DOI:10.5392/IJoC.2010.6.3.001

A Study on Quality Checking of National Scholar Content DB

Byung-Kyu Kim, Seon-Hee Choi, Jay-Hoon Kim, Beom-Jong You
Department of Knowledge Information Center
Korea Institute of Science & Technology Information, Daejeon, South of Korea

ABSTRACT

The national management and retrieval service of the national scholar Content DB are very important. High quality content can improve the user's utilization and satisfaction and be a strong base for both the citation index creation and the calculation of journal impact factors. Therefore, the system is necessary to check data quality effectively. We have closely studied and developed a web-based data quality checking system that will support anything from raw digital data to its automatic validation as well as hands-on validation, all of which will be discussed in this paper.

***Keywords**: Scholar Content, Data Quality, Data Quality Checking System*

1. INTRODUCTION

The quality management of DB in the nationwide private companies and the public areas are drawing more and more attentions, because the processing of incorrect data results in the degradation of satisfaction degree of data service and it becomes an obstacle that reduces the value of the utilization of the data. KISTI has built the academic paper DB for 452 academic societies in domestic science and technology areas while executing the academic society informatization support project since 1996, and provides the service through the Korea Society Community of Science and Technology and NDSL site[1][2].

Especially, this academic paper DB is used for the base DB to generate the citation index and the academic journal citation indexing, in order to build KSCI, SCI that is customized for Korean environment, consequently errors in the data are directly related to producing incorrect indexes [3].

Therefore, the quality of the data processing is very important and a system that can efficiently verify the accuracy of the quality is required. Many researches have been carried out both domestically and internationally for data quality management and evaluation, but the research and development tools for verifying the quality of correct data considering the factors and environments for academic paper DB are not sufficient. With this background, this paper presents the research, implementation, and application to the actual working environment for the process and the system of verification of the quality of the processed data for academic paper DB.

2. RELATED RESEARCH

There have been active researches on related areas where the importance of data quality management is emphasized.

Most of the earlier researches on the data quality management were the researches on the quality of the data values. The most representative case is the one of Bell lab of AT&T, in which they emphasize the importance of the data quality in 1994, and the accuracy, reliability, up-to-date, completeness, and consistency were presented as the quality evaluation measure for evaluating the data quality[4]. After that, people were getting interested in the structural quality such as fundamental data design structure. Lately the research has been heading toward the unified data quality management adding the concept of data management process.

In Korea, KDB (Korea Database Agency) developed the database quality evaluation guide in 1998 and database quality evaluation model in 2002, presented the expanded model in 2003, and revised the model in 2006 [5]-[7]. The quality evaluation model presented by KDB expands the quality of the data base including not only the quality of data and system but also the data management process.

Separately, KISTI has constructed standard XML of academic papers and has defined and revised an input instruction about the headlines in order to precisely process and control domestic academic paper DB [8]-[10]. In addition, KISTI has developed 2 systems: the input system is applying an input instruction and academic information administration system is managing the processing procedure, so that they have been utilized in data processing [11]-[13].

3. QUALITY CHECKING PROCESS

The process of verifying the data processing, DB establishment, and the data quality was designed as shown in the Fig 1. Based on the electronic files of the papers collected

This is an excellent paper selected from the papers presented at ICCC 2009.
[*] Corresponding author. E-mail : yourovin@kisti.re.kr
Manuscript received Apr. 05, 2010 ; accepted Jun. 30, 2010

through on-line base, data are processed to KISTI academic information XML, and business rule based automatic test and sampling manual test are executed repeatedly until the quality certification once the delivery is made through the effectiveness of XML.

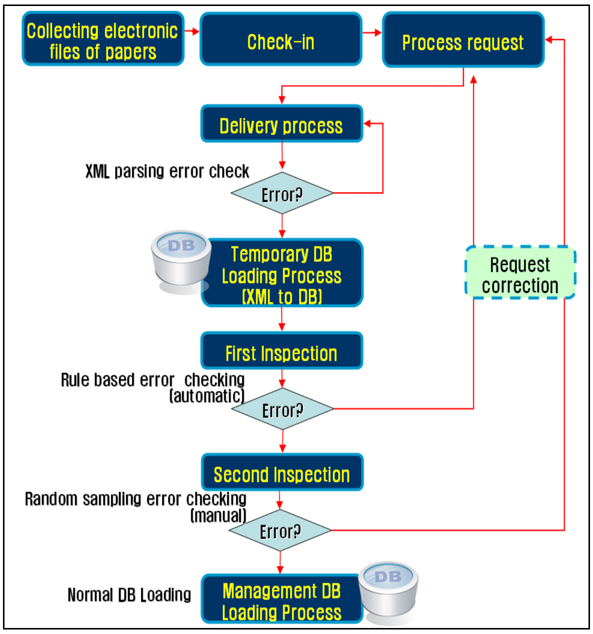

Collecting electronic files of papers
Check-in
Process request
Delivery process
XML parsing error check
Error?
DB
Temporary DB Loading Process (XML to DB)
Request correction
First Inspection
Rule based error checking (automatic)
Error?
Second Inspection
Random sampling error checking (manual)
Error?
Normal DB Loading
Management DB Loading Process
DB

Fig. 1. Quality Checking Process

The standard for checking the quality of the academic paper processing data for each step of quality inspection is defined as shown in the table 1.

Table 1. Type of errors for each major item (Article & reference documents)

Scholar Content (Meta & Original)		Reference	
Main Item	Error Type	Main Item	Error Type
Publishing-Institute	Institute Code	Original Information	LaTex
Publication	KOJIC		Original Omission
Volume	Volume Notation	Journal Name	Journal Name - Error
Article Name	Errata	Conference Name	Conference Name - Error
	LaTex	Volume	Notation Error
	Chinese Notation	ISSN	Notation Error
Author Name & Belongingness	Errata	Page Range	Errata
	Name Classification	Article Name	Errata
	Author Order	Author Name	LaTex
Author-additional Info.	Email	Language used	Errata
	Author - Classification	KOI	Name - Classification
Keyword	Errata	DOI	Language used
Abstract	Errata	ISBN	Search/Matching
	LaTex	Publishing- Institute	Search/Matching
	Chinese Notation	Publisher	Errata
Page	Beginning Page	Issuance Date	Errata
	Ending Page	Publication Region	Errata
Language used	Language Used	Patent Application	Errata
Reference	Reference Number	Standard Number	Errata
OriginalPDF	Meta Matching	Meeting Place	Errata
	Content Lost	Meeting Date	Errata
	Quality-Abnormality	Data Type	Classification-Error
Administration-Number	Administration-Number Citation	URL	Citation Error

4. QUALITY CHECKING SYSTEM DEVELOPMENT

The process for checking the quality of the academic paper processing data was implemented on the KISTI academic information unified management system[14]. Fig 2 is the online collection screen for electronic paper files which are loaded by journal publisher (society, association, research center).

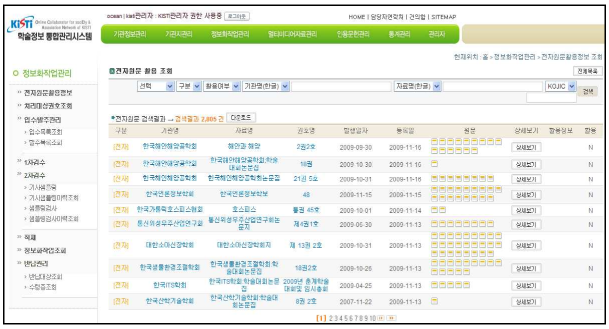


Fig. 2. Online collection screen for electronic paper files

Management functions to track and monitor the history of the manual sampling test are implemented as shown in Fig 3. And the business rule editor was implemented as shown in Fig 4,5 so that new type or error found from the manual test can be added, registered, and managed to the business rule for automatic test.

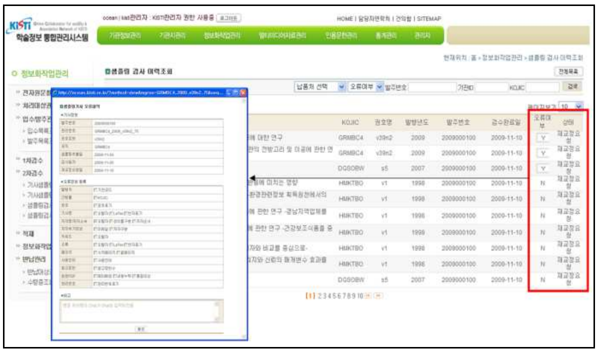
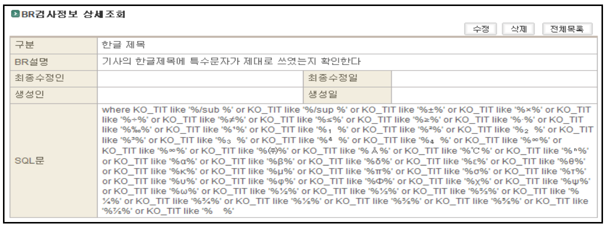


Fig. 3. Sample Quality Checking History

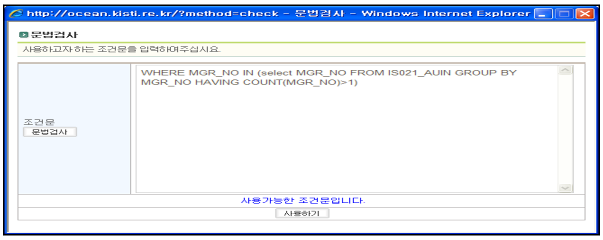


BR검사정보 상세조회

수정 삭제 전체목록

구분	한글 제목		
BR설명	기사의 한글제목에 특수문자가 제대로 쓰였는지 확인한다		
최종수정인		최종수정일	
생성인		생성일	
SQL문	where KO_TIT like '%/sub %' or KO_TIT like '%/sup %' or KO_TIT like '%±%' or KO_TIT like '%×%' or KO_TIT like '%÷%' or KO_TIT like '%≠%' or KO_TIT like '%≤%' or KO_TIT like '%≥%' or KO_TIT like '%·%' or KO_TIT like '%‰%' or KO_TIT like '%¹%' or KO_TIT like '%₁ %' or KO_TIT like '%²%' or KO_TIT like '%₂ %' or KO_TIT like '%³%' or KO_TIT like '%₃ %' or KO_TIT like '%⁴ %' or KO_TIT like '%₄ %' or KO_TIT like '%∞%' or KO_TIT like '%∞%' or KO_TIT like '%㈜%' or KO_TIT like '%Å%' or KO_TIT like '%℃%' or KO_TIT like '%ⁿ%' or KO_TIT like '%α%' or KO_TIT like '%β%' or KO_TIT like '%δ%' or KO_TIT like '%ε%' or KO_TIT like '%θ%' or KO_TIT like '%κ%' or KO_TIT like '%μ%' or KO_TIT like '%π%' or KO_TIT like '%σ%' or KO_TIT like '%τ%' or KO_TIT like '%υ%' or KO_TIT like '%φ%' or KO_TIT like '%Φ%' or KO_TIT like '%χ%' or KO_TIT like '%ψ%' or KO_TIT like '%ω%' or KO_TIT like '%½%' or KO_TIT like '%⅓%' or KO_TIT like '%⅔%' or KO_TIT like '%¼%' or KO_TIT like '%¾%' or KO_TIT like '%⅛%' or KO_TIT like '%⅜%' or KO_TIT like '%⅝%' or KO_TIT like '%⅞%' or KO_TIT like '% %'		

Fig. 4. Business Rule Editor



http://ocean.kisti.re.kr/?method=check - 문법검사 - Windows Internet Explorer

문법검사

사용하고자 하는 조건문을 입력하여주십시요.

조건문 [문법검사]

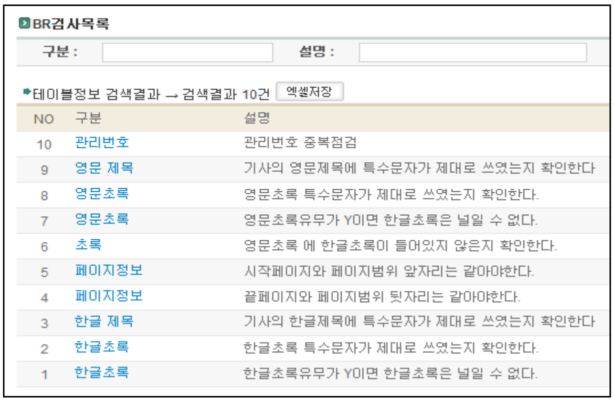
```
WHERE MGR_NO IN (select MGR_NO FROM IS021_AUIN GROUP BY
MGR_NO HAVING COUNT(MGR_NO)>1)
```

사용가능한 조건문입니다.

[사용하기]

Fig. 5. Business Rule Validation Checker

In result, the same type of error can be detected in the automatic testing stage from the next time so that the time and resources for manual test can be reduced gradually. Figure 6 shows the screen to manage frequent errors by the business rules based on these data. Therefore, the data are automatically checked by these business rules.



BR검사목록

구분 : 설명 :

테이블정보 검색결과 → 검색결과 10건 엑셀저장

NO	구분	설명
10	관리번호	관리번호 중복점검
9	영문 제목	기사의 영문제목에 특수문자가 제대로 쓰였는지 확인한다
8	영문초록	영문초록 특수문자가 제대로 쓰였는지 확인한다.
7	영문초록	영문초록유무가 Y이면 한글초록은 널일 수 없다.
6	초록	영문초록 에 한글초록이 들어있지 않은지 확인한다.
5	페이지정보	시작페이지와 페이지범위 앞자리는 같아야한다.
4	페이지정보	끝페이지와 페이지범위 뒷자리는 같아야한다.
3	한글 제목	기사의 한글제목에 특수문자가 제대로 쓰였는지 확인한다
2	한글초록	한글초록 특수문자가 제대로 쓰였는지 확인한다.
1	한글초록	한글초록유무가 Y이면 한글초록은 널일 수 없다.

Fig. 6. Business Rule List

5. CONCLUSION

The process for checking the quality of the academic paper processing data was implemented on the KISTI academic information unified management system. Fig 2 is the online collection screen for electronic paper files, data processing quality of the academic paper DB. The systematic process was designed for the series of process from collecting the electronic paper files, verification of the effectiveness of the processed data in XLM format, business rule based automatic test, to the sampling manual test and the corresponding system was implemented to establish the foundation for verifying the data quality accurately.

In the future, we plan to develop the quality evaluation indexes suitable for academic paper DB and apply them to this system. We will improve and complement the system by intense development of the deepness of the data processing and expanding the range of quality verification at the same time.

REFERENCES

[1] http://society.kisti.re.kr
Korea Society Community of Science and Technology

[2] http://www.ndsl.kr
NDSL(National Digital Science Links)

[3] http://ksci.kisti.re.kr
KSCI(Korea Science Citation Index Service)

[4] Fox. C. et al, "The Notion of Data and its Quality Dimensions", *Information Processing & Management*, Vol.30, No.1, 1994, pp.9-14.

[5] C.Y Lee,H.J Park, “A Case Study on Database Quality and Quality Factors”, *journal of Information Technology Applications and Management*, Vol.11, No.4, 2004, pp.209-225.

[6] DPC, “Data Quality Management Maturity Model”, DPC, 2006.

[7] DPC, “The Guideline for Data Quality Management”, DPC, 2006.

[8] KISTI, “XML based Academic Content and Association Technical Content Processing Guide”, KISTI, 2005.

[9] KISTI, “XML based Scholar Content Processing Guide”, KISTI, 2007.

[10] KISTI, “National Academic Content Processing Guide”, KISTI, 2010.

[11] Byung-Kyu Kim, “A Study on the System of Association Technology Information Management & Service”, *KOSTI 2004 Workshop*, 2004, pp.251-261.

[12] KISTI, “The Manual for ACMS1.0(Academic Content Management System)”, 2004

[13] KISTI, “The Manual for ACMS2.0(Academic Content Management System)”, 2005

[14] KISTI, “The Manual for national scholar content total management system”, 2007

Byung-Kyu Kim

He received the B.S., M.S in computer science from Chungnam National university, Korea in 2001, 2003 respectively. Since then, he has been with the Knowledge Information Center, Korea Institute of Science & Technology Information. His main research interests include metadata management system and applications for metadata processing.

Seon-Hee Choi

She received the B.S., M.S in computer science from Yonsei university, Korea in 1992, 1995 respectively. Since then, he has been with the Knowledge Information Center, Korea Institute of Science & Technology Information. Her main research interests include advanced information service and Science Citation analysis in Korea.

Jay-Hoon Kim

He received the B.S.in library and information science from Yonsei university, Korea in 1999 and the M.S. in business administration from Sungkyunkwan university, Korea in 2008. He has worked at KAIST for developing and managing NDSL (National Digital Science library) project from 1999 to 2005. Since 2006, he has been with the Knowledge Information Center, Korea Institute of Science & Technology Information. His main research interests include library consortia and content service for scholars.

Beom-Jong You
He is Principal Researcher of Department of Knowledge Resources at Korea Institute of Science & Technology Information (KISTI), Republic of Korea. He received Master and Ph.D. degrees in Library and Information Science from Chungnam National University, Korea. His research interests lie in the information science, knowledge bases and semantic technologies.