Postprint: A Scientific Knowledge Mapping Study of International Data Curation

Authors: Yu Chenlin

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] Data curation constitutes a crucial component of research data management within the information-driven scientific research environment. By systematically reviewing extant international research achievements, this study aims to foster a comprehensive understanding of data curation and provide references for domestic data management research. [Method/Process] Utilizing Web of Science as the data source, with literature searched through October 2016 using "data curation" as the search term, the retrieved publications serve as the research objects. Through bibliographic co-occurrence and co-citation analysis methods, and employing the CiteSpace III software tool, we mapped the knowledge domain of international data curation research. Content analysis was conducted to interpret, analyze, and summarize international data curation research across four dimensions: disciplinary distribution, research institutions, researchers, and knowledge base. [Results/Conclusions] International data curation research originated in 2000 and has since matured, forming specific research disciplines, institutions, and communities. The knowledge base of this research field primarily encompasses data description, integration and interlinking, data maintenance and value-added activities throughout the research process, data curation stakeholders, and new paradigms of library services.

Full Text

Preamble

ChinaXiv Partner Journal 【Academic Exploration】
Research on Mapping the Knowledge Domain of International Digital Curation
Yu Chenlin 1,2

Abstract

[**Purpose/Significance**] Digital curation is a crucial component of research data management in the e-Science environment. By systematically reviewing existing international research, this study aims to provide a comprehensive understanding of digital curation and offer references for domestic data management research. [**Method/Process**] Using the Web of Science as the data source, we retrieved literature published through October 2016 with "digital curation" and "data curation" as search terms. Based on co-occurrence and co-citation analysis methods and utilizing CiteSpace III software, we constructed a knowledge map of international digital curation research. Content analysis was employed to interpret, analyze, and summarize the research from four dimensions: disciplinary distribution, research institutions, authors, and knowledge base. [**Result/Conclusion**] International digital curation research began in 2000 and has now entered a mature stage, forming specific research disciplines, institutions, and communities. The knowledge base primarily consists of data description, integration and association, data maintenance and value-added activities in the research process, digital curation stakeholders, and new models of library services.

Keywords: digital curation; data management; research data; knowledge mapping
Classification Number: G250
Citation Format: Yu Chenlin. Research on Mapping the Knowledge Domain of International Digital Curation [J/OL]. Knowledge Management Forum, 2017, 2(3): 201-213 [citation date]. http://www.kmf.ac.cn/p/1/137/.

1 Introduction

With the development of E-Science, contemporary scientific research is characterized by data-driven exploration, where research data serves as the driving force of scientific activities. Science has entered a new paradigm of data-intensive research known as "big data science" [1]. In the era of big data, the connotation and characteristics of research data have changed fundamentally, featuring broad sources, diverse types, massive volumes, and real-time data flows—collectively termed scientific big data [2]. Traditional data management models are no longer adequate for managing research data, making data vulnerable to damage and contamination, preventing effective utilization and long-term preservation, and ultimately hindering contemporary scientific research. Scholars from various fields have conducted theoretical and practical explorations of research data curation based on their academic backgrounds. This paper systematically reviews international digital curation research to provide a holistic and comprehensive understanding of its overall landscape.

2 Definition of Digital Curation

The Digital Curation Centre (DCC) in the UK defines digital curation as the active and dynamic management of digital research data throughout its entire lifecycle for the purposes of maintenance, preservation, and value addition. Active management of research data aims to ensure its future research value and mitigate the risks of digital obsolescence. Curated data stored in trusted digital repositories can promote data sharing within the UK research community, reduce duplicated efforts in data creation, and enhance the long-term value of data by improving the availability of high-quality research [3]. The Joint Information Systems Committee (JISC) describes digital curation as a series of activities that maintain and utilize digital data and research outputs throughout their lifecycle to serve current and future users [4].

From an archival perspective, digital curation integrates digital preservation, digital library management, digital archiving, and stage-specific data management interventions into a cohesive whole. The term "digital curation" emerged because the concept of "digital archiving" had been misused in the field of information resource preservation, distorting its meaning and necessitating a new term to accurately describe the lifecycle management of digital resources [5]. The University of Illinois Graduate School of Library and Information Science proposes that digital curation involves active and continuous data management throughout the data lifecycle in academic research, science, and education activities. Through data certification, archiving, management, preservation, and description, it facilitates data discovery, long-term preservation, and value-added reuse [6].

In summary, digital curation is characterized by: (1) active, continuous, and uninterrupted data management throughout the entire research data lifecycle, ensuring a traceable and continuous chain of data management processes; (2) the purpose of maintaining and adding value to research data, ensuring its authenticity, reliability, and long-term availability to meet current and future usage needs; and (3) promoting the retrieval, discovery, sharing, and utilization of research data resources while reducing redundant construction of scientific resources.

3.1 Data and Methods

To comprehensively understand international digital curation research and avoid omitting important literature, this study selected data from the Web of Science (WOS) Core Collection database. We searched for "digital curation" OR "data curation" in titles or topics, with a time span of 1900-2016 and document types including article, editorial, letter, proceeding paper, and review. The search was conducted on October 31, 2016, and after deduplication and cleaning, we obtained 319 valid records.

The growth trend of international digital curation literature follows Price's law of exponential growth in scientific literature, with a goodness-of-fit R^2 of 0.974 (see

Figure 1 [Figure 1: see original paper]). International digital curation research began in 2000. From 2000-2005, the number of publications was small and development was extremely slow, indicating an initial stage. From 2006-2013, annual publications showed growth, with actual numbers exceeding theoretical values, marking a period of rapid growth. After 2013, actual publications fell below theoretical values, with the gap widening year by year, indicating that the field has entered a mature stage. The absolute number of annual publications continues to grow, with more than 40 papers published each year since 2013, reaching 62 papers in 2015.

This study employs scientific knowledge mapping, an interdisciplinary research method combining information visualization, applied mathematics, graphics, computer science, and scientometrics. It transforms massive amounts of literature data in scientific frontier fields into visual images, revealing the overall landscape, development trends, and structural characteristics that are difficult to obtain through personal experience alone. The specific analytical methods include co-occurrence analysis to identify research subjects in international digital curation and co-citation analysis to examine the knowledge base.

3.2 Research Subjects in Digital Curation

Using CiteSpace software's co-occurrence mapping analysis, we analyzed citing literature from three dimensions—disciplinary distribution, research institutions, and authors—to explore the research subjects in digital curation.

3.2.1 Disciplinary Distribution

As shown in Figure 2 [Figure 2: see original paper], computer science and library & information science have large node rings, indicating high publication volumes. The node rings consist of blue, green, and yellow colors, suggesting research across three time periods with long-term and sustained attention. Disciplines such as biochemical research methods, astronomy & astrophysics, computer science, imaging science & photographic technology, statistics & probability, geography, biochemistry & molecular biology, remote sensing, and genetics have nodes marked with purple circles, representing high centrality ($ $0.1) and important central positions in the network structure with significant influence.

Digital curation research is multidisciplinary, with both applied and basic disciplines actively addressing digital curation issues. This phenomenon arises primarily because: (1) research data is mainly generated by specific basic disciplines, originating from numerical records obtained through observation, detection, investigation, and comprehensive analysis in scientific research. With the information technology revolution of the 21st century, new scientific research methods and approaches have propelled data production into automated, sensor-based systems. Research data has disciplinary attributes, and basic disciplines often conduct digital curation research around specific projects to meet their particular needs for data curation within their knowledge systems. (2) Different

disciplines share common attributes in data management and services. Applied disciplines have strengthened the foundation of digital research and unified technical standards for research data, providing strong support for network infrastructure, information technology, policy guidance, and curation theory in digital curation.

Computer science research in digital curation focuses on artificial intelligence, information systems, interdisciplinary applications, software engineering, and theoretical methods, providing comprehensive technical support since 2001. Life sciences and biomedicine attach equal importance to digital curation research as computer science. With the emergence of new sequencing tools and technologies, genetic research generates massive amounts of genomic data, increasing the demand for gene data management to ensure timely updates, real-time maintenance, resource integration and association, long-term preservation, and effective access, thereby driving new scientific discoveries. Library and information science has published 84 papers, representing a substantial academic body with strong influence and serving as a major force in advancing digital curation research.

3.2.2 Research Institutions

Figure 3 [Figure 3: see original paper] shows that the University of North Carolina at Chapel Hill, University of Edinburgh, Purdue University, University of Glasgow, Johns Hopkins University, University of South Florida, and University of California, San Diego are particularly active in digital curation research.

Burst detection refers to significant changes in variable values over short periods and serves as a means to measure deeper transformations. Analyzing institutional bursts helps identify critical transition points in digital curation research. In 2007, the University of North Carolina at Chapel Hill published four papers on digital curation, focusing on talent cultivation and software tool development. Its School of Library and Information Science undertook the Digital Curation Curriculum (DigCCurr) project, which included training graduate-level professionals in digital curation and exploring curriculum design [7]; defining digital curation professionals and the skills and knowledge they should possess [8]. The Vidarch Project captured relevant information about data resources, achieving comprehensive annotation based on metadata and contextual relationships [9]; and developed the ContextMiner tool to help digital curators query, compile, and store data in databases [10].

The University of Edinburgh published four papers on digital curation between 2004-2007. Confronting the explosive growth of biological data, P. Buneman advocated for database curation to ensure data security and reliability [11]; P. Buneman also noted that archivists and curators focus on long-term preservation and reliable access to data resources, while researchers emphasize visualization, annotation, and association of data resources [12]; C. Rusbridge et al. believed that the establishment of DCC would better guide digital curation activities [13];

and M. McGinley called for incorporating digital curation into legal frameworks to effectively guide the openness or confidentiality of research data [14].

Purdue University published two papers on digital curation in 2008 . Purdue University Libraries, guided by library and archival principles and utilizing distributed institutional repository infrastructure, conducted explorations of discipline-specific research data management, providing practical case studies for digital curation research [15]. M. Y. Eltabakh developed a scalable database engine for biological databases to support unified data management for biological database systems, including annotation and storage of data and derived information, data querying, and tracking, thereby promoting research data management at Purdue University [16].

3.2.3 Author Analysis

As shown in Figure 4 [Figure 4: see original paper], the color changes in node rings reflect researchers' active periods. Based on the color changes in the time-sliced map, we categorize major researchers in digital curation into three generations, with 2006 and 2012 as the dividing points.

First-generation researchers' nodes are predominantly blue. With the advancement of e-Science, the demand for research data curation has continuously increased. The P. Buneman team advocated for and elaborated on the significance of digital curation and the establishment of digital curation centers; the P. Martin team developed integrated analysis tools for gene databases to support integrated data research. Second-generation researchers' nodes are mainly green, focusing on digital curation research in library science and computer science. The C. Prom team addressed digital curation education by leading the Digital Curation Curriculum (DigCCurr) and Closing the Digital Curation Gap programs to cultivate professionals; the L. Martinez-Uribe team studied the role positioning and service innovation of libraries in digital curation; and the S. Ross team developed text genre classification methods for automatic metadata extraction. Third-generation researchers' nodes are primarily yellow, focusing on fine-grained digital curation activities for specific disciplines. The Á. Sánchez-Ferrer team proposed specific digital curation requirements based on biological gene needs; the W. Los team established digital curation to promote open data resource sharing; the C. Jandrasits team highlighted the importance of digital curation in nanotechnology; the B. Stvilia team studied digital curation and data quality requirements from a genetics perspective; and the J. Bhate team described data curation measures such as interactive quality control and cross-curation implemented by the International Molecular Exchange Consortium (IMEx Central).

3.3 Knowledge Base of Digital Curation Research

Figure 5 [Figure 5: see original paper] reveals that the document co-citation network consists of eight main clusters. Based on cited and citing documents

and cluster labels, we interpret the research content and core viewpoints of each cluster, finding that research can be broadly categorized into: new value of digital curation for research activities, hardware and software infrastructure construction, application in specific disciplines, digital curation stakeholders, and library service models.

3.3.1 New Value of Digital Curation for Research Activities

Table 1 lists cited and citing documents from Cluster #3 (scientific data), illustrating the new value of scientific data for research activities. These studies examine how digital curation achieves data maintenance and value addition, involving research workflows and management of data sharing and publication. Scientific research is data-driven and collaboratively open; data sharing can support research reproducibility and verification, ensure research results are publicly accessible, facilitate new research using existing data, and enhance research innovation [17].

As the scientific community's awareness of the potential value of small research data deepens [18], C. Borgman used habitat ecology as an example to demonstrate how digital libraries can support data management for "small science" disciplines using embedded sensor networks, addressing the tendency of small research data to be heterogeneous, personally managed, or unsaved and unmanaged [47]. Despite massive data generation creating a "data deluge," data sharing has only emerged in a few fields. C. Tenopir et al. (2011) surveyed 1,329 scientists about data sharing practices and perceptions, finding that the primary barriers to data sharing are insufficient time and funding, followed by lack of open platforms, standards, and policies [19]. M. H. Cragin et al. conducted the Data Curation Profiles project to study data sharing issues from researchers' perspectives, analyzing data sharing behaviors across three dimensions: what data to share, when to share, and with whom [20]. P. Borgman analyzed what data should be shared, by whom, under what conditions, why, and what efforts are required, providing guidance for data policy formulation and practice [17].

M. J. Costello proposed replacing data sharing with data publication, constructing a data citation and access system to motivate environmental and biological scientists to publish research data and solve data availability issues [21]. R. R. Downs and R. S. Chen designed interdisciplinary data submission workflows to facilitate data submission for cross-domain researchers [22].

3.3.2 Hardware and Software Infrastructure for Digital Curation

Infrastructure for digital curation includes platforms supporting digital curation and software technologies enabling data integration and association. Table 2 lists cited and citing documents from Cluster #2 (biologist-centric software), focusing on infrastructure construction for digital curation, including curation software development, platform building, service system construction, and best practice exploration.

The open-source digital repository software Fedora describes complex relationships between digital objects, providing a foundation for organizations to manage and preserve digital resources [23]. iRODS (integrated Rule-Oriented Data System) data grids help users efficiently and easily manage various data resources [24]. The UK Office for Library and Information Networking summarized a service framework for digital curation, identifying key stakeholders, analyzing their responsibilities, rights, and collaboration methods, defining data management objectives (preservation, access, and reuse), and establishing mechanisms, processes, and practices to achieve these goals [25]. Purdue University Libraries built a collaborative structure for embedded research services in the e-Science environment, offering research data management services including data description, standards for types and formats, collection, organization, archiving, and preservation [26]. The University of Colorado Boulder Libraries' involvement in domain-specific scientific data curation demonstrated that libraries' advantages in professional talent, infrastructure, and information services would facilitate digital curation activities [27]. These library explorations have become best practices in digital curation.

Table 3 lists cited and citing documents from Cluster #6 (annotation), focusing on data integration and association through digital curation. These studies aim to build large-scale, knowledge-enabled scientific data networks to facilitate in-depth mining and effective interpretation of the connotations and relationships among various resource objects in research data.

The Microarray Gene Expression Data Society developed standards for microarray data, specifying the minimum information required for microarray experiment interpretation [28], promoting data exchange among international genomics laboratories and public databases. C. A. Ball reviewed microarray data standards, which standardized annotation descriptions and exchange formats for microarray experiment data, assisted in constructing microarray databases and developing data analysis tools, promoted sharing of high-quality gene expression data, and paved the way for standardization in genetic research [29]. S. A. Sansone proposed using technical means and incentive mechanisms to promote bioscience data interoperability, enhancing scientific communities' full utilization and open sharing of research data [30]. D. Howe argued that the emergence of biological research data management and biocuration could resolve the contradiction between growing demands for high-quality data and limited, outdated data management practices [31]. B. M. Good et al. built biomedical semantic network links in Wikipedia using Semantic Wiki Links, directly embedding them in the Wikipedia editor to compute semantic relationships in article contexts and enhance semantic presentation for user query and discovery [32].

3.3.3 Application of Digital Curation in Specific Disciplines

Digital curation has been extensively applied in biological sciences, cheminformatics, and bioinformatics. Table 4 lists cited and citing documents from

Cluster #0 (database), demonstrating digital curation applications in biological sciences. These studies focus on data curation activities based on domain ontologies and metadata, enabling standardized description and classification of biological data for computer processing.

With the rapid development of next-generation gene sequencing technologies, genomics and transcriptomics have entered the high-throughput sequencing era, generating massive nucleotide sequence data in laboratories and gene databases. However, inconsistent description and storage formats for nucleotide sequence data severely hinder academic exchange and resource sharing. The Gene Ontology unified standardized gene function annotation and description [33]; life science research databases adopted Gene Ontology to annotate research data; the Universal Protein Resource (UniProt) provides integrated, high-quality, accessible protein resource data for scientific communities [34]; and PlasmoDB integrates experimental and computational data on malaria parasites, associating genomic localization and transcript information to facilitate queries for malaria researchers [35]. Standardized data description, annotation, and storage formats facilitate new discoveries, and unified gene ontology terminology enables integration of high-quality data resources and discovery of evidence for gene interactions [36].

Table 5 lists cited and citing documents from Cluster #1 (QSAR modeling), illustrating digital curation applications in cheminformatics. These studies focus on curation activities throughout the research data modeling process, exploring relationships between data, extracting information, and discovering knowledge based on mathematical principles. Quantitative Structure-Activity Relationship (QSAR) modeling, as a primary method in cheminformatics, quantitatively describes relationships between chemical structures and their activities [37].

Establishing research data aggregation mechanisms and models, such as the Aggregated Computational Toxicology Resource (ACToR), Kyoto Encyclopedia of Genes and Genomes (KEGG), and genotype-phenotype databases, addresses the low efficiency of data utilization caused by multi-source and heterogeneous data. As e-Science advances and data-driven research develops, data quality directly determines research success or failure. Chemical data modeling analysis processes adopt standard specifications [38] and define analysis stages to ensure QSAR model analysis validity [39]. Given that predictive toxicology data sources involve broad disciplines and flexible data representations, F. Xin argued that digital curation can ensure high-quality computational foundations for predictive toxicology and advance the field [40]. A. J. Williams and S. Ekins advocated for digital curation in chemistry databases to guarantee data quality and promote scientific progress [41].

Table 6 lists cited and citing documents from Cluster #5 (bioinformatics), demonstrating how digital curation supports new research models in bioinformatics. J. Bellenson noted that microarray chip technology applications in identifying carcinogens and environmental hazards have shifted toxicology research paradigms from hypothesis-driven to data-driven experiments [42], making data

increasingly important for scientific research. W. Tong et al. pointed out that ArrayTrack integrates toxicology data storage, analysis, and visualization functions to support toxicology research progress and new discoveries [43].

3.3.4 Digital Curation Stakeholders and Library Services

Table 7 lists cited and citing documents from Cluster #4 (digital curation), identifying digital curation stakeholders. These studies focus on role positioning, responsibility allocation, and collaboration among stakeholders.

The U.S. National Science Board (NSB) published "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," highlighting management-level recognition of long-term digital data collection management and advocating for data management research and education to support post-2000 scientific research. Based on data management requirements at different stages, it proposed role positioning for data services across different institutions and departments to achieve collaborative data management service goals [44]. As participants in information resource management, libraries expand and extend data services by positioning management roles and responsibilities, researching technical standards and data lifecycle theories, aiming to play important roles in research data management and scientific research. H. R. Tibbo examined digital curation from a social science perspective, arguing that while digital curation development relies on computer technology support, social sciences provide more guidance for long-term curation of data assets [45].

Table 8 lists cited and citing documents from Cluster #7 (science), describing library explorations under new research models. These studies focus on library digital curation service models. L. Lyon noted that with the "information shift," libraries need to examine their institutional goals and service scope in the data-driven research environment [46]. P. Hswe and P. Hswe discussed the necessity and participation models for libraries in data management from the perspectives of staffing, infrastructure, and service positioning in academic libraries, indicating that new professional roles would emerge in libraries to meet data management needs [47]. G. S. Choudhury developed digital curation services at Johns Hopkins University based on existing institutional repositories and other infrastructure, emphasizing the roles of new positions such as data scientists and digital humanities experts in comprehensively supporting university research data management [48]. L. M. Delserone described how the University of Minnesota Libraries collaborated with institutional repositories and information technology departments to jointly plan and construct digital curation infrastructure, while configuring professional talent teams to meet library data management and service requirements and building a "science librarian team" [49]. L. Lyon summarized ten stages of library digital curation services based on the Research360 institutional research lifecycle model: data management requirements, planning, informatics infrastructure, citation, training, licensing, identification, storage, access, and impact [46].

4 Conclusion

With the information technology revolution of the 21st century, the scientific research paradigm has shifted toward data-intensive approaches, jointly driving the rise of digital curation research. Analysis and interpretation of international digital curation research indicate that research subjects are multidisciplinary. Life sciences and biomedicine conduct digital curation research around specific projects based on their disciplinary knowledge systems, while applied disciplines such as computer science and library science study generic research data infrastructure and technical standards based on common data characteristics. Research institutions are primarily concentrated in Europe and America, with the University of North Carolina at Chapel Hill, University of Edinburgh, and Purdue University being particularly active and influential in digital curation. In contrast, China's research on digital curation is relatively weak, though the School of Information Management at Wuhan University has conducted in-depth investigations and analysis on international digital curation professional talent cultivation, demonstrating strong influence. Scholarly collaboration is insufficiently close, lacking stable, high-quality research teams. The knowledge base of digital curation concentrates on new value for research activities, hardware and software infrastructure construction, applications in specific disciplines, digital curation stakeholders, and library service models. This systematic review of international English literature on digital curation aims to provide insights and references for domestic digital curation research.

References

[1] Wu Jinhong, Chen Yongyue, Hu Muhai. Research on Quality Control Models in Scientific Data Curation under e-Science Environment [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(3):
[2] Guo Huadong, Wang Lizhe, Chen Fang, et al. Scientific Big Data and Digital Earth [J]. Chinese Science Bulletin, 2014 (12): 1047-1054.
[3] What is digital curation [EB/OL]. [2017-04-10]. http://www.dcc.ac.uk/digital-curation/what-digital-curation.
[4] BEAGRIE N, POTHEN P. Digital curation: digital archives, libraries and e-Science seminar [EB/OL]. [2017-04-10]. http://www.ariadne.ac.uk/issue30/digital-curation/.
[5] CUNNINGHAM A. Digital curation/digital archiving: a view from the National Archives of Australia [J]. The American archivist, 2008, 71(2): 530-573.
[6] MURAKAMI Y. Metal fatigue: effects of small defects and nonmetallic inclusions [M]. Amsterdam: Elsevier,
[7] LEE C A, TIBBO H R, SCHAEFER J C. DigCCurr: Building an International Digital Curation Curriculum & the Carolina Digital Curation Fellowship Program [EB/OL]. [2017-04-10]. http://chinesesites.library.ingentaconnect.com/content/ist/ac/2007/00002007
[8] LEE C A, TIBBO H R, SCHAEFER J C. Defining what digital curators do and what they need to know: the DigCCurr project [EB/OL]. [2017-04-10].

http://dl.acm.org/citation.cfm?id=1255183.
[9] SHAH C, MARCHIONINI G. Capturing relevant information for digital curation [EB/OL]. [2017-04-10]. https://ils.unc.edu/vidarch/Shah-JCDL2007poster.pdf.
[10] SHAH C, MARCHIONINI G. ContextMiner: A tool for digital library curators [EB/OL]. [2017-04-10]. https://ils.unc.edu/vidarch/Shah-JCDL2007demo.pdf.
[11] BUNEMAN P, CHENEY J, TAN W C, et al. Curated databases [EB/OL]. [2017-04-10]. http://dl.acm.org/citation.cfm?id=1376918.
[12] BUNEMAN P. The Two Cultures of Digital Curation [EB/OL]. [2017-04-10]. http://www.inf.ed.ac.uk/teaching/courses/ad/lectures04/buneman.pdf.
[13] RUSBRIDGE C, BURNHILL P, ROSS S, et al. The digital curation centre: a vision for digital curation [EB/OL]. [2017-04-10]. http://ieeexplore.ieee.org/abstract/document/1612461/.
[14] MCGINLEY M. The legal environment of digital curation–a question of balance for the digital librarian [EB/OL]. [2017-04-10]. https://link.springer.com/chapter/10.1007%2F978-3-540-74851-9_{62}?LI=true.
[15] WITT M. Institutional repositories and research data curation in a distributed environment [J]. Library trends, 2008, 57(2): 191-201.
[16] ELTABAKH M Y, OUZZANI M, AREF W G, et al. Managing biological data using bdbms [EB/OL]. [2017-04-10]. http://ieeexplore.ieee.org/abstract/document/4497631/.
[17] BORGMAN C L. The conundrum of sharing research data [J]. Journal of the American Society for Information Science and Technology, 2012, 63(6): 1059-1078.
[18] BORGMAN C L, WALLIS J C, ENYEDY N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries [J]. International journal on digital libraries, 2007, 7(1/2): 17-30.
[19] TENOPIR C, ALLARD S, DOUGLASS K, et al. Data sharing by scientists: practices and perceptions [J]. PloS one, 2011, 6(6): e21101.
[20] CRAGIN M H, PALMER C L, CARLSON J R, et al. Data sharing, small science and institutional repositories [J]. Philosophical transactions of the Royal Society of London A: mathematical, physical and engineering sciences, 2010, 368(1926): 4023-4038.
[21] COSTELLO M J. Motivating online publication of data [J]. BioScience, 2009, 59(5): 418-427.
[22] DOWNS R R, CHEN R S. Designing submission and workflow services for preserving interdisciplinary scientific data [J]. Earth science informatics, 2010, 3(1/2): 87-97.
[23] LAGOZE C, PAYETTE S, SHIN E, et al. Fedora: an architecture for complex objects and their relationships [J]. International journal on digital libraries, 2006, 6(2): 124-138.
[24] HEDGES M, HASAN A, BLANKE T. Curation and preservation of research data in an iRODS data grid [EB/OL]. [2017-04-10]. http://ieeexplore.ieee.org/abstract/document/4426919/.
[25] LYON L. Dealing with data: roles, rights, responsibilities and relationships. consultancy report [EB/OL]. [2017-04-10]. http://opus.bath.ac.uk/412/.
[26] BRANDT D S. Librarians as partners in e-research: Purdue University

Libraries promote collaboration [J]. College & research libraries news, 2007, 68(6): 365-396.
[27] LAGE K, LOSOFF B, MANESS J. Receptivity to library involvement in scientific data curation: a case study at the University of Colorado Boulder [J]. portal: libraries and the academy, 2011, 11(4): 915-937.
[28] BRAZMA A, HINGAMP P, QUACKENBUSH J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data [J]. Nature genetics, 2001, 29(4): 365-371.
[29] BALL C A, SHERLOCK G, PARKINSON H, et al. Standards for microarray data [J]. Science, 2002, 298(5593): 539-539.
[30] SANSONE S-A, ROCCA-SERRA P, FIELD D, et al. Toward interoperable bioscience data [J]. Nature genetics, 2012, 44(2): 121-126.
[31] HOWE D, COSTANZO M, FEY P, et al. Big data: the future of biocuration [J]. Nature, 2008, 455(7209): 47-50.
[32] GOOD B M, CLARKE E L, LOGUERCIO S, et al. Building a biomedical semantic network in Wikipedia with Semantic Wiki Links [J]. Database, 2012, 2012: bar060.
[33] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene Ontology: tool for the unification of biology [J]. Nature genetics, 2000, 25(1): 25-34.
[34] APWEILER R, BAIROCH A, WU C H, et al. UniProt: the universal protein knowledgebase [J]. Nucleic acids research, 2004, 32(S1): D115-D119.
[35] BAHL A, BRUNK B, CRABTREE J, et al. PlasmoDB: the Plasmodium genome resource. a database integrating experimental and computational data [J]. Nucleic acids research, 2003, 31(1): 212-215.
[36] GOERTSCHES R H, HECKER M, KOCZAN D, et al. Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-β-1b treatment in relapsing remitting MS [J]. Pharmacogenomics, 2010, 11(2): 147-161.
[37] Zhou Xibin, Han Wenjing, Chen Jing, et al. Application and Research Progress of Several QSAR Modeling Methods in Chemistry [J]. Computers and Applied Chemistry, 2011, 28(6): 761-765.
[38] FOURCHES D, MURATOV E, TROPSHA A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research [J]. Journal of chemical information and modeling, 2010, 50(7): 1189-1204.
[39] TROPSHA A. Best practices for QSAR model development, validation, and exploitation [J]. Molecular informatics, 2010, 29(6/7): 476-488.
[40] FU X, WOJAK A, NEAGU D, et al. Data governance in predictive toxicology: a review [J]. Journal of cheminformatics, 2011, 3(1): 24.
[41] WILLIAMS A J, EKINS S. A quality alert and call for improved curation of public chemistry databases [J]. Drug discovery today, 2011, 16(17): 747-750.
[42] SCHENA M. DNA microarrays: a practical approach [M]. Oxford: Oxford University Press, 1999.
[43] TONG W, CAO X, HARRIS S, et al. ArrayTrack—supporting toxicogenomic research at the US Food and Drug Administration National Center for Toxicological Research [J]. Environmental health perspectives, 2003, 111(15):

1819.
[44] PRYOR G, DONNELLY M. Skilling up to do data: whose role, whose responsibility, whose career? [J]. International journal of digital curation, 2009, 4(2): 158-170.
[45] TIBBO H R. Placing the horse before the cart: conceptual and technical dimensions of digital curation [J]. Historical social research, 2012, 37(3): 187-200.
[46] LYON L. The informatics transform: re-engineering libraries for the data decade [J]. International journal of digital curation, 2012, 7(1): 126-138.
[47] HSWE P. Data management services in libraries [EB/OL]. [2017-04-10]. http://pubs.acs.org/doi/pdf/10.1021/bk-2012-1110.ch007.
[48] CHOUDHURY G S. Case study in data curation at Johns Hopkins University [J]. Library trends, 2008, 57(2): 211-225.
[49] DELSERONE L M. At the watershed: preparing for research data management and stewardship at the University of Minnesota Libraries [J]. Library trends, 2008, 57(2): 202-210.

Note: Figure translations are in progress. See original paper for figures.