Indian Journal of Science and Technology, *Vol 8(23), DOI: 10.17485/ijst/2015/v8i23/79229, September 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Study on the Problem Analysis and Improvement Plan of the Data Quality Management System of National R&D Data

Sang Gi Lee[1], Byeonghee Lee[2] and Hanjo Jeong[2*]

[1]School of Computer Science, University of Seoul and KISTI, Korea; sklee@kisti.re.kr
[2]NTIS Center, KISTI, Korea; bhlee@kisti.re.kr, hanjo.jeong@kisti.re.kr

Abstract

National Science and Technology Information Service (NTIS) is constructed and operated to provide a comprehensive data provider service of national R&D data for improving the performance of the national R&D projects and disseminating the co-utilization of the national R&D data nationwide in Korea. The national R&D data is collected from 17 government departments and agencies. With the spread of the Government 3.0 paradigm, the demand for data sharing and utilization is being increased. However, the immature of the data quality management process for the public data hinders the progress on the data openness. This research presents the problems of the data quality management via a study on the current data quality management process/system of the national R&D data and proposes an improvement plan for the data quality management. In addition, DQC-M (Database Quality Certification-Management), which is a representative evaluation and certification model for assessing the quality and maturity of data quality management, is used to find insufficiency through the evaluation of the quality and matureness of the data quality management of NTIS service.

Keywords: Data Openness, Data Quality Management, DQC-M, National R&D Data

1. Introduction

Traditionally, data quality management is an important factor to satisfy the customers and manage the business processes in the enterprise applications[1,2]. Also, most of data quality management has been performed in the areas of data warehouse and decision making as such applications require high quality data[3–5]. Data quality management becomes one of the critical aspects that should be performed in the all business and government applications[6,7]. With the arise of cloud computing and big data framework, the information systems and database management systems are based on more distributed structure. Moreover, the size and scope of data has dramatically increased along with the increase of the number of data sources. Such evolution makes the problem of data quality insurance more complex and controversial[8–10]. Recently, the Semantic Web technologies and big data analytics are used to ensure and improve the data quality by constructing a ontological framework with a vocabulary[11–13]. Also, the big data analytics are used to improve the data quality by finding patterns and creating rules based on the patterns[14,15].

NTIS (National Science and Technology Information Service)[16] is created to collect and manage the national R&D data produced by 17 Korean government ministries and agencies in a centralized place. This centralized data management makes convenient for the analysis and utilization of the national R&D data. It also assists to improve the efficiency and transparency of national R&D projects and prevent duplicate R&D funding, thereby contributing to the enhancement for the National Science and Technologies. Therefore, the data quality assurance is one of the most important things to ensure. Table 1 shows the expected problems that can be occurred if the data quality of the national R&D data is declined.

In this research, the NTIS data quality management performing on the lifecycles of the data from collecting to service is analyzed to find the *problems and improvement plans. Also, the maturity evaluation of the data quality*

[*] *Author for correspondence*

Table 1. Expected problems when the data quality is declined

R&D Data Type	Data Utilization	Problems when data quality is declined
R&D Project Information	- Supports policy decision and new project discovery - Provides indicators of National Science and Technology statistics	- Cause problems in R&D-project evaluation and R&D-funding reorganization - Make errors in the indicator estimation
Participant Researcher Information	- Co-utilization of participant researcher information of national R&D projects - Assurance of specialty and transparency of the evaluation committee selection	- Decrease the co-utilization of the participant researcher information - Cannot ensure the fairness in the evaluation committee-selection process
R&D Output Information	- Establishment of a virtuous cycle on national R&D-project conduct - Assurance of transparency of R&D outputs	- Lower the utilization ratio of R&D output data and its transparency
Facilities & Equipments Information	- Establishment of a virtuous cycle on national R&D-project conduct - Assurance of transparency of R&D outputs	- Decrease the sharing ratio of the facilities and equipment

management is performed, and the plans for consistently improving the maturity level are also presented in this paper.

2. Problems in National R&D Data Quality Management

As shown in Figure 1, there are lots of obstacles in data quality management of NTIS, as the R&D data is collected from 17 different government departments and agencies. To overcome and solve the problems, a data quality management framework would be established as follows: 1) the data quality management should be considered from the raw data as the data is from many sources; 2) A monitoring and tracking system is required to monitor and trace the errors in national R&D data; 3) An advanced and standardized data quality management framework requires to be developed and disseminated to the data management systems in the 17 government departments and agencies.

Issues and problems in data quality management of the national R&D data found from an analysis and the maturity evaluation on the current data quality management system are described as follows:

2.1 Needs of Online Collaborative Environment for Business Rule Management

Business rules are established and managed by corresponding process admins in NTIS to ensure the data

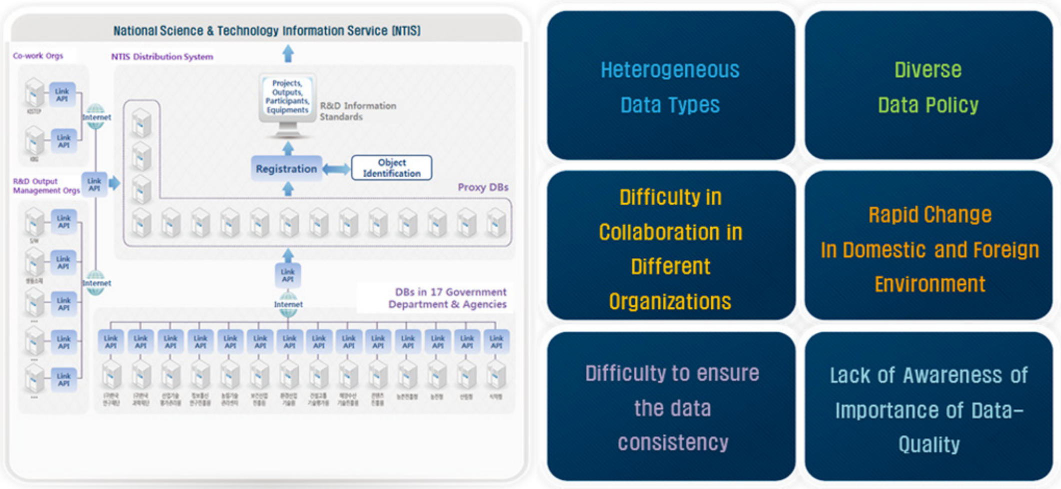
National Science & Technology Information Service (NTIS)
Co-work Orgs
NTIS Distribution System
Projects, Outputs, Participants, Equipments
R&D Information Standards
Registration
Object Identification
Proxy DBs
R&D Output Management Orgs
Link API
Internet
DBs in 17 Government Department & Agencies
Heterogeneous Data Types
Diverse Data Policy
Difficulty in Collaboration in Different Organizations
Rapid Change In Domestic and Foreign Environment
Difficulty to ensure the data consistency
Lack of Awareness of Importance of Data-Quality

Figure 1. Difficulties in data quality management of NTIS.

quality. However, the business rules are still stored and managed as offline documents. Thus the timely update, sharing, and applying of business rules is unrealizable.

2.2 Needs of Data Quality Management for Free-Text Data

National R&D data consists of structured and unstructured data, and currently the data quality management system in NTIS deals with only the structured data. The data quality management needs to take into account of the unstructured data as well to stimulate the openness and sharing of the national R&D data and to strengthen the protection of the personal information in the free-text data.

2.3 Needs of Tracing of Data Error Corrections

In NTIS, data errors are found from a monitoring system of data linkage system. The data errors are checked and corrected by a person in charge of the linkage system directly to the data. In this reason, it is untraceable to find when, who, and how the data errors are corrected.

2.4 Needs of Analysis on Data Errors

As the data quality management becomes focusing on processes rather than data, a fundamental analysis is required to prevent the data errors from the beginning.

2.5 Needs of Intensifying Data Security and Contents Filtering

Some of national R&D data is in free-text format and the data is requited to be filtered since it can contain wrong data and harmful data such as illegal file-sharing sites and illegal sales sites of firearms, drugs, and alcohol.

2.6 Need to Secure the Data Quality from the Data Sources

NTIS collects the national R&D data from 17 different government departments and agencies. Thus, there is a limitation on improving the data quality even if we strengthen the data quality management in NTIS. To correct the fundamental errors, the data quality management in the data sources is required to be strengthened as well.

3. Improvement Plans for the Data Quality Management System

This research mainly presents the problems of the data quality management via a study on the current data quality management process/system of the national R&D data. A maturity evaluation is performed by DQC-M (Database Quality Certification-Management)[17], and proposes an improvement plan for solving the issues and problems in the current data quality management.

3.1 Online Collaborative Environment for Business Rule Management and Sharing

Business Rules take an important role of the data quality management as they represent the rules for checking the data errors. The business rules subject to be often modified with the environment changes, which can be caused by institutional changes. The collaborative online environment for sharing and managing business rules are urgent. It is because the national R&D projects are operated and managed by 17 different government ministries and agencies, and it is difficult to apply the modified business rules in real time without such environment.

Figure 2 shows a unified business rule management system proposed in this research. The system uses the national R&D information standards currently having 389 items, and the business rules are created for the items. A database is created for the unified business rules and also a repository is created for the verification rules for verifying the business rules. The business rules can be classified into two types: one is the rules that can be verified in the unified management system, and the other is the rules that cannot be verified. The most of the business rules are classified to the former one. However, some of the business rules cannot be verified in the system since they require the data that resides in a data-source system for the verification. As in Figure 2, the system also provides two different types of APIs: one is the BR inquiry APIs and BR verification APIs. The BR inquiry APIs provide unified business rules according to an R&D data type and its detailed fields. It also provides whether the business rule can be verified from the unified system. The BR verification APIs are created to provide the verification on the fly.

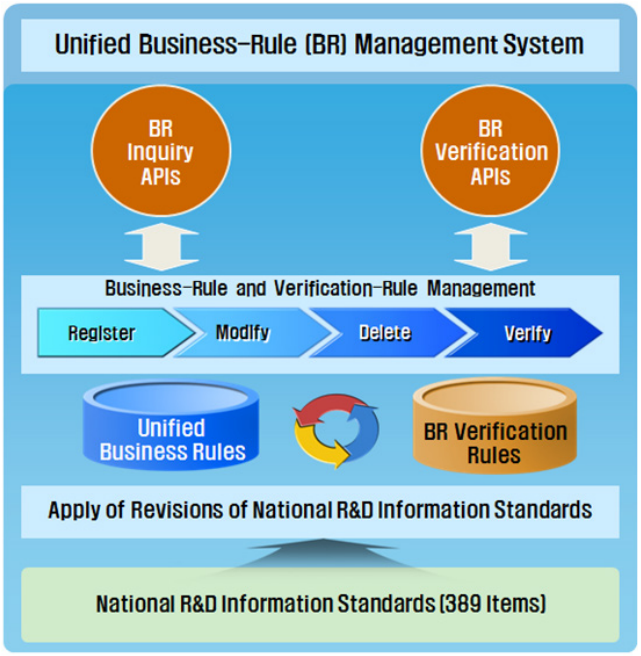
Unified Business-Rule (BR) Management System
BR Inquiry APIs
BR Verification APIs
Business-Rule and Verification-Rule Management
Register
Modify
Delete
Verify
Unified Business Rules
BR Verification Rules
Apply of Revisions of National R&D Information Standards
National R&D Information Standards (389 Items)

Figure 2. A unified Business Rule (BR) management system.

Table 2 specifies the signature of Business-Rules inquiry and verification APIs, and the verification API’s signature is only provided in the case of the business rules that can be verified in the unified system. If a verification API’s signature exists for a business rule, it can be verified through a verification API in real time with the signature. For example, the following inquiry can be posed to get a list of business rules registered for the basic information of R&D projects on a certain year. It returns a list of corresponding business rules in a JSON (JavaScript Object Notation) format. The example shows a returning business rules that require checking if the values of the basic information fields are not null as the basic information of R&D projects are necessary.

- Request: http://ntis_BR_API_URL/BRSearch?Information_Type=’basicInfo’&subject_field=’basisYear’
- Response:http://ntis_BR_API_URL/BRSearch?{“BR_List”:{“BR”[{“BR_Code”:”PR001”,”BR_Specification”:”NOT_NULL_Check_Is_Required”,”Is_Verification_API_Exist”:”Y”},...]}}

Table 2. Signature of Business Rules (BR) inquiry and verification API

API Type	Input	Output

BR Inquiry API	- Information_Type
	- Subject_Field
		- BR_List {
		BR [{
		BR_Code,
		BR_Specification,
		Is_Verification_API_Exist
		}]
		}

BR Verification API	- BR_Code - Subject_Year	- Required_Field_Verification -Date_Type_Verification

3.2 Data Error and Refinement Tracking System

The national R&D data in NTIS is collected from the 17 government departments and agencies. Thus the quality of the source data is significant for maintaining the data quality in NTIS. Figure 3 shows the current data-collection and -Refinement system with daily feedback.

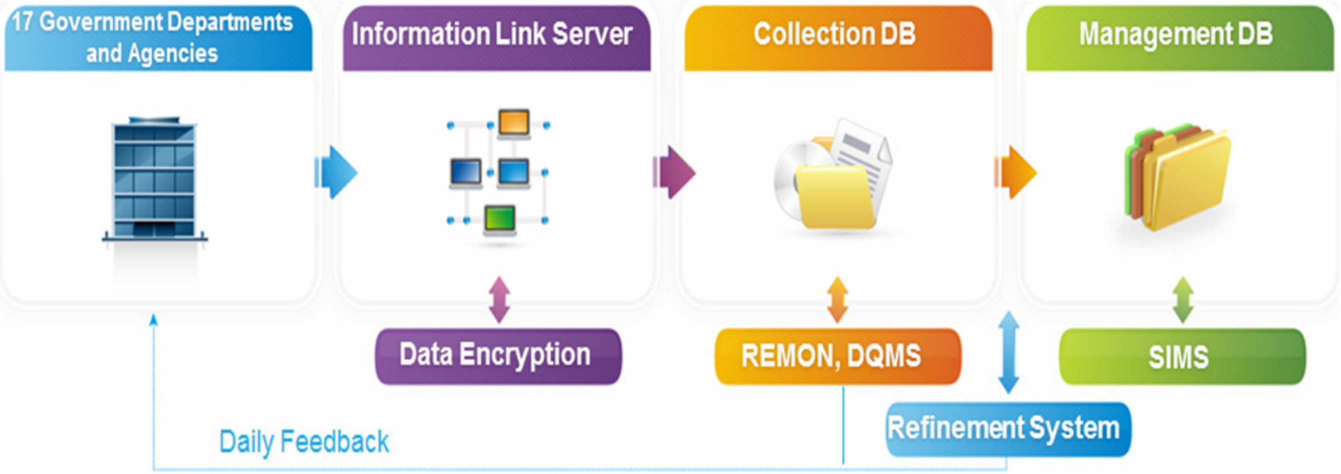

Figure 3. NTIS data-collection and -refinement system.

However, who, when and how the data is refined are still untraceable in the current system. The data errors are untraceable because the data errors found and corrected from the refinement system. The data refinement is separately managed from the various systems such as the information-linkage monitoring system, the data refinement and transfer monitoring system, and the data quality management system. Therefore, a systematic tracking system for the data error corrections is required from the collection status to transfer status.

The information-linkage monitoring system represents the data errors with details of the error information. The data errors tagged as 'not-processible' indicates the fatal errors that should be corrected before it goes to next process. However, it is difficult to correct the errors because the history of the data error corrections is not managed in the system. To solve such problems, the history of the data collection and refinement processes is required to be managed with the information of whom, when, and how the data is modified. Also, the unique identification codes require to be generated for each national R&D data, and the modification information should be stored along with the identification codes.

3.3 Free-Text Processing System

The summary of R&D projects is one of the most important fields for detecting similar and duplicated projects. However, useless and repeated text data is often found in the summary field. The summary should be more than 25 letters and less than 2000 letters in Korean, and the many of R&D-project participants do not fill the summary field elaborately since it is not a requirement. To cope with the problem, we used methods and algorithms shown in Table 3.

Table 3. Verification algorithms for the summary of R&D projects

Applied Algorithm	Description
Useless-Data Detection Method	Detects and extracts the useless data
Repeated-Sentence Detection Algorithm	Detects and extracts the repeated and blank sentences (Threshold is 80%)
Levenshtein Distance Algorithm	Detects and extracts the repeated words and phrases (Threshold is 95%)

3.3.1 Useless-Data Detection Method

A useless-term dictionary is established by a complete inspection of the national R&D project data in NTIS,

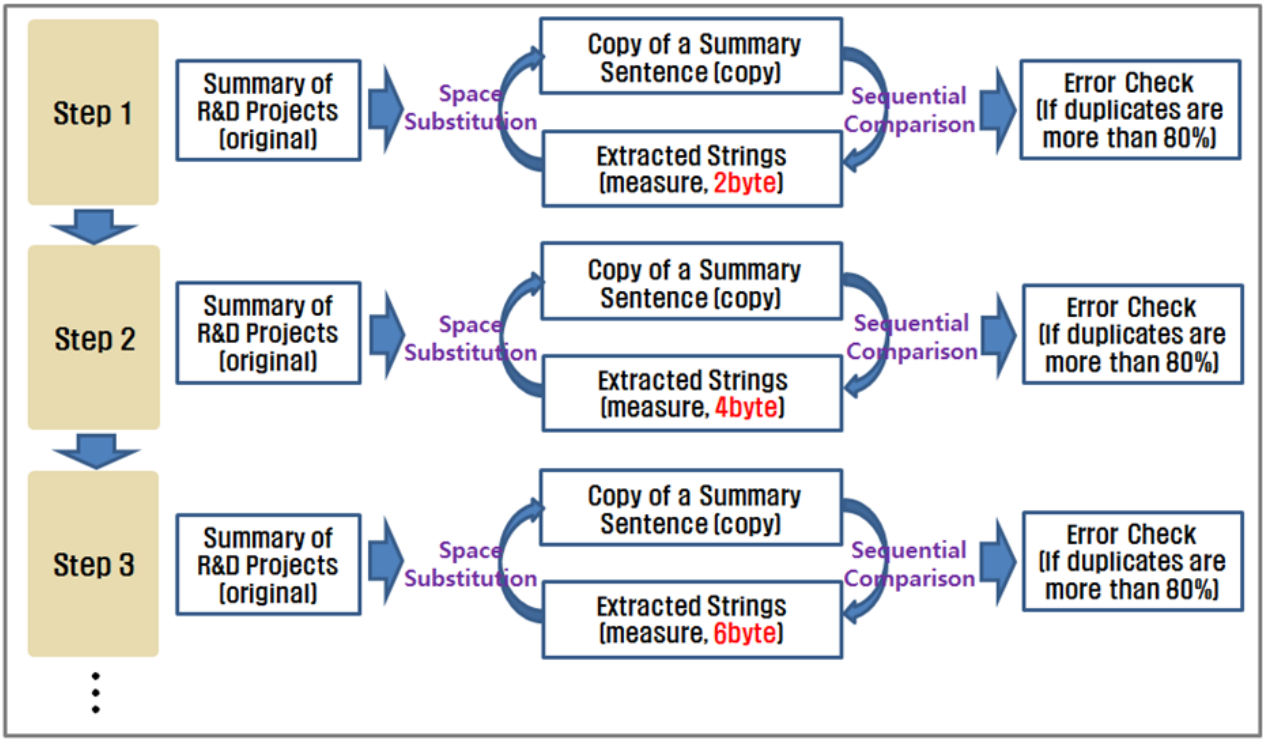
Step 1
Summary of R&D Projects (original)
Space Substitution
Copy of a Summary Sentence (copy)
Extracted Strings (measure, 2byte)
Sequential Comparison
Error Check (If duplicates are more than 80%)
Step 2
Summary of R&D Projects (original)
Space Substitution
Copy of a Summary Sentence (copy)
Extracted Strings (measure, 4byte)
Sequential Comparison
Error Check (If duplicates are more than 80%)
Step 3
Summary of R&D Projects (original)
Space Substitution
Copy of a Summary Sentence (copy)
Extracted Strings (measure, 6byte)
Sequential Comparison
Error Check (If duplicates are more than 80%)
⋮

Figure 4. The process of repeated-sentence detection.

which are about 430 thousand. In the summary field, terms are removed if the terms are found in the useless term dictionary.

3.3.2 Repeated-Sentence Detection Method

Figure 4 shows the process of the repeated sentence detection method, and it finds repeated sentences by using incremental extractions of duplicated character strings. At first, a copy of a summary sentence is created from the original summary text, and the end part of the copied sentence is extracted with a unit of 2 bytes. The extracted part assigns to the measure string and compares it with the copied sentence. If the copied sentence has an exact match with the measure at the end, deletes the matched part. Such process is iterated by extending the measure by 2 bytes until the measure becomes identical to the copied sentence. Finally, compare the remaining copied sentence with the original sentence. If the value of [original] / [copy] is bigger than 5, i.e., there are duplicated strings more than 80%. Then, the original sentence is regarded as a repeated one. Table 4 shows the results of the repeated sentence detections of the national R&D projects in 2014, whose total is 53,492, and 3,659 R&D project data has at least more than one repeated sentence.

Table 4. Repeated-sentence extractions in the national R&D projects in 2014

Summary Type	Number of R&D Projects Containing Repeated Sentences
Research Goals	398
Research Contents	3,231
Expected Effects	30
Total	**3,659**

3.3.3 Levenshtein Distance Algorithm

For the duplicated terms and phrases, we used Levenshtein distance algorithm to extract the duplicated parts. Levenshtein distance algorithm is created to measure a similarity between two character strings[18]. Levenshtein distance is determined as some character exchanges to make two strings identical. For example, if two strings, s_1 and s_2, are both 'test', the distance is zero. If the s_2 is 'tent', the distance is one since the character 's' should be changed to 'n' to make the two strings identical. Table 5 shows the results of the Levenshtein distance algorithm in the summary of the national R&D projects in 2014, and 3,445 projects containing repeated strings are found. However, it takes too much time to run the Levenshtein distance algorithm, and the extracted repeated strings should be checked by manually to make sure that the extracted strings are useless repeated one.

Table 5. Repeated-sentence extractions in the national R&D projects in 2014

Summary Type	Number of R&D Projects Containing Repeated Strings
Research Goals	2,707
Research Contents	712
Expected Effects	26
Total	**3,445**

4. Improvement Plans for the Maturity of NTIS Data Quality Management

The evaluation of the maturity of the NTIS data quality management system using the unified business rule management system is performed through the DQC-M. The NTIS data quality management system earned the validity level 3 for accuracy and consistency, which is one of the top levels amongst the public agencies. However, the usage part such as usability, accessibility, timeliness, and security earned level 2. The improvement plans for improving such parts are described in next subsections.

4.1 Usability Improvement

In general, usability in data quality management system is evaluated by the degree of coverage and depth of the data required for organization needs. Thus, such data requirement management is also a part of the usability. The data requirement is collected and managed by a call center system in NTIS. A simple requirement can be processed immediately by the priority of the request. However, a complex requirement might require a system change or a new development, and it can be implemented in a future development. Such complex request cannot be managed by the call-center system, and it is required to be linked with a task tracking system in NTIS to monitor the status of the requesting process.

4.2 Accessibility Improvement

Accessibility represents the users' ease of access to the data, and it can be categorized as ease of use and ease of search. The user interface, online help and customer support in a system mainly affect the level of the ease of use. The level of ease of search can be determined by the search capabilities of the system. In NTIS, a unified User-interface concept and CSS (Cascading Style Sheet) are applied to provide unified and standardized views, and a search engine is employed for entire service to provide consistent results. However, the most of NTIS service is developed for the internet explorer browser to use ActiveX mainly for security and DRM. This makes a bad influence on the accessibility by making the users using other internet browsers difficult to access. Also, user interfaces should be more intuitive and simple, and customer training is also required to improve the level in terms of the ease of use.

4.3 Timeliness Improvement

In general, timeliness represents non-functional requirements such as response time and recency. To improve the timeliness, periodical enhancement of DB performance is required by DB tuning and managing efficient DB schema while satisfying the users' ad-hoc requirements. Data would also be managed efficiently and effectively through the information lifecycle management. In NTIS, the response time is set as in no more than 3 seconds and SQL & DB tuning is performed once in every year. However, the information lifecycle management is still necessary since the size of the national R&D data is growing fast. Figure 5 shows how to manage and optimize the performance of DB through a life cycle of data using an information lifecycle management system. As shown in Figure 5, the size of inactive data is becoming large as the time is being elapsed. Managing the inactive data in a separate DB is one of the most important factors for the information lifecycle management system. The management of the inactive data enables us to keep DB in an optimized size thereby reducing the storage cost and enhancing the DB performance. Also, the inactive data is also required to be provided for ad-hoc requests.

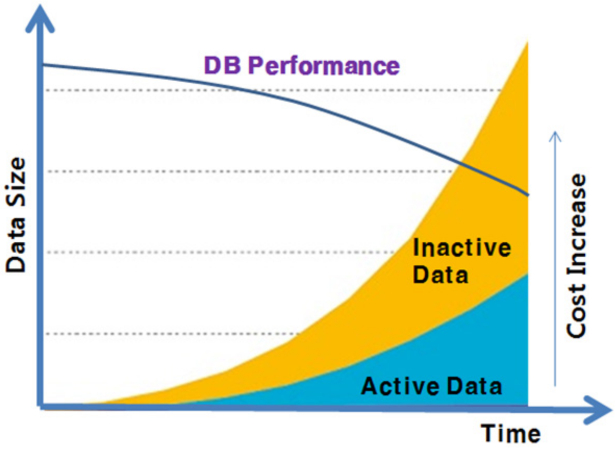

Figure 5. Relationships between data increment and db performance.

5. Conclusions

In this paper, we analyzed the problems in data quality management in the centralized NTIS data collection and services for the national R&D data. We also presented a unified business rule management system for improving the data quality management process with a free-text analysis methods and algorithms for improving the data quality of the summary data of the national R&D projects. Furthermore, the maturity of NTIS data quality management system is evaluated by using DQC-M and analyzed the maturity level and its insufficiency for the next high level. Lastly, we proposed improvement plans for achieving such high maturity by supplementing such insufficiencies. In future research, the remained problems such as the free-text data processing and metadata standardizations will be tackled to improve the data quality management process as in[19,20].

6. Acknowledgment

This research was supported by Maximize the Value of National Science and Technology by Strengthen Sharing/Collaboration of National R&D Information funded by the Korea Institute of Science and Technology Information (KISTI).

7. References

1. Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. Journal of MIS. 1996; 12(4):5–34.
2. Thomas CR. The impact of poor data quality on the typical enterprise. Communications of the ACM. 1998; 41(2):79–82.
3. Shankaranarayanan G, Zhu B. Data quality metadata and decision making. 2012 45th IEEE Conference on Hawaii International System Science (HICSS); 2012. p. 1434–43.
4. Redman TC. Data quality management past, present, and future: Towards a management system for data. In Handbook of Data Quality. Berlin, Heidelberg: Springer; 2013. p. 15–40.
5. Khan N, Iqbal S, Mahboob T. A comparative study of data quality management in data ware houses. 2015.
6. Lee G, Kwak YH. An open government maturity model for social media-based public engagement. Government Information Quarterly. 2012; 29(4):492–503.
7. Jung SH, Jeong DH. A study on the influence factors in data quality of public organizations. KIPS Transactions on Software and Data Engineering. 2013; 2(4):251–66.
8. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR). 2009; 41(3):16.
9. Otto B, Hüner KM, Österle H. Toward a functional reference model for master data quality management. Information Systems and e-Business Management. 2012; 10(3):395–425.
10. Fan W, Geerts F. Foundations of data quality management. Synthesis Lectures on Data Management. 2012; 4(5):1–217.
11. Baumgärtel P, Lenz R. Towards data and data quality management for large scale healthcare simulations. Proceedings of the International Conference on Health Informatics; 2012. p. 275–80.
12. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Talaei-Khoei A. Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. International Journal of Medical Informatics. 2013; 82(1):10–24.
13. Fürber C, Hepp M. Using semantic web technologies for data quality management. Handbook of data quality. Berlin, Heidelberg: Springer; 2013. p. 141–61.
14. Kwon O, Lee N, Shin B. Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Management. 2014; 34(3):387–94.
15. Saha B, Srivastava D. Data quality: The other face of big data. 2014 IEEE 30th International Conference on Data Engineering (ICDE). ; 2014. p. 1294–7.
16. National Science and Technology Information Service (NTIS). Available from: http://www.ntis.go.kr/
17. Database Quality Certification-Management (DQC-M). Available from: http://www.dqc.or.kr/
18. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady. 1966; 10(8).
19. Lee SG, Chae CJ, and Hong EK. Quality management model of atypical science and technology big data based on data profiling and regular expression. International Journal of Contents. 2014; 14(12):486–93.
20. Lee SG, Hong EK. The definition of standard metadata and its quality management model for facilitation of scientific and technical big data sharing. International Journal of Applied Engineering Research. 2014; 9(20):7959–70.