big data and
cognitive computing

MDPI

Article

Treatment of Bad Big Data in Research Data Management (RDM) Systems

Otmane Azeroual

German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, 10117 Berlin, Germany; azeroual@dzhw.eu; Tel.: +49-30-206417738

Received: 10 September 2020; Accepted: 15 October 2020; Published: 18 October 2020

Abstract: Databases such as research data management systems (RDMS) provide the research data in which information is to be searched for. They provide techniques with which even large amounts of data can be evaluated efficiently. This includes the management of research data and the optimization of access to this data, especially if it cannot be fully loaded into the main memory. They also provide methods for grouping and sorting and optimize requests that are made to them so that they can be processed efficiently even when accessing large amounts of data. Research data offer one thing above all: the opportunity to generate valuable knowledge. The quality of research data is of primary importance for this. Only flawless research data can deliver reliable, beneficial results and enable sound decision-making. Correct, complete and up-to-date research data are therefore essential for successful operational processes. Wrong decisions and inefficiencies in day-to-day operations are only the tip of the iceberg, since the problems with poor data quality span various areas and weaken entire university processes. Therefore, this paper addresses the problems of data quality in the context of RDMS and tries to shed light on the solution for ensuring data quality and to show a way to fix the dirty research data that arise during its integration before it has a negative impact on business success.

Keywords: research data management systems (RDMS); research data life cycle; big data; poor quality of information; data integrity and quality; institutional decision making

1. Introduction

Research data are essential resources and are becoming more and more extensive. This has to do with the fact that the way researchers work has changed and more and more data is being digitized and stored. However, precisely because research data are not only the basis of scientific cognitive processes, but are also related to other data, professional research data management systems (RDMS) is becoming increasingly important. Management systems of research data offer both research institutions and librarians a new opportunity to support the research process [1]. RDMS is not just a scientific discipline in computer science, and research institutions must provide organizational structures and processes as well as pragmatic solutions (hardware and software resources) in order to implement the first simple tasks of the RDMS [2]. RDMS, data curation, digital library and digital storage are very often used interchangeably in the literature [3]. As well as some institutions, in addition to RDMS, institutional repositories have been used as the basis for data storage, while others are experimenting with more extensive data description environments despite the diversity of existing workflows [4].

The explosion in data volumes is becoming a serious problem for research institutions [5]. In addition to the increase in pure data volume, the number of sources, users and the handling of data also increase the speed requirements. In order to keep control and to get as much benefit from the information as possible, in addition to infrastructural measures, measures to ensure data security and data quality are required. Almost all institutions rate good data quality as an existential basis for their own business. However, only a few invest the time and resources. Most of the time, a bad data basis in individual departments are either accepted reluctantly or, in the worst case, is not even noticed.

Poor quality of research data quickly leads to serious problems: for example, if complaints about incorrect research information accumulate, the handling of complaints causes noticeable costs and, in the worst case, even dissatisfied decision-makers move.

The facility receives up to 1 million lines of research information from several different systems every day. Dealing with large amounts of data is part of the day-to-day operations of the research institutions. The correctness and reliability of the data in the RDMS play a crucial role here. Because RDMS aims to collect, capture, store, track, and archive all the data generated in scientific projects and experiments [2].

In order to turn research data into scientifically broadly usable data with added social value, institutions have agreed to set up an RDMS. Data quality is also of crucial importance in order to prepare and save research data for interdisciplinary reuse or to make it usable for non-scientific dimensions such as economic exploitation, needs for society as a whole or cultural significance. Many authors in the literature (e.g., [6–8]) deal explicitly with safeguarding data quality in different information systems and databases. Little attention is paid to this in the area of RDMS, although the variety of methods and procedures for generating, processing and disseminating research data is increasing rapidly. The more decisions are made on the basis of digital research data, the more important questions about their origin and quality become, both from individual researchers and from universities, scientific organizations, sponsors and research-supporting infrastructure facilities (e.g., libraries, data centers, etc.). When assessing the quality of data, it should be made clear that the decision about its high quality or uselessness is always made by the potential user. This decision can only be meaningfully made by him if the development of the data is documented. The quality of the metadata plays the decisive role, i.e., who has; when; for what purpose; what and with what. Good metadata is not a guarantee of the quality of the data itself. It is imperative to subject the data to quality control yourself. Two typical main problems in the field of RDMS from practice are as follows: System failures can result in sporadic incomplete data during automated data generation. These gaps must be dealt with in order to avoid systematic bias. Occasionally, the professional interpretation of analysis results is incorrect if the process of data generation is not analyzed and conclusions are drawn on the basis of inferences. To counteract data quality problems, a systematic approach should be developed. Incorrect research data have a direct impact on the operational process and thus the success of the facility. Institutions must address the quality of their research data as a strategic task. Too often they have so far concentrated on mere repair work for acute problems. Higher data quality increases customer satisfaction and customer loyalty, and ultimately also sales.

This is the reason why this paper is being prepared and therefore the process of ensuring data quality in institutions must be viewed holistically. However, certain aspects that are not strictly necessary for the fulfillment of the project requirements are only discussed thematically. The topic of data quality and the handling of incorrect data in data quality control is the theoretical focus of this paper, because the management of research data is a major challenge for research institutions and their scientists today, and huge amounts of digital research data are produced in universities in a variety of forms at high speed [9].

This paper deals with the treatment of bad data in data quality control in the RDMS. It is important to illustrate the importance of the quality of data in the RDMS, what exactly is bad data and how incorrect data can affect research institutions. The aim is to use a case study to show methods of action and approaches when dealing with bad data in the data quality management process and thus to answer the following research question “how can RDMS users ensure their data quality?”.

The novelty with this work is to present a framework for the treatment of bad big research data for the institutions, which was used by the author to improve the quality problems in practice and thus reached 75% of the quality of research data. It enables research institutions and their scientists to monitor ongoing problems with the data quality in the RDMS and to intervene if necessary. Regular data cleansing ensures that the planned measures can be continuously checked and strengthened or adapted. One way to get representative results is to monitor the dirty data going on by ensuring its quality. The research data are examined from various points of view such as correctness, completeness and consistency.

The presentation of the solution on paper offers new insights into the subject of data quality in relation to the RDMS. Using this developed solution, valuable knowledge can be generated for the RDMS community or the users (institutions and their scientists). In order to convert this knowledge into value in a stable and sustainable manner, targeted data quality management is required. Continuous monitoring of data quality and the transfer of methodological knowledge are essential measures that should be implemented in every facility.

2. Research Data Management Systems (RDMS)—What Are They? Why Do We Need Them?

The scope of research data has increased in recent years solely through the development of new computer technologies and storage media. This is not only due to the fact that scientists collect large amounts of research data, so that their processing, storage and evaluation can only be guaranteed by highly developed computer technology. It is also related to the fact that the way researchers work has changed and more and more research data are being saved because more and more storage space is available. These stored data are not only the basis of scientific knowledge processes, they are also related to other data and serve as a new database on the basis of which new knowledge can be gained in research [10].

It is not easy to define what can be called research data, because in science, depending on the research area, people speak of raw data, measurement data, empirical data or source data. The term research data is data that is used in the context of scientific projects e.g., through digitization, source research, experiments, measurements or surveys. Nevertheless, the term research data has prevailed in German-speaking countries, which should include all data that are developed in the course of a research project. According to the OECD (www.oecd.org), research data are "*factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings*" [11]. The German Research Foundation (DFG, www.dfg.de) understands research data to be data (such as measurement data, texts, simulations, audio files, etc.) that represent an essential basis for scientific work. Because there are a variety of different scientific disciplines, research data can also be very diverse.

RDMS is currently an important topic for institutions and their scientists. Scientific institutions generate a lot of digital research data. A comprehensive RDMS ensures that this data can also be used by other scientists. Networking and cooperation help everyone involved to better master the tasks and challenges within their own facilities. Scientists usually only seek support and information when faced with a specific problem. This could be the case, for example, when a data management plan (DMP) is in place, or when research data is mandatory as part of a publisher's publication. The aim of developing an RDMS is to sensitize scientists preventively and to inform them about new developments. RDMS should become an integral part of the research process.

Research data must be managed in an appropriate manner so that reuse is possible for researchers. Structured handling of research data can only be facilitated with an RDMS. As in Figure 1, the process can be seen in an RDMS. With an RDMS, research institutions can collect their research activities and results (e.g., publications, employee information, projects, financial data, etc.) and integrate them into an RDMS. This data can then be automatically output in various forms on the RDMS website. Such a system, which interweaves many individual sources of information and administrative processes in the field of research management, is not only a provider of information, but also a work tool—on the one hand for scientists and on the other hand for members of the university administration who are entrusted with the management of research metadata stored in the RDMS.

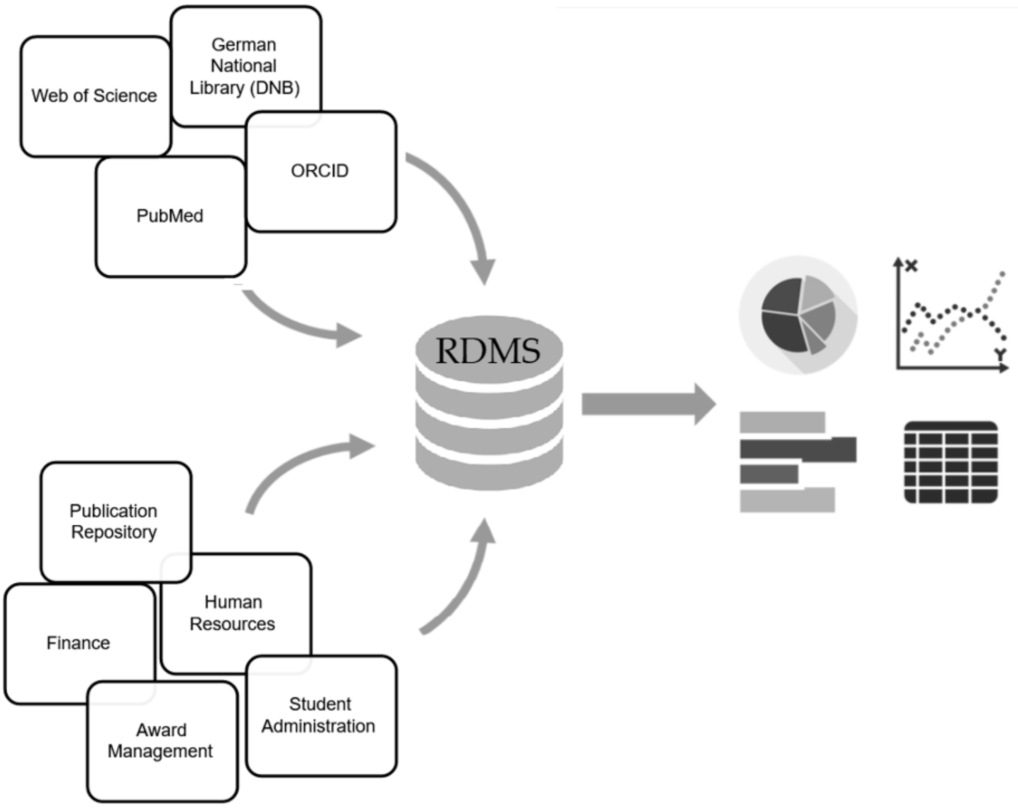

Figure 1. Example of an research data management systems (RDMS) in research institutions.

RDMS describes the structured handling of digital research data [3]. This is to ensure that research results obtained are also reusable. This includes the various steps in the data lifecycle as well as overarching tasks such as organizational, legal and financial matters. This makes it easier to work according to good scientific practice and meets the requirements of third-party funding providers, which increasingly require transparency about the data generated by research. In addition, working with your own and third-party data is made easier and precious time is saved.

According to a general definition, all measures that serve to be able to use data are included in the RDMS. A distinction must first be made between tasks during the research process that relate directly to the specific research work and that are relevant to the processing of a research question, and secondly, tasks that have to be performed after the end of the project, such as data archiving and/or ensuring the long-term availability of the research data to ensure their continued use in new research contexts. General models for structuring and systematizing the RDMS are designed to describe which specific tasks arise in the various work phases and which work steps are involved in the specific setup of the examination materials. There are various models for describing the life cycle of research data [12]. So-called data life cycle models describe the administration and processing of the examination materials as a cycle, in which firstly the different phases of the data management process and the research process run in parallel, and secondly the phase of providing the research data, i.e., the final phase of a research project, directly in the planning phase for new one's research project passes [1]. In its most detailed form, six essential steps characterize this cycle:

1. Data creation: This step includes the planning of the research project and the data collection. If requested by the third-party sponsor, a data management plan should also be drawn up in this phase. A DMP describes the structured handling of the research data generated in the course of a project. It contains information about the project, the responsibilities and which subject-specific standards/guidelines have to be observed. In addition, a statement is made about the amount

of data, file formats and their archiving and making available. If the DMP is kept up-to-date, it makes it easier to share and reuse the project data and reduces the risk of data loss.

2. Data processing: The research data received are processed, possibly digitized and anonymized. A comprehensible system of assigning file names and entering metadata makes it easier to find the data later. The data must also be backed up.
3. Data analysis: After evaluating and interpreting the results, the data can be published. This is usually done in the form of a publication.
4. Data archiving: For long-term archiving, suitable file formats must be used, an archive selected and the associated metadata stored. Both archival and general data repositories are suitable for archiving, which may also be subject to a charge depending on the provider.
5. Data access: In addition to the article, many scientific journals require the research data to be published in a suitable repository. Therefore, access rights must be set for archived data. Furthermore, the copyrights and the reusability of the data should be clearly defined.
6. Data reuse: Many data records allow re-examination under other aspects. This can be done by the data creator as well as by other scientists. Both new results can be obtained (at this point the data life cycle closes) and old findings can be checked and confirmed. Through an efficient exchange and effective reuse of research data, support can be provided in order to make the research data discoverable and interpretable as a scientific result and also to use it in teaching [13].

The mentioned steps of the research data cycle reveal an important effect of the RDMS [14]. While research is currently being organized in projects that are characterized by a start and end point and are thus self-contained, the systematic engagement with digital sources draws attention to the fact that each individual project is part of a continuously progressing process of knowledge, in which results of ongoing research are also the preparatory work for future research. In order to be able to master the complex tasks in the interaction between research and RDMS, the collaboration of researchers with experts from the fields of information science in the information infrastructure facilities on-site or at the national and international level is necessary.

To sum up, good research needs a good database. With the advancing digitalization, the amount of research data from science increases inflationary. This data is of outstanding value for interdisciplinary research because it enables large-scale comparative evaluations and meta-analyzes. The responsible and transparent handling of research data makes the scientific results reproducible and is therefore an essential part of good scientific practice. The RDMS is committed to making research data findable, accessible, interoperable, reusable (FAIR) and, whenever possible, openly accessible.

3. Data Quality—Success Factor of Research Institutions

The topic of data quality is becoming increasingly significant, especially in connection with the increasing flood of data in research institutions. The increase in collected data and its sources presents institutions with new and difficult challenges. The reason for this is not least the intention to combine the collected data into a homogeneous amount, to put it in a logical context and, consequently, to be able to evaluate and present it for research-related decisions. If this research data is incorrect, complete or uniform, this can have considerable consequences for the institution. Poor data quality is not only reflected in sales losses—it is not uncommon for additional costs or reputational losses to arise.

With the increasing amount of data and the number of internal and external source systems (such as personal and project databases, publication databases, etc.), it becomes increasingly difficult to meet the requirements for data acquisition and transformation processes. An increase in the amount of data can lead to more quality errors, both with manually recorded data and with automated data collection processes [15]. The type and number of users can also have an impact on data quality [15]. For example, if a facility relocates its business to the Internet and scientists and/or internal employees can change their data in the research-internal databases via the web, the susceptibility to incorrect data also increases. The research data basis is gaining new importance for institutions, their success and, in the long run, for the well-founded basis of profitable decisions. In reality, however, this trend shows that there is a large deficit precisely with this research data basis. Universities and research institutions are faced with the challenge of managing the growing volume of data technically and using it efficiently in their business processes. The data collection, storage, management, use and evaluation are usually very special and ideally oriented close to the business processes of the facility. The research landscape results in a multivariate data processing culture that is difficult to generalize and standardize.

Numerous definitions of data quality are found in the existing literature. The definition of the term data quality refers to its suitability to fulfill certain purposes. "*Data that is suitable for a particular use by data users*" [16,17] are therefore considered. The suitability of the data for use is the focus of the data quality approach by [18]. In the literature, the concept of the authors is one of the most cited concepts for describing and evaluating data quality. It aims to identify characteristics of data quality from the user's perspective and is based on an empirical survey among IT users. The approach assumes that the data user can best assess and evaluate data quality. The concept of fitness for use applies [16].

In this paper, data quality, following the majority of the literature, is defined according to the user-related approach. Accordingly, data is of high quality if it meets the needs of the user. That means, however, also that data quality can only be assessed individually. Quality is a relative and not an absolute property. The quality of the data can therefore only be assessed in relation to its respective use. In the literature, in addition to the term data quality, one often also finds the term information quality. Since the same topic is usually dealt with by both terms and so far, there is no uniform delimitation of the terms data and information, it seems permitted to consider the terms as largely synonymous. The term data quality is used in this paper.

The data quality refers to the quality of the data stored in an RDMS. High quality databases are very important for achieving business goals. In order to achieve high data quality, the first question to be asked is what a research institution wants to achieve with the research data collected. Research success is largely based on reliable research data. However, research institutes often have no control over their data volumes. Having the right technology is just one way to improve data quality in RDMS. It is much more necessary to use the right methodology, because without methodology there is no success. With data quality measures, research institutions are able to gradually and sustainably improve the quality of their research data.

4. Bad Data—Its Emergence in RDMS

Conversely, from the definition of data quality, the following can be defined via bad data: bad data or dirty data refers to data in a data storage system that does not meet the required data quality requirements at a given point in time. Bad data can negatively affect decisions and have a negative impact on results [5]. Good research data becomes dirty research data if it contains incorrect information and thus falsifies reports or even generates serious errors in a system. These falsifications and errors occur because the research data used:

- are not available if required (missing research data),
- contain incorrect, different or fuzzy information (incorrect or inaccurate research data),
- are in the wrong field (inappropriate research data),
- do not meet the target standard of the system (non-compliant research data),
- occur twice (double research data),
- bad and inconsistent formats contain insufficient entries (bad input of research data),
- or record research data in different languages, fonts or units of measurement in the RDMS.

The causes of bad data can be very different in nature. Just like the evaluation of data quality, the statement about when research data are invalid is very complex and subjective. With an increase in different users, user groups, systems and interfaces in the research process, the probability of bad data occurring also increases. In addition to technical applications and processes, all users who process data generated in research institutions themselves can form a potential source of errors. In this case, the cause can be, for example, employees of the facility or their researchers while recording research data in the RDMS. The first errors can arise when collecting research data in the RDMS. If mistakes are made while the data is being recorded, this can have a variety of negative effects. Typing errors, misunderstandings or missing values are just a few examples of errors in the collection of research data.

From a technical point of view, there are a large number of errors. This can be incorrect article keys and descriptions, author data as well as third-party funded project data, etc. These errors can be found in RDMS using column checks. Typical inspections are:

- Zero values in the required columns
- Numerical values that fall below/exceed a threshold
- Column values that are unexpectedly long/short
- Compliance with required schemes or entries in input patterns
- Match a value from a list of incorrect values where a list of correct values would be too long
- Spelling errors
- etc.

For RDMS, high data quality is not only something that the operator desires, but one of the main criteria that determine whether the project is successful and the statements made are correct. The occurrence of incorrect research data not only influences the RDMS and thus partly directly on the operational process, but can also have far-reaching latent consequences. This not only affects research institutions where this incorrect research data occurs, but it can also occur if they already deal with bad data.

The greatest monetary impact here is usually the cost of strategic wrong decisions by the management due to incorrect data. The time required to validate and interpret inconsistent key figures also falls into this category. The problem when planning solutions is then to decide which investments to improve the data quality are worthwhile. So, one has to weigh costs against costs and have to find out how budgets can be generated to improve data quality and what problems need to be considered in the context of limited facility budgets. The following is a summary of the top reasons to answer the question "why is it worth investing in data quality management?":

1. Data in general is a fully-fledged asset that controls automation and should therefore be maintained.
2. Reporting can be better interpreted if one knows what the quality of the information shown is.
3. Data quality problems can damage the image.
4. Only those who have their research data under control can effectively control their facility.
5. Data quality problems cause unnecessary rework during process execution.
6. Data quality problems cause prominent projects to fail.
7. Data quality problems annoy employees.
8. Data quality problems cause wrong decisions.

5. Dealing with Big Bad Research Data—Best Practice Framework

One of the most important points when dealing with integrated data sources, which were previously heterogeneous and distributed, is the checking of errors. As already mentioned, errors or bad research data can severely restrict the ability to analyze the data and lead to incorrect results that ultimately allow wrong decisions or conclusions to be drawn. The correct handling of research data of unknown quality is therefore extremely significant.

In this context, the RDMS should discover errors and be able to correct them. This process is divided into two sub-processes: data profiling and data monitoring. During data profiling, domain experts use tools to examine the database. Aids are provided by statistics such as maxima and minima or frequency distributions of attribute values or zero values. Pattern recognition is also an important tool in profiling. With the help of pattern recognition, recurring orders or regularities in data can be analyzed. For example, errors in orders such as the date of birth of the publication authors can be recognized, since they always have the same typical patterns as (DD/MM/YYYY, YYYY-MM-DD, etc.).

Finally, the monitoring fulfills the aim of supervising the measures taken to correct errors or eliminate error sources. The cleaning of faulty research data is referred to as data cleaning, data cleansing or data scrubbing. This process is in turn divided into two sub-processes. In the first part, simple errors that only affect individual data records are cleaned up. In the second phase, cross-tuple errors are then considered and eliminated.

At the start of data cleaning, all data values are converted into standardized formats. This process does not yet correct any errors, but it simplifies the further processing of the data and facilitates error correction [19]. A format is defined for each attribute at the beginning of this step. To make it easier to compare textual data, in many cases all letters are replaced by capital letters. Furthermore, textual data can be freed from the first data errors by automatic spellchecks. The removal of stop words ("the", "it", "and", etc.) and the return of words to their basic form (stemming) increasingly simplifies the ongoing work with text-based attributes. Ultimately, general abbreviations can still be replaced by their full spelling.

Publication data is also reduced to standardized formats. Author names, titles, publishers and page details, etc. usually consist of several components. An author's name, for example, consists of salutation, title, first name and last name, into which it can be broken down. Publication data also consists of various components. In addition to the title and year, there are also volume, issue and pages. These components must first be divided and then normalized as described above, so that, for example, the abbreviation "Vol." is written out for volume.

Ultimately, standard formats can be selected for information such as the date of birth of the researchers or project amounts. For example, dates are changed from 07.07.07 to 07.07.2007. Project amounts are converted into desired currency amounts based on the current exchange rate. The conversion of data values into uniform dimensions is crucial for a uniform understanding.

A big problem in RDMS is missing values or outliers. Missing values can be individual values (zero values), but entire tuples, partial relations or entire relations can also be missing. In all respects, missing values are detrimental to the meaningfulness of the data if the special information of this data is needed. Zero values can usually be found by manually or automatically checking the amount of data. Incomplete data value distributions that suggest missing partial relations can be determined using profiling tools [20].

Missing numerical values are often supplemented with imputation. Imputation cannot establish actual value. Rather, the statistical analysis of other values allows imputation for drawing a conclusion about the approximate data value. This is done, for example, by calculating the average values of related entries or by using far more complex techniques using data relationships. A far more secure tool for adding missing values is the use of reference tables. These help in checking publication data, third-party funded project data, financial data and addresses of employees in institutions with their telephone numbers or bank details. In addition to adding missing values, reference lists can also be used to check the consistency of the other data, e.g., reference tables for third-party funded project data contain lists of all project names, project start, project end and year.

If the information has consistency problems or is incomplete, it can be corrected or completed using the information in the reference table. With regard to the address data of the authors, geocoding/reverse geocoding offers an alternative to the reference list. Using a geocoding web API, companies such as Google make their geographical database available in order to draw conclusions about the locations of customers or companies. For this, geographic and statistical data of the environment are used to compensate for insufficient address information or to limit or even determine the possible location.

A crucial point is the elimination of possible duplicates that can occur when merging heterogeneous data sets. To do this, two tasks must be performed so that duplicates can be reliably removed. First, duplicates must be recognized as such, and in the second step for merging the multiple entries, inconsistencies between the duplicates must be recognized and corrected.

Dealing with the bad research data in RDMS has shown that no data set is excluded from errors and negligence in data storage or human error when creating data sets leads to a potential for errors. It was also explained that heterogeneity must be considered before the possible integration of research data. For these reasons, a clear, structured procedure is necessary in order to clean research data in RDMS from disruptive factors and to raise it to the highest possible data quality level.

In order to achieve the highest possible data quality when integrating research data into RDMS, a framework is now displayed as a flow chart (see Figure 2) and shows the main points of the basic steps for data preparation of the research data. Frameworks are based on already known methods of processing data for use in databases. In addition, the developed framework tries to specify a clearly structured workflow of research data with the unknown format and data quality measure for integrable input data that have a format conforming to the target system.

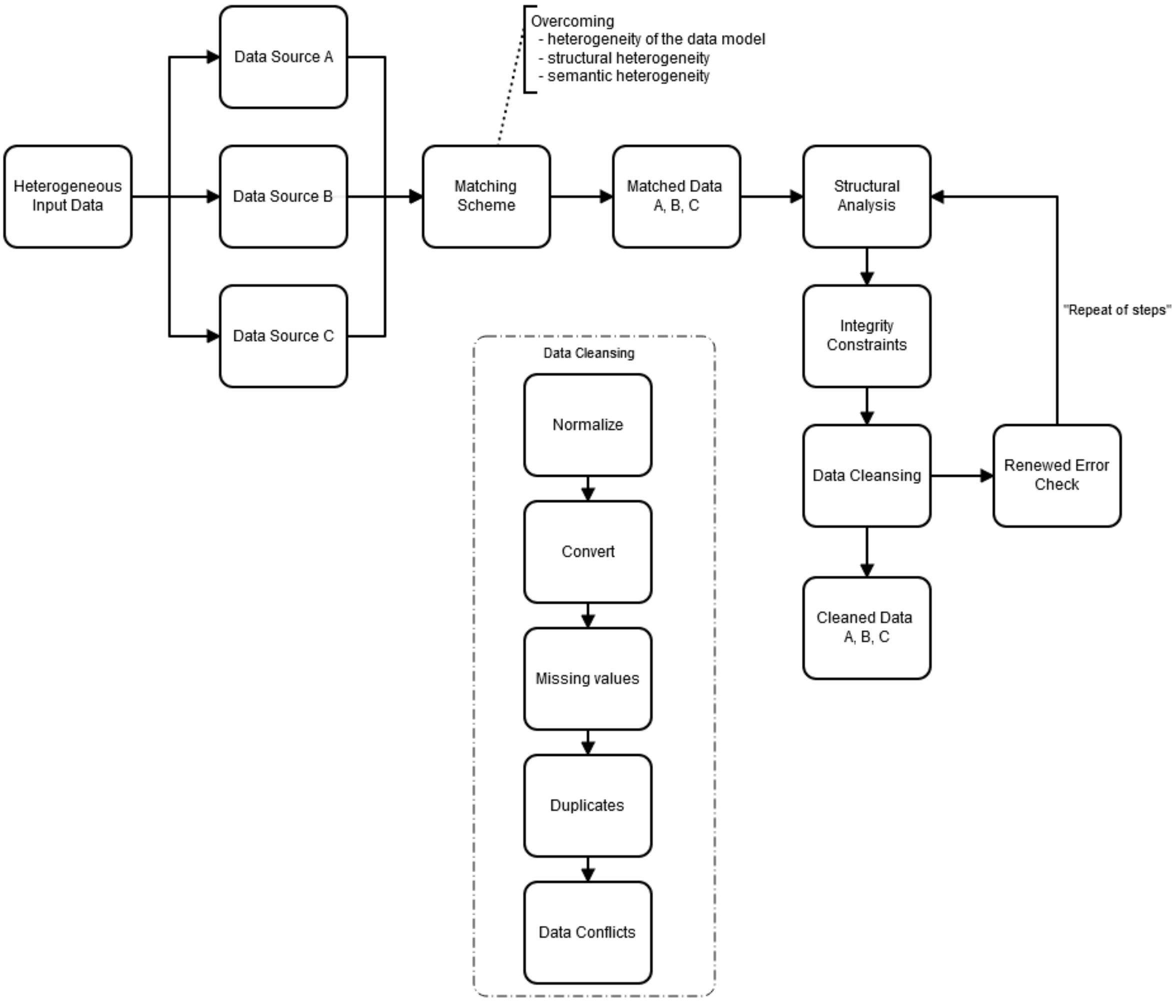

Figure 2. Framework for the elimination of bad research data.

In the progress graph, the framework can be described by four work processes for the homogenization and error correction of research data. These four work processes are grouped into two rough sub-processes. First of all, the elimination of heterogeneity so that the research data is subsequently available in a uniform form, which makes it easier to apply error correction processes to the data. In addition, the research data can only be integrated into the target system (RDMS) in an adapted form. The second sub-process, which includes the structural analysis, integrity conditions and data cleansing work processes, is used to find and eliminate errors so that the input data, which were previously of unknown quality, are brought into the most ideally usable data quality.

6. Conclusions

The aim of this paper was to introduce the importance of data quality in RDMS to the reader and to examine the handling of incorrect research data in the process of data quality management (the data quality check) and to demonstrate approaches. It becomes clear that data quality is becoming increasingly important in connection with research-related decisions. The level of data quality is increasingly a major component of economic success. Faulty, inaccurate and missing research data thus have a direct impact on the usability of the research data. It also shows that data quality in the facility cannot only be defined by the quality and amount of reactive measures and systems. The topic of data quality is not new territory for universities and research institutes and is nevertheless mostly perceived only as an annoying addition to large data and is therefore usually treated as an annoying consequence. Poor data quality not only creates technical errors, but can have far more far-reaching consequences.

With the help of the developed framework for the treatment of bad research data in the RDMS, incoming amounts of data can be filtered as a preventive measure. The framework offers the possibility of minimizing reactive measures through automated checks in order to create the scope for the actually required more in-depth data quality measures.

In conclusion, it can be stated that data quality cannot be treated thoughtlessly and, as described, should be consolidated in the overall strategy of the research institution. In order to avoid the mostly cost-intensive reactive measures, a holistic data quality management process is required. This makes it possible to introduce quality into scientific institutions as an overall goal for research data and to guarantee it permanently.

Work based on this paper could further improve the data cleansing model through new approaches in order to also correct the previously unsolved data errors. One possibility would be to use fuzzy logic algorithms to improve the detection of bad research data (such as duplicates).

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Surkis, A. Research data management. *J. Med. Libr. Assoc.* **2015**, *103*, 154–156. [CrossRef] [PubMed]
2. Heuer, A. Research Data Management. *It-Inf. Technol.* **2020**, *62*, 1–5. [CrossRef]
3. Tammaro, A.M.; Casarosa, V. Research Data Management in the curriculum: An interdisciplinary Approach. *Procedia Comput. Sci.* **2014**, *38*, 138–142. [CrossRef]
4. Amorim, R.C.; Castro, J.A.; Rocha da Silva, J.; Ribeiro, C. A comparison of research data management platforms: Architecture, flexible metadata and interoperability. *Univ. Access Inf. Soc.* **2017**, *16*, 851–862. [CrossRef]
5. Azeroual, O. Data Wrangling in Database Systems: Purging of Dirty Data. *Data* **2020**, *5*, 50. [CrossRef]
6. Batini, C.; Barone, D.; Mastrella, M.; Maurino, A.; Ruffini, C. A Framework and {A} Methodology for Data Quality Assessment and Monitoring. In Proceedings of the 12th International Conference on Information Quality, Cambridge, MA, USA, 9–11 November 2007; pp. 333–346.
7. Aljumaili, M.; Karim, R.; Tretten, P. Metadata-based data quality assessment. *VINE J. Inf. Knowl. Manag. Syst.* **2016**, *46*, 232–250. [CrossRef]
8. Haegemans, T.; Snoeck, M.; Lemahieu, W. A theoretical framework to improve the quality of manually acquired data. *Inf. Manag.* **2019**, *56*, 1–14. [CrossRef]
9. Pinfield, S.; Cox, A.M.; Smith, J. Research Data Management and Libraries: Relationships, Activities, Drivers and Influences. *PLoS ONE* **2014**, *9*, e114734. [CrossRef] [PubMed]
10. Kindling, M.; Schirmbacher, P. Die digitale Forschungswelt als Gegenstand der Forschung. *Inf.-Wiss. Prax.* **2013**, *64*, 137–148.

11. OECD. OECD Principles and Guidelines for Access to Research Data for Public Funding. Available online: http://www.oecd.org/sti/sci-tech/38500813.pdf (accessed on 27 August 2020).
12. Cox, A.M.; Kennan, M.A.; Lyon, L.; Pinfield, S. Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 2182–2200. [CrossRef]
13. McDonald, J.P. Toward more effective data use in teaching. *Phi Delta Kappan* **2019**, *100*, 50–54. [CrossRef]
14. Tang, R.; Hu, Z. Providing Research Data Management (RDM) Services in Libraries: Preparedness, Roles, Challenges, and Training for RDM Practice. *Data Inf. Manag.* **2019**, *3*, 84–101. [CrossRef]
15. Azeroual, O.; Lewoniewski, W. How to Inspect and Measure Data Quality about Scientific Publications: Use Case of Wikipedia and CRIS Databases. *Algorithms* **2020**, *13*, 107. [CrossRef]
16. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
17. Tayi, G.K.; Ballou, D. Examining Data Quality. *Commun. ACM* **1998**, *41*, 54–57. [CrossRef]
18. Lee, Y.W.; Strong, D.M.; Wang, R.Y. 10 Potholes in the Road to Information Quality. *IEEE Comput.* **1997**, *30*, 38–46.
19. Azeroual, O.; Saake, G.; Abuosba, M. Data Quality Measures and Data Cleansing for Research Information Systems. *J. Digit. Inf. Manag.* **2018**, *16*, 12–21.
20. 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]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.