International Journal of Information Management 41 (2018) 50–56

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Analyzing data quality issues in research information systems via data profiling

Otmane Azeroual[a,b,*], Gunter Saake[b], Eike Schallehn[b]

[a] German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117, Germany
[b] Otto-Von-Guericke-University Magdeburg, Department of Computer Science, Institute for Technical and Business Information Systems Database Research Group, P.O. Box 4120, 39106 Magdeburg, Germany

ARTICLE INFO

Keywords:
Current research information systems
CRIS
Research information systems
RIS
Research information
Data sources
Data quality
Extraction transformation load
ETL
Data analysis
Data profiling
Science system
Standardization

ABSTRACT

The success or failure of a RIS in a scientific institution is largely related to the quality of the data available as a basis for the RIS applications. The most beautiful Business Intelligence (BI) tools (reporting, etc.) are worthless when displaying incorrect, incomplete, or inconsistent data. An integral part of every RIS is thus the integration of data from the operative systems. Before starting the integration process (ETL) of a source system, a rich analysis of source data is required. With the support of a data quality check, causes of quality problems can usually be detected. Corresponding analyzes are performed with data profiling to provide a good picture of the state of the data. In this paper, methods of data profiling are presented in order to gain an overview of the quality of the data in the source systems before their integration into the RIS. With the help of data profiling, the scientific institutions can not only evaluate their research information and provide information about their quality, but also examine the dependencies and redundancies between data fields and better correct them within their RIS.

1. Introduction

In recent years, the requirements for research reporting of all institutions have risen massively. Both internal institution (presidium, departmental and specialist management) and external (relevant ministries, interested public) need comprehensive information on the research activities of the institutions for planning and control purposes. Research activities include research information such as Projects, third-party funds, patents, partners, prices and publications, etc. This research information is stored and managed in a corresponding research information system. A research information system is used when the information system from different sources brings together research information via an integrated ETL process and makes it possible to analyze it using various output and analysis functions. With the introduction of RIS, 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. Furthermore, RIS users can not have any extra effort in the survey of their research activities.

The growing volumes of data and the increasing number of source systems can lead to possible data errors, duplicates, missing values, incorrect formatting and contradictions in RIS. In addition, it serves to analyze the quality of the data in the source systems prior to their integration into the RIS. The sooner quality defects are controlled and remedied, the better. For this purpose, the aim of this paper will be to present the possible methods of data profiling applicable to research information in order to allow facilities to carry out a detailed analysis and evaluation on the existing data.

2. Research information and research information systems

In addition to teaching, research is one of the core tasks of universities. Information about task fulfillment and services in this area must be available with less time and more reliably. For the research area, data and information are mainly collected to map the research activities and their results, and to administer the processes associated with the research activity. This may include information on research projects, their duration, participating researchers and related publications. This information is also called research information or research data. The categories of research information (RI) with different key figures and indicators can be seen from Table 1:

However, not only the data itself, but also the structure of the data

* Corresponding author at: German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117, Germany.
E-mail addresses: Azeroual@dzhw.eu (O. Azeroual), Saake@iti.cs.uni-magdeburg.de (G. Saake), Eike@iti.cs.uni-magdeburg.de (E. Schallehn).

https://doi.org/10.1016/j.ijinfomgt.2018.02.007
Received 30 January 2018; Accepted 24 February 2018

Table 1
Research Information (RI) with different Key Figures and Indicators (Azeroual et al., 2018).

RI	Key Figures, Indicators
Person	Name, Title, Gender, Nationality
Publications	Authorship, Peer-Reviewed, Citations
Projects	Number, Duration, Financing, Cooperation
Third Party Funds	Procurement TPF, Disbursed TPF, Competitive TPF
Awards	Prizes
Patents	Patent Family, Spin-Offs, Patent Number
Institutional Information	University Unit, Employment Relationship, Staff Category, Department

and the linking of the individual research information among each other are central, in order to be able to aggregate and evaluate information on different organizational levels for different purposes (levels can mean a professor, a chair, an institute, a department, or a faculty).

Processes and systems that store and manage the research information correspond to Research Information Systems (RIS or CRIS for Current Research Information System).

A RIS is a **central database** that can be used to collect, manage and deliver information about research activities and research results.

The following figure (see Fig. 1) shows an overview of the integration of research information from a university into the research information system and the architecture of RIS.

The building blocks of RIS architecture can be seen as a three-stage process:

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

The *Data Access Layer* contains the internal and external data sources (operational systems). This level contains, for example, databases from the administration or publication repositories of libraries, identifiers such as e.g. ORCID or bibliographic data from the Web of Science or Scopus, etc. A filling of these data sources takes place via classical ETL process into the RIS. The *Application System Layer* contains the research information system and its applications, which merge, manage and analyze the data held at the underlying level. The *Presentation Layer* 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 the business intelligence tools. In addition to various reporting options, portals and websites of the facilities can also be filled here.

Orthogonal to the layers described, there are the Infrastructural Services, the overarching services for all information systems, such as authentication (LDAP), authorization, single sign-on, etc.

Offers for the standardized collection, storage and exchange of research information in RIS are the Research Core Dataset (RCD) Data Model and the Common European Research Information Format (CERIF) Data Model. These two models describe the entities and their relationship to each other.

RIS provide a holistic view of research activities and results at a research institution. They map the research activities not only institutions but also researchers current, central and clear. By the central figure in the system a working relief is possible for the researchers. Data is entered once with the RIS and can be used multiple times, e.g. on

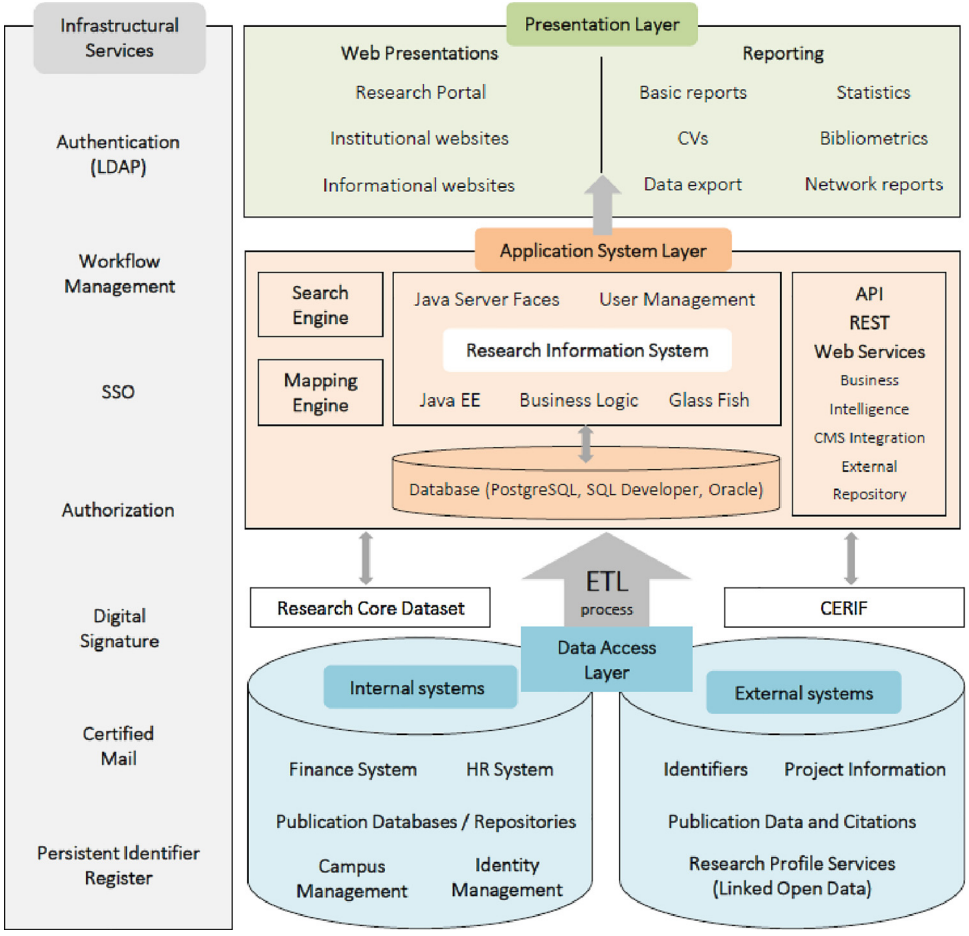


Fig. 1. Research Information System (RIS) (Azeroual and Abuosba, 2017).

websites, for project applications or reporting processes. A double data management and with it an additional work for the users should be avoided. Another objective is to establish research information systems as a central instrument for the consistent and continuous communication and documentation of the diverse research activities and results. Improved information retrieval helps researchers looking for collaborators, companies in the allocation of research contracts, to provide the public with transparency and general information about their institution.

3. Data profiling

On the basis of the collection and integration of different internal and external data sources (such as different databases, text files or XML files, etc.) of the facilities in research information systems, there is a wide variety of data errors that must be further processed by the RIS. Some of these errors are as follows:

- Missing values (characteristic completeness)
- Incorrect information caused by input, measuring or processing errors (characteristic correctness)
- Duplicates in the dataset (characteristic redundancy-free)
- Inconsistently represented data (characteristic consistency)
- Logically contradictory values (characteristic consistency)

The quality of the source data has a direct influence on the quality of the RIS. To avoid structural and content-related data quality problems, the analysis and cleansing of the data takes place within the data integration. Data integration is the transfer of operational data from the various legacy systems to the RIS. The filling of RIS takes place via the ETL process. The term stands for **extraction**, **transformation** and **load** and is understood as a data procurement process. His goal is to clean up the data from different structures and to standardize them in order to permanently store them in the RIS. To load the data into a RIS, they must first be extracted. Since the information is usually loaded from several source systems, the data of the respective systems must be coordinated with each other. Then the data is loaded into a RIS. This process is illustrated by the following figure (Fig. 2):

To successfully review the quality of the data and to decide whether data problems in source systems can be improved or resolved through corrections in the context of the ETL process. But data profiling can be used to better understand the structure of data sources and to detect and automatically correct potential errors. These detected errors are not fixed in the data but it only corrects the belonging metadata and these then form the data quality problems to be solved.

Data profiling is a new term and is used as a synonym for data analysis. Data profiling is an automated process for analyzing existing data (Olsen, 2003). Different methods or techniques for systematic analysis provide information about the structure, content and quality of the data collection in order to obtain and gain an accurate picture of the current state, such as (Olsen, 2003):

- Definition of allowed data values.
- Anomalies and outliers within columns.
- Relationships between columns (primary key, dependencies, etc.).
- Columns in which possible date formats and e-mail addresses etc. can be found.

To analyze the research information in RIS, data profiling offers three types of analysis (see Fig. 3). Behind each of these three types are different data profiling methods. These depend mainly on the way the data is analyzed: within a column (“attribute analysis”), in dependencies of columns (“functional dependency”) or in dependencies of attributes/columns in different tables (“reference analysis”).

Attribute Analysis: Gets general and detailed information about the structure, contents of a table, and all relationships between different tables, the columns and values that appear in the table. Here are the key questions to answer about the data for each of the following aspects (Apel, Behme, Eberlein, & Merighi, 2015):

1. Analysis of the data structure: (Do the data correspond to the corresponding metadata?)
2. Analysis of the data content: (Are the data values complete, correct and up-to-date, is the data standardized according to the applicable rules?)
3. Dependency Analysis: (Does the data in all columns and tables have the required mapping to the specified key relationship? Are there derived relationships across columns, tables, and databases? Are there redundant data?)

Behind these core questions, a lot of information is found in the following areas (Apel et al., 2015):

3.1. Attribute name analysis

The attribute name analysis refers to the attribute name. Here, attribute names should match the data type and content of the data (for example, the “Author ID” can be given a numeric value).

3.2. Data type analysis

This analysis evaluates the data type of the attribute. Are in a varchar data type field, e.g. only numbers exist, it may be beneficial to change the data type for more efficient processing. This makes it

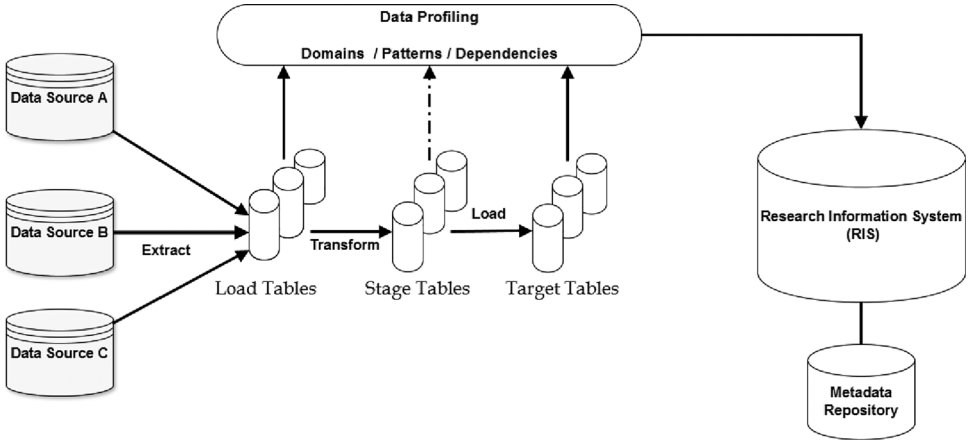


Fig. 2. Data Integration into RIS.

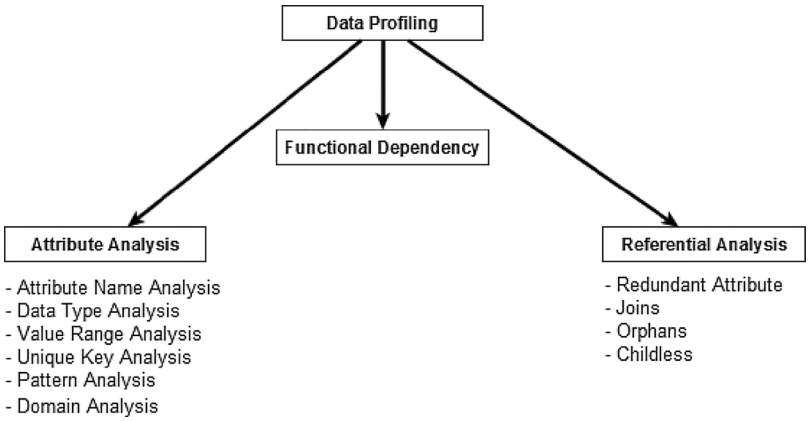
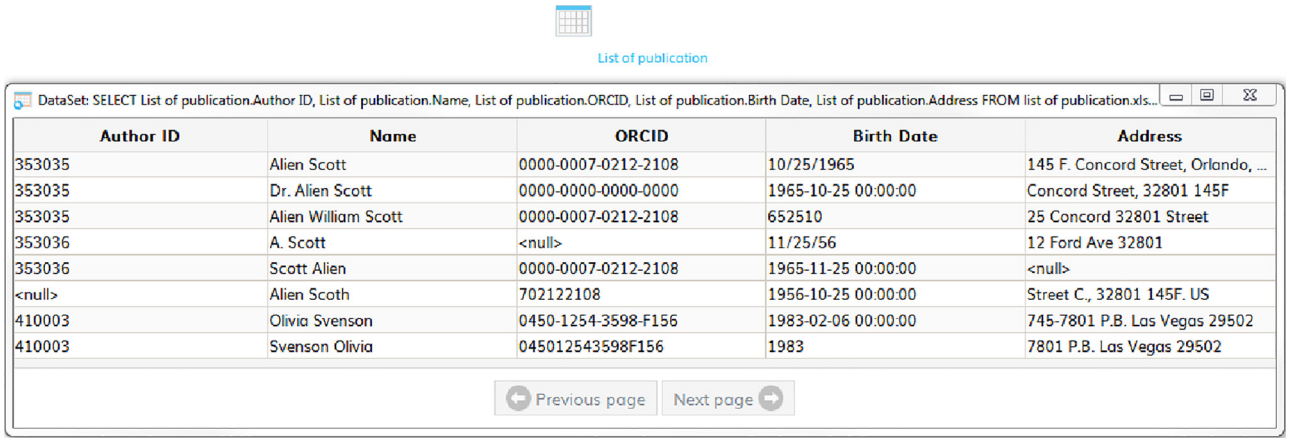


Fig. 3. Data Profiling Analysis Types.



Author ID	Name	ORCID	Birth Date	Address
353035	Alien Scott	0000-0007-0212-2108	10/25/1965	145 F. Concord Street, Orlando, ...
353035	Dr. Alien Scott	0000-0000-0000-0000	1965-10-25 00:00:00	Concord Street, 32801 145F
353035	Alien William Scott	0000-0007-0212-2108	652510	25 Concord 32801 Street
353036	A. Scott	<null>	11/25/56	12 Ford Ave 32801
353036	Scott Alien	0000-0007-0212-2108	1965-11-25 00:00:00	<null>
<null>	Alien Scoth	702122108	1956-10-25 00:00:00	Street C., 32801 145F. US
410003	Olivia Svenson	0450-1254-3598-F156	1983-02-06 00:00:00	745-7801 P.B. Las Vegas 29502
410003	Svenson Olivia	045012543598F156	1983	7801 P.B. Las Vegas 29502

Fig. 4. Example of a Publication List.

possible to define a rule that ensures that all values have the same data type. The goal of data type analysis is to find metrics such as Minimum, maximum or lengths, thereby allowing e.g. Discrepancies in storage are found out.

3.3. Value range analysis

In this analysis, different statistical key figures are used to analyze the data (minimum, maximum, mean, frequency distribution, standard deviation, etc.).

3.4. Unique key analysis

This analysis is about finding nulls or duplicates. These two are dangerous to any evaluation and process, so it is necessary to create a rule that ensures that all the values entered are neither duplicate nor contain null values.

3.5. Pattern analysis

In this analysis, patterns or general representations are identified by an analysis of the attributes. First the values are searched for possible patterns and then these values are identified with the filtered out patterns and put into relation. The percentages calculated in this way provide information on the validity of the samples. Then, rules can be created to resolve any detected issues. The possible recognized patterns are e.g. date formats, creation of e-mail addresses and telephone numbers, etc.

3.6. Domain analysis

The domain analysis provides information about possible values/value ranges that occur frequently. As an example, here is a column “Marital status” and “Gender”. After examining these columns, it is determined that the occurring values are found among the following: “single”, “married” or “divorced” and “M” or “W”. This domain can be used to derive rules and limit the allowed values. In addition, such a rule facilitates aggregation and increases the accuracy of the data.

The following Figs. 4–8 demonstrates a practical example of a publication list and analyze its data structure and data content using the data cleaner tool.

3.7. Functional dependency

This analysis determines the dependencies between individual columns. Here it is found out which attributes can be calculated or derived by other values. For this purpose, if-then rules are checked with a high confidence score.

Table 2 shows an example of the contents of the publication list in which the attribute “Publication Year” depends on the attribute “Title”.

3.8. Referential analysis

In this analysis, connections are made between multiple objects (in different relations). It filters out for more specific dependencies. Here, the expressions are used to identify the objects to be checked. The terms are Redundant Attributes, Joins, Orphans and Childless. Using the expressions of this analysis, reference rules can be specified or calculated.

Table 3 shows an example of the reference analysis. “Publications

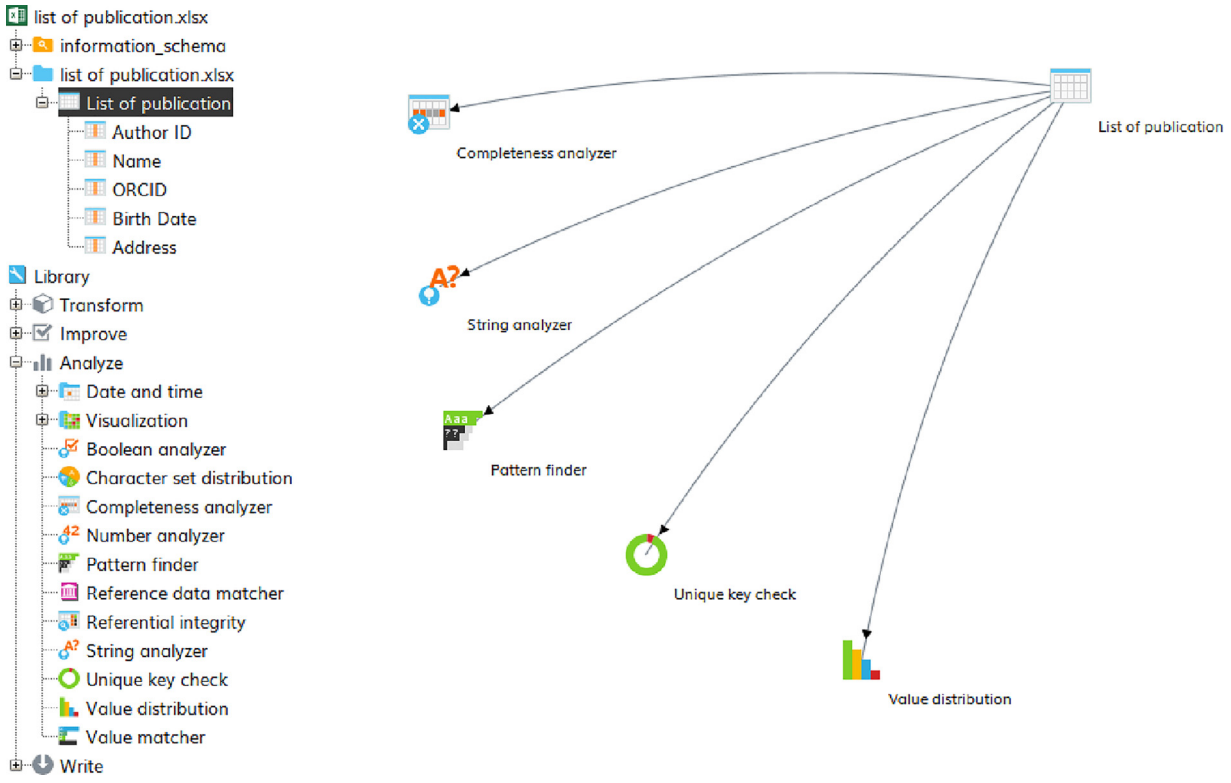


Fig. 5. Example of a Publication List in Data Cleaner Tool.

String analyzer

(5 columns)

	Author ID	Name	ORCID	Birth Date	Address
Row count	8	8	8	8	8
Null count	1	0	1	0	1
Blank count	0	0	0	0	0
Entirely uppercase count	0	0	2	0	0
Entirely lowercase count	0	0	0	0	0
Total char count	42	103	120	104	182
Max chars	6	19	19	19	37
Min chars	6	8	9	4	17
Avg chars	6	12,875	17,143	13	26
Max white spaces	0	2	0	1	5
Min white spaces	0	1	0	0	3
Avg white spaces	0	1,25	0	0,5	3,714
Uppercase chars	0	18	2	0	24
Uppercase chars (excl. first letters)	0	8	0	0	10
Lowercase chars	0	73	0	0	61
Digit chars	42	0	103	80	59
Diacritic chars	0	0	0	0	0
Non-letter chars	42	12	118	104	97
Word count	7	18	7	12	33
Max words	1	3	1	2	6
Min words	1	2	1	1	4

Fig. 6. String Analyzer.

List" is the child object that inherits from "Title and Year of Publications" the parent object.

A reference analysis of these two objects would show that the title "Databases" from the table "Publication List" is (orphan) and the titles "Project Management", "Data Analytics" and "Cloud Computing" from the table "Title and Year of Publications" are (childless). It would also display a link in the Title column.

Based on these results, you can derive reference rules that determine the cardinality between the two tables ("Publication List" and "Title and Year of Publications").

4. Discussion

Data profiling identifies the problematic data and automates metadata while allowing for the correction of typical data errors in the data. The scientific institutions can profile their source data to recognize structures, relationships and data rules. In addition, attribute analysis, reference analysis, functional dependency analysis, or profile data can be performed using custom rules.

In order to monitor the quality of the data in the RIS, the following developed meta-process flow can be used as a basis for the serving facilities and serve as a model or help to show how to analyze them in the facilities for data errors in RIS fixes and improves.

The following figure introduces the just mentioned meta-process flow for analyzing and improving data quality in RIS (Fig. 9).

At the beginning of the process flow, external and internal data sources of a device are collected by the management or technical personnel and this source data is profiled and all information containing the data is recognized. Then they derive some of the data rules and then use these derived rules to derive corrections. These fixes are performed

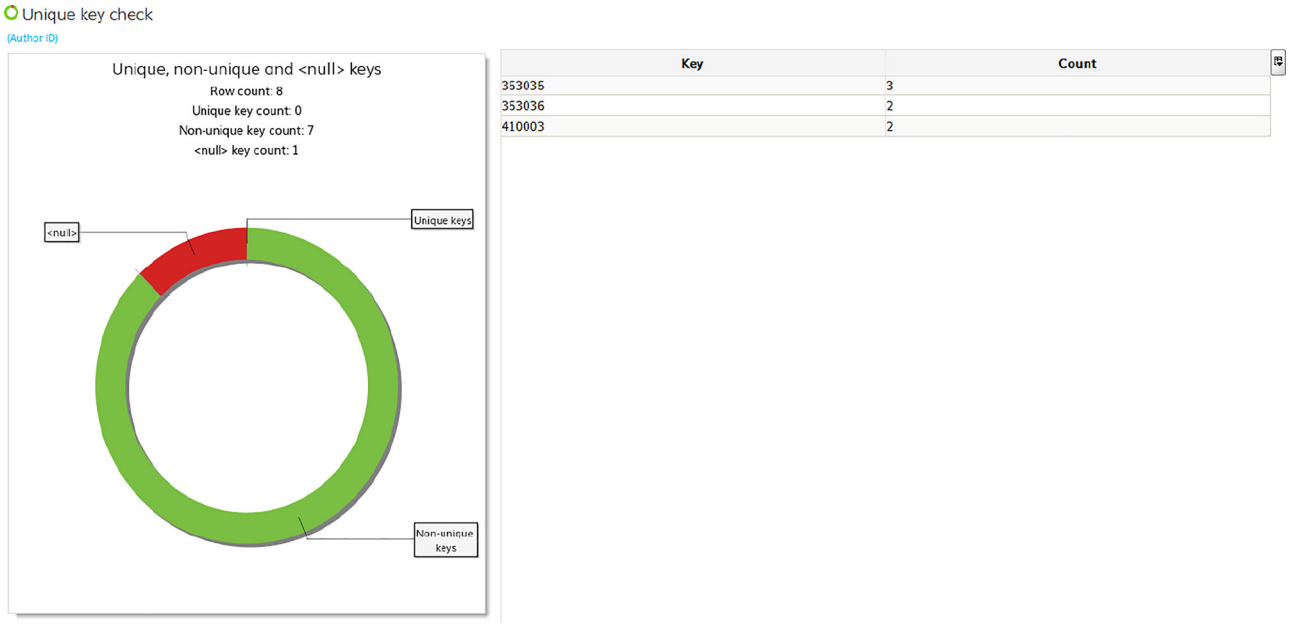


Fig. 7. Unique Key Check.

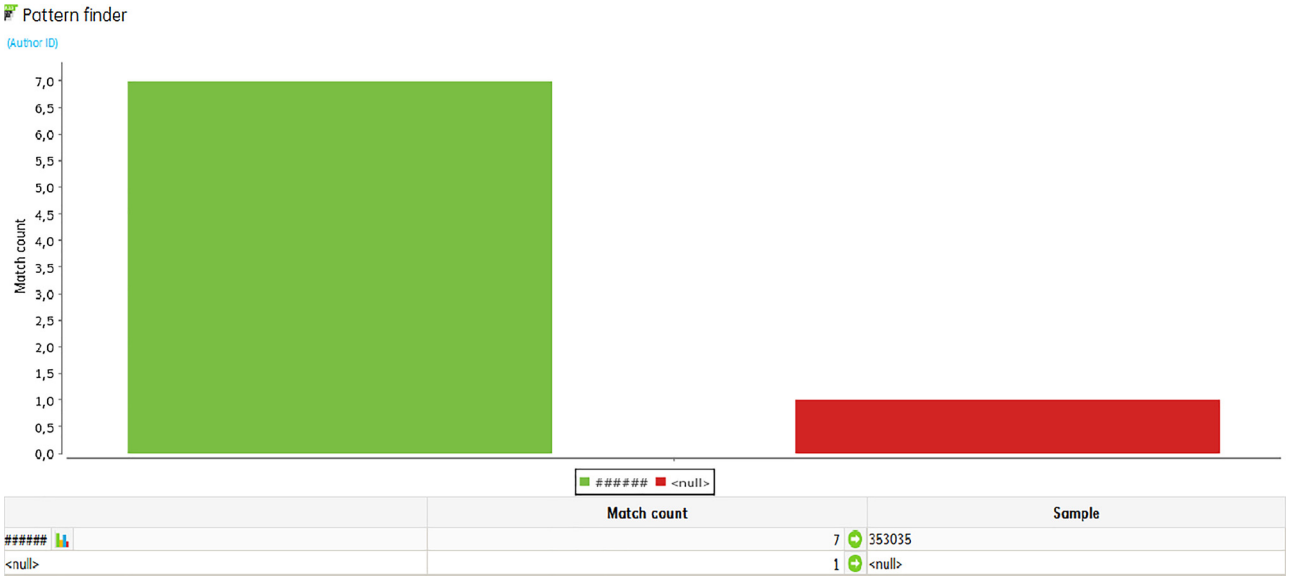


Fig. 8. Pattern Finder.

using the cleansing method to clean up the data in the target. Finally, the data loaded in the RIS will be presented. By means of portals, reporting and other front-end applications, the information coming from the system is visualized. Here, the processed information and analyzes are made available to the user in a clear form through various application components.

Table 2
Example for the Functional Dependency of a Publication List.

Author ID	Name	ORCID	Birth date	Title	P_Year
353035	Alien Scott	0000 – 0007-1222-2301	1965	Data Integration	2011
400015	Virginia Mic	0000 – 0123-1201-0111	1985	Big Data	2015
410003	Olivia Svenson	0450-1254-3598-F156	1983	Databases	2017

Table 3
Example for the Reference Analysis.

Author ID	Name	ORCID	Birth date	Title	P_Yea-r	
353035	Alien Scott	0000 – 0007-1222-2301	1965	Data Integration	2011	(Child)
400015	Virginia Mic	0000 – 0123-1201-0111	1985	Big Data	2015	
410003	Olivia Svenson	0450-1254-3598-F156	1983	Databases	2017	

Title	P_Year	
Data Integration	2011	(Parents)
Project Management	2014	
Big Data	2015	
Data Analytics	2015	
Cloud Computing	2016	

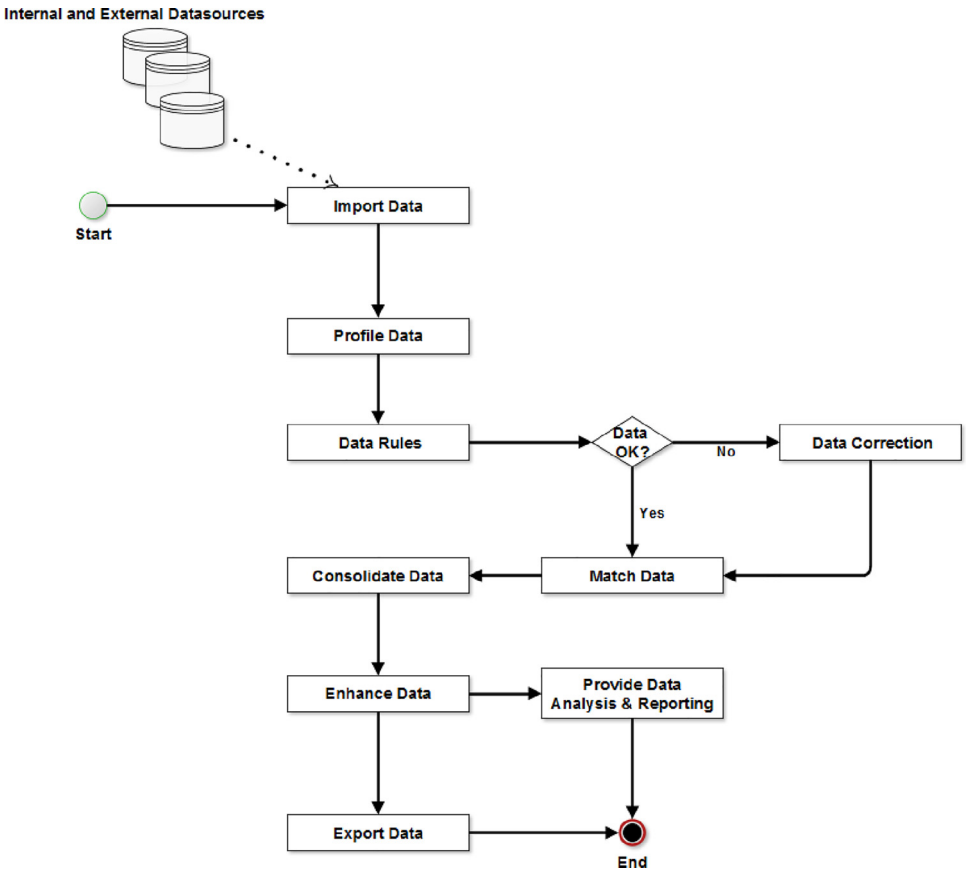


Fig. 9. Process Flow Diagram of Analyzing and Improving Data Quality in the RIS.

5. Conclusion and outlook

After data has been transferred to the RIS, different analyzes can be made. This includes assessing completeness, identifying patterns, hard duplicates, nulls, showing differences, and validating attributes. To identify redundant data sets that are not fully equivalent, data profiling methods have been developed. These methods were presented in this paper to detect, analyze and correct the data errors that have occurred in research information systems. Data profiling is seen as an important component in improving data quality before data can be integrated into a RIS. The institutions must have used data profiling as early as possible in order to get to know their data for the first time and derive DQ rules from this familiarization, which are then targeted and evaluated in the course of data quality measurements. Tools help with the implementation of data profiling. Profiling tools evaluate the actual content, structure and quality of the data. To do so, they check the relationships between data that exists within records, as well as relationships between data between records. With the appropriate tools, abnormalities and data errors can be detected even from the largest amounts of data, which can then be reformulated into additional data quality rules together with the responsible department staff or data administrator.

Data Profiling tools are primarily commercial and available for both small application contexts and comprehensive application suites for data quality and data integration. In recent years, a market for data profiling is also developing as a service. The use of tools for data profiling is worthwhile, since they significantly reduce the resource requirements. Especially with repeated use, the effort is much lower than without use of tools. In addition, obtained data profiling results can be used quickly and easily.

References

Apel, D., Behme, W., Eberlein, R., & Merighi, C. (2015). *Successfully control data quality. practical solutions for business intelligence projects* (3rd Ed.). Dpunkt Verlag Revised and Expanded Edition.

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

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

Olsen, J. (2003). *Data quality – the accuracy dimension.* San Francisco: Morgan Kaufmann Publishers.

Otmane Azeroual is a researcher at the German Institute for Higher Education Research and Science Studies (DZHW) in Berlin. After studying Business Information Systems at the University of Applied Sciences (HTW) Berlin, he began his Ph.D. in Computer Science at the Institute for Technical and Business Information Systems (ITI), Database Research Group of the Otto-von-Guericke-University Magdeburg and at the Department of Computer Science and Engineering of the University of Applied Sciences (HTW) Berlin.