The Current Situation of Domestic Data Curation through Document analysis

Guo likang

Library of China West Normal University,
Nanchong, Sichuan, 637002, China
E-mail: lkguosc@sina.com

Abstract

To get to know the current situation of data curation in China and offer the study with reference, the method of bibliometrics which is targeted at the CNKI-collected data permanent to the data curation is applied to conduct analyses to the statistic from year distribution of document, distribution of periodical, research subject, keywords frequency, author and fund-raise projects.

Keywords: data curation; specialized works; statistical analysis; biblimetrics;

In the information age as the most essential, active, and influential strategic resource, scientific data obvious calls for innovations on science and technology. And the theory and practice of data monitoring is put forward and evolves with the background of that doing a good job in collecting, arranging, storing, and evaluating scientific data is the primary step of data reusing and sharing. Data monitoring, also called data curation keep manage and run through the whole live span of digital resource, stressing appreciation of material and active management. Nowadays, it has become a hot spot for library and intelligence agent, and academic circles home and abroad start to conduct extensive experiments on data curation no matter on theory and practice, yielding to a large amount of academic achievements. With quantitative and qualitative studies on literatures on data curation, the paper reveals the current situation of data curation in China, offering reference for building research and service on data curation.

1. Data Origin and Mean of analysis

With CNKI as the tool for searching data, select cross-base searching, then insert “Data Curation + Digital Curation + Data monitering + data management+ Data planning” at the blank for keywords searching and subject searching, combine the result of documents from the different searing methods and reduce the irrelevant and the overlapping, we are to have 110 texts (expired by Feb. 22th,2016). The result will be used for analyses by bibliometrics.

2. Result Adding Up and Analyses

2.1 Year Distribution of Literature

One important indicator for the development of a field of study is the year distribution of issued literature on the field. By adding up and analyzing the data of year distribution, we can reveal how the research achieves progress on time scale.

Table 1: Year Distribtuion of Literatures

Year	2011	2012	2013	2014	2015	2016	Total
Amount of issued text	2	10	24	42	31	1	110
Proportion（%）	1.8	9.1	21.8	38.2	28.2	0.9	100

We found from the analysis that the earliest papers about data curation are the two published in 2011, one of which is *Data Curation, New exploration in American College Libraries* published on *College Library Information* by Yang helin,2011 and the other is an translated text *New Age New Role: Supervising in the Data Preservation* by Tan rong,2011 issued on *Books and Intelligence*. Showing increasing amount of publication on data curation from the chart of Year distribution suggests that domestic researches on the field is on departure stage and it requires relevant staffs to strengthen the theory and application of data curation, making it useful in the administrating and providing service for data.

2.2 Distribution of Periodicals Containing Papers on Data Curation

By looking up all the types of literary on data curation, there are one hundred and six periodicals, three academic dissertations, and one patent. So, periodical is the major source of data curation achievements, here is the Table 2 showing the distribution of the studies at periodicals.

Table 2: Distribution of Periodicals Containing Papers on Data Curation

Serial Num.	Name	Amount of texts
1	*Information and Documentation Services*	12
2	*Library & Information*	11
3	*Research on Library Science*	10
4	*Library Journal*	9
5	*Library and Information Service*	8
6	*Journal of Academic Libraries*	6
7	*Library*	5
8	*Information Studies: Theory & Application*	4
9	*New Century Library*	4
10	*Library Tribune*	4
11	*New Technology of Library and Information Service*	3
12	*Journal of Library Science in China*	3
13	*Journal of Academic Library and Information Science*	2

14	*Journal of the National Library of China*	2
15	*Journal of Library and Information Sciences in Agriculture*	2
16	*Information Science*	2
17	*Library Development*	2
18	*Journal of Intelligence*	2
19	*Journal of Library Science*	2
20	*Library Work in Colleges and Universities*	1
21	*Sci-Tech Information Development & Economy*	1
22	*Journal of Ningbo Institute of Education*	1
23	*Library Work and Study*	1
24	*Library Theory and Practice*	1
25	*Journal of Information Resources Management*	1
26	*Journal of Medical Informatics*	1
27	*Science and Technology of West China*	1
28	*Chinese Journal of Medical Library and Information Science*	1
29	*Master*	1
30	*e-Science Technology & Application*	1
31	*Sichuan Archives*	1
32	*Chian Education Network*	1

It can be seen from Chart 2 that all the 106 papers on data Curation are published on thirty-two periodicals in total, of which twenty-six belong to the intelligence type, showing that library and information is the major group to study data curation. And of the twenty six periodicals, there are 16 constituted by Library of CSSCI, Information *and Philology,* reflecting the information specialists' concern about data curation and the general high quality of the papers.

2.3 Subject Analysis

As subject reflect the main feature of what the text should be involved, it can point for us the condition, orientation and nature of the study, and define the goal and what will happen in the future. Through analyzing the title, abstract, part of the text, we will conduct the subject analysis on four perspectives:

1. On theory: the definition, nature, function, academic views, current situation and future tendency of data curation will be involved in discussion.

2. On investigating projects and application of the theory in actual use: to launch investigations on projects and activities involving data curation both home and abroad and make comparisons.

3. On management education and training: about the role, management, education, career planning of the staff engaging in the data curation.

4. On technical and platform: to concern about the technology for saving, operating data curation, building system and platform.

Table 3: Distribution of the Data Curation Literatures Themes

Subject of Analysis	Amount of text	Percentage（%）
Theoretical	59	53.6
Projects and Practice	23	20.9
Education and training in relation to data management	17	15.5
Technology and platform	11	10.0
Total	110	100

Seen from distribution of subjects, theoretical analyses occupy 53.6 percent in the total number of literary, which is the main target in recent years. With regard to projects and practice analysis, the main focus is devoted on the foreign projects, yet worthy to catch notice some scholars have published papers on domestic practice of data curation, for example *Practice of Data Curation Service Aimed at Pure English Teaching* by Wang caihong, *Data Obtained from Scientific Yields Curation System Exploration and Practice——Set Ji Nan University as Model* by Tang jingqian, *Social and Scientific Data Sharing and Service—— Modeled after Social and Scientific Data Exchange Platform, Fu Dan University* by Zhang jilong. They suggest that scholars' studies on data curation turn to the actual aspect from abstract. And data management education, training, technology, and platform researches also receive attention from the experts.

2.4 Analysis of Keywords Frequency

The words that predominant the research of a discipline throw lights on the hot spot of the field and its changes also points out where the research leads to. Statistics shows that there are 471 key words emerging in total 110 papers on data curation. After sorting out words with similar or overlapping meaning, using Excel to calculate the occurrence of keywords, words that enjoy a frequency over three times are shown in Table 4.

Table 4: Key words with over 3 occurance in the data curation literatures

Keywords	Frequency	Keywords	Frequency
数据监护	34	metadatas	6
scientific data	32	long tail data	5
university library	26	curriculum design	5
数据监管	20	The United States	5
institutional repository	18	technology and platforms	5
数据管理	18	*数据策管*	4
data preservation	15	library and information science.	4
library	15	digital uration	3
data curation	14	role definit	3
resource sharing	13	disclosed fetch	3
research data	11	*数据策展*	3

education	11	data service	3
high school	9	*数据管护*	3
Information lifecycle	9	data science	3
big data	8	information resources	3
e-science	7	scholarly commons	3
subject service	6	knowledge service	3

The result from statistics in Table 4 demonstrates the features of data curation in our country:

1.Chinese translation of Data Curation includes *数据监护*、*数据监管*、*数据管理*、*数据策管*、*数据策展*、*数据管护* and so on, of which *数据监护* is the most widely-used, having an occurrence of thirty four.

2. At the age of big data, e-science, Studies including on the living span, opening for fetch and restore, data sharing of data of science, data for scientific researches, long-tail data, and providing services to customers by institutional database arouse interest from scholars.

3. Data saving and platform constructing is hot issue for experts.

4. To educate the library employees working with data, and what contents should be presented, their responsibility and position takes up attentions.

5. It also is found that some of the keywords are the names of foreign college library held projects for data curation, such as LOCKSS, DataStar, DigCCurr, DRCC, EIDCSR, D2C2, etc. Investigations of these services and projects offer lessons for domestic researches. College library, according to the statistics is the top among the institutions that provide data curation

2.5 Analysis of co-authoring

Scale of co-authoring refers to the proportion between the number of thesis that is coauthored and the total number in a periodical in certain years. Degree and rate of author partnership are indicative to the level of writers who work for some periodicals or discipline playing to their intelligent potential and once the degree is higher, the fuller the potential is played. Shown in the Table 5, in 110 texts there has existed 211 authors, papers that is co-authored of 64, resulting to co-authored degree of 1.98, rate of co-authored reaching 58.2%, a comparative high figure. This is because the job can only be performed through joint efforts by library staff, computer specialist and researchers. Their partnership serves as important means for realizing scientific yields complementing to each other's advantages, increasing intellectual exchanges and sharing, and it can boost the productive capacity of scientific employers, promote the quality and influence of the achievements. With the deepening of data curation service and analysis, collaborations among all disciplines and departments are in need as well as a strengthened partnership from between experts in the following working, making a brighter progress for data curation service and research.

Table 5: Analysis of co-authoring

Number of Authors	Number of text	Percentage (%)
One	46	41.8
Two	45	40.9
Three or above	19	17.3
Total	110	100

2.6 Analysis of fund-raised essays

There are 44 essays receiving 67 kinds of fund to support (for some may get more than one resources of fund) out of 110 texts on data curation. Number of national funded is 16, provincial funded 26, local funded 10, research funded 1, college funded 11, and association funded 3. Distribution of fund-raised essays on year scale is shown in Table 6.

Table 6: time distribution of fund-financed literatures

Year	2012	2013	2014	2015	2016	Total
Number of Text	2	11	18	12	1	44

According to the Table, funded projects on data curation has been on rise since 2012.Essays on data curation amounting to forty percent of the total volume of funded essays suggests its importance to authorities interested from various levels, and it also shows attention the experts pay to claim for raising funds and active part they play in asking supports from the supervisors, which help to form better condition for data curation service and study, to enhance their quality and level.

3. Summary

Conclusions drawn from the statistics which studies on the data curation essays contribute to are made as followed: Both publication of essays and funded projects on data curation are on the climbing track and it indicates all the more attention the scholars and authorities interested pay to the data curation researches and service. Then, In addition to stressing studies on the field, the experts also resort heavily on the places like database with the aim of making efforts to realize data curation. Moreover, care is taken to the collecting, storing and reusing scientific projects of major influence, as well as to the long-tail data generated by individual or group research having little or even no funds.

Reference

[1] Wu Minqi .（2012）. Digital Curation: An Emerging Research Field of Library and Information Science,*Library Journal*, (3), 8-12

[2] Yu Haiyan,Wei Junchao.(2014). Investigation and Analysis on University Data Curation Projects Abroad,*Library and Information Service*, 58（22）,38-47

[3] Wang Fang,Shen Jinhua. (2014). Advances in Data Curation Abroad: Research and Practice,*Journal of Library Science in China,* (7),116-128

[4] Qiu Junping, Zu Xuan, Guo Lilin**,** Xiao Ting-ting. (2015). Research status and developing trend of institutional repository of visualization analysis, *Information Studies:Theory & Application*, 38(1), 12-17

[5] Chinese Social Sciences Research Evaluation Center. Chinese Social Sciences Citation Index（2014-2015）the source journals directory. [DB/OL]. [2016-01-19].http://cssrac.nju.edu.cn/news_show.asp?Articleid=569

[6] Ma Zikun, Peng Lijuan.(2013). Analysis on Research Status and Development Trends of Domestic Institutional Repository Based on Content Analysis, *Library Tribune,* 33(5),28-32

[7] Wang Caihong, Gao Xinlin, Gao Xuyang.(2015). Big Data Guardianship Services Practice Based on Full Engish Education Project. *Information and Documentation Services*,(6),83-86

[8] Tang Jingqian;Yang Helin;Wang Xiaoqiang. (2015). Design and Practice of MetadataCuration System: A Case Study of Jinan University Library. *Library Tribune*, (5):75-83

[9] Zhang Jilong,Yin Shenqin,Zhang Yong,Guo Yaodong,Zhang Ying. (2015). Social ScienetificData Sharing and Serving – An Example of Fudan University Social Scientific DataPlatform. *Journal of Academic Libraries*, (1):74-79

[10] Li Wenlan,Yang Zuguo.(2005). Analysis on Frequencies of Keywords in Chinese Information Science Journals Papers. *Information Science*, 23（1）,68-70，143

[11] Zhao Yanzhi.(2015). The Long Tail Data and Its Guardianship in Scientific Research. *Information and Documentation Services*, (3):22-25

[12] Zhu Dan.(2005). Statistics Analysis for the Thesis of Quotation Analys Research from 1989 to 2004.*Information Science*, 23(9):1353-1356,1397

[13] Qiu Junping,Wen Fangfang.(2011). Correlation Analysis on the Relationship between the Scientific Collaboration Degree among Authors and the Output of Scientific Research.*Science & Technology Progress and Policy*, 28（5）：1-5

[14] Zhou Weihua ,Zhu Yihong.(2012). Library network information service of domestic research situation-The empirical analysis based on CNKI research paper. *Research on Library Science*, (2): 6-11

[15] Zhao Yanzhi.(2015). Curation of Long-Tail Science Data in the Library:Case Study of the UIUC.*Journal of the National Library of China*,（3）:83-84

About the author:

Guo likang, Researcher of the library of China West Normal University;

Address：Library of China West Normal University, No.1 Shida Road, Shunqing District ,

Postcode: 637002, Nanchong, Sichuan, China,

E-mail: lkguosc@sina.com