International Journal of Information Science and Management
Vol. 21, No. 4, 2023, 189-213
DOI: https://doi.org/10.22034/ijism.2023.1977986.0 / DOR: https://dorl.net/dor/20.1001.1.20088302.2023.21.4.9.5

Original Research

Measuring Data Quality of Theses and Dissertations in the Data Preparation Stage of Registration Systems

Mohammad Javad Ershadi
Associate Prof., Information Technology Department, Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran.
Corresponding Author: Ershadi@irandoc.ac.ir
ORCID iD: https://orcid.org/0000-0002-7006-7580

Amirmahan Mohseni
B.Sc., Industrial Engineering Department, Islamic Azad University, Central Tehran Branch, Tehran, Iran.
amirmahan1999@gmail.com
ORCID iD: https://orcid.org/0009-0003-8533-1502

Seyed Taghi Akhavan Niaki
Professor, Industrial Engineering Department, Sharif University of Technology, Tehran, Iran.
niaki@sharif.edu ORCID iD: https://orcid.org/0000-0001-6281-055X

Received: 09 November 2022
Accepted: 13 March 2023

Abstract

Today, academic research plays a very influential role in the economic development of countries. These researches are often recorded and disseminated in the form and structure of theses and dissertations in scientific institutes. The better the quality of this data in the systems that collect and distribute it, the more it can be used and exploited by organizations and businesses. Therefore, providing this data requires proper monitoring to put the output of the recording and dissemination process in good condition. This paper offers a framework for evaluating theses and dissertation data quality. In the framework, the data inconsistency coding structure is introduced and presented in Word and PDF files and in the form of metadata (bibliographic information). The approaches presented in data quality methodologies (TDQM and DWQ) are also used to provide solutions to improve data quality in the provisioning phase. At this stage, approaches such as owner attribution to data or process, root cause analysis, process control, and continuous monitoring are considered. The focus group method determines the operational strategies for quality improvement. Finally, process-oriented techniques, such as quality control checklists and image processing, and data-driven approaches, such as data cleansing, are localized and developed in this section to improve the quality of theses/dissertation documents. The provided improvement solutions were categorized into two different groups. Guiding the user in the "Theses/Dissertations" registration process is identified as a process-driven category. On the other hand, introducing a specific format for "Theses/Dissertations" files and resolving the quality issues of PDF files were among the data-driven solutions.

Keywords: Data Quality, Information Quality, Quality Control, Thesis/Dissertations Registration System, Incompatible Analysis, Data Preparation.

Introduction

In today's world, the quality of data and information is one of the most critical concerns of organizations and has become very influential in information systems. The rapid growth of data warehouses and direct access of managers and users to various data sources, on the one hand, and the need of users for quality information, on the other hand, have increased the need to pay attention to data quality and its awareness. In the studies, researchers identified data quality as one of management information systems' six main dimensions (MIS) dimensions. Research Information Systems (RIS), one of the most well-known management information systems in research, has recorded, controlled, and disseminated research. Today, research management and policy-making are inconceivable without them. Hence, the quality of data and research information in a RIS and its evaluation and measurement are management strategies.

The "Ganj" system is one of the critical systems in disseminating research outputs in Iran. For this reason, improving the quality of search results of this system is one of the basic strategies to enhance its performance. The root of the quality problems observed in the results obtained from the search process, as well as the increase in user satisfaction with the "Ganj" system, depends on the correct and quality operation of the registration system, which is at the beginning of the registration, organization, and dissemination process. Based on the quality plan of the registration system, the discrepancies observed in the documents are identified, and the researcher is informed. In the current situation, the inconsistencies observed by the staff of the Registry and the provision of information in the process of registration of "Thesis/Dissertations" documents are not structured. This means that after observing any discrepancies in the documents registered by the graduates, each employee reports the discrepancies in the registration system in the space allocated based on their knowledge of the desired status of the documents. On the other hand, in the existing structure of the registration system, there is no specific category for registering inconsistencies in "Theses/Dissertations" documents. This situation brings the following problems in terms of process and quality management.

• The current situation does not allow proper analysis of why the documents are returned due to non-compliance with the standards. This problem results in an incorrect analysis of the cause of inconsistency and identification of the frequency of each cause, an inability to provide solutions to improve the process, and the registration system cannot be presented more efficiently and effectively.

• It is impossible to fully train new staff (if needed) and provide a framework for developing the system's capacity to control more credentials. This means that in the current situation, if the system's capacity increases and new users need help, each new person must be next to an educator. Only if the educator has a lot of accuracy and patience can he educate the new person about the quality problems and the reasons for returning them. This method is not systemic and can lead to errors and forgetting some causes.

In addition to the above, standardization and analysis of the causes of inconsistencies in the registration system will make the integration of registration and indexing processes more straightforward in the long run and lead to the development of large systems. It also facilitates more detailed analyzes of the status of different universities in terms of the quality of the documents they register. This study aims to consider the many benefits of standardizing causes in retrieving "Theses/Dissertations" documents by providing a standard framework for the discrepancies observed in the documents of graduates of the entire country, providing a proper

analysis of the discrepancies and their extent.

Research background

Previous works emphasized that data quality generally addresses areas of semi-structured, unstructured, and multimedia data types. These studies focus on semi-structured and unstructured data in terms of data quality to a strong historical relationship between data quality and database design and structure. While biased, integrated approaches focus primarily on structured data that provides the most information resources in organizations (Avenali Batini, C., Bertolazzi, P. & Missier, 2008). Today, semi-structured and unstructured data are much more common as organizational resources. Knowledge management, web protection, and GIS are essential research areas in the context of data quality. Rahman, Mannan, Hossain, Zaman and Hassan (2018) examined data quality techniques for semi-structured and unstructured data. Improving data quality techniques for unstructured and semi-structured data in these areas requires more correlation between different specialties, prolonging data quality improvement (Taleb, Serhani, M. A. & Dssouli, 2018).

Concerning the various dimensions of information quality, concerns about the quality and reliability of information are increasing among online searchers, information professionals, and the general public (Russell-Rose, Chamberlain, J. & Azzopardi, 2018). Based on this principle, information professionals assist clients in understanding the limitations of data sources, fostering critical thinking about the data, and assessing the accuracy and validity of the information collected. Therefore, quality improvement is a vital issue today and needs to be considered an essential strategy for organizations. Throughout the data production chain, from the database producer to the end-user, building collaboration and a comprehensive commitment to the system to improve quality continuously is essential and inevitable.

The relationship between data and process quality is a significant part of research because of organizations' relevance and diversity of business process characteristics. Different effects of data quality have been studied at three specific levels: operational, tactical, and strategic (Batini, Cappiello, Francalanci & Maurino, 2009). Data quality and its relationship with the quality of services, products, business operations, and consumer behavior in many general terms have been studied in this research. Standard and common data quality models such as Wang and Strong (1996) and the Data Warehouse Institute model in 2006, as well as the framework presented by Vaziri, Mohsenzadeh and Habibi (2017) and Khosroanjom, Ahmadzade, Niknafs and Mavi (2011), are studies in the field of data quality that considered various aspects[1]. Also, today, models such as the model presented by Kwon, Lee and Shin (2014) for the metadata quality in the organization are available. Another main framework in the background of the research is the extensive classification proposed by Ershadi, Jalalimanesh and Nasiri (2019) for different types of discrepancies in the registration system.

In addition to the above, another group of studies has viewed data quality measurement as a goal and presented various reports on measuring its aspects. Customer domain and data measurement are currently considered by some researchers (Peltier, Zahay, D. & Lehmann, 2013; Ershadi & Omidzadeh, 2018; Petrović, 2020; Sharma, 2020). Training metadata is another practical aspect that Ochoa and Duval (2006) developed as a measuring model. This model has received more attention in recent research focusing on metadata (Nikiforova, 2020). A similar framework is provided in real estate registry data to assess the quality of the database, where categories such as relevance, accuracy, and cost of poor-quality data are considered

(Heinrich et al., 2009).

Some solutions are related to data improvement, while others focus on production and processes, aiming to provide improvement. Solutions such as guiding the dissertation registration process users fall into the process-oriented category. Process-oriented strategies for improving data quality have been considered by Azeroual, Saake, Abuosba and Schöpfel (2020), Falge, Otto and Österle (2012), and Michelberger, Mutschler and Reichert (2011). On the other hand, introducing a specific format for dissertation files and resolving the quality issues of pdf files fall into data-driven solutions. In this regard, Azeroual et al. (2020) considered solutions to address inconsistencies in research data management. Such solutions were also presented by Sidi, Panahy, Affendey, Jabar, Ibrahim and Mustapha (2012) and Batini, Cabitza, Cappiello & and Francalanci (2008).

Determining strategies for improving data quality may be considered a managerial policy of an organization (Azeroual, Ershadi, Azizi, Banihashemi & Abadi, 2021) and needs strategic tools such as the SWOT approach. For electronic health report (HER) data Kapsner et al. (2019) provided a comprehensive framework that includes solutions for innovative medicine. Elouataoui El Alaoui, I., El Mendili, S. and Gahi (2022) developed a big data quality framework using weighted metrics. Also, data mining as a technical method can help data quality experts provide effective improvement solutions (Edris Abadi, Ershadi & Niaki, in Press).

Regardless of the strategy adopted to improve data quality, after studying the research background, a set of common models and methodologies in data quality and information were identified and categorized, introduced in Table 1.

Table 1
Data and Information Quality Models

Abbreviation	Main name	References
TDQM	Total Data Quality Management	Wang (1998)
DWQ	The Data Warehouse Quality Methodology	Jeusfeld, Quix & Jarke (1998)
TIQM	Total Information Quality Management	English (1999)
AIMQ	A Methodology for Information Quality Assessment	Lee, Strong, Kahn and Wang (2002)
CIHI	Canadian Institute for Health Information Methodology	Long & Seko (2005)
DQA	Data Quality Assessment	Pipino, Lee and Wang (2002)
IQM	Information Quality Measurement	Eppler & Muzenmaier (2002)
ISTAT	Istat Methodology (Italian National Bureau of Census)	Falorsi & Righi (2008)
AMEQ	Activity-Based Measuring and Evaluating for Product Information Quality Methodology	Su & Jin (2007)
COLDQ	Cost-Effect of Low Data Quality	Loshin (2001)
DaQuinCIS Data	Data Quality in Cooperative Information Systems	Scannapieco, Virgillito, Marchetti, Mecella and Baldoni (2004)
QAFD	Methodology for The Quality Assessment of Financial Data	De Amicis, Barone, & Batini (2006)
CDQ	Comprehensive Methodology for Data Quality Management	Batini et al. (2008)
EHR Data Quality Framework	Data Quality Framework for Electronic Health Record	Kapsner et al. (2019)
Big Data Quality Framework	An Advanced Big Data Quality Framework Based on Weighted Metrics	Elouataoui et al. (2022)

On the other hand, in each of the methodologies mentioned in Table 1, various methods for improving data quality are summarized in Table 2.

Table 2
Comparison of Data Quality Models Based on Improvement Steps

Abbreviation	Cost estimation	Assign the owner to the process	Assign the owner to the data	Strategy selection	Identifying the root causes of errors	process control	Design data recovery solutions	Process redesign	Improvement management	Continuous monitoring
TDQM	+	+	+	+	+			+	+	+
DWQ	+		+	+	+		+		+	
TIQM	+	+	+	+	+		+	+		+
DQA					+					
ISTAT				+	+		+	+		
AMEQ					+					+
COLDQ	+			+	+	+	+	+		+
DaQuinCIS Data				+	+					
CDQ	+	+	+	+	+	+	+	+		

Although some studies mentioned the use of control charts to monitor the data involved in research information systems (Ashtarian Esfahani, Ershadi & Azizi, 2020) or provided root analysis (Ershadi & Ershadi, 2018), measuring and analyzing these data in a structured way that can improve data production and processes leading to management decisions has not been considered. Also, the distinguishing feature of data quality assessment in research information systems compared to other systems is that the controls and their results are in the form of text and do not have a specific and predetermined structure. This makes it more challenging to classify qualitative characteristics into standard structures such as accuracy, completeness, and relevance and reinforces the need to provide a native structure for evaluation. For example, the quality control expert of a dissertation text file inserts the defects observed in the file in a few phrases in the system, making it very difficult to map it with data quality standards.

Therefore, this research aims to provide a framework for measuring and improving the data quality of the Thesis/dissertation. Other sub-goals are presented as follows. To give a proper statistical analysis of the quality of scientific or research documents, a correct and predetermined structure is needed that is also reproducible. In addition, the framework presented in this research can be the basis of data quality research in other databases such as journals, books, and conferences. The current paper identifies and classifies discrepancies and provides statistical analysis of the results of these discrepancies and executive guidelines. In the next section, the research method is addressed.

Materials and Methods

As mentioned in the previous section, providing a clear and codified framework for evaluating and analyzing the quality of Theses/Dissertations documents is critical for developing and improving the data quality of this field of work. The approach used in this study, which is a continuation of many previous studies to evaluate and improve data quality, is based

on the Total Data Quality Management (TDQM) framework, which, if properly implemented, can lead to continuous data quality improvement in the organization (Wang & Stuart, 1990).

Research approach

The TDQM approach supports continuous data quality improvement by following four steps: planning, measuring, analyzing, and improving (Figure 1). This framework has been the basis of many data quality improvement programs and is still used in many studies today (Cahyono & Sucahyo, 2020; Liu, Mutschler & Reichert, 2020). Each of the main stages of TDQM shown in Figure 1 includes various activities, the most important of which are mentioned below.

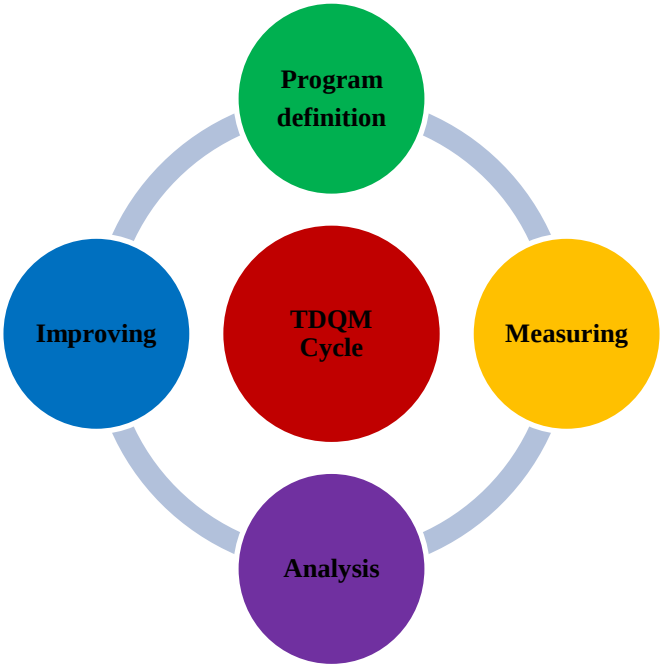

Figure 1: TDQM Cycle (Wang & Strong, 1996)

At the Program definition stage, the main framework for evaluating data quality, indicators, and how to assess them is determined. Then, in the Measuring stage, the data quality is measured based on the framework and actions of the previous step. The data quality measurement results are analyzed in the third step. The extent of the differences between the data quality indicators and their predefined position is determined using appropriate techniques at this stage. Finally, in the fourth step, data quality improvement programs are codified. At this stage, the actions taken are essential in two ways: (1) changing the value of the data and (2) changing the data production processes.

Introducing the focus group method

One of the data collection methods is in mixed and qualitative designs of focus groups. Focus groups are organized discussion sessions, a group of people who can be considered the focus of a thematic discussion. Then, through group interviews, these people represent their opinions and experiences. Focus groups are superior to other research methods because their primary purpose is to explore people's attitudes, feelings, beliefs, experiences, and reactions that are not understood by other methods such as observation, personal interviews, and survey questionnaires.

Focused group discussions are a helpful way to bring people together to create new insights and ideas about the research topic. The benefits of using focus groups include quickly gaining insights on viewing and discussing an issue with complementary or contrasting perspectives. Participants are invited to share their views, and as they can hear and reflect on what others in the group are saying, they can also respond to what they have heard. This open and discursive approach to data collection can create a consensus in the group or create tensions and differences between group members when interacting (Robinson, 2019). Unlike interviews, the researcher assumes an environmental role in a focused group discussion rather than a pivotal role (Nyumba, Wilson, Derrick & Mukherjee, 2018).

The focus group method consists of four main steps:

1. Research design
2. Data collection
3. Analysis
4. Results

Research design

This process begins with identifying the primary purpose and defining the main objectives of the research. Based on the research goals, a list of questions is prepared to guide each focus group discussion session.

Data collection

The primary data collection methods during group discussions include recording audio and video, taking notes, and observing participants. However, each of these methods has different advantages and disadvantages, and researchers must consider specific topics of discussion when choosing a data collection method.

Analysis

At this stage, observational and qualitative data are examined and analyzed, and the answers given are categorized by content analysis and Ethnography Analytic Techniques.

Results

After analyzing the data, the results should be combined into a coherent report. The report can be presented as a narrative. In this report, in addition to the critical quotations of the participants, the information of the participants, such as gender, age, and level of education, should also be recorded.

Research sample

To implement the provided structure in quality control of the Theses/Dissertations, 10,000 registered documents were extracted as research samples. Quality appraisal of these documents was implemented based on the provided framework, and results were analyzed.

Research steps

The current research is based on the four-step TDQM approach and the following framework. In the first step, a framework for measuring metadata quality is defined. After determining the standard framework for evaluating the quality of documents in the

"Theses/Dissertations" registration system in the country, the structure provided by experts in science and technology of data quality and its validity is evaluated. Then, in the second step, based on the framework defined in the first step, the quality of data and metadata is measured in the registration system. The results obtained in the second step are analyzed in the third step. Finally, suggestions for improving data quality based on the performed analyses are presented in the fourth step. Figure 2 summarizes the method on which this research is based.

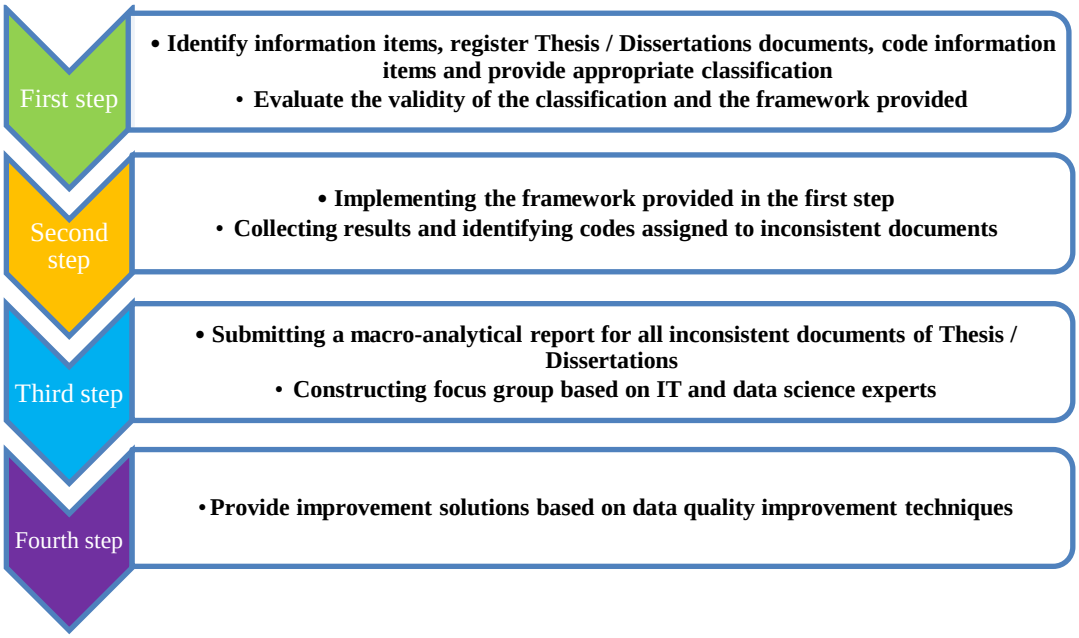

Figure 2: Research Steps to Measure and Analyze the Quality of "Thesis / Dissertations" Documents

In the continuation of this research, the results will be shown in four steps in the fourth section.

Results

Based on the four steps in the third section, this section will show the results in four sub-sections corresponding to each step.

Providing a macro framework to classify inconsistent materials

In the current situation, the registration and information management department staff first controls the metadata registered by the user in the registration system. Then, the PDF file and, finally, the Word file are checked. Accordingly, the discrepancies observed in the "Theses/Dissertations" documents are divided into these three general categories. In other words, the inconsistencies observed in the registration system either refer to the information entered in the system by the user (metadata) (first category) or to the Word file uploaded (second category) or the PDF submission file (third category). Hence, inconsistencies are divided into three main categories. In each of the three metadata groups, inconsistencies in the information items registered by the user can be categorized. Figure 3 illustrates the categorization described. Metadata contains information registered in the registration system by users and includes various information items such as name, surname, name of the supervisor, etc. As shown in Figure 3, the second level deals with information items. Errors may occur by the user in any information item; in other words, the type of discrepancy observed depends on

it (Level 3). For example, the user may misspell his name in the bibliographic information section of the registration system (metadata) or may not enter his full name. So, this discrepancy is related to metadata at the first level. Also, the "name" information item is the second level of this discrepancy. The third level, "no suffix inserting," will be the type of inconsistency. Next, an email is sent to the user, indicating the discrepancy observed. Accordingly, the last part of the structure of inconsistencies is dedicated to the email sent to the user. In other words, to standardize the notification process to the user registering for the Theses/Dissertations, a standard text is sent to the user to edit their document based on the email description after selecting the type of inconsistency.

As seen in the structure of Figure 3, in the first level, the structure of a data-determining inconsistency or metadata of the "Theses/Dissertations" document is examined. This level only specifies inconsistency in the metadata or data category (Word or PDF files). At the second level, the information item will be the basis for classification from the discrepancies observed in the "Theses/Dissertations" metadata in the registration system. Therefore, at the next level, various information items of the registration system must be specified.

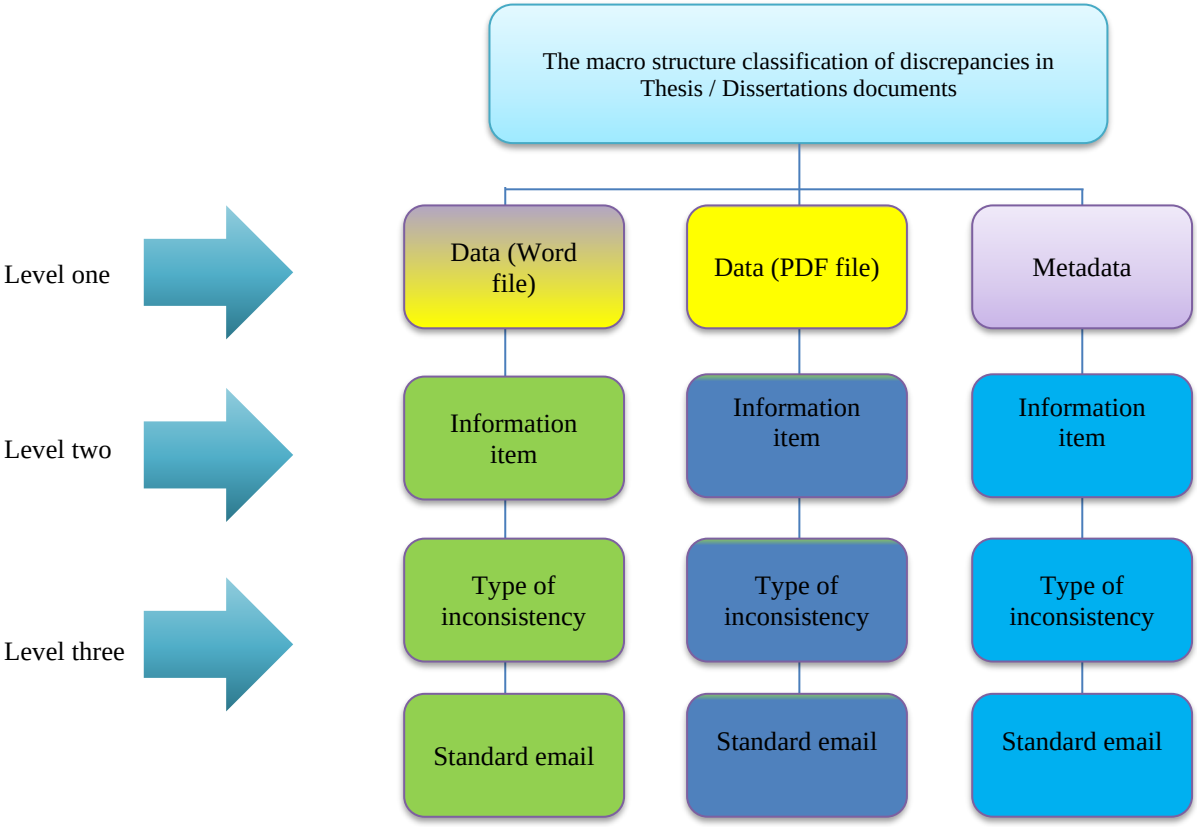

Figure 3: Macrostructure of inconsistent cases in the document registered in the " Theses/Dissertations " registration system

Figure 4 shows the broad categorization of the information item in "Theses/Dissertations" metadata. Similarly, in the data recorded in the registration system (Word and PDF files), the second level of inconsistency classification includes various information items in this data.

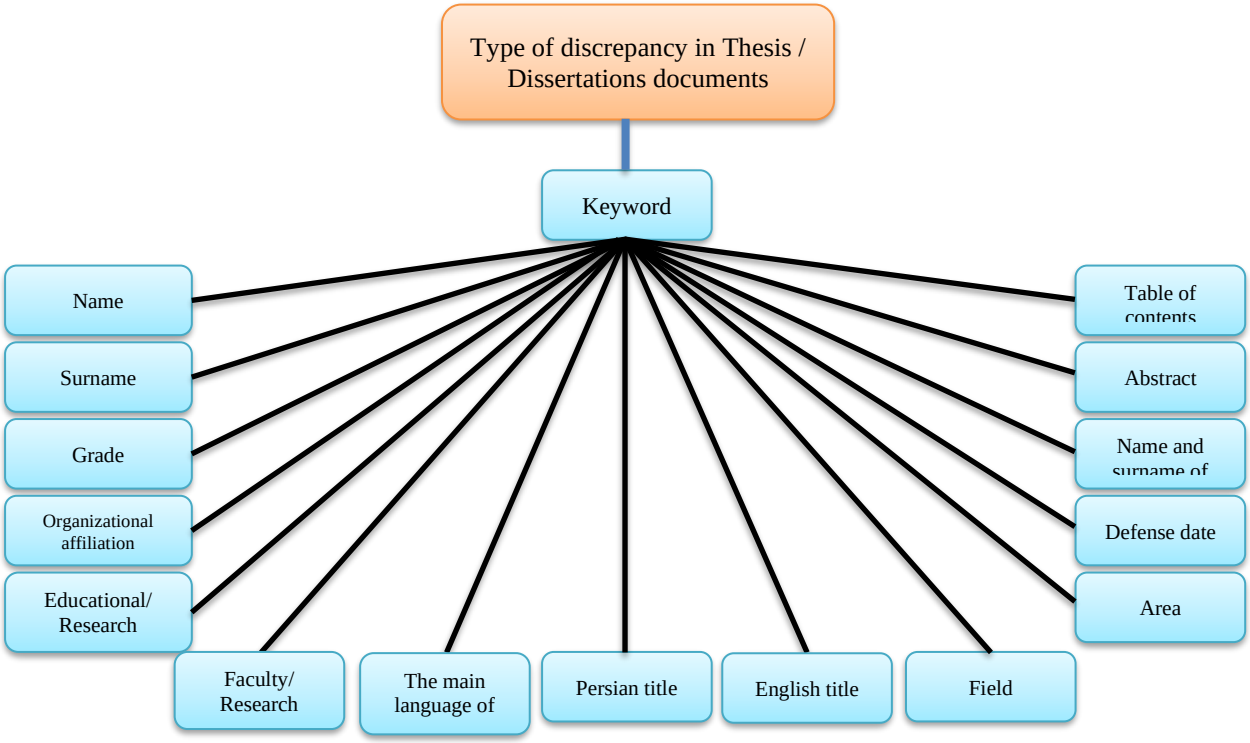

*Figure 4: Extensive structure of the second level (information item) in **metadata** inconsistencies in the documents registered in the "Thesis/Dissertations" registration system.*

Figure 5 shows the broad structure of these discrepancies.

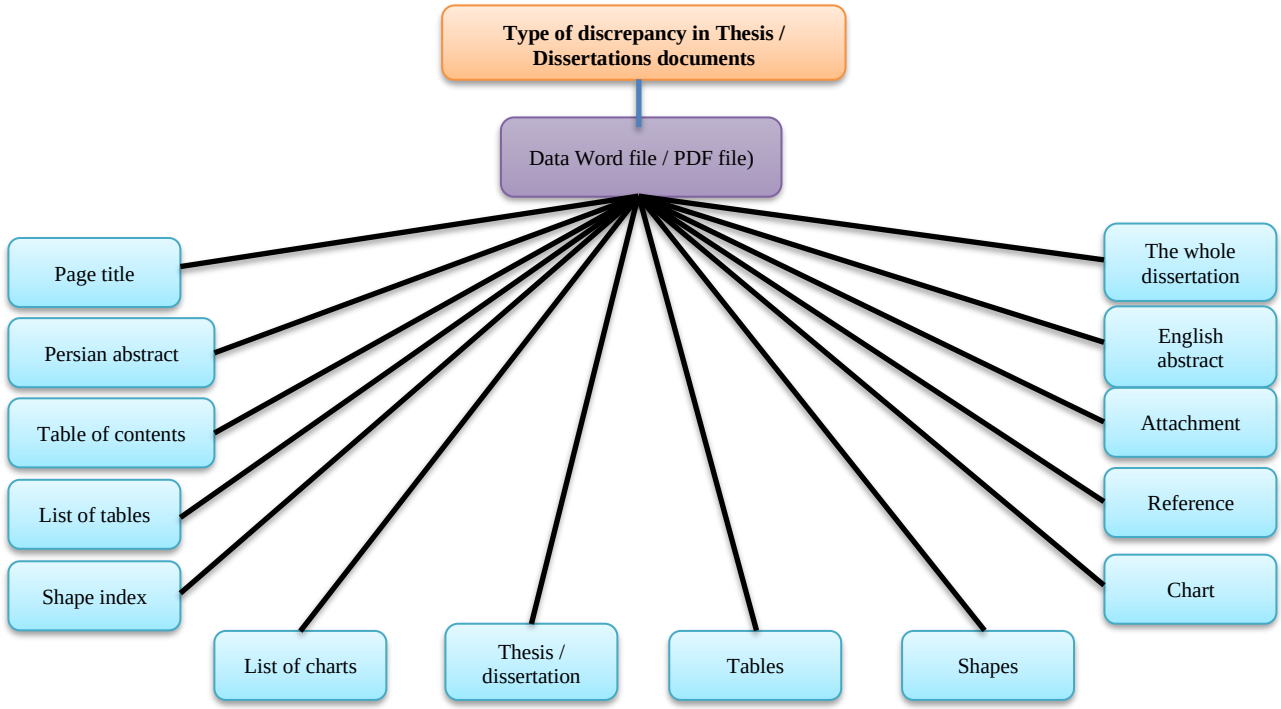

*Figure 5: Extensive structure of the second level (information item) in **data** inconsistencies in the documents registered in the "Thesis / Dissertations" registration system*

Mismatch coding structure

To improve the access of the experts controlling the registration system to the standard structure of inconsistencies, the codes given to each inconsistency are provided in a standard form so that the code can be assigned easily. As shown in Figure 3, the discrepancies are presented in three categories: metadata, Word, and PDF. For this reason, the codes of these three categories start with the numbers 1, 2, and 3, respectively. The general structure of this coding can be seen in Figure 6.

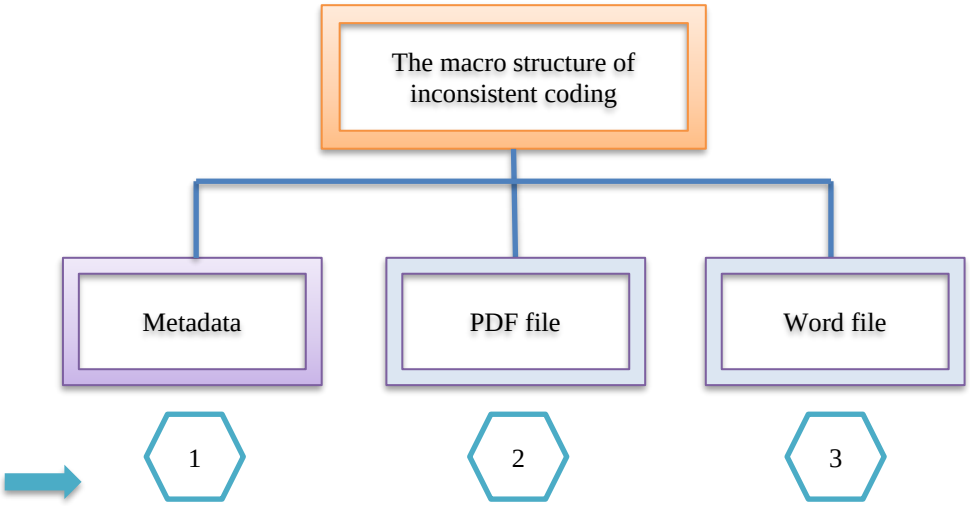

Figure 6: Macrostructure of coding inconsistent items in " Thesis / Dissertations " documents

On the other hand, the second digit is related to information items, where an attempt has been made to provide a suitable structure for the second digit. Figures 7 and 8 show the structure of the second digit of the discrepancies for the metadata (initial digit 1), the PDF file (initial digit 2), and the Word file (initial digit 3), respectively.

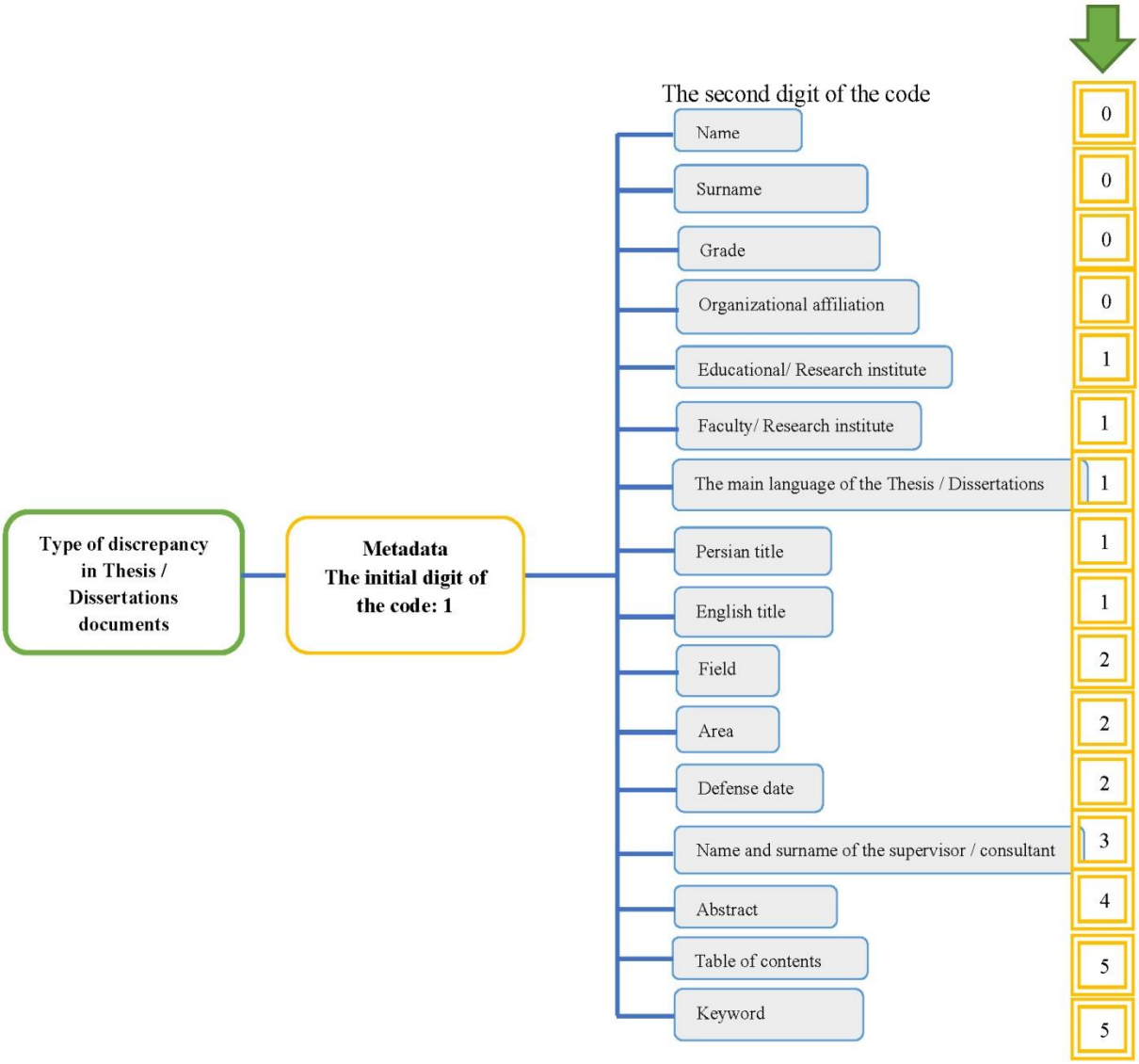

Figure 7: Coding structure of the second digit in cases of inconsistencies in the metadata of the "Thesis / Dissertations" registration system

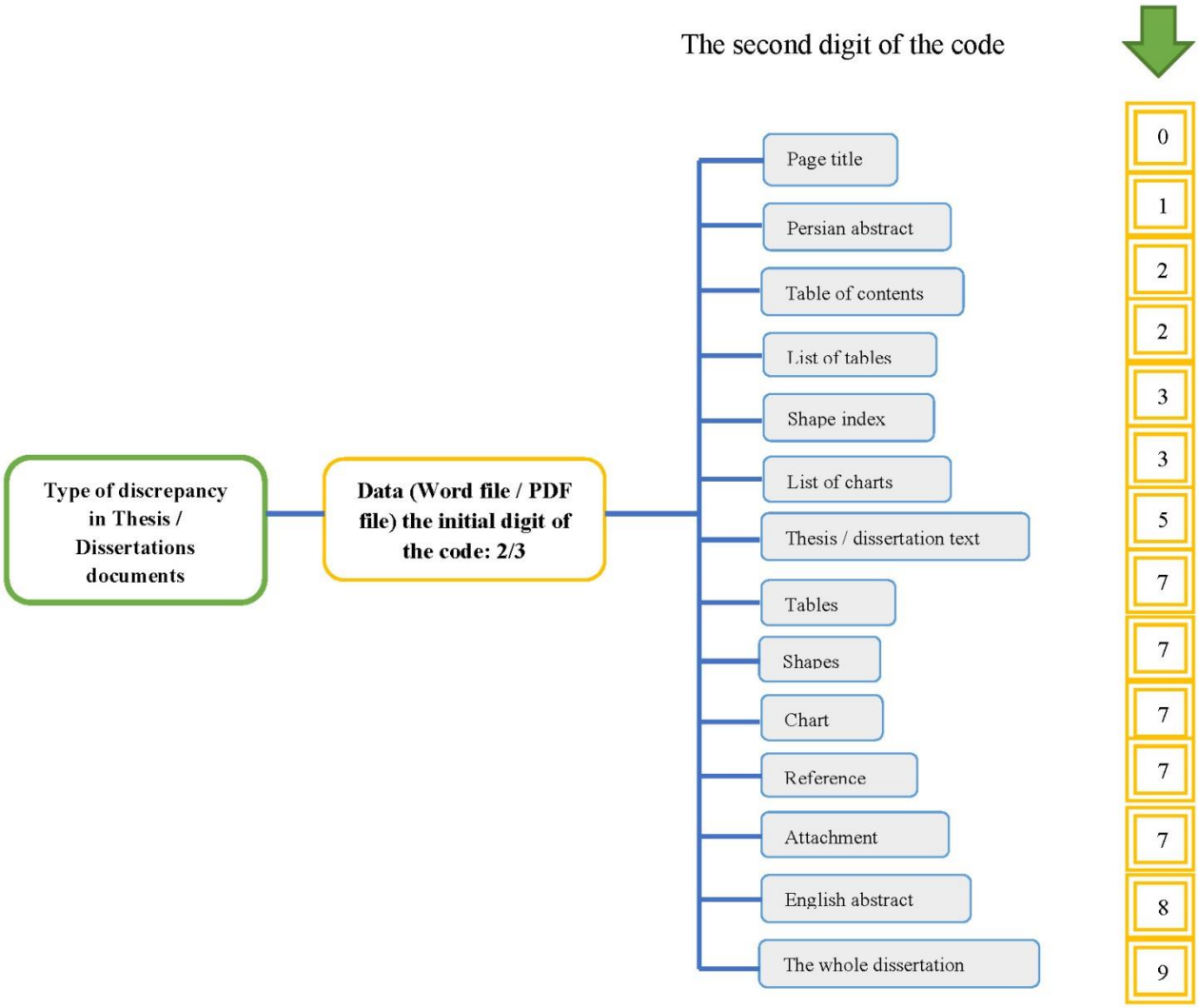

Figure 8: The coding structure of the second digit in cases of inconsistencies in the PDF and Word files of the " Thesis / Dissertations " registration system

After determining the first and second digits of the mismatch code structure, the third digit starts and continues one after the other. Given that data quality analysis focuses on the first two digits, the structure of the third digit (due to limited article size) is not mentioned. A working group of data quality/metadata experts is formed to evaluate the validity of the prepared framework. Iran's Science and Technology Information Deputy, Director of Information Organization and Analysis, Director of Information Registration and Provision, and three key experts working to control "Theses/Dissertations" documents form this working group. Each working group member evaluates and reviews the prepared draft from various technical and practical aspects. Ease of use for registration system experts, ease of understanding for the user of the registration system, the ability to analyze and summarize, and the completeness and comprehensiveness of the codes and structure provided are the main criteria in reviewing the validity of the proposed structure. Finally, after finalizing the structure and the codes, the second step of the research, namely the implementation of this structure, began, which is presented as follows.

Data and metadata quality evaluation of "Theses/Dissertations"

In this step, control experts use the standard framework presented in the previous section at the operational level. In this regard, the expert defines the standardized inconsistencies of the registration system based on the coding so that experts can use the new structure when evaluating each "Theses/Dissertations" document. The increased validation structure of the discrepancies is used experimentally over two months. At the end of the evaluation phase of each "Theses/Dissertations" document, if the document is considered inconsistent, one or more of the standardized codes in the second step are selected and registered in the system. These codes can be reported at any time desired by the control expert, providing a good platform for analyzing the results in other research steps. In the third step of this research, an analysis of the results obtained in the second step is presented.

Analysis of quality assessment results

After receiving the information and the codes selected for returned documents (10,000 documents) (the structure of this code as shown in section 4.1), each code's frequency is first obtained. Then, each code's observation percentage is obtained by dividing the number of views by the total number of repeats of all codes. To increase the accuracy in analyzing the results of data quality evaluation, the results in this section are presented in two subsections. In the first subsection, regardless of the type of degree in a specialized field such as engineering or humanities, a macro-analysis of all data and metadata is provided. Then, as the specialized field of the graduates significantly affects the quality of the registered documents, another analysis appropriate to the specialized field of the documents is presented. One of the essential advantages of this analysis is that instructions based on the specialized domain of users can be provided when entering information into the registration system. For example, a guide given to an engineering student may differ from a science student who uses Latex software to write the text of their dissertation.

In what follows in subsection 4.4.1, the results of the macro analysis of the documents are reported.

Results of quality assessment analysis in all inconsistent documents

Data and metadata codes are entirely put together with no separations. Also, no particular categorization is performed on the documents (such as specialized fields). Table 3 shows the observed frequency of all codes separately for each code. Figure 9 provides a better comparison of the information obtained from Table 3. It should be noted that for the visibility of the repetitive codes and to compare them properly, only the codes accounted for more than one percent of all the codes presented in this diagram.

Table 3
Frequency and Percentage of Observed Causes of Inconsistency in All Documents

Row	Code name	Frequency	Observation percentage	Row	Code name	Frequency	Observation percentage
1	125	963	9.26	41	115	50	0.48
2	200	707	6.79	42	130	48	0.46
3	280	575	5.53	43	101	47	0.45
4	300	540	5.19	44	354	46	0.44
5	256	463	4.45	45	225	45	0.43

Row	Code name	Frequency	Observation percentage	Row	Code name	Frequency	Observation percentage
6	140	443	4.26	46	353	44	0.42
7	380	442	4.25	47	128	43	0.41
8	158	440	4.23	48	135	42	0.40
9	147	408	3.92	49	328	42	0.40
10	223	296	2.84	50	155	41	0.39
11	351	288	2.77	51	284	41	0.39
12	119	263	2.53	52	230	37	0.36
13	251	240	2.31	53	333	36	0.35
14	220	219	2.10	54	143	34	0.33
15	355	211	2.03	55	255	34	0.33
16	148	189	1.82	56	129	33	0.32
17	141	177	1.70	57	121	32	0.31
18	203	168	1.61	58	210	32	0.31
19	142	154	1.48	59	352	32	0.31
20	303	147	1.41	60	156	31	0.30
21	146	137	1.32	61	325	30	0.29
22	320	137	1.32	62	287	29	0.28
23	323	131	1.26	63	124	28	0.27
24	391	124	1.19	64	126	25	0.24
25	105	108	1.04	65	201	25	0.24
26	136	88	0.85	66	292	25	0.24
27	133	81	0.78	67	330	25	0.24
28	117	80	0.77	68	257	23	0.22
29	116	75	0.72	69	301	23	0.22
30	228	70	0.67	70	213	21	0.20
31	291	68	0.65	71	294	21	0.20
32	293	61	0.59	72	384	21	0.20
33	393	60	0.58	73	111	19	0.18
34	233	59	0.57	74	102	18	0.17
35	394	58	0.56	75	122	18	0.17
36	254	57	0.55	76	387	18	0.17
37	252	56	0.54	77	238	17	0.16
38	253	56	0.54	78	104	16	0.15
39	123	55	0.53	79	153	16	0.15
40	115	50	0.48	80	157	15	0.14

The results presented in Figure 9 show that the most inconsistency is related to the quality of metadata (information recorded in the system by the user) (Code 125). In other words, the defense date problem and its inconsistency with the "Theses/Dissertations" title have been seen in more than 9% of the returned documents. There are four other data codes in the five principal codes used in inconsistent documents, except for the first code related to the user's defense date (PDF and Word files). Code 200 (second rank), the discrepancy between the Persian title page in the PDF file and the information registered in the system, has been seen in 6.67% of the returned documents. Code 280 (third rank) refers to the English title page. Inconsistency of this page with the information entered in the system has been seen in 3.53% of inconsistent documents. In 5.19% and 4.45% of the returned documents, Codes 300 and 256, the Persian title page problem in the Word file, and the presence of white pages in the dissertation have the fourth and fifth ranks, respectively.

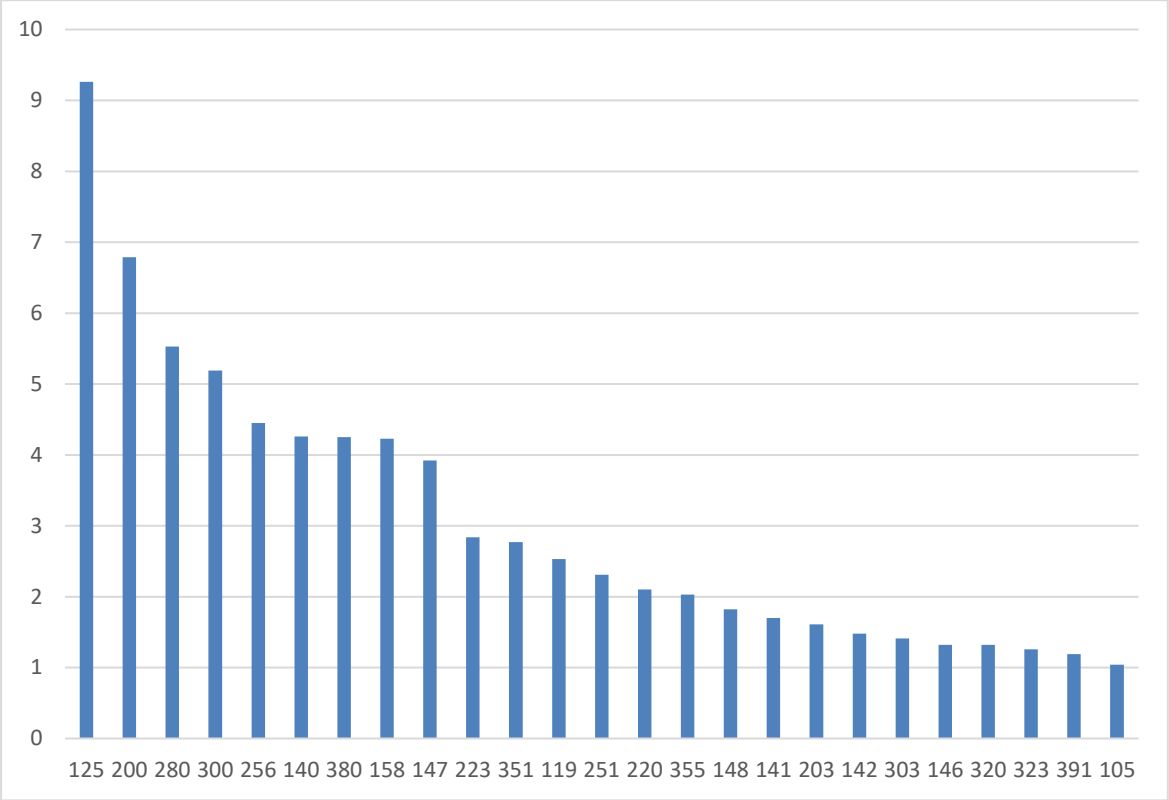

Figure 9: Percentages of different codes for inconsistencies in all documents

Quality assessment analysis in inconsistent documents by specialized fields

This section analyzes the reasons for returning "Theses/Dissertations" documents separately in specialized fields such as medical sciences, engineering, etc. Based on the performed analyses, the specialized group presents some key results separately (Table 4). Among the 10,406 codes used in documents returned to users, 93 (1%) belong to medical sciences, 2154 (21%) to engineering, 60 (1%) to veterinary, 5747 (55%) to humanities, 1039 (10%) are related to science, 593 (6%) are related to agriculture, and 720 (7%) are related to art. In the veterinary field, codes 256, 125, 158, 200, and 300 have the highest frequency among the selected codes. White pages in the PDF file, the contradiction of the registered defense date with the title page, registration of more than one keyword in a field, and inconsistency of the Persian title page in the PDF and Word files are the most common reasons for the discrepancies. 10, 8, 7, 7, and 7 are the percentage of inconsistent documents with this type of error.

In the field of humanities, codes 125, 200, 280, 300, and 140 have the highest frequency among the selected codes. The inconsistency of the recorded defense date with the title page is the most frequent. Inconsistency of Persian and English title pages in PDF files, inconsistency of Persian title page in the Word file, and carelessness in registering the abstract are the most common reasons for contradiction, representing 9, 7, 6, 5, and 5 percent of inconsistent documents, respectively.

Table 4

Comparison of the Percentage of Repetitive Codes Viewed by Specialized Field

Specialized field	Veterinary		Humanities		Basic Science		Engineering		Medical Sciences		Agriculture		Art	
Repetitive codes with inconsistencies	code	Percentage	code	Percentage	code	Percentage	code	Percentage	code	Percentage	code	Percentage	code	Percentage
	25	10	12	9	39	10	12	11	20	8	15	9	12	10
	12	8	20	7	25	6	20	7	25	7	12	7	28	7
	15	7	28	6	12	6	30	6	14	5	25	7	20	7
	20	7	30	5	20	6	28	6	11	4	20	4	30	5
	30	7	14	5	15	5	25	5	12	4	14	4	35	5

In basic science, codes 391, 256, 125, 200, and 158 have the highest frequency among the selected codes. Writing the dissertation in another format, the presence of white pages in the PDF file, the discrepancy between the registered defense date and the title page, the discrepancy between the information registered in the system and the Persian title page in the PDF file, and finally, inserting more than one keyword in a field are the most common reasons for contradiction. 10, 6, 6, 6, and 5 percent of inconsistent documents have these errors, respectively.

In engineering, codes 125, 200, 300, 280, and 256 have the highest frequency among the selected codes. The discrepancy between the date of defense registered with the title page, inconsistency between the information registered in the system and the Persian title page in the PDF file, inconsistency between the information registered in the system and the Persian title page in the Word file, contradiction between the English title page in the PDF file, presence of white pages and finally inconsistency of the date of defense recorded with the title page are the most common reasons for inconsistencies, with 11, 7, 6, 6, and 5 percent of inconsistent documents having these types of errors respectively.

In medical sciences, codes 200, 256, 140, 119, and 125 have the highest frequency among the selected codes. Unlike other areas, metadata codes have a larger share of the five most common codes in this area. The discrepancy between the Persian title page in the PDF file, the presence of white pages in the PDF file, the word "abstract" at the beginning of the abstract registered in the system, the discrepancy between the English title of the Theses/Dissertations in the system and the title page are the most common reasons for the contradiction. 8, 7, 5, 4, and 4 percent of inconsistent documents have these errors.

Codes 158, 125, 256, 200, and 140 have the highest frequency among the selected codes in agriculture. Inserting more than one keyword in a field, the inconsistency of the defense date with the information of the Theses/Dissertations title page, the presence of white pages in the PDF file, the inconsistency of the Persian title page in the PDF file with the system information, and finally the existence of the word "abstract" at the beginning of the abstract registered in the system, are the most common reasons of inconsistencies, with 9, 7, 7, 4, and 4 percent of inconsistent documents having these types of issues, respectively.

In the field of art, codes 125, 280, 200, 300, and 351 have the highest frequency among the selected codes. The discrepancy between the defense date and the information on the Theses/Dissertations' title page is the highest. Incompatibility between the English and Persian title pages of the PDF file with the system information, the inconsistency of the Persian title

page in the Word file with the system information possesses the next highest percentage. Finally, incomplete Word file text is the most common reason for contradiction, with 10, 7, 7, 5, and 5 percent of inconsistent documents having these errors, respectively.

Also, 6,105 data codes (a problem in files) and 4,301 metadata codes (information registered in the system) have been used 10,406 times of using codes. In other words, 59% of the discrepancies are related to the attached files and 41% to the information recorded in the system. As mentioned earlier, the inconsistency of the defense date is the biggest problem observed. Besides, carelessness in abstract registration (code 140) and inserting more than one keyword in a field (code 158) have the highest repetition in the observed inconsistencies. In the following, some key results are presented based on the analysis performed separately by the expert group (Table 4).

Providing improvement solutions

As mentioned in the second section, various data quality models include strategies such as cost evaluation, process owner allocation, data owner allocation, strategy selection, process root cause identification, process control, data improvement solutions design, process redesign, improvement management, provide continuous monitoring to improve data quality. On the other hand, as shown in Section 4.4, the statistical analysis performed on the causes of discrepancies in dissertation documents in different departments shows the diversity of discrepancies. In this section, we put the knowledge extracted from the statistical analysis of discrepancies in the rows of the table and standard approaches to improve data quality in the columns of this table to demonstrate the results of proposed corrective measures in improving data quality in the body of the table. These actions are based on focus group meetings (described in Section 3) involving experts in information technology, data science, and research information systems. Table 5 contains the results.

Table 5
Standard Data Quality Improvement Solutions

Row	Repetitive codes	Code description	Titles of standard data quality improvement solutions (based on data quality management models)									
			Cost estimation	Assigning the owner to the process	Assigning the owner to the data	Strategy selection	Identifying the root causes of errors	Process control	Design data recovery solutions	Process redesign	Improvement management	Continuous monitoring
1	125	The problem of the date of defense and its inconsistency with the title page	Evaluating the costs of mechanically controlling the content of the title page using techniques such as image processing, OCR	Determining the appropriate and specialized owner for the registration process	Determining a suitable and specialized owner for the data and metadata registered in the system	To prevent the recurrence of this inconsistency, use process-oriented strategies in the registration process	Users' carelessness when registering information in the registration system, Insufficient familiarity of users with system information fields	Using checklists in the data quality control process	Using machine methods to control data quality Use of error-free approaches in data recovery	Using ToolTip to guide users better	Forming quality circles to provide system improvement solutions	Collecting and analyzing reports of discrepancies Provide continuous quality control reports

Row	Repetitive codes	Code description	Titles of standard data quality improvement solutions (based on data quality management models)									
			Cost estimation	Assigning the owner to the process	Assigning the owner to the data	Strategy selection	Identifying the root causes of errors	Process control	Design data recovery solutions	Process redesign	Improvement management	Continuous monitoring
2	200	The discrepancy between the title page in the PDF file and the information registered in the system (Meta Data)	Evaluating the cost of machine extraction of metadata from title page information.	Determining the appropriate and specialized owner for the registration process	Assigning two suitable and specialized owners for the data and metadata registered in the system in terms of accuracy and being relevant	To prevent the recurrence of this inconsistency, use process-oriented strategies in the registration process	The inattention of users when registering information in the registration system, insufficient knowledge of users when completing information items in the system	Using checklists in the data quality control process	Using the feature of Recommendation Systems in the registration system when completing information items by users use of online editors in textual information items	Design the sampling process in the final control in the process	Forming quality circles to provide system improvement solutions	Collecting and analyzing reports of discrepancies Provide continuous quality control reports
3	280	Inconsistency of the English page with the information entered in the system	Evaluating the costs of mechanically controlling the content of the title page using techniques such as image processing, OCR	Determining the appropriate and specialized owner for the registration process	Assigning two suitable and specialized owners for the data and metadata registered in the system in terms of accuracy and being relevant	Data-driven strategy in improving the quality of information related to the English file title page	Users' carelessness when registering information in the registration system, Insufficient familiarity of users with system information fields	Using checklists in the data quality control process	Using machine methods to control data quality	Using ToolTip to guide users better	Forming quality circles to provide system improvement solutions	Collecting and analyzing reports of discrepancies Provide continuous quality control reports
4	300	problems of the Persian title page in the Word file	Feasibility of using Pdf software to convert Word file to Pdf with desired quality	Determining the appropriate and specialized owner for the registration process	Determining a suitable and specialized owner for the data and metadata registered in the system	A data-driven strategy is chosen to improve inconsistency. Because this discrepancy is related to the quality of the data	Postgraduate students are unfamiliar with the principles of compiling and editing dissertation files, and Insufficient accuracy in compiling the final files of dissertations	Using checklists in the data quality control process	Developing guidelines for controlling title page inconsistencies as well as training data quality control staff	Improve the dissertation registration process for the use of Pdf builder software Using ToolTips to better guide users in compiling Pdf files in the registration system	Forming quality circles to provide system improvement solutions	Collecting and analyzing reports of discrepancies Provide continuous quality control reports
5	256	the presence of white pages in the dissertation	Feasibility of using Pdf software to convert Word file to Pdf with desired quality	Determining the appropriate and specialized owner for the registration process	Determining a suitable and specialized owner for the data and metadata registered in the system	Data-driven strategies such as data cleansing and automated whiteboard removal solutions	Postgraduate students are unfamiliar with the principles of compiling and editing dissertation files	Mechanical control of dissertation data Using quality control checklists	Use of error-free approaches in data recovery	Using ToolTip to guide users better	Forming quality circles to provide system improvement solutions	Collecting and analyzing reports of discrepancies Provide continuous quality control reports

Discussion

In the "Theses/Dissertations" registration process, the non-standardization of identified discrepancies and the lack of proper classification for discrepancies make statistical analysis of qualitative metadata problems complex and lead to improbable root analysis of observed errors. Therefore, in this study, the inconsistent structure observed in the registration system after standardization is placed as a tree structure in the registration system. The ultimate goal of this

study was to improve the quality of data (or metadata) of "Theses/Dissertations" within the country.

To provide improvement, two general strategies proposed by Batini et al. (2009), Glowalla, Balazy, Basten and Sunyaev (2014), and Günther, Colangelo, Wiendahl and Bauer (2019) were pursued. These strategies are (1) data-driven strategies and (2) process-driven strategies. Data-driven strategies improve data quality by directly modifying data values. For example, the quality of obsolete data is improved by updating it from another database and replacing it with updated data. Process-driven strategies also enhance quality by redesigning production and data transformation processes.

In the current work, some supervisory strategies were improved to propose improvement strategies, while the others included production and processing steps. The improvements included solutions such as guiding the user in the "Theses/Dissertations" registration process or the process-driven category. Process-driven strategies to improve data quality were considered in the works by Falge et al. (2012), Michelberger et al. (2011), Ershadi and Nabizadeh (2022), and Azeroual et al. (2020). On the other hand, introducing a specific format for "Theses/Dissertations" files and resolving the quality issues of PDF files were among the data-driven solutions. In this regard, Azeroual et al. (2020) considered solutions to address inconsistencies in research data management. Such solutions have been presented by Sidi et al. (2012), Batini et al. (2009), and Ershadi, Rajabi, Shirani and Rezaee (2016).

The main customized improvement actions defined for the quality of theses/dissertations through standard data quality models are as follows.

- Evaluating the costs of mechanically controlling the content of the title page using techniques such as image processing, OCR
- Determining the appropriate and specialized owner for the registration process
- Determining a suitable and specialized owner for the data and metadata registered in the system
- Data-driven strategies such as data cleansing and automated whiteboard removal solutions
- Using checklists in the data quality control process
- Use of error-free approaches in data recovery
- Using ToolTip to guide users better
- Forming quality circles to provide system improvement solutions
- Collecting and analyzing reports of discrepancies
- Providing continuous quality control reports

Conclusions

Based on the TDQM approach, this research presented a framework for evaluating and analyzing and finally improving the quality of "Theses/Dissertations" data and metadata. After implementing this framework in the "Theses/Dissertations" registration system in Iran, the results were comprehensively analyzed, based on which quality improvement strategies were proposed. The wide range of specialties in various dissertations, theses, and perspectives of researchers in each field makes it challenging to quickly analyze the recorded information's quality. As this study showed, the quality of the information recorded (at the metadata level) and the information contained in the uploaded files of "Theses/Dissertations" differed in each

field. This implied that it was necessary to define strategies for each area of expertise to improve the quality of the information in "Theses/Dissertations" at the macro level. In future research, using data mining techniques on the results obtained from inconsistent codes, more complete analyses can be performed on the quality problems of documents by a university, field, area, or other information items.

As inconsistent codes are written using data quality techniques and approaches, mapping between these codes and data quality dimensions such as completeness, accessibility, and evaluation of code status from a data quality approach will be significant in future research. On the other hand, correlations between inconsistent codes are among the issues not addressed in this work. It is suggested that this indicator be examined in the future. The practical results of this work are noteworthy because if the document control staff see a code, they can predict that another inconsistency is likely to occur and consider the contradiction in the quality control process.

Acknowledgments

We would also like to show our gratitude to Dr. Azadeh Fakhrzadeh and Mr. Ali Asghar Hojjatpanah for sharing their pearls of wisdom with us during this research and participating in focus group meetings.

Endnote

1. For more information, refer to Ershadi's (2019) book

References

Ashtarian Esfahani, A., Ershadi, M. J. & Azizi, A. (2020). Monitoring indicators of research data using I-MR control charts. *Iranian Journal of Information Processing and M*anagement 35 (4), 957-933. https://doi.org/10.35050/JIPM010.2020.025 [in Persian]

Avenali, A., Batini, C., Bertolazzi, P. & Missier, P. (2008). Brokering infrastructure for minimum cost data procurement based on quality-quantity models. *Decision Support Systems* 45 (1), 95-109. https://doi.org/10.1016/j.dss.2007.10.012

Azeroual, O., Ershadi, M. J., Azizi, A., Banihashemi, M. & Abadi, R. E. (2021). Data quality strategy selection in CRIS: Using a hybrid method of SWOT and BWM. *Informatica*, 45(1), 65-80. https://doi.org/10.31449/inf.v45i1.2995

Azeroual, O., Saake, G., Abuosba, M. & Schöpfel, J. (2020). Data quality as a critical success factor for user acceptance of research information systems. *Data*, 5 (2), 35. https://doi.org/10.3390/data5020035

Batini, C., Cabitza, F., Cappiello, C. & Francalanci, C. (2008). A comprehensive data quality methodology for web and structured data. *International Journal of Innovative Computing and Applications* 1 (3),205-218. https://doi.org/10.1504/IJICA.2008.019688

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41 (3), 1-52. https://doi.org/10.1145/1541880.1541883

Cahyono, S. H. & Sucahyo, Y. G. (2020). Pengukuran Kualitas Data Menggunakan framework total data quality management (TDQM): Studi Kasus Sistem Informasi Beasiswa Universitas Indonesia Data Quality Assessment Using the TDQM Framework: A Case Study of University of Indonesia (UI) Scholarship Information System. *Jurnal IPTEK-KOM (Jurnal Ilmu Pengetahuan dan Teknologi Komunikasi)*, 22 (2), 193-206. https://doi.org/10.17933/iptekkom.22.2.2020.193-206

De Amicis, F., Barone, D. & Batini, C. (2006). An analytical framework to analyze dependencies among data quality dimensions. In *ICIQ* (pp. 369-383).

Edris Abadi, R., Ershadi, M. J. & Niaki, S. T. A. (in Press). A clustering approach for data quality results of research information systems. *Information Discovery and Delivery*. https://doi.org/10.1108/IDD-07-2022-0063

Elouataoui, W., El Alaoui, I., El Mendili, S. & Gahi, Y. (2022). An advanced big data quality framework based on weighted metrics. *Big Data and Cognitive Computing*, 6(4), 153. https://doi.org/10.3390/bdcc6040153

English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. J. Wiley & Sons.

Eppler, M. J. & Muenzenmayer, P. (2002, November). Measuring information quality in the web context: a survey of state-of-the-art instruments and an application methodology. In *Proceedings of the Seventh International Conference on Information Quality ICIQ* (pp. 187-196).

Ershadi, M. J. & Ershadi, M. M. (2018). Implementation of failure modes and effects analysis in detergent production companies: A case study. *Environmental Quality Management* 27 (3), 89-95. https://doi.org/10.1002/tqem.21531

Ershadi, M. J., Jalalimanesh, A. & Nasiri, J. (2019). Designing a metadata quality model: case study of registration system. *Iranian Journal of Information Processing & Management* 34 (4): 1528-1499.

Ershadi, M. J. & Nabizadeh, M. (2022). Providing a structural methodology for measuring and analyzing the quality of theses and dissertations in the country. *Iranian Journal of Information Processing and Management*, 37(3), 667-694. https://doi.org/10.35050/JIPM010.2022.256 [in Persian]

Ershadi, M. J. & Omidzadeh, D. (2018). Customer validation using hybrid logistic regression and credit scoring model: A case study. *Quality - Access to Success*, 19 (167), 59-62. Retrieved from https://www.researchgate.net/profile/Mohammad-Ershadi-3/publication/329671299_Customer_validation_using_hybrid_logistic_regression_and_credit_scoring_model_A_case_study/links/5da1d84345851553ff8c1288/Customer-validation-using-hybrid-logistic-regression-and-credit-scoring-model-A-case-study.pdf

Ershadi, M. J., Rajabi, T., Shirani, F. & Rezaee, N. (2016). Application of root-cause analysis on quality problem solving of research information systems: A case study on dissemination system of theses and dissertations (GANJ). *Iranian Journal of Information Management*, 1 (1), 89-75. Retrieved from https://www.aimj.ir/article_50658_d3bd1b73f795d1dbaa1206ffd6bb7c84.pdf?lang=en [in Persian]

Falge, C., Otto, B. & Österle, H. (2012, January). Data quality requirements of collaborative business processes. In *2012 45th Hawaii International Conference on System Sciences* (pp. 4316-4325). IEEE. Retrieved from https://silo.tips/download/data-quality-requirements-of-collaborative-business-processes#

Falorsi, P. D. & Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34(2), 223-234. Retrieved from https://www.istat.it/en/files/2016/10/Falorsi-engSURVEY_METH.pdf

Glowalla, P., Balazy, P., Basten, D. & Sunyaev, A. (2014, January). Process-driven data quality management--An application of the combined conceptual life cycle model. In *2014 47th Hawaii International Conference on System Sciences* (pp. 4700-4709). IEEE.

Günther, L. C., Colangelo, E., Wiendahl, H. H. & Bauer, C. (2019). Data quality assessment for improved decision-making: A methodology for small and medium-sized enterprises. *Procedia Manufacturing* 29, 583-591. Retrieved from https://publica-rest.fraunhofer.de/server/api/core/bitstreams/e24462fd-a7b2-4597-9548-73a5a4d70978/content

Heinrich, B., Klier, M. & Kaiser, M. (2009). A procedure to develop metrics for currency and its application in CRM. *Journal of Data and Information Quality (JDIQ)* 1 (1), 1-28. https://doi.org/10.1145/1515693.1515697

Jeusfeld, M. A., Quix, C. & Jarke, M. (1998). Design and Analysis of Quality Information for Data Warehouses. In: Ling, TW., Ram, S., Li Lee, M. (eds) *Conceptual Modeling – ER '98. ER 1998*. Lecture Notes in Computer Science, vol 1507. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-49524-6_28

Kapsner, L. A., Kampf, M. O., Seuchter, S. A., Kamdje-Wabo, G., Gradinger, T., Ganslandt, T. & Prokosch, H. U. (2019). Moving towards an EHR data quality framework: the MIRACUM approach. In *German Medical Data Sciences: Shaping Change–Creative Solutions for Innovative Medicine* (pp. 247-253). IOS Press.

Khosroanjom, D., Ahmadzade, M., Niknafs, A. & Mavi, R. K. (2011). Using fuzzy AHP for evaluating the dimensions of data quality. *International Journal of Business Information Systems* 8 (3), 269-285. https://doi.org/10.1504/IJBIS.2011.042409

Kwon, O., Lee, N. & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management* 34 (3), 387-394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

Lee, Y. W., Strong, D. M., Kahn, B. K. & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management*, 40(2), 133-146. https://doi.org/10.1016/S0378-7206(02)00043-5

Long, J. A. & Seko, C. E. (2014). A cyclic-hierarchical method for database data-quality evaluation and improvement. In *Information quality* (pp. 52-66). Routledge.

Loshin, D. (2001). *Enterprise knowledge management: The data quality approach*. Morgan Kaufmann.

Liu, Q., Feng, G., Zhao, X. & Wang, W. (2020). Minimizing the data quality problem of information systems: A process-based method. *Decision Support Systems* 137. 113381. https://doi.org/10.1016/j.dss.2020.113381

Michelberger, B., Mutschler, B. & Reichert, M. (2011). Towards process-oriented information logistics: Why quality dimensions of process information matter. *Lecture Notes in Informatics* (EMISA 2011), (pp.107-120). Bonn: Gesellschaft für Informatik.

Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. *Baltic Journal of Modern Computing* 8 (3), 391-432. https://doi.org/10.22364/bjmc.2020.8.3.02

Ochoa, X. & Duval, E. (2006). Quality metrics for learning object metadata. In *EdMedia+ Innovate Learning* (pp. 1004-1011). Association for the Advancement of Computing in Education (AACE).

Peltier, J. W., Zahay, D. & Lehmann, D. R. (2013). Organizational learning and CRM success: a model for linking organizational practices, customer data quality, and performance. *Journal of Interactive Marketing* 27(1), 1-13. https://doi.org/10.1016/j.intmar.2012.05.001

Petrović, M. (2020). Data quality in customer relationship management (CRM*):* Literature review. *Strategic Management*, 25 (2), 40-47. https://doi.org/https://doi.org/10.5937/StraMan2002040P

Pipino, L. L., Lee, Y. W. & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211-218. https://doi.org/10.1145/505248.506010

Rahman, M. S., Mannan, M., Hossain, M.A., Zaman, A. H. & Hassan, H. (2018). Tacit knowledge-sharing behavior among the academic staff: Trust, self-efficacy, motivation and big five personality traits embedded model. *International Journal of Educational Management*, 32 (5): 761-782. https://doi.org/10.1108/IJEM-08-2017-0193

Robinson, J. (2019). Focus groups. In: Atkinson, P., Delamont, S., Cernat, A., Sakshaug, J. W. and Williams, R. A. (eds.) *SAGE Research Methods: An Encyclopedia.* SAGE. http://dx.doi.org/10.4135/9781526421036821959

Nyumba, T. O., Wilson, K., Derrick, C. J. & Mukherjee, N. (2018). The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution*, 9(1),20-32. https://doi.org/10.1111/2041-210X.12860

Russell-Rose, T., Chamberlain, J. & Azzopardi, L. (2018). Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management*, 54 (6), 1042-1057. https://doi.org/10.1016/j.ipm.2018.07.003

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M. & Baldoni, R. (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, 29(7), 551-582. https://doi.org/10.1016/j.is.2003.12.004

Sharma, S. (2020). Big data analytics for customer relationship management: A systematic review and research agenda. In *Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4* (pp. 430-438). Springer Singapore.

Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H. & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management* (pp. 300-304). IEEE. Kuala Lumpur. [DOI:10.1109/InfRKM.2012.6204995]

Su, Z. & Jin, Z. (2007). A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In *Information Quality Management: Theory and Applications* (pp. 190-220). IGI Global. https://doi.org/10.4018/978-1-59904-024-0.ch009

Taleb, I., Serhani, M. A. & Dssouli, R. (2018). Big data quality assessment model for unstructured data. *In 2018 International Conference on Innovations in Information Technology (IIT)* (pp. 69-74). IEEE. AL AIN UAE. https://doi.org/10.1109/INNOVATIONS.2018.8605945

Vaziri, R., Mohsenzadeh, M. & Habibi, J. (2017). Measuring data quality with weighted metrics. *Total Quality Management & Business* Excellence, 30(5-6), 708-720. https://doi.org/10.1080/14783363.2017.1332954

Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2), 58-65. https://doi.org/10.1145/269012.269022

Wang, R. Y. & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12 (4), 5-33. https://doi.org/10.1080/07421222.1996.11518099

Wang, R. Y. & Stuart, E. M. (1990). A polygen model for heterogeneous database systems: *The source tagging perspective.* In *Proceedings of the 16th International Conference on Very Large Data Bases* (pp. 519-538). San Francisco, CA, United States. Retrieved from http://web.mit.edu/tdqm/www/tdqmpub/polygenmodelAug90.pdf