Postprint: Research on the Theoretical Framework for Data Quality Assurance Strategies in Web Archiving

Authors: Wang Wenling, Yunpeng Qu

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

Abstract

[Purpose/Significance] Data quality assurance constitutes a critical task that permeates the entire web archiving workflow, determining the success of web resource archiving endeavors. [Method/Process] Through analysis, research, and comparison of quality assurance strategies and methods from preservation institutions both domestically and internationally, this study proposes a strategic theoretical framework for data quality assurance. [Results/Conclusion] The framework adopts a data-centric approach, establishing a series of business standards and operational protocols, leveraging existing software tools to implement full-process data quality inspection, while supplementing these efforts with team building, operational environment maintenance, and authorized website backup acquisition to ensure the procurement of high-quality archived data.

Full Text

Research on the Theoretical Framework of Web Archiving Data Quality Assurance Strategies

Wang Wenling[1], Qu Yunpeng[2]

[1]National Library of China, Beijing 100081
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract

[Purpose/Significance] Data quality assurance is a critical component of web archiving that permeates the entire workflow and determines the success or failure of web resource preservation efforts. **[Method/Process]** Through analysis, research, and comparison of quality assurance strategies and methods employed by preservation institutions both domestically and internationally, this paper

proposes a strategic theoretical framework for data quality assurance. [**Result/Conclusion**] This framework is data-centered, establishing a series of business standards and operational specifications. It utilizes existing software tools to conduct comprehensive data quality inspections throughout the entire workflow, supplemented by team building, maintenance of operational environments, and authorized acquisition of website backups to ensure the acquisition of high-quality archived data.

Keywords: web archiving, quality assurance, quality inspection
Classification Number: G251
Citation Format: Wang Wenling, Qu Yunpeng. Research on the theoretical framework of web archiving data quality assurance strategies [J/OL]. *Knowledge Management Forum*, 2018, 3(2): 106-115 [citation date]. http://www.kmf.ac.cn/p/131/.

1 Introduction

In the practice of web archiving, various data quality issues frequently arise, including missing website content files, inability to display multimedia content, and formatting errors. Without rigorous quality control measures for these problems, important information may be lost, resulting in low data quality and even mission failure. According to the "Web Archiving Life Cycle Model" published by the Internet Archive' s technical team in 2013, quality inspection and analysis occupy the inner loop of the web archiving lifecycle, serving as a crucial step between the previous collection-storage-organization phase and the subsequent collection-evaluation-selection phase. This step determines the direction of future work, making data quality assurance one of the key factors affecting the success of web archiving.

Generally, three major evaluation metrics define archived data quality: appearance integrity, interactive completeness, and data consistency. High-quality web archives refer to the complete capture of knowledge content from target websites within the shortest possible time, while preserving the visual content and browsing experience. Different archiving institutions establish varying quantitative metrics based on their specific collection requirements and budget constraints to define applicable quality evaluation standards. Web archiving data quality assurance encompasses the measures and methods employed by archiving institutions to ensure that collected web resources meet predetermined quality standards, including both automated execution and manual intervention, covering the entire workflow before, during, and after collection.

Given the limitations of web archiving technology and the complexity of web resources, perfect archiving is unrealistic. Numerous international web archiving institutions, including the International Internet Preservation Consortium (IIPC), the Library of Congress, and the French National Library, have implemented corresponding data quality assurance initiatives. These organizations

research data quality assurance issues based on their specific web archiving needs, develop tailored quality assurance strategies and methods, and attempt to solve quality problems and improve collection data quality to varying degrees. The primary objective of this study is to investigate quality assurance practices at domestic and international web archiving institutions, summarize and deeply analyze their methods and approaches, and propose a universal theoretical framework for data quality assurance strategies.

In 2014, B. R. Ayala and colleagues from the University of North Texas conducted a survey on web archiving quality assurance practices under IIPC sponsorship. The survey targeted IIPC member institutions and some non-member organizations through document analysis, email exchanges, conferences, and face-to-face communications, covering web archiving data quality issues, institutional attitudes toward quality inspection, quality assurance methods and approaches, and solutions to various quality problems. Survey results revealed that the vast majority of institutions conduct quality inspection concurrently with web resource collection, with fewer than 5% never performing quality checks, and 11.1% implementing full-process quality control during collection. This demonstrates that web archiving institutions attach great importance to quality assurance, considering data quality among the most critical issues in web archiving.

---

2 Quality Assurance Practices at Domestic and International Institutions

2.1 International Web Archiving Institutions

2.1.1 Drexel University Drexel University conducts web archiving focused on higher education resources using Archive-It as its preservation tool. Quality control is primarily achieved through manual inspection by staff to ensure the availability of important content. Staff use Excel forms to record seed collection status, then examine basic issues in Archive-It' s automatically generated quality assurance reports, such as whether collection volume is excessive, whether data has entered storage queues, and compliance with robots.txt protocols. Subsequently, staff identify seed errors (whether collection is misdirected to other websites) and embedded document issues (missing content or display failures). After completing these fundamental quality controls, staff make necessary modifications and perform patch collection or re-collection to ensure archived content matches the original website.

2.1.2 French National Library In a 2006 work report, the French National Library described various quality issues encountered in their web archiving legal deposit work and shared their quality control methods. They argued that quality inspection methods for massive automatically collected data should be determined based on data volume and structure. For broad-domain collection,

where data is disordered, the primary quality inspection tasks involve checking, describing, and validating data to ensure accurate storage and preservation. Main methods include collecting and analyzing collection log reports, checking general technical environments and software operation status, and sampling collected data to verify extractability and accessibility.

For focused collection with limited seed lists or regular collection of specific websites, where resources are relatively orderly, systematic verification and inspection should be conducted. They developed a tool component for more refined automated inspection: the first module removes invisible characters from URLs and performs deduplication; the second module validates URL effectiveness, checks whether URLs have been archived, detects robots.txt protocols, and analyzes website geolocation; the third module automatically compares seed lists with existing collection log reports, which is particularly useful for quality inspection during collection.

2.1.3 Bentley Historical Library, University of Michigan The Bentley Historical Library (BHL) operates two web archiving projects: the University Archives and Records Program (UARP) and Michigan Historical Collections (MHC). In 2011, they published quality control guidelines and procedures for these projects, which have undergone multiple revisions. The guidelines analyze potential content and technical issues in web archiving and propose detailed quality control procedures and operational specifications: (1) confirm quality control objectives and inspect WAS (Web Archiving Service, Archive-It' s predecessor) quality control tool reports; (2) confirm successful initiation and completion of archiving processes; (3) verify collection settings correctness and metadata accuracy; (4) determine whether particularly important content is missing (defined as content indispensable for understanding the website' s main content or key functions, without needing to concern individual missing images, audio, video, or text unless crucial for research value); (5) resolve prominent quality issues by adjusting collection settings, contacting website owners, or re-collecting; (6) document the entire quality control process in detail.

The collection team found that automated quality assurance processes significantly improved collection quality but did not noticeably enhance replay quality.

2.1.4 Library of Congress The Library of Congress, in collaboration with the Internet Archive, has experimented with semi-automated quality assurance. Although manual operations remain necessary, certain stages have been fully automated. Their web archiving collection process based on collection frequency includes: (1) **Pre-collection**: Only the seed' s homepage is collected to detect issues with seed lists or SURT format (seed list attachment files), enabling real-time adjustments; (2) **Collection**: Collection proceeds according to established frequency, with any detected issues reported to the web archiving team within 24 hours. After collection, CDX files (indexes of collected URLs) and WAT files (metadata files for all WARC files) are generated; (3) **Automated quality as-**

surance: Browser analysis simulation uses the PhantomJS browser emulator and Wayback replay software to replay important seeds, capture website snapshots, and record HTTP response codes to generate per-page reports. These reports extract lists of missing files requiring re-collection, which are added to collection software for supplementary collection. Pig scripts simultaneously analyze WAT index files for link analysis, classify external links by type, and check whether all external link objects, especially embedded objects (such as CSS and JS files), have been collected. Hadoop tools compare collected resources in CDX files with external link objects to identify uncaptured resources for supplementary collection candidate lists; (4) **Supplementary collection**: Identify whether quality issues are replay or collection problems and collect needed supplementary content; (5) **Manual quality assurance**: Data administrators browse archived content in proxy mode, inspect log reports, and verify whether required content has been fully collected.

2.2 Domestic Web Archiving Institutions

2.2.1 Peking University The Institute of Network and Distributed Systems at Peking University developed the Web Infomall system early on, dedicated to archiving, organizing, and providing services for Chinese Internet web pages. By September 2013, the system had preserved 8.5 billion web pages totaling 73TB. Compared to other web archiving projects, Web Infomall employs a relatively simple collection strategy, storing only static information from web pages using a self-developed Tianwang storage format with incremental archiving. The collected web information is distributed free of charge under common public licenses to research institutions and serves as a corpus for data mining research and applications, providing data sources for four derivative data service systems. Under the premise of meeting project objectives, the static-only archiving strategy significantly reduces collection difficulty and failure rates. The structured Tianwang storage format' s fault tolerance further ensures collection data quality.

2.2.2 National Library of China As China' s only IIPC member, the National Library of China began web information resource collection and preservation experiments in 2003 and established the National Library Internet Information Resource Preservation and Protection Center in 2009. Through over a decade of exploration and practice, large-scale collection has been achieved, with total web navigation and resource collection reaching 114.73TB by the end of 2016. The library conducts two types of web archiving: website collection (primarily for government and organizational websites) using breadth-first strategy, and topical collection (for major events like the 19th Party Congress) using depth-first strategy. Both types have detailed resource selection principles, standards, and seed importance ranking criteria. For data inspection, tool software is fully utilized for replay checks, with corresponding inspection mechanisms, workflows, and operational specifications established, including inspection methods and sampling rates.

Currently, the National Library focuses on mobilizing capable domestic public libraries to participate in joint web resource construction, thus establishing detailed collection strategies and strict quality assurance specifications. Quality inspection remains primarily manual through log analysis and replay checks to ensure uniform collection data quality.

2.3 Comparative Analysis of Quality Assurance Practices

Drexel University' s quality control occurs post-crawling through manual analysis of collection logs and software quality reports. The French National Library employs different quality assurance strategies for broad-domain and topical collections: log analysis, sampling checks, and software/hardware status checks for broad collection, and custom-developed tools for semi-automated quality control for topical collection. The Bentley Historical Library has formulated detailed quality assurance workflow specifications at the strategic level to minimize quality issues caused by individual factors. The Library of Congress extends quality assurance to all collection stages, adding pre-collection phases, monitoring crawling processes, using software replay with automatic quality issue recording, achieving a high degree of automation throughout the quality control process. Peking University primarily considers the timeliness characteristics of web resources, aiming to quickly archive network resources locally with relatively low data quality requirements limited to format checks.

As shown in Table 1 , due to varying requirements in collection needs, budget constraints, and timeliness, each archiving institution employs different methods and invests varying time and effort in quality assurance work. All institutions have selected measures appropriate for their project characteristics. Evidently, more complex and sophisticated quality assurance measures yield higher data quality.

Table 1 Comparison of Quality Assurance Practices

Institution	Process/Specification	Automation Level	Timeliness Guarantee
Drexel University	Log analysis, quality report analysis	Post-collection	—
French National Library	Pre/post-collection log analysis, sampling checks	—	—
Bentley Historical Library	Quality report analysis	—	—
Library of Congress	Pre/mid/post-collection automated replay, log analysis	High	Yes

Institution	Process/Specification	Automation Level	Timeliness Guarantee
National Library of China	Log monitoring, replay checks	—	—

Note: "—" indicates information not mentioned in available materials, not absence of measures.

3 Strategic Framework for Web Archiving Data Quality Assurance

Without considering specific project requirements or costs, and in pursuit of the highest possible archived data quality, this paper proposes a strategic theoretical framework for web archiving data quality assurance (see Figure 1 [Figure 1: see original paper]). This data-centered framework establishes a series of business standards and operational specifications, utilizes existing software tools for full-process data quality inspection, and supplements these with team building, environment maintenance, and authorized website backup acquisition to ensure high-quality archived data.

Figure 1 Framework for Web Archiving Data Quality Assurance Strategy

3.1 Establishing Strict Collection Business Standards and Operational Specifications

As the investigation reveals, web archiving quality assurance relies primarily on manual operations, while quality control experts possess varying backgrounds, technical levels, and proficiency. To avoid data quality issues caused by human factors, unified business standards and strict operational specifications should be established. Recommended standards include:

(1) Data Quality Standards. The previously mentioned concept of high-quality web archives represents an idealized qualitative description. Archiving institutions should balance three factors: complete preservation of knowledge content, complete preservation of visual content and browsing experience, and rapid completion of collection tasks. Quantifiable and operable standards should be established based on collection objectives and requirements, forming the foundation of all quality assurance work.

(2) Data Format Standards and Metadata Specifications. Common web archiving formats include WARC, ARC, and KW, among which WARC is both an international and national standard and the preferred format in

web archiving, gradually replacing ARC and KW industry standards. Archival website object metadata specifications should be developed to facilitate future management of website objects and archived data, including website title, main content tags, collection time, capacity, URL quantity, and geographic location, where geographic location and content tags can be used to filter whether websites meet collection requirements.

(3) Software Usage Standards. Web archiving requires software tools including collection software, distributed storage software, antivirus software, and replay software. Corresponding usage standards should specify selection scope, versions, and standard configurations. For example, Heritrix and Wget are the most commonly used collection software, with WarcCreate as a desktop user-oriented tool. In practice, Heritrix may be designated as the sole collection software with strict rule limitations to ensure data consistency.

(4) Seed Selection and Ranking Criteria. For both domain and topical collections, seed selection can reference well-known statistical rankings like Alexa or authoritative existing website lists. Establishing selection standards ensures original website data quality. After seed selection, seeds meeting requirements should be prioritized based on certain characteristics, with collection time, frequency, and scope set for each seed website. Ranking criteria enable more targeted and orderly web archiving.

(5) Crawler Default Configuration. Depth and breadth collection are the most commonly used strategies. Collection teams should provide default configurations for both strategies based on business needs, making minimal modifications for specific seed collection requirements to reduce subjective configuration errors and improve data quality.

(6) Various Operational Specifications. As web archiving involves significant manual intervention, establishing executable workflows and specifications for each procedural step reduces randomness and error probability, including pre-collection workflow specifications, log inspection procedures, virus screening protocols, and software replay quality inspection workflows.

3.2 Conducting Comprehensive Data Inspection

Direct data quality inspection is the most effective quality assurance measure and the core of quality assurance work, which should permeate the entire web archiving workflow. Based on timing, data inspection can be divided into pre-collection, during-collection, and post-collection checks, each with different purposes and content.

3.2.1 Pre-Collection Inspection Pre-collection analysis of target website structure and content is crucial for successful collection. Analysis enables timely strategy adjustments to avoid quality issues caused by multi-domain names and external links, helping quality control experts determine whether website content meets collection requirements and complies with robots.txt protocols. Pre-

collection generally uses generic collection strategies and configurations, with content not written to local files but only crawler logs obtained for analysis.

During log analysis, quality control experts must identify incomplete, inaccurate, or unsuccessful collection results, determine non-compliant collections, and identify root causes. This may involve confirming crawling settings, reviewing crawl reports and logs, examining target website content, layout, features, and source code, and recording any technical limitations, crawler protocol exclusions, or other issues preventing accurate capture. Pre-analysis should determine collection depth, breadth, and frequency; identify script-generated resources causing collection failures; detect crawler traps; determine which servers host required content; establish collection rules; and verify seed list completeness and accuracy.

3.2.2 During-Collection Inspection During-collection inspection primarily ensures normal crawler operation and successful task completion. Tasks include: (1) real-time monitoring of crawler operation for technical issues like memory overflow, crawler traps, or network problems; (2) regular log checks to verify correct collection of required content, identify unforeseen pre-collection issues, and assess crawler setting rationality. These checks enable timely problem resolution and ensure successful collection.

3.2.3 Post-Collection Inspection Post-collection quality inspection 主要包括: (1) data validation to verify compliance with established formats; (2) virus scanning; (3) checking whether archived files can reproduce original website appearance through log analysis and software replay.

(1) Data Validation. The first step is checking compliance with data standards and performing integrity verification. Most institutions adopt the international WARC format as the standard, which can be validated using tools like JHove2 and WARC Tools.

(2) Virus Scanning. While not mandatory, virus scanning addresses the reality of contemporary web characteristics. From a business perspective, data quality has two standards: meeting predetermined objectives and reflecting objective reality. If archiving serves users, resources must be virus-free; if for preservation, virus presence may be tolerated. Common antivirus software (Kaspersky, Avira) can be used, but understanding each software' s handling mechanism is crucial—some remove virus files from WARC files, some delete entire WARC files causing data loss, and some overly sensitive to scripts cause false positives.

(3) Log Analysis. Experienced experts can identify technical issues like automatic redirects and external references through log analysis, discovering problems based on file quantities and task duration. Special attention should be paid to server response errors and timeouts, as crawler abandonment after multiple failed attempts causes information omission. Such resources require analysis and potential inclusion in supplementary collection lists.

(4) Software Replay. Using specialized replay software to render archived content enables manual verification of website completeness, link validity, and interactive completeness through clicking and viewing. Due to manual operation requirements, this method is difficult for massive collections and can only be performed through sampling as a supplement to log analysis.

3.3 Implementing Semi-Automated Quality Assurance

To improve automation and reduce manual workload, increasing web archiving software tools integrate quality assurance modules, with numerous specialized auxiliary tools available. Proper utilization enables quality control experts to achieve more with less effort.

3.3.1 Crawler Software Heritrix is the most widely used crawler software in web archiving, providing robust support for pre-collection, during-collection, and post-collection quality control. Heritrix offers powerful task configuration with dozens of granular collection rules covering scope, protocols, content types, and output formats, supplemented by regex filtering, directory depth filtering, redirect filtering, and SURT filtering for compliant URLs. Flexible rule combinations enable highly targeted collection strategies for complex tasks.

Heritrix provides a console for real-time monitoring of task progress, speed, runtime, threads, and queue status, issuing alerts for access timeouts, startup failures, or parsing errors. Detailed reports record resource quantities, types, capacities, and domains, enabling task administrators to analyze collection status per domain and identify failure causes for rule refinement.

3.3.2 Collection Software Packages **Web Curator Tools (WCT)** is an open-source web archiving tool suite including collection authorization management, task scheduling, quality inspection, data validation, and metadata management. Using Heritrix as its crawler, WCT provides full-process quality control. Its most notable feature is a graphical quality inspection tool displaying collected resource URLs in a tree structure rooted at seed URLs, with statistics including total URLs, successful collections, failures, and object sizes. Experts can prune unnecessary resources or import missing URLs/files, with WCT automatically updating archives upon saving.

NetArchiveSuite, developed jointly by the Danish Royal Library and Statsbiblioteket, is a complete web archiving package for planning, scheduling, and collection. It provides ViewProxy, a dedicated replay inspection tool combining browser emulators for archived resource browsing, collecting missed URLs for supplementary collection, and directly re-collecting lost URLs.

3.3.3 Auxiliary Tools **Monitrix** is a front-end monitoring and analysis software specifically designed for Heritrix 3, offering: (1) real-time task monitoring with visual graphics and statistics; (2) timeline generation showing per-unit

statistics including data volume, URL counts, new hosts, and completed hosts; (3) detailed statistics browsing including host counts, URL counts, warning counts, and virus detection; (4) single-host analysis including first/last access times, HTTP response code pie charts, virus check charts, MIME type distributions, and subdomain lists.

3.3.4 Specialized Quality Inspection Software **Jhove2** is a renowned open-source format validation software widely used in long-term preservation, supporting WARC standard document analysis and validation since version 2.1.0. **JWAT (Java Web Archive Toolkit)** provides a graphical interface for reading and validating WARC/ARC documents. **Warc tools**, IIPC-funded, offers multiple scripts for WARC document processing including validation, automatic summarization, and ARC-to-WARC conversion. **Wayback Machine**, developed by the Internet Archive, indexes and replays WARC/ARC documents with a user search interface. **OpenWayback**, its Java version led by IIPC, implements most Wayback Machine functions and is currently the mainstream replay software.

3.4 Other Strategies

3.4.1 Strengthening Web Archiving Team Building As web archiving quality assurance relies primarily on manual work, quality control experts' professional capabilities directly affect outcomes. Qualified experts should master Internet-related knowledge including data transmission technology, web development, and network hardware, plus strong mathematical, logical reasoning, and programming abilities. Facing rapid Internet technology development, continuous team training and capability enhancement provide talent guarantees for quality assurance.

3.4.2 Maintaining Operational and Network Environments Web archiving success depends not only on software functionality but also on hardware and network environments. Maintaining robust operational environments is prerequisite for high-quality archiving. Collection teams should establish strict server and hardware management regulations, utilize network hardware monitoring equipment and software, and conduct regular inspections to provide optimal environments.

3.4.3 Direct Acquisition of Website Resource Backups Web crawlers are inherently flawed archiving technology, simulating human browsing without complete intelligence, thus never perfectly presenting original websites. If preservation institutions cooperate with resource owners to obtain direct website backups—including backend databases, embedded resources, and dynamic scripts—after resolving intellectual property issues, this would represent the "ultimate" solution. Though currently impractical at scale, this approach merits exploration for small-scale trials.

3.5 Summary and Recommendations

Through investigating and analyzing quality assurance measures at domestic and international web archiving institutions, this paper proposes a universal theoretical framework for archived data quality assurance. This framework is not based on specific project requirements nor considers manpower and material costs, serving as a general reference framework. Among the three major quality metrics—appearance integrity, interactive completeness, and data consistency—this framework emphasizes the first two while giving less consideration to data consistency (collection timeliness). D. Denev et al.' s SHARC framework employs quality-focused collection timing strategies, increasing collection frequency for rapidly changing pages to ensure data consistency. Archiving institutions should select specific methods from this framework based on project objectives, requirements, and budget constraints.

---

4 Conclusion

Web archiving currently faces two major challenges: intellectual property issues and crawler technology limitations. China' s web archiving legal deposit legislation remains absent. From a web culture preservation perspective, relevant legislation should be promoted to enable preservation institutions to bypass technical restrictions like robots.txt and comprehensively preserve website content. Before legal resolution, web crawling remains the primary collection method. Future efforts should enhance crawler capabilities to solve rich application encapsulation collection problems, thereby improving collection quality.

References

[1] BRAGG M, HANNA K. The Web Archiving Life Cycle Model[EB/OL]. [2018-03-12]. https://archive-it.org/static/files/archiveit_{{{life}}*{{cycle}}}*{model}.pdf.
[2] Wang Wenling, Qu Yunpeng. Preliminary study on web archiving data quality issues[J]. *Digital Library Forum*, 2018(4): 8-13.
[3] AYALA B R, PHILLIPS M, KO L. Current quality assurance practices in Web archiving[EB/OL]. [2018-02-05]. https://digital.library.unt.edu/ark:/67531/metadc333026/m2/1/high_{r
[4] ANTRACOLI A, DUCKWORTH S, SILVA J. Capture all the URLs: first steps in Web archiving[EB/OL]. [2018-03-01]. http://palrap.pitt.edu/ojs/index.php/palrap/article/view/67/370
[5] ILLIEN G. Sketching and checking quality for Web archives: a first stage report from BnF[EB/OL]. [2016-05-05]. http://bibnum.bnf.fr/conservation/bnf-qualityforwebarchives-feb06.pdf.
[6] SHALLCROSS M. Quality assurance for the Bentley Historical Library Web archives: guidelines and procedures[EB/OL]. [2018-03-01]. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/94162/BHL{WebArchivesQA}-v3-20130909.pdf.

[7] Yan Hongfei, Huang Lian' en, Xie Zhengmao, et al. Web Infomall: a large-scale Web archiving system[C]//Proceedings of the Web Resource Collection and Digital Resource Long-term Preservation Symposium. Beijing: National Library of China Publishing House, 2013.
[8] National Library of China. National Library of China 2017 year-book[EB/OL]. [2018-03-12]. http://www.nlc.cn/dsb_{footer}/gygt/ndbg/nj2017/201712/P020171220578252 1
[9] Heritrix[EB/OL]. [2018-03-12]. https://webarchive.jira.com/wiki/spaces/Heritrix/overview.
[10] NetArchiveSuite[EB/OL]. [2018-03-12]. https://sbforge.org/display/NAS/NetarchiveSuite.
[11] JHOVE2[EB/OL]. [2018-03-12]. https://bitbucket.org/jhove2/main/wiki/Home.
[12] CLARKE N. Java Web archive toolkit[EB/OL]. [2018-03-18]. https://sbforge.org/display/JWAT/Overview
[13] Hanzo. WARC Tools project[EB/OL]. [2018-03-18]. http://netpreserve.org/projects/warc-tools-project/.
[14] Wayback machine[EB/OL]. [2018-03-18]. http://wayback.archive-it.org/.
[15] OpenWayback[EB/OL]. [2018-03-18]. https://github.com/iipc/openwayback/wiki.
[16] DENEV D, MAZEIKA A, SPANIOL M. The SHARC Framework for data quality in Web archiving[EB/OL]. [2018-03-12]. https://domino.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/AuthorEditorIndividualView/0de8d19ced5a8ae7c1257849005270a3/$FIL vldbj.pdf.

Author Contributions:
Wang Wenling: Responsible for data collection, analysis, and paper writing.
Qu Yunpeng: Proposed the research idea and revised the paper.

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