Governance and Value-Added Utilization of Academic Library Patron Data: A Case Study of the 2017 Peking University Reading Report (Postprint)

Authors: Wu Yaping, Bie Liqian, Zhou Chunxia, Zhao Fei, Wang Cong

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] This paper proposes a framework for reader data utilization to guide libraries toward more standardized and systematic collection, organization, preservation, and mining of reader data, thereby achieving its value-added utilization. [Method/Process] By introducing the concepts of data curation and lifecycle-based data service extension, a lifecycle-based reader data curation framework is established. Based on this framework, and taking the "2017 Peking University Reading Report" as a case study, the paper analyzes how to conduct reader data curation for both libraries and readers. [Results/Conclusion] Reader data curation can, on the one hand, help libraries clarify patterns of reader visits and borrowing, identify core reader groups, and understand resource utilization, thereby enabling rational optimization of various resources and services; on the other hand, by recommending reading resources and services to readers, it can help them improve their reading quality and fully utilize the library, thus providing a valuable reference for the standardized and in-depth development of reader data curation and value-added utilization.

Full Text

Preamble

Reader Data Curation and Value-added Utilization in University Libraries: A Case Study of the 2017 Peking University Reading Report

Wu Yaping, Bie Liqian, Zhou Chunxia, Zhao Fei, Wang Cong
Peking University Library, Beijing 100871

Abstract

[**Purpose/Significance**] This paper proposes a reader data utilization framework to guide libraries in more systematically collecting, organizing, preserving, and mining reader data to achieve value-added utilization. [**Method/Process**] By introducing data curation and data lifecycle-based service expansion concepts, the paper establishes a lifecycle-based reader data curation framework. Using this framework, it analyzes how to curate reader data for both libraries and readers through the case study of the *2017 Peking University Reading Report*. [**Result/Conclusion**] Reader data curation helps libraries clarify reader visitation and borrowing patterns, identify core reader groups, and understand resource utilization, thereby enabling rational optimization of resources and services. Simultaneously, by recommending resources and services to readers, it helps improve their reading quality and fully utilize library resources, providing valuable reference for the standardized and in-depth development of reader data curation and value-added utilization.

Keywords: university library; reader data; data curation; value-added

1. Introduction

Library operations are inherently connected with readers, whose needs and preferences guide library optimization and development. Big data expert Xu Zipei noted at the launch of his book *Digital Civilization* that data is leading humanity toward a high-definition society and reshaping civilization [1], suggesting that data enables deeper societal understanding. Similarly, in the library field, data helps us more clearly perceive reader needs and the effectiveness of library resources and services. In the big data environment, libraries have elevated their appreciation for reader data value to new heights. The IFLA Trends Report explicitly states that librarians need a better and broader understanding of copyright, data protection, and privacy [2]. Libraries have gradually established data librarian positions with titles such as data librarian, data services librarian, data visualization analyst, data management librarian, data repository librarian, and research data specialist [3]. Innovative applications based on reader characteristic and behavioral data have permeated various services, such as analyzing reader borrowing, periodical browsing, and electronic resource access behaviors to support decision-making in resource development, layout, and space optimization. In recent years, under the development concepts of smart libraries and personalized libraries, libraries have placed greater emphasis on reader data analysis and presentation, launching various forms of data reading services including reading reports, big data reports, personal reading bills, and real-time data display screens. These represent practical implementations of reader data curation and value-added utilization. Building on this work, this paper proposes a reader data curation framework to promote more standardized and in-depth development of reader data utilization.

The term "curation" originated in Western cultural heritage preservation, where in museology it refers to "curating"—planning, selecting, and exhibiting—and "curator" designates personnel who conceive, organize, and manage the preservation and exhibition of fragile and valuable artifacts [4]. In the e-Science research environment, "Digital Curation" was first proposed in 2001. The UK's Digital Curation Centre (DCC) defines it as all activities involved in managing, maintaining, and preserving data throughout its lifecycle to achieve data value-added [5]. In 2002, Microsoft Chief Researcher and Turing Award winner J. Gray introduced "data curation," noting that ephemeral data has irreplaceable and non-reconstructible characteristics that must be preserved, such as daily temperature records that cannot be recovered if not documented [6]. Domestic scholars typically translate this as data curation, data guardianship, or data stewardship to express its meaning of enabling data reuse and value-added through selection, organization, and storage, thereby maintaining data viability, presentability, and comprehensibility [7]. The data here encompasses digital assets, digital data, digital research data, raw materials, datasets, etc. [8]. This paper adopts "data curation" as the translation.

Reader data, as a type of ephemeral data, plays a crucial role in optimizing library resources and services and must be timely preserved, managed, and value-added. Data value-added involves extensive collection of domain-specific data, followed by validation and correlation, with correlation results displayed through visualization technology to achieve data value-added while protecting the data [9]. The data curation process encompasses key steps of data value-added and serves as an important pathway. As owners and providers of information resources, libraries must not only collect, organize, and mine information but also artistically present information products according to user habits or specific themes [10], fulfilling the "curator" role. In libraries' gradually developing big data reading services, data must also possess scientific rigor, real-time availability, controllable lifecycle activities, and dynamic maintainability of data management—qualities for which data curation serves as an important guarantee of service quality [11].

2. Related Research

2.1 Data Curation as an Important Pathway to Value-Added

Annual reading reports, personal reading bills, reading lists, and real-time data display screens all represent manifestations of reader data curation. Annual reading data reports have gradually matured, with nearly 100 libraries having published 2017 reading data reports as of January 26, 2018 [12]. These can be categorized by perspective into comprehensive reading reports and personalized reading reports. Comprehensive reports are library-centered with strong holistic and group characteristics, providing overall understanding of resource utilization and reader reading patterns. Personalized reports are reader-centered, describing individual reading behaviors with stronger reader presence and participation. Beyond year-end statistics, reader data reports demonstrate regularization and

refinement. For example, Shenyang Normal University Library [13] released summer vacation reading data, Heilongjiang University Library [14] released winter vacation reading data, and Shenyang Normal University Library also pushes monthly popular borrowing lists [15]. Many libraries have even launched real-time data display screens. User groups have been segmented from the entire reader population to freshmen, graduates, alumni, etc. For instance, Peking University Library launched "Freshman Special: When Newcomers Meet the Library" [16] for new students, "Book • Time" graduation memorial card sets for graduates supporting online generation and sharing of electronic reading memorial cards [17]; Xiamen University Library launched "Carp • Time: A Gift for Graduates" [18]; and Sichuan University Library launched "A Unique Love Letter from the Library for Graduates" [19]. These personalized reading data displays record readers' reading footprints, are highly collectible, and deeply welcomed by readers.

With advances in digital media display technology and display equipment such as LCD and LED screens, real-time data displays have become a new library service promotion method. Shanghai Library uses visualization and multimedia display technology to intuitively and real-time present business statistics, providing decision support for libraries while serving as a reading promotion method for big data applications [20]. Scholars have proposed that libraries could even establish dedicated personalized areas for readers, building personalized knowledge models based on borrowing history so that after swiping their cards, readers can see potentially interesting books and their basic information and shelf status [21]. As a new service form, big data reading services represent manifestations of reader data curation—collecting and processing data based on established goals, mining data connotations, and artistically presenting them through data visualization to guide libraries in better optimizing resources and services and help readers improve their reading quality.

2.3 Data Visualization Technology Maturing and Playing a Pivotal Role in Data Curation

Data visualization is a key step in data curation. Today, people have lost interest in monotonous and conservative presentation methods, expecting more intuitive and efficient information presentation [22]. Over 80% of human information is obtained through the visual system [23], and "seeing is believing" demonstrates that people trust information acquired visually. Data visualization provides an easy-to-read and aesthetically pleasing visual experience, more efficiently conveying data connotations and now applied across multiple fields. Visualization variables including position, shape, direction, color, texture, grayscale, and size continue to expand [24]. As an efficient information display form, visualization technology has permeated multiple library business scenarios: (1) In collection resource guidance, intuitively displaying library resource layout, mostly used for new student orientation and daily guidance; (2) In resource revelation, extracting resource utilization and reader reading patterns to form holistic under-

standing; (3) In resource retrieval, achieving visualization of collection resource association representation, query process, and search results [26]; (4) In business statistics, visual display facilitates more efficient management decisions, such as Xiamen Library's business data visualization platform that centrally displays reader information and collection usage, providing scientific basis for collection procurement, digital collection development, and weeding [27]. Additionally, there are applications in library chronicle visualization based on Timeline [28] and citation analysis [29].

Today, libraries value data visualization more and have established dedicated positions to provide more professional services. Duke University, UC Berkeley, and Harvard University have all established data visualization librarian positions requiring capabilities in data extraction, cleaning, transformation, analysis technology, and data visualization [30]. Comprehensive data visualization technologies can be categorized into standard 2D/3D techniques for low-dimensional data such as bar charts, line charts, and pie charts, and new technologies for multi-dimensional, large datasets including geometric projection, image-based, pixel-oriented, and hierarchical techniques [31]. Data visualization tools have also developed rapidly, including entry-level tools like Excel, open-source tools like E-Charts and D3.js, professional tools like R and Gephi, and commercial tools like Tableau. With continuous upgrades, visualization tools feature stronger operability, real-time capability, dynamism, rich display forms, and support for multiple data formats. Visualization technology can more intuitively, quickly, and accurately capture data connotations and will play a key role in reader data curation.

3. A Lifecycle-Based Reader Data Curation Framework

Data curation encompasses all activities of managing, maintaining, and preserving data throughout its lifecycle to achieve value-added. The data lifecycle is a cyclical process from data generation, processing, and publication to reuse. Data lifecycle-based service expansion can be divided into three categories: initial data processing, data reprocessing, and knowledge extraction [32]. Combining data curation connotations with lifecycle-based data service expansion, this paper proposes a library reader data curation framework comprising three stages and eight modules.

3.1 Initial Data Processing Stage

The initial data processing stage forms the foundation and guarantee for value-added, with its content and framework shown in Figure 1 [Figure 1: see original paper]. The data collection stage aims to comprehensively gather reader data, which includes four aspects: (1) Reader group data such as electronic resource access, helping objectively reflect development trends and identify popular resources from a holistic perspective; (2) Personalized data containing reader IDs such as borrowing records, enabling personalized services; (3) Bibliographic data

in OPAC systems, more detailed in describing book features as important extensions of book attributes in reader behavioral data; (4) Reader campus card data, as important extensions of reader identity information in behavioral data.

The data preprocessing module after collection is crucial for value-added utilization, mainly comprising four components: (1) Data cleaning, including filling missing data through third-party sources (e.g., matching and supplementing "department unknown" records in borrowing data using campus card data to minimize information uncertainty) and deleting irrelevant data (e.g., removing binding records mixed in borrowing records); (2) Data standardization, unifying different representations of the same attribute value (e.g., standardizing both full and abbreviated department names to full names); (3) Data reduction, deleting redundant attributes (e.g., keeping only department name while removing department code); (4) Feature creation, creating new features based on analysis objectives (e.g., creating faculty/school fields based on department fields to explore borrowing characteristics across faculties). Data storage runs through the entire process, preserving both raw data and processed results.

3.2 Data Reprocessing Stage

The data reprocessing stage represents important value-added realization, with content and framework shown in Figure 2 [Figure 2: see original paper]. Based on standardized heterogeneous, multi-source data from the initial processing stage, the first step is establishing data connections to increase data correlation, mainly including two aspects: On one hand, integrating reader behavioral data (campus card data, borrowing records, library entry records, reservation records, renewal records, etc.) based on reader ID to expand reader characteristics and comprehensively reconstruct reader behaviors in the library; on the other hand, integrating reader behavioral data with OPAC bibliographic data based on book ID to expand book features in reader behavioral data.

After establishing multi-dimensional data connections, general statistical analysis can be conducted from dimensions like time and resource type to identify development trends and discover popular authors and resources, or cross-analysis across multiple attributes can reveal favorite authors among male/female readers and resource utilization differences across faculties. Data analysis and visualization complement each other: data analysis endows visualization with value and meaning, while visualization helps extract value and knowledge from data [33]. When employing data visualization, principles and objectives must first be clarified, and appropriate methods and tools selected accordingly, ensuring final results feature high "ink-ratio," clear perspectives, combined dimensions, appropriate comparisons, and dynamic interactivity, thereby fully utilizing human visual bandwidth to enhance users' thinking capacity and understanding efficiency [34].

3.3 Knowledge Extraction Stage

The knowledge extraction stage represents the highest level of value-added, with content and framework shown in Figure 3 [Figure 3: see original paper]. Knowledge extraction is the prerequisite for knowledge services. In big data environments, knowledge services must synchronize with explosive data growth and socialization trends, focusing on fragmented information, user behaviors, and relationships, targeting real-time data, unstructured data, and machine data generated thereby, and invoking internal, external, and public information to make forward-looking data judgments [35]. Reader data curation should better utilize internal library data, other campus departments' data, and publicly available online data. Beyond internal data, on one hand, campus data such as course grades and research achievements should be integrated to mine more correlations—for example, ACRL reports indicate students using library services more frequently (borrowing, database access, interlibrary loan) often achieve higher academic success (course grades, GPA) [36]; on the other hand, bibliographic data should be obtained through multiple channels to expand book features, as bibliographic data serves as a book's "business card" and window for users, and single sources cannot comprehensively describe a book's characteristics. The Web 2.0 resource sharing philosophy has brought increasing online resources, with rating, tagging, and review data from book e-commerce sites like Amazon and review communities like Douban serving as important supplements to library OPAC bibliographic data.

Obtaining large reader datasets through data fusion forms the foundation for knowledge extraction, with data mining technology as an important means. Currently, data mining technology is mainly applied to reader analysis research, resource development optimization, intelligent services, automated information processing, and personalized services [37]. Reader data curation services require deeper mining of data connotations through association mining to find readers with similar reading interests and similar content resources; through cluster analysis to conduct refined services for more specific reader and resource groups; through social network analysis to identify key readers and resources for optimal reading promotion strategies; and through recommendation systems to achieve precise personalized reading recommendations.

4. Reader Data Curation Results and Presentation

The *2016 Peking University Reading Report* [38] was the first annual reading report launched by Peking University Library, systematically reviewing reading resources and services and analyzing/ presenting library utilization characteristics. The *2017 Peking University Reading Report* [39] optimized content and form based on its predecessor, enhancing standardization, comprehensiveness, and aesthetics as an important achievement in reader data curation and value-added utilization. Building on the curation framework and processes in Section 3, the 2017 report primarily conducted curation for two aspects: library service optimization and reader reading optimization.

4.1 Curation for Library Service Optimization

Curation for library service optimization aims to clarify reader visitation and borrowing patterns, identify core reader groups, and understand resource utilization to help libraries rationally optimize resources and services.

4.1.1 From Reader Visitation Perspective Reader visitation is no longer limited to physical library visits, with homepage visits and WeChat follows serving as important channels. 12:00-12:59 is the peak hour for average daily visits, with variations across periods—physical visits significantly increase during exam weeks and decrease during holidays (see Figure 4 [Figure 4: see original paper]), enabling libraries to allocate human resources more scientifically based on reader density. In WeChat official accounts, the collection search column receives the most clicks and users, demonstrating that WeChat is not only a promotional tool but also a popular functional resource discovery portal.

4.1.2 From Resource Utilization Perspective Databases, multimedia electronic resources, and more convenient resource search/access methods like discovery systems and self-service are increasingly popular. From 2013-2017, "Weiming Academic Search," electronic resource full-text downloads, and multimedia resource online searches and on-demand playback continued to rise, making electronic resources and related services increasingly important. Reading devices upgrade rapidly, with Kindle series readers currently best meeting reader needs, leading reservation lists. Self-service also plays an important role, with self-service borrowing accounting for 50.34% of total borrowing, exceeding manual borrowing. Branch libraries play important roles in the literature guarantee system, with increasing numbers opening to all readers. In 2017, branch libraries' total external borrowing reached over 80,000 volumes, with the School of Information Science and Technology branch ranking first in cross-department borrowing at 50%. Different collection methods for different literature types align with actual borrowing patterns, demonstrating rationality—open-shelf borrowing exceeds closed-shelf, storage library, and interlibrary borrowing (see Figure 5 [Figure 5: see original paper]). The proportion of closed-shelf and storage library borrowing in 2017 was significantly higher than in 2016 due to the library's collection reorganization for the east library renovation project, which compressed open-shelf space, with data reflecting increased closed-shelf, storage library, and interlibrary borrowing.

4.1.3 From Reader Characteristics and Reading Habits Perspective Humanities and social sciences faculty readers are the primary user groups of library spaces and physical resources. Humanities books show stronger interdisciplinary penetration, occupying larger proportions across all faculties' borrowing subject categories. In terms of visits, humanities and social sciences readers account for 30% and 21% of visitors respectively, together comprising half of all visitors; in borrowing, they account for 66% (humanities 44%, social sciences

22%). The School of Foreign Languages has the highest total visits and borrowing volume, while the School of Marxism has the highest per capita borrowing and proportion of readers with borrowing behavior. Readers' interdisciplinary reading trends are evident (see Figure 6 [Figure 6: see original paper]), showing that each faculty's readers engage with multiple disciplines beyond their own.

4.2 Curation for Reader Reading Optimization

Curation for reader reading optimization aims to recommend resources and services to help readers improve reading quality and fully utilize the library. It begins by introducing library resources, services, and operating hours to all readers, then presents popular resources including top borrowed Chinese and foreign books, most reserved books, popular e-books, and popular textbooks, as well as favorite authors among male/female readers (see Figure 7 [Figure 7: see original paper]), and finally displays 2017 reading activities on a timeline including humanities lectures, film lectures, Peking University's first reading marathon, "Book Sounds Distant" World Book Day series, and cultural workshops, attracting readers to review lecture videos and follow future series, stimulating learning interest while promoting reading resources.

Current practice in reader data curation and value-added utilization primarily focuses on initial data processing and data reprocessing stages. Through curation for both library service optimization and reader reading optimization, it helps libraries conduct panoramic scans of resources and services to support decision-making in collection development, service layout, and staffing, while helping readers comprehensively understand library resources and services and discover high-quality, popular resources to enhance their reading quality. The limitation is that practice in the knowledge extraction stage remains limited, with real-time data-driven, refined service segmentation, and personalized push being future optimization and application directions for reading reports.

Modern information technology has driven libraries toward smart library development since the beginning of this century. The arrival of the artificial intelligence era will inevitably drive current smart library services to meet social service demands with intelligent characteristics, evolving toward intelligent libraries [40]. Reader data curation and value-added utilization are important means to enhance library intelligence characteristics in the big data environment. Based on reading report practice, this paper preliminarily establishes a lifecycle-based reader data curation framework.

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