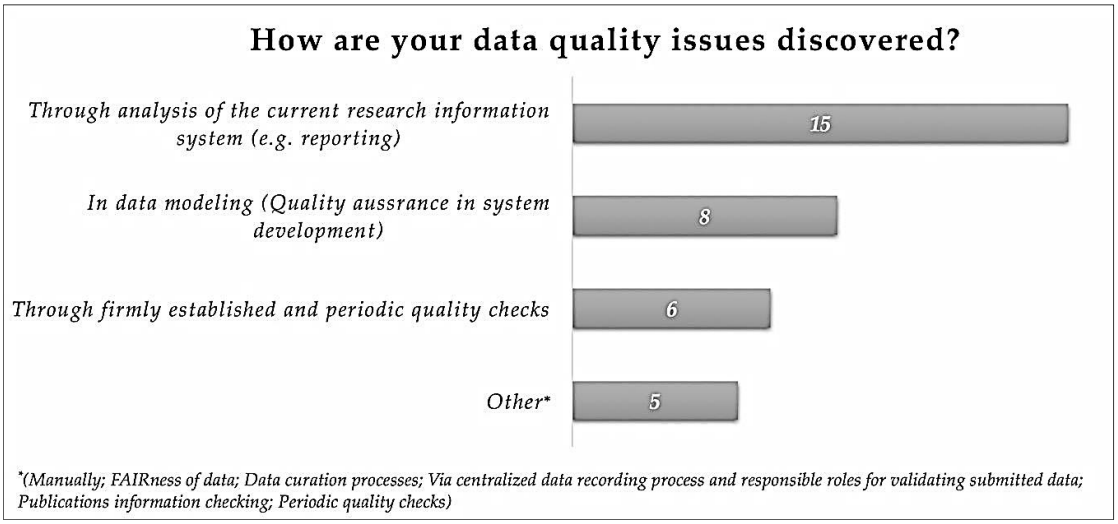

Figure 4. Discovery of data quality issues in CRIS (N = 17).

Most experts indicated that they discovered data problems through an analysis of the system, with the help of business intelligence tools, through reporting, etc. Half of the respondents identified problems via data modeling and formal quality assurance during the development of the CRIS, while one third of the experts control the quality of their data through well-established, regular quality checks. Other methods seem less important, such as manual quality checks and formalized data curation processes.

After the "how" of quality control, a complementary question was asked about the "when", i.e., the point of data processing when the quality control generally takes place. Figure 5 presents the results. Again, we can observe a main yet not unique response.

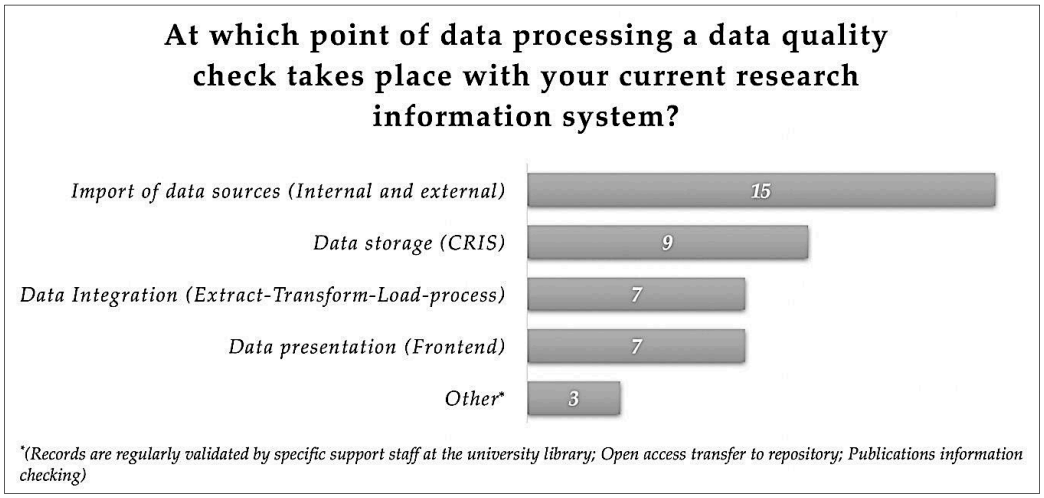

Figure 5. Point of data processing for data quality check in CRIS (*N* = 17).

Most experts indicate that their quality check takes place at the interface of the CRIS with other, internal or external systems, when ingesting data into the CRIS. Half of the experts mention three other "check points", the data storage, the data presentation (frontend) and the ETL process. Other respondents said that they control the quality of export to the open access repository, that they specifically check the control of publication records (input), and that all records are regularly validated by trained librarians, without specifying which records.

The methods of handling the data quality issues in CRIS are varied and depend on the importance of the problem, the context of the application, and the cause of the problem. The scientific literature discusses above all four dimensions which can contribute to high data quality in CRIS, i.e., completeness, correctness, consistency and timeliness, which have proven to be easy to measure, represent a particularly representative illustration of the reporting for the CRIS users and lead to an improved basis for decision-making. The sample of experts were asked which one they would use to test and measure quality in their system (see Figure 6).

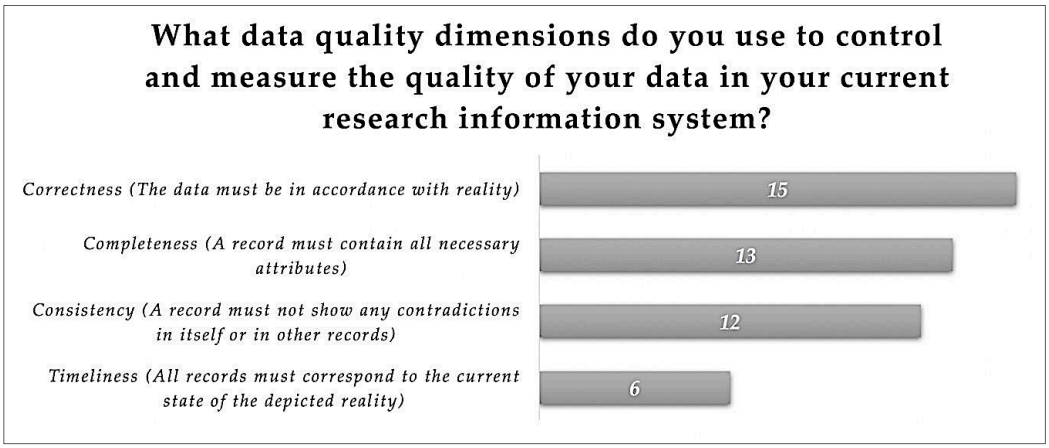

Figure 6. Data quality dimensions for checking and measuring data quality in CRIS (*N* = 17).

The respondents considered correctness, completeness and consistency as the most useful criteria to control data quality, more than timeliness. Two other questions assessed the way in which the institutions deal with quality issues in their CRIS. The first question was about the measures and methods they carry out for quality assurance and quality improvement of CRIS data (see Figure 7).

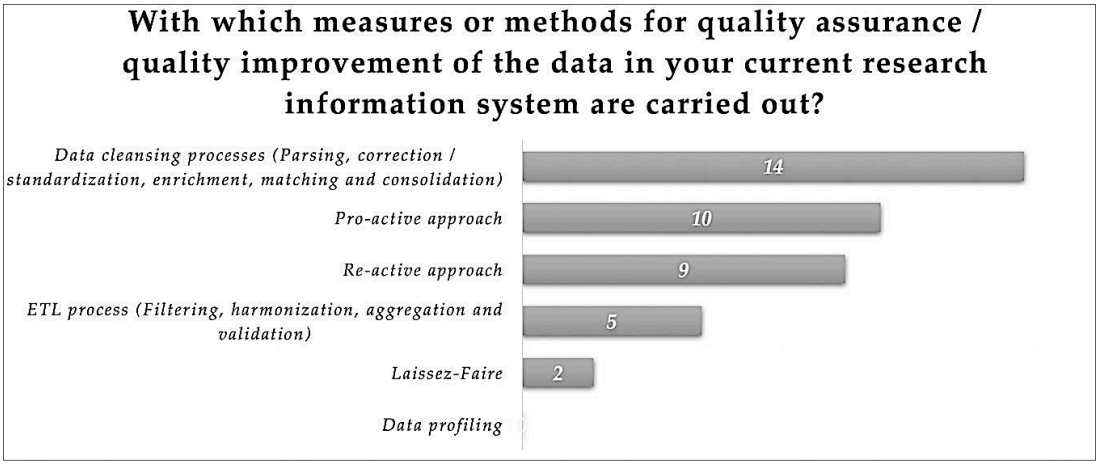

Figure 7. Measures and methods for quality assurance and quality improvement of CRIS data (*N* = 17).

According to the experts' answers, the preferred approach to quality assurance are various data cleansing processes, such as parsing, enrichment or consolidation, followed by pro-active and re-active methods. Other methods seem less important; in particular, ETL processes (filtering, etc.), laissez-faire and data profiling, but similar to quality dimensions (see Figure 6), we must keep in mind that most respondents provided two or three answers, confirming that they consider CRIS data quality to be a multidimensional concept which needs more than one approach.

The last question was about the perceived importance of securing data quality (Figure 8). While all experts stated that data quality is a problem, most of them considered that this issue is not an important problem, but only a "low problem", less important and less urgent. Nevertheless, nearly half of the respondents indicated that quality is indeed an important issue which needs awareness, attention and action.

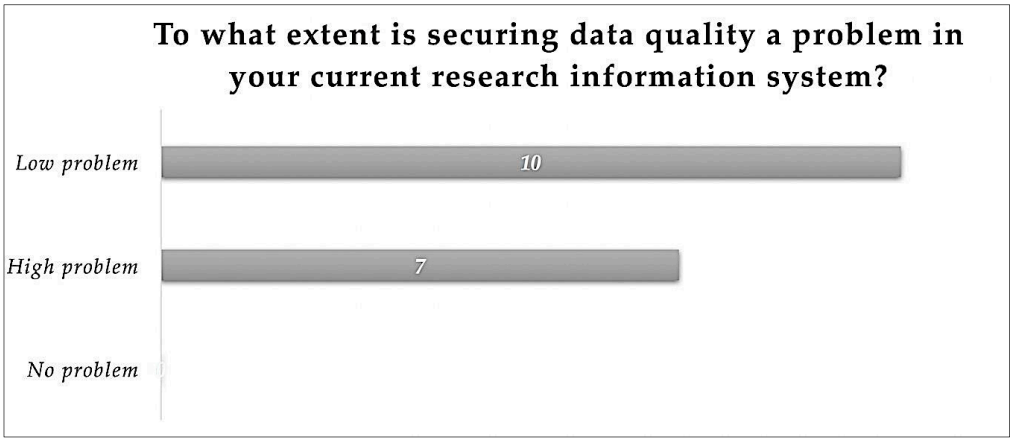

Figure 8. Securing data quality is a problem in CRIS (*N* = 17).

5. Discussion

The results from the last question may be surprising—less than half of the sample of CRIS administrators and experts consider data quality to be an important problem. This is quite different compared to former studies, where quality issues were considered as a crucial topic for CRIS development and management ([14]; see above, Section 2.3), and it probably reflects the users' satisfaction with their CRIS, their quality in terms of reliable functioning and trustworthy results, as the CRIS technology has reached a certain level of maturity. However, we must keep in mind that this may be the particular opinion of a small group of experts deeply committed to this technology, as project leaders, system administrators, etc.

Even if the small sample is not representative, some aspects can be compared to other studies, including the unpublished results of our own representative survey with German universities and research organizations [29]. For instance, the perceived importance ranking of the different data quality dimensions is quite similar in both studies, with more attention paid to data correctness and completeness than to consistency and timeliness. Also, the preferred approaches to quality management were ranked in a similar way, with data cleansing and pro-active approaches on the first places. However, the German sample ranked data profiling higher than re-active and ETL approaches; but without a larger investigation, it is not possible to speak about potential "international differences" between quality management, and we did not find any evidence about this in other published studies.

In the following, we will discuss two aspects which need further attention, data cleansing and data monitoring.

5.1. Data Cleansing

Due to the collection, transmission and integration of different internal and external data sources of the universities and research institutes in current research information systems, data quality issues, as mentioned above, have to be overcome. An important step can be data cleansing or data scrubbing, which corrects the existing data to provide a good overview of the CRIS data. Data cleansing involves various methods or procedures for removing or correcting data errors in current research information systems. For example, the errors may include incorrect, redundant, incomplete, inconsistent or incorrectly formatted data.

Data cleansing is done in five consecutive processes (see Figure 9) and is critical for data quality in current research information systems [9,11].

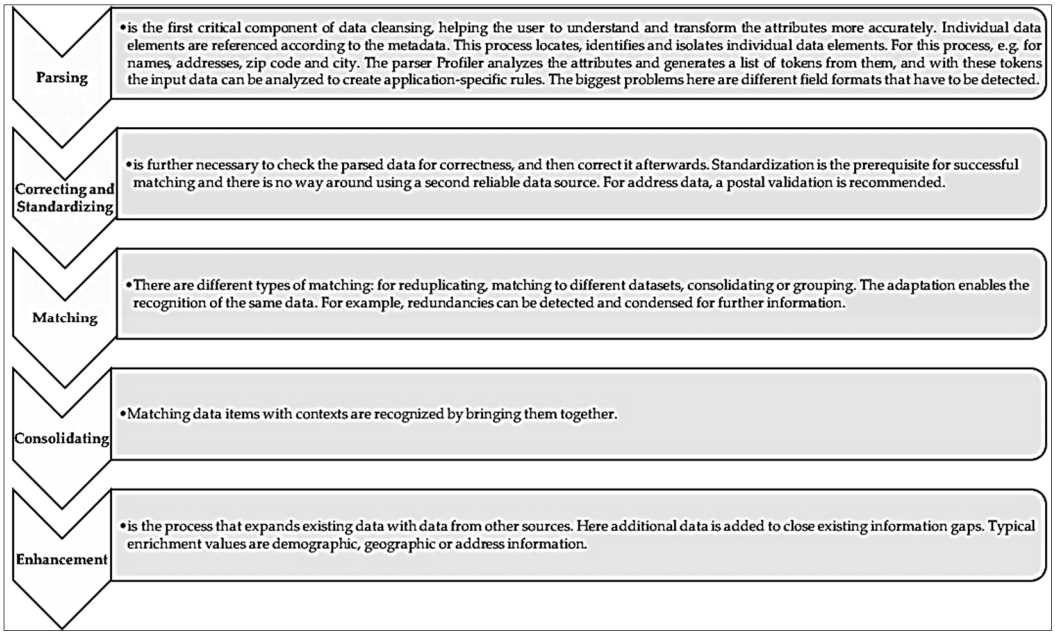

Figure 9. Data cleansing processes.

All these mentioned steps are essential for achieving and maintaining maximum data quality in current research information systems [9,11]. Errors in the collection, transmission and integration of multiple internal and external data sources in a CRIS are eliminated by the data cleansing [9,11].

Data cleansing creates optimal data quality in terms of completeness, correctness, consistency and timeliness. With this, the data cleansing process lays the foundation for content of higher information value and better-quality analysis of research information in an organization. Figure 10, below, illustrates

an example of identifying records with bad names in a publication list to show how the data cleansing process can improve the quality of CRIS data.

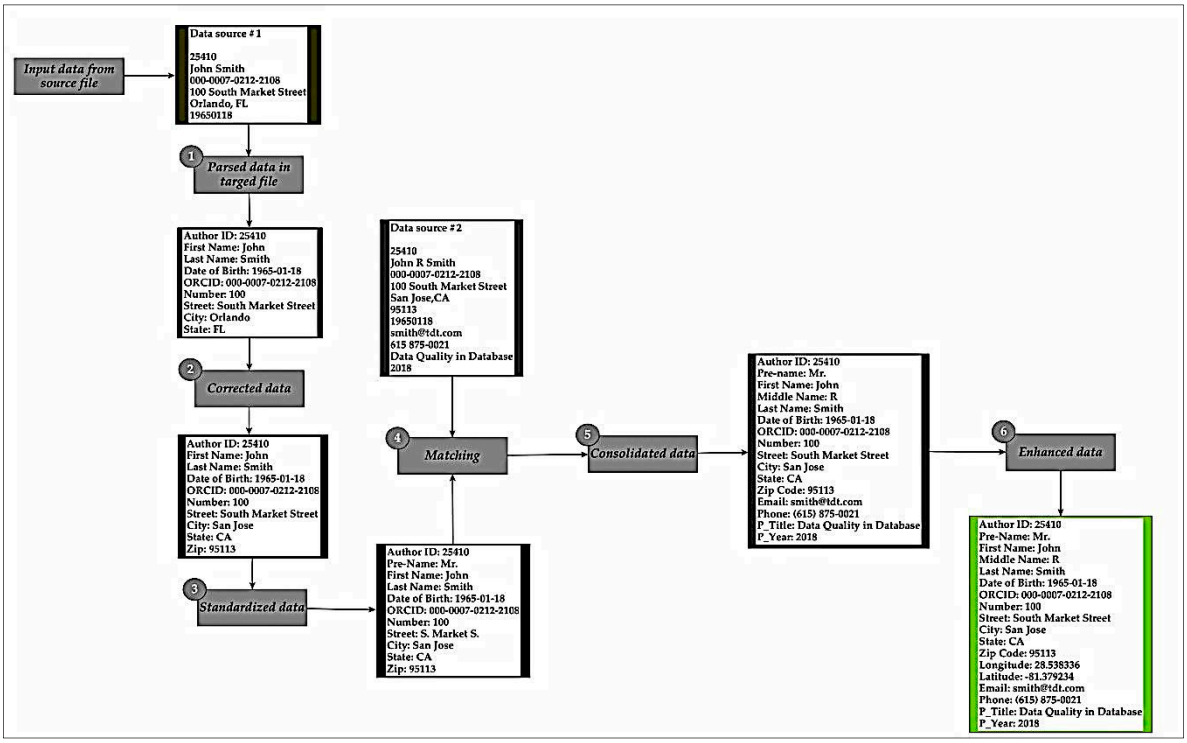

Figure 10. Examples of data cleansing processes in the context of CRIS.

One special question is raised by quality standards. So far, the current quality standards of CRIS appear to be shaped by communities and conditioned by the expected outcomes. Do CRIS report research activities and outcomes in a reliable and consistent way? Are these reports compliant with laws (intellectual property, public research, ethics, etc.) and with funding bodies' criteria, such as the distinction between public and private research, different types of communication (open access, cf. plan S), etc.? In particular, research management expects 100% reliable information, and errors have to be fixed "manually", because of their potential negative effect on evaluation. Regarding the CERIF data model, this means the data monitoring must focus especially on the fundamental entities of the research environment (project, person, organization), the entities representing research results (publication, patent, product) and their relationships, more than on other entities like infrastructures, etc.

5.2. Data Monitoring

After identifying an error, it is not enough to just clean it up, which can often entail a time-consuming process. To avoid the recurrence of errors, the aim should be to find the causes of these errors, resolve them, and undertake continuous monitoring. Only a continuous review of data quality ensures that high-quality data remains complete, current and consistent [30]. Typical usage scenarios of data monitoring include reviewing new data, re-evaluating data quality metrics, or permanently validating business rules in the system.

Data monitoring describes a proactive approach that enables early detection of data quality issues and continuous collection and monitoring of data quality metrics. Through this ongoing data monitoring, the facility is able to provide information about its data quality status at any time, as well as increasing the confidence in its existing data. The cycle of data quality monitoring in scientific institutions can be described in three steps (see Figure 11).

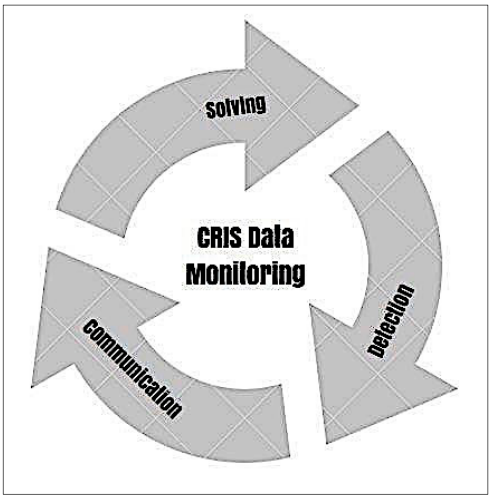

Figure 11. CRIS data monitoring cycle.

The first step is to detect data quality issues as part of ongoing data quality monitoring. The second step is to communicate the issue to all stakeholders who need this information to perform their tasks. The final step is solving the detected data quality problem. This, in turn, is followed by ongoing monitoring to reflect progress in the implementation and effectiveness of the measures.

Data monitoring is certainly an effective method, so it is advisable to establish a permanent data quality team in universities or research institutes that can coordinate and monitor the individual measures of the current data quality project and systematically identify errors through reproducible data analyzes [31]. Figure 12 illustrates a process flow for managing and monitoring problem records in the context of CRIS.

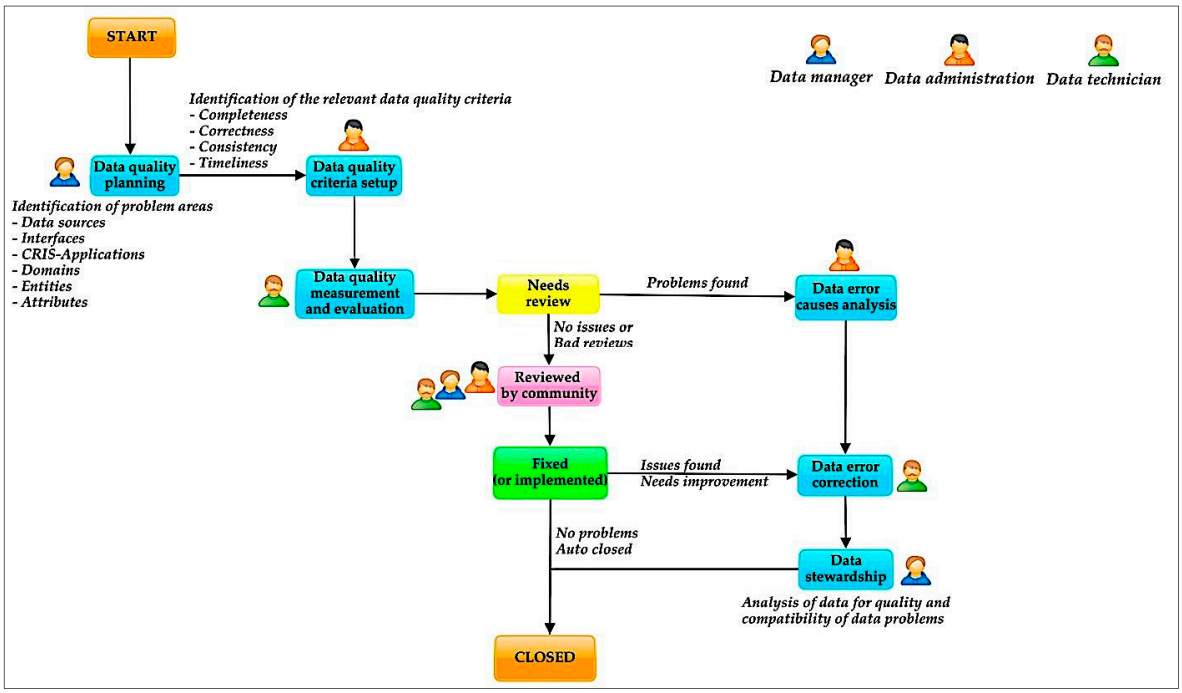

Figure 12. Process flow for managing and monitoring problem records in the context of CRIS.

To monitor the quality of the data in CRIS, the following developed iterative process flow (see Figure 13) can be used as a basis for the institutions using CRIS and should serve as a guide to demonstrate how to analyze, detect, fix, and improve data quality issues in CRIS in the institutions.

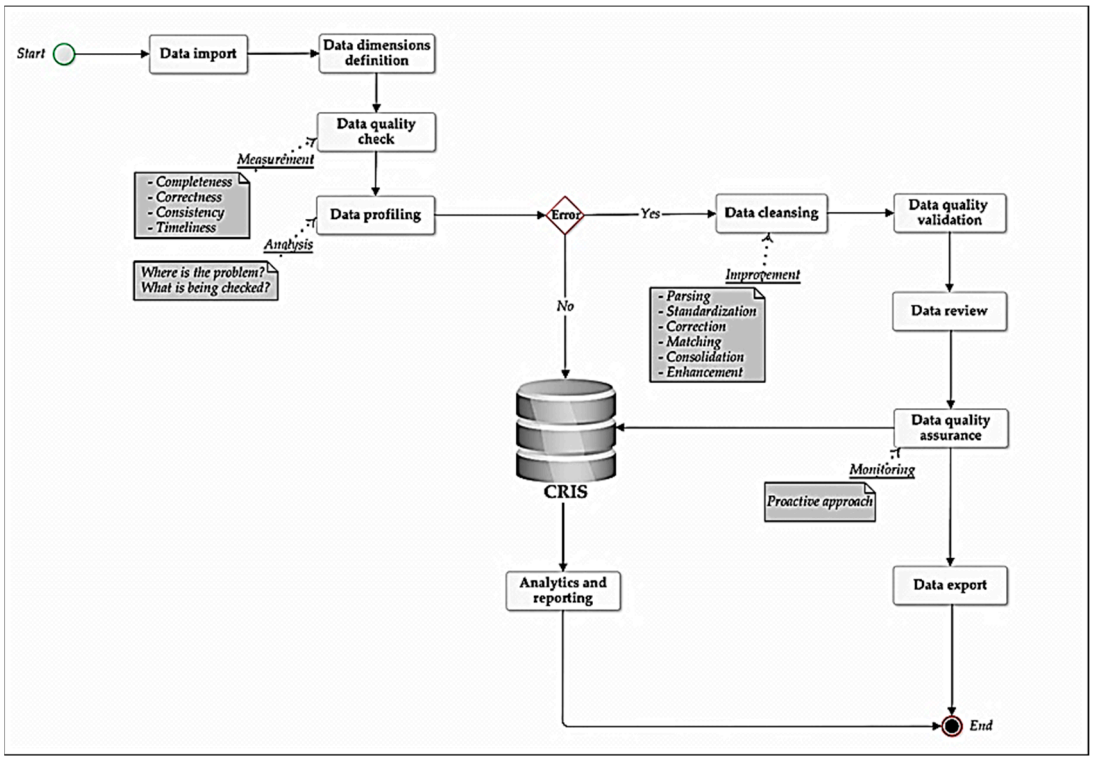

Figure 13. Iterative process flow for improving and monitoring the quality of CRIS data.

Figures 12 and 13, as iterative processes, provide a permanent assurance of data quality in current research information systems, as erroneous data in a collection are a fundamental challenge for managers and IT, quality assurance and many other fields (see also [9,12,13]). To accomplish the most difficult steps, such as analyzing and correcting data errors in the two processes, the data quality team should begin by agreeing on where the problems are greatest and the impact of the missing or erroneous data. For this purpose, the inventories are analyzed and adjusted with a data profiling and data cleansing tool to determine incomplete, incorrect or duplicated research information. Data stewards play a decisive role in the success of a data quality offensive. They come from a specialist department and know the processes and challenges in the creation and use of research information. A data steward sets rules on how data are generated, maintained, and deployed in the departments. Stewards are also responsible for the current monitoring and compliance of data integrity and the continuous adaptation of quality procedures. Importantly, data stewards measure and monitor ongoing data quality improvements. In parallel, institutions should develop a culture in which high data quality plays a significant role. Not only the data steward and the IT department, but also employees from the specialist departments need to be aware of the importance of reliable information and how to maintain and optimize the data.

We mentioned above the question as to data quality standards in the field of CRIS. Speaking about data cleansing and monitoring necessarily raises another, fundamental question about data quality. Following the usual definitions and models, data quality means "fitness for use by data consumers", a dynamic and multidimensional concept, correlated with user satisfaction and system acceptance, and conditioned by accessibility, interpretability and relevance for the user or "data consumer". Yet, the problem with CRIS is that they mobilize different user groups, with different standards, expectations, needs, skills and practices. The present investigation was conducted with project leaders and system administrators, committed to CRIS and involved in the (mainly European) euroCRIS network, with papers, communications, working groups etc. However, a CRIS community in a given institution is more heterogeneous, made up of research managers, project leaders, data officers, administrative staff,

librarians, people from the IT department, etc., and, last but not least, by the scientists themselves, with their own divisions (disciplines, infrastructures, early career scientists vs. tenured professors, etc.).

It seems difficult, therefore, to define simple data quality standards or criteria. A system administrator may be satisfied with "her/his" CRIS performance and willing to recommend it to other institutions, while other users, e.g., the technical staff or the involved librarians may be more critical regarding the system's ease of use; the focus of research managers will be on the output (reporting) quality, compliant (or not) with the requirements of authorities and funding bodies while scientists may pay more attention to the quality of the information about their output, projects, awards, affiliations, etc., including their identity and the spelling of their name. Research on CRIS often lacks a differential approach to needs, usage and information literacy of specific user groups; at least, the link is missing with other surveys and studies on the data behavior and literacy of different communities.

This situation excludes simple and single approaches to data quality management. The multiple user groups define the data "fitness for use" in their own way and in relation to their own and specific practice and needs. Also, a global model of data quality management is required, covering the whole process (and even the data models and sources outside of the CRIS) and including the whole range of methods, techniques and tools of data quality management. To raise awareness on this situation and to contribute to an appropriate, global approach is the main objective of this paper.

6. Conclusions

The ultimate goal of research information systems is "providing the best evidence base on which to support high quality decision making" [32]. High data quality is one condition necessary to attaining this goal.

The larger the data sets become by electronic data processing, the more important good data quality and information quality becomes. The quality of the data is becoming increasingly important in a wide range of institutions. Data that have been modeled for a specific operational use and have a high level of heterogeneity needs to be adjusted and controlled in order to be used as the basis for such analyzes in the CRIS context. Basically, the data errors in CRIS should be resolved through quality assurance as early as possible in the data flow. These typical quality errors are caused by spelling errors or incorrect and duplicate entry in the data. The new techniques and methods of data cleansing and data monitoring presented in this paper can eliminate these data quality errors within the CRIS and optimize the entire business process in CRIS. Since the data are constantly changing and a one-time cleanup is not sufficient, data must be maintained permanently, and their quality continuously monitored. Such continuous improvement and enhancement of data quality in CRIS requires specific measures such as "pro-active approach" [9,11].

When it comes to implementing data cleansing and data monitoring, open source tools (such as *Quadient DataCleaner*[5]) help to identify, improve and monitor data errors (like spelling mistakes, inconsistencies and other noticeable problems, etc.) from large amounts of data that can be adjusted together with the responsible CRIS employee in the specialist departments. The two proposed and practically used software tools work completely autonomously and hypothesis-free, so they do not need manually specified filters or search criteria. In addition, quantitative analysis provides well-founded corrective guidance, allowing for effective, automated, or at least software-supported, resolution of data errors.

Based on the presented results, a larger and representative survey has been conducted with German universities and research organizations. Further research is needed for a better understanding and modelling of the relationships between quality, satisfaction, acceptance and perceived usefulness in different environments, with different user groups and with different solutions. More detailed insight is needed, too, regarding the appropriateness of specific tools and techniques in different

[5] https://datacleaner.org/.

environments and regarding different data errors. What is the impact (feedback loop) on the data models and quality standards of internal and external data sources providing input for the CRIS? Can the merging of CRIS and other systems (e.g., institutional repositories) offer solutions for the data quality management? What about the outsourcing of CRIS and the partial or total transfer of data processing and quality management in the cloud? What is the impact on problems with quality issues, acceptance and satisfaction? The CRIS landscape is rapidly evolving, with new technologies and requirements and with a steadily increasing number of CRIS projects. More research and empirical evidence are needed for better solutions, improved project management and appropriate user training and assistance.

Author Contributions: Both O.A and J.S authors jointly contributed to the design and implementation of the study, to the analysis of the results and to the writing of the manuscript.

Funding: The research was funded by the German Center for Higher Education Research and Science Studies (DZHW) and by the German Federal Ministry of Education and Research (BMBF) and the 16 Länder governments in the context of the project "Helpdesk to facilitate implementation of the Research Core Dataset"[6]. One part of the work received funding from the European Institute of Social Sciences and Humanities (MESHS Lille) and from the Regional Council (Conseil Régional Hauts-de-France), as part of the research project "D4Humanities"[7].

Acknowledgments: The authors wish to express their thanks to all respondents and their institutions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stempfhuber, M. Information quality in the context of CRIS and CERIF. In Proceedings of the 9th International Conference on Current Research Information Systems (CRIS2008), Maribor, Slovenia, 5–7 June 2008.
2. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
3. Miller, H. The multiple dimensions of information quality. *Inf. Syst. Manag.* **1996**, *13*, 79–82. [CrossRef]
4. English, L.P. *Information Quality Applied: Best Practices for Improving Business Information, Processes, and Systems*; Wiley: Hoboken, NJ, USA, 2009.
5. English, L.P. *Seven Deadly Misconceptions about Information Quality*; Information Impact International, Inc.: Brentwood, TN, USA, 1999.
6. Evans, B.; Druken, K.; Wang, J.; Yang, R.; Richards, C.; Wyborn, L. A data quality strategy to enable FAIR, programmatic access across large, diverse data collections for high performance data analysis. *Informatics* **2017**, *4*, 45. [CrossRef]
7. Schöpfel, J.; Prost, H.; Rebouillat, V. Research data in current research information systems. In Proceedings of the 13th International Conference on Current Research Information Systems (CRIS2016), St Andrews, Scotland, 9–11 June 2016.
8. Hahnen, H.; Güdler, J. Quality is the product is quality—Information management as a closed-loop process. In Proceedings of the 9th International Conference on Current Research Information Systems (CRIS2008), Maribor, Slovenia, 5–7 June 2008.
9. Azeroual, O.; Saake, G.; Abuosba, M. Data quality measures and data cleansing for research information systems. *J. Digit. Inf. Manag.* **2018**, *16*, 12–21.
10. Azeroual, O.; Saake, G.; Wastl, J. Data measurement in research information systems: Metrics for the evaluation of the data quality. *Scientometrics* **2018**, *115*, 1271–1290. [CrossRef]
11. Azeroual, O.; Saake, G.; Abuosba, M. Investigations of concept development to improve data quality in research information systems. In Proceedings of the 30th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), Wuppertal, Germany, 22–25 May 2018; Volume 2126, pp. 29–34.
12. Azeroual, O.; Abuosba, M. Improving the data quality in the research information systems. *Int. J. Comput. Sci. Inf. Secur.* **2017**, *15*, 82–86.

[6] http://kerndatensatz-forschung.de/.
[7] https://d4h.meshs.fr/.

13. Azeroual, O.; Saake, G.; Schallehn, E. Analyzing data quality issues in research information systems via data profiling. *Int. J. Inf. Manag.* **2018**, *41*, 50–56. [CrossRef]
14. Mugabushaka, A.M.; Papazoglou, T. Information systems of research funding agencies in the "era of the big data". The case study of the research information system of the European Research Council. In Proceedings of the 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 6–9 June 2012.
15. Ilva, J. Towards reliable data—Counting the Finnish open access publications. In Proceedings of the 13th International Conference on Current Research Information Systems (CRIS2016), St Andrews, Scotland, 9–11 June 2016.
16. Krause, J. Current research information as part of digital libraries and the heterogeneity problem. In Proceedings of the 6th International Conference on Current Research Information Systems (CRIS2002), Kassel, Germany, 29–31 August 2002.
17. Lopatenko, A.; Asserson, A.; Jeffery, K.G. CERIF—Information retrieval of research information in a distributed heterogeneous environment. In Proceedings of the 6th International Conference on Current Research Information Systems (CRIS2002), Kassel, Germany, 29–31 August 2002.
18. Chudlarský, T.; Dvořák, J. A national CRIS infrastructure as the cornerstone of transparency in the research domain. In Proceedings of the 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 6–9 June 2012.
19. Lingjærde, G.C.; Sjøgren, A. Quality assurance in the research documentation system frida. In Proceedings of the 9th International Conference on Current Research Information Systems (CRIS2008), Maribor, Slovenia, 5–7 June 2008.
20. Johansson, A.; Ottosson, M.O. A national current research information system for Sweden. In Proceedings of the 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 6–9 June 2012.
21. Rybiński, H.; Koperwas, J.; Skonieczny, A. OMEGA-PSIR—A solution for implementing university research knowledge base. In Proceedings of the 21st EUNIS Annual Congress, Dundee, Scotland, 10–12 June 2015.
22. McCutcheon, V.; Kerridge, S.; Grout, C.; Clements, A.; Baker, D.; Newnham, H. CASRAI-UK: Using the CASRAI approach to develop standards for communicating and sharing research information in the UK. In Proceedings of the 13th International Conference on Current Research Information Systems (CRIS2016), St Andrews, Scotland, 9–11 June 2016.
23. Jörg, B.; Höllrigl, T.; Sicilia, M.A. Entities and identities in research information systems. In Proceedings of the 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 6–9 June 2012.
24. Ribeiro, L.; de Castro, P.; Mennielli, M. *EUNIS—EUROCRIS Joint Survey on CRIS and IR*; Final Report; ERAI EUNIS Research and Analysis Initiative: Paris, France, 2016.
25. Biesenbender, S.; Hornbostel, S. The research core dataset for the German science system: Challenges, processes and principles of a contested standardization project. *Scientometrics* **2016**, *106*, 837–847. [CrossRef]
26. Van den Berghe, S.; Van Gaeveren, K. Data quality assessment and improvement: A Vrije Universiteit Brussel case study. In Proceedings of the 13th International Conference on Current Research Information Systems (CRIS2016), St Andrews, Scotland, 9–11 June 2016.
27. Berkhoff, K.; Ebeling, B.; Lübbe, S. Integrating research information into a software for higher education administration—Benefits for data quality and accessibility. In Proceedings of the 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 6–9 June 2012.
28. Nevolin, I. Crowdsourcing opportunities for research information systems. In Proceedings of the 13th International Conference on Current Research Information Systems (CRIS2016), St Andrews, Scotland, 9–11 June 2016.
29. Azeroual, O.; Schöpfel, J.; Saake, G. Implementation and user acceptance of research information systems. An empirical survey of German universities and research organisations. *Data Technol. Appl.* submitted.
30. Apel, D.; Behme, W.; Eberlein, R.; Merighi, C. *Datenqualität Erfolgreich Steuern. Praxislösungen für Business Intelligence-Projekte*, 3rd ed.; Dpunkt Verlag: Heidelberg, Germany, 2015.

31. Hildebrand, K.; Gebauer, M.; Hinrichs, H.; Mielke, M. *Daten- und Informationsqualität, Auf dem Weg zur Information Excellence*; Springer Vieweg: Wiesbaden, Germany, 2015.
32. Haak, L.; Baker, D.; Probus, M.A. Creating a data infrastructure for tracking knowledge flow. In Proceedings of the 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 6–9 June 2012.