IIC 2021

TRANSITION OR TRANSFORMATION OF LIBRARIES DUE TO COVID PANDEMIC: LESSONS TO LEARN

(Proceedings of the 4th I-LISS International Conference - IIC 2021)

Editors in Chief

Prof Dong-Geun Oh

Prof B Ramesh Babu

Editors

Dr S K Asok Kumar

Dr P Rajendran

Associate Editors

Dr A Bagavathi

Mr P J Y Joysingh

Co-Host by

The Tamil Nadu Dr Ambedkar Law University, Chennai

In association with

JISTaP

Keimyung University
South Korea

Institute of Culture Convergence Archiving & Graduate School of Records and Archives Management
Jeonbuk National University

Organised by

International Library and Information Science Society

THE ROLES OF LIBRARIAN AND ARCHIVIST IN AI AND BIGDATA ERA: FOCUSED ON DATA QUALITY MANAGEMENT

Jeong Ho Na[1]* Jin Sol Lim[1]* Hyo-Jung Oh[2]*

Introduction

As artificial intelligence (AI) technology has shown high growth rate and business feasibility recently, the data underlying the technology is drawing attention. According to IDC(International Data Corporation)'s "2021 Global Artificial Intelligence Market Forecast Report", the AI market will grow 17.5% annually and total sales will reach $554.3 billion by 2024[1]. Reflecting this trend, the U.S. is setting data construction as its main goal to secure the base of artificial intelligence technology in AI R&D strategy. In addition to Canada, the EU, the U.K. and China are also putting forward data-building strategies to lead AI research. Similarly, the Korean government is striving to develop the data industry by selecting a '*data dam*' project to advance AI technology as one of the key challenges of the 2020 Digital New Deal policy[2][3].

However, as the types and amounts of data collected and produced by governments and private companies exponentially have been increased, a lot of problems related to the data quality have been arisen. According to Garter™report "Measuring the Business Value of Data Quality",40 percent of the expected value of all business plans in the organization is not achieved. This is a problem caused by poor data quality during the planning and execution stages. Low quality data reduces operational efficiency, increases excessive processing costs, and affects productivity [4].Even though data quality management(DQM) is the essential requirement to reduce losses from low quality data and increase the efficiency of data analysis and AI technologies, most companies focus only on the analysis of data, but do not consider quality management important and academic research is also insufficient.

In this study, our ultimate goal is a consideration of the capabilities of librarians and archivists on DQM and reconsideration of the prestige of librarians and archivists in the era of AI and big data. To this end, we conducted literature surveys on some requirements for DQM and details in DQM procedures. Comparisons with the roles of big data analysts, librarians and archivists were conducted. In addition, we derived the implications through the cases of big data curation where librarians and archivists played a major role.

Relevant literature

Most of studies related to DQM said that data production increases and its types are also diversifying as technology advances. For this reason, it was argued that DQM is becoming important to secure reliable data. Dravis (2004) said that in order to maximize the benefits of organizations investing in BI (Business Intelligence) projects, data quality should be managed. He argued that in order to secure the quality of data, the data quality strategy should be established considering six factors: context, storage, data flow, work flow, stewardship, and continuous monitoring [5]. Viscusi et al. (2014) discussed a quality-based framework for compliance

1 *Graduate School of Archives and Records Management, Jeonbuk National University, Korea*
2 *Dept. of Library & Information Science, Jeonbuk National University, Korea*
**Institute of Culture Convergence Archiving & BK21 Program for Homo D-Biblos*

assessment of open government data using public data provided by local public administration in Italy. The study collected public data provided by 50 administrative sites among Italian local institutions. After that, they classified the collected data by data provider and evaluated the quality in terms of completeness, accuracy, timeliness, and compliance. As a result, about 40% of websites operated local public administrations did not satisfy completeness, and more than 55% of the datasets are Word/PDF/Power Point, formats not suitable for machine readable. It suggested that the quality management of public data is not being done properly [6].Ghasemaghaei & Calic (2019) investigated effect of processing and quality of big data on firm decision-making using the Organizational Learning Theory and the data quality framework. As a consequence, the utilization of big data does not significantly impact the quality of firm decision-making However, it is fully mediated through data quality and data diagnosticity, thus, they emphasized the need for DQM prior to the utilization of big data [7].

According to a study discussing the work of the librarians and archivists in the data era, the competencies required of librarians and archivists are changing as the information environment changes. Accordingly, previous studies emphasized that librarians and archivists should have necessary competencies and skills as information professionals in line with the changed demands. Park et al. (2018) analyzed the technology substitutability for each task of public library librarians in the era of the 4th industrial revolution and discussed the roles of librarians required in the present age by deriving core tasks of future librarians. They said that the competency of ICT technologies affecting most librarian works should be strengthened and librarians should be performed the role of a data scientist and the role of a service creation foundation element proposer [8]. Klapwijk (2018) argues that data science has appeared as the times have become data-intensive and knowledge of computer programming, software engineering, and statistics are included in data science, but the most important is data cleaning and data preparation. he said that librarians should collect, preserve, and manage scholarly datasets as data curators, and accept specialized skills to add new value [9].Larson (2020) examined the digital preservation needs of government big data from the perspective of archival theory. she emphasized that archivists need technical knowledge and a critical point of view to capture the context and records of big data. because archivists are called to intervene in the digital preservation of government big data as big data becomes a key part in transactions and decision-making processes[10].

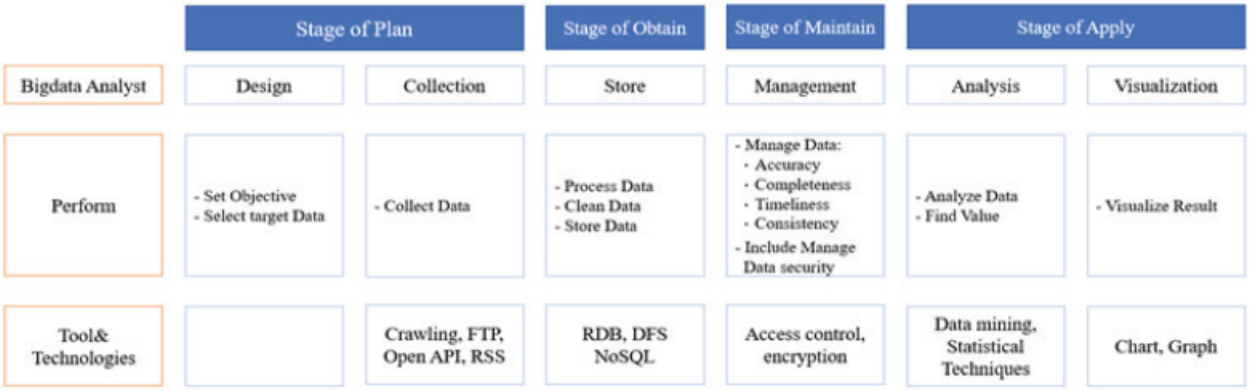
Researches so far have raised issues related to low-quality data and said that the demand for DQM is rising. Even though most of the studies designed and built a quality management framework or system to manage data quality and evaluated its effectiveness, there have been few studies considering the role that information experts can play to contribute to DQM.

Therefore, this study examines the changes in the skills and competencies required for librarians and archivists, who are traditional information professionals, as new types of data are generated. and we identify the role that can contribute to DQM through comparison with big data analysts.

Bigdata Analysis and Data Quality ManagementProcedure

3.1 Big data Analysis

Big data is datasets whose size is beyond the ability of typical database to capture, store, manage and analyze [11], it should be performed at each stage of the information life cycle in 4 processes: *Plan, Obtain, Maintain,* and *Apply* as shown in the <Fig. 1>. Big data analysts are responsible for producing value-added results through the process of designing, collecting, storing, managing, analyzing and visualizing them.



<Fig 1> Bigdata process procedures performed by bigdata analyst

Big data analysts, as shown in the <Fig.1>, define the goals they want to solve and determine the data needed to solve the problem at the design stage. In other words, big data analysts formulate and hypothesize problems arising from organizations as data analysis problems, determine the data required for data analysis, and identify resources such as manpower, applications, and technologies to implement selected analytical models [12]. Furthermore, big data analysts deal with unstructured, semi-structured, and structured data as well as data of various sizes ranging from terabytes to zettabytes. In hence, they need specialized skills for data analysis like statistics and design techniques.

3.2 Data Quality Management

Researches on data quality were being conducted in a variety of ways, and as the utilization of data increased, the definition of data quality was also changing. Among them, ISO 8000-2 defines data quality as "degree to which a set of inherent characteristics of data fulfills requirements". Systematic and continuous management is required to maintain and enhance this data quality. So, DQM means "coordinated activities that direct and control the organization in relation to data quality" and includes attributes such as accuracy, integrity, consistency, validity, timeliness and accessibility [13].The UK Government has argued that data quality assessments should be performed at each stage of the information lifecycle[14] and NIA(National Information Society Agency) of Korea also systematized the DQM process to be applied at each stage of POSMAD information life cycle in 4 processes[15], as shown in the <Fig. 1>.

Data quality should be considered at all stages in the process of big data analysis. First, in the *planning* stage, an organization or institution establishes a DQM plan for quality management. In this stage, the organization selects DQM targets, defines quality management goals, establishes data quality diagnosis and improvement plans, standardization measures, and linked data quality assurance plans, and then allocates personnel and resources required for quality management.

The *obtain* stage is a stage of building databases by storing the collected data. quality management in the *obtain* stage can improve data quality by standardizing data, managing outputs from data construction, eliminating unnecessary items and performing data cleaning such as filtering, converting, refining, and reducing according to the standardization plan established in the *planning* stage.In addition, it is possible to prepare the basis for data quality diagnosis and quality improvement activities in the *maintain* stage by grasping the status of the collected data.

The *maintain* stage is a phase that includes data quality diagnosis and improvement in

the process of operating the database managed by the organization. This stage includes activities such as quality management of linked data, quality diagnosis and improvement of the quality management target selected in the *planning* stage, database product check, maintaining integrity and timeliness due to data change, and notification of database changes to stakeholders.

Lastly, in the *apply* stage, the organization identifies and improves data quality problems that arise when internal and external users utilize the data collected or generated by organization. In addition, they reflect the actions according to the data quality diagnosis and evaluation in the data quality goals and management targets defined at the *planning* stage. In particular, data created by an organization may be inappropriate data for use, and data corruption may occur during the service process, so data quality feedback management is required. When managing feedback, it should be defined in advance whether the scope of feedback to be accepted should be limited only to simple data errors or including difficulties encountered when users use data.

Big data has different quality problems for each data life cycle. Therefore, for DQM, organizations must identify factors that impede data quality at each stage of the life cycle and prepare countermeasures to maintain and improve data integrity, uniqueness, consistency, timeliness, validity, and accuracy. In addition, in order to encourage the use of data, it is necessary to continuously manage the quality of data through user feedback and quality evaluation while securing the reliability of the data.

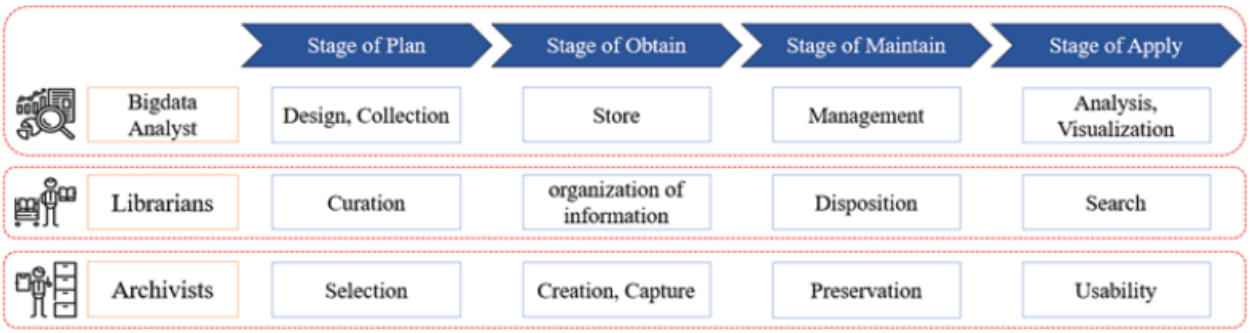
The roles of Librarian and Archivist for Data Quality Management

Librarians and Archivists

Librarians are generally a profession that serves to collect, analyze, organize, accumulate, preserve, and use materials for users in centering on libraries. Meanwhile, in the big data era beyond the information age, librarians are expected to play the role of an information manager who understands the characteristics of data and collects data needed by users. Also, they are required to analyze the collected data, evaluate usefulness, organize or accumulates data, and mediates data in the most appropriate form when data needs occur.

Records management is a set of activities to ensure that the authenticity, reliability, integrity, and usability of records are not compromised in all processes from creation to disposition and preservation. ISO 15489-1 defines records management as "field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records"[16]. Recently, as the management and preservation of datasets are included in the field of records management, archivists are required to have specialized knowledge to collect and preserve data for a long period of time.

Indeed, both librarians and archivists have traditionally played a role in dealing with information sources. As times and their tasks have been changed, the size of the information they handle increases, and the types of information are also diversifying. In particular, the information that they handle is expanding from simple document type to unstructured data type. With such a background, it is considered that both occupations can contribute greatly to DQM in that they manage records including documents and data according to a specific information life cycle and aim to preserve and utilize information.



<Fig 2>The roles of Big data analyst, librarian, and archivist in Data processing procedures

4.2 Specialist for Data Quality Management

Based on the analysis in Chapter 3, The <Fig. 2> summarizes the roles of big data analysts, librarians, and archivists that can be performed in DQM procedures.

Librarians can contribute to DQM by performing roles such as data curation, information organization, and retrieval in the data processing procedures. Librarians can help to derive meaningful results while discovering and preserving trustworthy data sources in the *planning* stage. In the *obtain* and *maintain* stages, unlike big data analysts who simply store data, the librarian can take charge of organizing the collected data so that it is easy for users to access it. Especially librarians excel at creating index or constructing metadata in order to increase the effectiveness of data use, in terms of librarian's own job. Finally, in the *apply* stage, librarians suggest information access points to archive collected and organized data or improve user's ability to utilize information through information services such as information provision, education, and guidance, and mediate information sources to obtain information. It is possible to increase the data utilization of the requestor.

The remarkable examples of big data curation by librarians, there are library big data platforms such as 'Data for Library' and 'Solomon' built and operated by the National Library of Korea. In fact, librarians were sharing and analyzing library big data through big data platforms. The analysis results and data obtained from the platform were used as the basis for setting collection management, program planning, and policy direction for the acquisition. In addition, library big data was provided to users as a book information service [17]. Furthermore, spatial usage analysis based on user activity big data logs in library can be conducted [18].

In addition, university libraries such as UCLA Library, UCSD Library, and University of Melbourne Library provide users with data curation services such as data management, access improvement, data protection, citation and documentation, and creation of collaboration opportunities through sharing for research data[19][20][21].In short, librarians acquired and created reliable data, manage and share the collected data, and use it to create meaningful values while preserving data quality. In addition, it can be confirmed that the librarian has established a virtuous cycle system that recycles library data utilization results through the library big data platform.

<Fig. 2> shows archivists can also play a significant role in DQM. Like librarians, archivists in *planning* stage can provide guidelines for selecting valid data or contribute to identifying the location of data and determining whether it can be gathered. In the *obtain* stage,

archivists are possible to ensure authenticity, reliability, accessibility by creating and capturing data according to the guidelines presented in the *planning* stage. Besides, they can be responsible for classifying and organizing the collected data according to the context, and at the same time creating descriptions for searching.

At the *maintain* stage, data is appraised, selected, and disposed of in a way that ensures public reliability while maintaining the integrity of the constructed data. In other words, archivists maintain and preserve data that is judged to be of continuous value through data appraisal. But they discard data that have not reached certain criteria through reappraisal. This will help archivists improve data quality. In the *apply* stage, archivists perform migration of data format to the new version to guarantee usability and keep the audit trails for proving that records have not changed from unauthorized use, modification, and destruction. They also can provide various retrieval tools such as thesaurus and catalogs. Thus, archivists will increase the usability of data by providing efficient access without compromising the integrity of the data.

As examples of data collection and construction by the archivist, Han et al. [22]attempted to establish a disaster safety information archive to systematically collect, preserve, manage, and utilize disaster safety record information resources for preemptive response and prevention against disasters. To this end, they analyzed the management status of disaster safety record information resources produced by the disaster safety-related institutions. As a result of the study, they said that the subjects of data creation and distribution differ according to the type of disaster safety record information resource. when collecting and classifying data, the characteristics of institutions that produce, distribute, and manage record information resources should be reflected. In addition, Han said that the classification system and metadata are different for each institution, so it should be standardized to maintain consistency and conformity. Through this process, it is possible to prepare a disaster response plan through the analysis of the constructed data and to increase the satisfaction of users by subdividing the service by purpose of use.

In addition, Yoo & Oh [23] compared acquisition methods, status and period of web records of domestic and foreign disaster archives such as PAHO (The Pan American Health Organization) and OASIS (Online Archiving & Searching Internet Sources) to suggest an automatic method for acquiring web records of disaster archives. The researcher analyzed life cycle of disaster issues that occurred in Twitter and Internet newspapers over the past 10 years and proposed an automatic acquisition method and acquisition cycle of web records suitable for each model. In order to maintain data quality in the automatic collection process, they urged that measures be taken to filter out duplicate and unnecessary records, and that exceptional abrupt factors due to political and social issues should be checked.

Looking at the above case for disaster archive construction, archivists analyzed location, characteristics, and current status of data, and suggested improvements and guidelines for disaster record acquisition and archive construction. It can be confirmed that archivists can secure reliable data and perform quality management of the acquired data.

Conclusion

This study compared the tasks of librarians and archivists with those of big data analysts in the data processing procedures. As a result, we try to establish the vision of librarians and archivists in the era of the 4th industrial revolution and consider the roles of librarians and archivists to contribute to DQM.

First, we analyzed the role of big data analysts and the DQM procedures based on the data life cycle in four stages: plan, obtain, store, and apply. Big data analysts were designing, collecting, storing, managing, analyzing and visualizing data according to the data life cycle. They were taking various measures to secure reliable data at each stage in order to maximize the value to be derived with big data. DQM should also be performed at each stage.

Second, we analyzed the traditional roles of librarians and archivists in terms of information professions. Librarians and archivists have been playing the role of dealing with information sources according to their respective purposes. Especially, it has been confirmed that librarians and archivists need skills and competencies to meet new information needs with the development of technology and changes in information media consists of various types of data.

Based on the comparison with big data analysts, the roles that librarians and archivists can contribute to DQM were identified and confirmed through actual case studies. Librarians can collect, preserve, and organize data through data curation in the planning, and provide access points for retrieval to increase the utilization of data. Examples of this were the case of using the library big data platform and the research data curation service implemented by the university library. In the case of archivists, guidelines for the acquisition of valid data can be presented at the *planning* stage. And data are created and captured according to the guidelines during the *obtain* stage and organized and classified according to the production context to ensure the authenticity, reliability, and integrity of the data. In the *maintain* stage, the quality of data can be improved by appraising the value and usefulness of data and disposing of data according to the results. As an example of data curation performed by archivists, there was a study that analyzed the disaster issues that appeared on disaster safety record information resources of disaster safety-related institutions and SNS to identify the characteristics and management status of data, and suggested improvements and acquisition guidelines for archive construction.

As a result of this study, librarians and archivists have many similarities with big data analysts in the way they handle information sources and perform appropriate tasks for DQM in real cases. We hope that this will set up the capabilities of librarians and archivists in the age of AI and big data and enhance their stature as data experts.

References

1. IDC Forecasts Improved Growth for Global AI Market in 2021 [Website]. (2021). Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS47482321.
2. Brandusescu, A., Iglesias, C., Robinson, K., Alonso, J. M., Fagan, C., Jellema, A., & Mann, D. (2017). Open Data Barometer: global report, 4th edition. World Wide Web Foundation, 35 sider.
3. South Korea's Digital New Deal [Website]. (2020). Retrieved from https://thediplomat.com/2020/06/south-koreas-digital-new-deal/
4. Friedman, T & Smith, M. (2011). Measuring the Business Value of Data Quality. Stamford, CT, USA: Gartner.
5. Dravis, F. (2004). Data Quality Strategy: A Step-by-Step Approach. In *ICIQ* (pp. 27-43).
6. Viscusi, G., Spahiu, B., Maurino, A., &Batini, C. (2014). Compliance with open government data policies: An empirical assessment of Italian local public administrations. *Information polity*, *19*(3, 4), 263-275.
7. Ghasemaghaei, M., &Calic, G. (2019). Can big data improve firm decision quality? The

role of data quality and data diagnosticity. *Decision Support Systems*, *120*, 38-49.
8. Park, T. Y., Han, H. J., Oh, H. J., & Yang, D. (2018). A study on the librarian's key tasks of the era of the 4 th Industrial Revolution. *Journal of Korean Library and Information Science Society*, *49*(2), 327-356.
9. Klapwijk, W., IFLA. (2018). Big Data Special Interest Group: A concept framework for data science in libraries, https://www.ifla.org/publications/node/92282?og=10123.
10. Larson, E. (2020). Big Questions: Digital Preservation of Big Data in Government. *The American Archivist*, *83*(1), 5-20.
11. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
12. NIA. (2014). Big Data Utilization Step-by-step business procedures and technology utilization manuals (version 1.0), Seoul, Korea: Ministry of Future Creation and Science.
13. ISO 8000-2:2020 Data quality – Part 2: vocabulary.
14. The Government Data Quality Framework [Website]. (2020). Retrieved from https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework.
15. NIA. (2018). Open Government Data Quality Management Manual v2.0, Seoul, Korea: National Information Society Agency.
16. ISO 15489-1:2016 Information and documentation – Records management – Part 1: Concepts and principles (2nd ed.).
17. On, J., & Park, S. H. (2020). Big Data Analysis for Public Libraries Utilizing Big Data Platform: A Case Study of Daejeon Hanbat Library. *Journal of the Korean Society for information Management*, *37*(3), 25-50.
18. Kim, T. Y., Gang, J. Y., & Oh, H. J. (2019). Spatial usage analysis based on user activity big data logs in library. *Library Hi Tech*, 38(4), 678-697.
19. Research Data Curation [Website]. (N.d.). Retrieved from https://library.unimelb.edu.au/Digital-Scholarship/research_data_curation
20. Research Data Curation [Website]. (N.d.). Retrieved from https://library.ucsd.edu/research-and-collections/data-curation/
21. Data Management & Curation Services [Website]. (N.d.). Retrieved from https://www.library.ucla.edu/support/publishing-data-management/scholarly-communication-services/data-management-curation-services
22. Han, H. J., Park, T. Y., Oh, H. J., & Kim, Y. (2017). A study on improvement and analysis of records management status for disaster safety archives in online environment. *Journal of Korean Library and Information Science Society*, *48*(2), 187-213.
23. Yoo, H. S., & Oh, H. J. (2018). Acquisition Methods for Disaster Archives Based on the Issue Life Cycle Model. *Journal of the Korean Society for information Management*, *35*(2), 115-139.