Research on Key Technologies and Development Trends of Next-Generation Institutional Repositories (Postprint)

Authors: Cui Haiyuan, Sun Chao, Pengcheng Luo

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

Abstract

[Purpose/Significance] By investigating the research status and service requirements of next-generation institutional repositories both domestically and internationally, analyzing their key technologies and functional characteristics, and proposing development trends, this study provides recommendations for the construction and development of next-generation institutional repositories. [Method/Process] Through literature review, this paper summarizes the technology and functional development trends of institutional repositories, and based on relevant research findings, analyzes 11 key technologies, standards, and protocols for next-generation institutional repositories. Finally, drawing upon research and practical experience, it proposes development trends for next-generation institutional repositories in terms of framework, functionality, and service objectives. [Results/Conclusion] The development trends of next-generation institutional repositories include: transitioning from institutional academic repositories to institutional information infrastructure; from self-archiving to automatic submission; from standalone platforms to integration and development with research management systems; from academic achievement management platforms to academic resource service centers; from academic achievement data retrieval to big data semantic research support; from achievement archiving to new academic communication communities; and from applied bibliometric indicators to establishing a new academic evaluation system.

Full Text

Research on Key Technologies and Development Trends of Next-Generation Institutional Repositories

Cui Haiyuan, Sun Chao, Luo Pengcheng
Peking University Library, Beijing 100871

Library and Information Service, Vol. 63, No. 19, October 2019

Abstract

[**Purpose/Significance**] By investigating the research status and service demands of next-generation institutional repositories both domestically and internationally, this paper analyzes key technologies and functional characteristics, proposes development trends, and provides recommendations for constructing next-generation institutional repositories. [**Method/Process**] Through literature research, we summarize the technological and functional development trends of institutional repositories. Based on relevant research findings, we analyze 11 key technologies, standards, and protocols for next-generation institutional repositories. Finally, through research and practical experience, we propose development trends for next-generation institutional repositories in terms of framework, functionality, and service objectives. [**Result/Conclusion**] The development trends of next-generation institutional repositories include: transforming from institutional academic archives to institutional information infrastructure; shifting from self-archiving to automated submission; evolving from independent platforms to integration and development with research management systems; transitioning from scholarly output management platforms to academic resource service centers; moving from academic output data retrieval to big data and semantic research support; changing from output archiving to new academic communication communities; and progressing from applied metrics to establishing entirely new academic evaluation systems.

Keywords: Institutional Repository; Next-Generation Institutional Repository; Key Technology; Development Trend

Classification Number: G251.7

Driven by the open access movement, global institutional repositories (IRs) have developed rapidly. As of March 2019, the number of IRs registered on the Directory of Open Access Repositories (OpenDOAR) increased from 88 in December 2005 to 3,996 [1]. However, compared with the widespread acceptance of open access concepts, the rapid growth in IR numbers, and the relentless efforts of global open access advocates, institutional repositories have yet to play their expected role in the scholarly communication system. Commercial databases and publishers remain the primary channels for academic exchange, and the goal of breaking the commercial monopoly on scholarly communication through open access has not been achieved. There remains a significant gap in quality and influence among IRs worldwide.

In 2016, MIT, a pioneer in IR development, celebrated that its IR had achieved a 44% deposit rate for faculty publications since the launch of its OA policy. In the same year, Oregon State University and the University of Nebraska-Lincoln

achieved deposit rates exceeding 40%. However, U.S. university IRs have long maintained deposit rates below 50%. In 2016, the University of California's IR deposit rate was only 25%, while other U.S. universities without deposit mandates had even lower rates. Meanwhile, the scholarly communication environment has been transformed by the widespread application of new technologies such as big data, cloud computing, ubiquitous networks, virtual reality, and artificial intelligence. In this new ecosystem, how can IRs play an effective role? Can global IRs leverage new technologies to form synergistic strength? How can IRs meet new challenges and seize new opportunities? Research and development of next-generation institutional repositories have become an inevitable choice.

2 Literature Review

2.1 Research and Practice on IR Function and Service Enhancement

Both domestically and internationally, there have been active explorations into IR function expansion and value-added services. In China, Ma Jianxia [3] proposed trends in IR content construction and service design, such as implementing mandatory deposit policies, adopting flexible access strategies, simplifying deposit procedures, integrating into user information environments, achieving economies of scale through IR consortia, providing knowledge auditing and capability analysis functions, offering long-term preservation services, and ensuring sustained technical and team support. Zhang Xiaolin [4] identified three future development trends for IRs: supporting non-textual information storage and utilization, supporting educational and research activities, and supporting institutional strategic knowledge management, along with a series of potential service functions. Liu Wei, Zhu Zhongming, Wu Zhiqiang, et al. [5-8] researched content visualization knowledge graphs, image retrieval, audio-visual resource support, and 3D model retrieval technologies for application in IRs to provide value-added services. Zhang Wangqiang et al. [9] simplified user submission processes and enabled automatic archiving through interoperability protocols.

Hong Kong University of Science and Technology's IR centers on scholars, showcasing their achievements and using visualization technology to construct collaborator networks while linking scholars' ScopusID, ResearcherID, and ORCID to comprehensively display academic trajectories [11]. The University of Hong Kong's IR associates papers with scholars and funders [12]. The Hong Kong Polytechnic University's IR displays detailed bibliometric indicators for papers, including Scopus citation counts, Web of Science citation counts, access frequency, download counts, and Altmetric information [13].

Internationally, research on IR function expansion and service enhancement has been more extensive and in-depth. L. Sterman et al. [14] used rich visualization tools to provide enhanced statistical and bibliometric data services and proactively pushed information services to authors based on access patterns. A. Cocciolo [15] studied and compared Web 2.0 applications for user engage-

ment in IRs, finding that Web 2.0 technologies help increase user interest and participation. J. Richard [16] explored how interoperability can facilitate IR development, potential use cases, and implementation methods. Research and implementation on integrating IRs with research management systems have become a trend. The University of Hong Kong's research management system extends its IR by adding a CRIS module to DSpace, implementing a CERIF-compatible DSpace-CRIS system [17]. King's College London [18] and Queen's University [19] have added IR modules with open access functionality to their research management systems. The University of St Andrews achieves interoperability between its IR and research management system through APIs [20].

COAR believes that leveraging a network of over 3,000 repositories distributed globally to share research outcomes can provide a comprehensive view of global research and enable every institution and scholar to participate in the worldwide scientific and scholarly research network. Establishing additional services such as standardized usage metrics, peer review, and social networks on top of this repository network will facilitate IR development and change the current situation where commercial publishers dominate the scholarly communication system [21].

2.2 Next-Generation Institutional Repository Research

In April 2016, the Confederation of Open Access Repositories (COAR) established a Next Generation Repositories Working Group to investigate new user requirements and propose new functions and technical solutions. In November 2017, COAR released the working group's research report "Next Generation Repositories—Actions and Technical Recommendations," which introduced research findings and proposed recommendations for applying new technologies, standards, and protocols to next-generation IRs to help repositories integrate into the web environment and play a greater role in the scholarly communication ecosystem.

The vision for next-generation IRs is to make repositories the foundation of a distributed, networked scholarly communication system that is more research-centric, open, and supportive of innovation, while being collectively managed by the academic community. An important component of this vision is that repositories will provide open access to diverse research outputs, support broad dissemination of scholarly achievements, and gain formal recognition in research assessment processes. The report describes 11 new functions and related technologies, standards, and protocols for developing new services including social networking, peer review, notifications, and usage statistics: Exposing identifiers; Declaring licenses at a resource level; Discovery through navigation; Interacting with resources (annotation, commentary, and review); Resource transfer; Batch discovery; Collecting and exposing activity metadata; User identification; User authentication; Exposing standardized usage metrics; Long-term preservation [22].

Led by COAR, research and application of next-generation IR functions, technologies, and development have become a hot topic in the IR field. On September 4, 2018, 11 major European research funding agencies from France, the United Kingdom, the Netherlands, Italy, and other countries, with support from the European Research Council (ERC), jointly signed a new open access initiative—cOAlition S (hereinafter referred to as "Plan S"). The core principle of Plan S states: "From January 1, 2020, all research projects funded by the aforementioned 11 countries and the European Research Council must publish their research results in fully open access journals or platforms." As an OA2020 initiative, Plan S is accelerating the pace of open publishing globally and changing the traditional scholarly publishing landscape [23]. In November 2018, the Wellcome Trust and the Bill & Melinda Gates Foundation joined Plan S, updating their open access policies: from January 1, 2020, all funded project papers must be open access and available through PMC and Europe PMC, and they will no longer cover article processing charges for hybrid open access journals [24].

From December 2-4, 2018, the 14th Berlin Open Access Conference was held in Germany, with funding agencies, research and education institutions, and libraries from 37 countries participating to further coordinate and promote immediate and comprehensive open access policies. Participants unanimously agreed to ensure authors retain copyright, strive for immediate open access to all papers, establish short-term transitional agreements to convert subscription journals to open publishing (at no additional cost and adjustable with market transformation), and called on publishers to work with the international research community to achieve immediate open access to all papers. Representatives from China's National Natural Science Foundation, National Science and Technology Library, and Documentation and Information Center of the Chinese Academy of Sciences released a position statement at the conference, clearly expressing China's support for OA2020 and Plan S, and supporting immediate open access to publicly funded research papers [25].

Plan S implementation guidelines require IRs to be registered or applying for registration in OpenDOAR. Additionally, they must follow these standards: Provide automated deposit functionality; Preserve full text in XML format according to standards such as JATS; Provide high-quality metadata in standard interoperable formats, including publication DOI, deposit version (AAM or COR), open access status, and deposit version license; Comply with cOAlition S metadata standards; Provide open APIs allowing others (including machines like search engines) to access content; Provide quality assurance mechanisms to link full-text content with core abstracting and indexing services (e.g., PubMed); Ensure long-term reliable operation; and Provide help desk services [26].

Plan S has already influenced the development direction of IRs. COAR responded to Plan S by supporting it in next-generation IR technical specifications, while also raising concerns and suggestions regarding some technical standards: Automated deposit solutions are not yet mature and should not be

mandatory; XML format is too resource-intensive—repositories should comply with full-text access standards such as Signposting rather than mandating XML storage; Open API requirements are too vague and impractical—partial recommended API suggestions should be provided; Requirements for allowing OpenAIRE to harvest data should be added; and Help desk services are provided by most institutional websites and need not be mandatory for IRs [27]. DSpace version 7.0, under development and planned for release in 2019, is already enhancing functions such as automated deposit, XML/JATS support, and more diverse API services according to Plan S [28].

3 Technology and Functional Characteristics of Next-Generation Institutional Repositories

Next-generation institutional repositories transcend the traditional definition of IRs as "platforms for digitizing, collecting, and preserving complete institutional scholarly outputs" [29], elevating the definition to: the foundation of a distributed, global, networked new scholarly communication system that is research-centric, provides open value-added services for research needs, and supports diverse research outputs. On this foundation, next-generation IRs can enable various value-added services including academic evaluation, peer review, and academic social networking, comprehensively supporting open scholarship and research innovation.

3.1 Functions, Technologies, Standards, and Protocols of Next-Generation IRs

COAR proposes that next-generation IRs should incorporate 11 technologies, standards, and protocols. Figure 1 [Figure 1: see original paper] compares the architecture of next-generation IRs with existing IRs [22]. In next-generation IRs, integrating new technologies such as cloud computing, search, and content management enables the design of a new IR cloud infrastructure that provides more service protocols, standards, and services to support the development of additional value-added services for IRs.

Figure 1. Architecture Comparison Between Next-Generation and Existing IRs

(1) Exposing Identifiers. When accessing academic portals such as IRs, users can easily locate target web pages, bibliographic record links, and author identities. However, due to different content presentation methods across portals, it is difficult for data harvesting services like search engines. How can users and machines conveniently locate and cite IR resources and enable smoother data exchange? Unique identifiers for data (metadata and outputs) in IRs provide a viable solution.

Signposting is a method for making the scholarly web more machine-friendly. It uses typed links to distinguish recurring patterns in academic portals. For

any media type resource, typed links are provided in HTTP link headers. For HTML resources, they are also provided in HTML link elements. Signposting uses typed links (in HTTP link headers, HTML elements, or ResourceSync elements) to identify recurring patterns in academic portals. Signposting can support automatic discovery of various resources related to scholarly objects, including bibliographic descriptions, persistent identifiers, license information, authors, or various resources that are part of the object. Adopting Signposting on websites allows machines to locate academic portal content in a unified manner, facilitating data interoperability. The HTTP header method offers many benefits: it can be used for any media type resource, not just HTML. Therefore, images, datasets, PDFs, etc., can all use the same method to clarify patterns. Headers can be accessed using HTTP HEAD requests that return only transaction metadata without content, allowing retrieval of headers for large resources such as big datasets or high-resolution images without actually downloading them. Similarly, HTTP HEAD requests can be used to obtain headers for restricted content, including paywalled articles [30].

(2) Declaring Licenses at a Resource Level. How can both users and machines clearly understand the intellectual property licensing status of IR resources? Adding clear license identifiers in content organization and license information in HTTP links is an effective solution. Through Signposting methods, adding intellectual property licenses such as Creative Commons Copyright Licenses to HTTP links is one solution.

(3) Discovery through Navigation. IRs contain rich data types. One metadata record may correspond to PDF and/or HTML versions of papers, one or more supporting datasets, book or table attachments, etc. To help machines accurately identify data objects and achieve precise search and navigation, providing data link relationships and data types in HTTP links is an effective solution. Through Signposting methods, adding a set of web resource information to HTTP links is a solution.

(4) Interacting with Resources (Annotation, Commentary, and Review). Numerous studies have proven that providing user interaction functions can increase user engagement. By integrating third-party social media services that allow users to annotate, comment, and review, IRs can function as academic communication centers, promoting researcher discussion and collaboration.

ActivityStreams 2.0 is a method for describing interactions with resources, including comments, likes, and shares. Interactions are expressed as JSON-LD using the ActivityStreams 2.0 vocabulary. Although this vocabulary targets general social network activities, it can be extended with academic vocabularies [31]. The Web Annotation Model and Web Annotation Protocol are specialized methods for expressing annotations (including comments and reviews) and the protocols for creating and managing them. Annotations use RDF-based vocabularies for expression and can be presented as JSON-LD. This protocol is HTTP-based and follows REST design principles [32]. The International Image Interoperability Framework (IIIF) is a protocol supporting image interoperabil-

ity APIs for image reuse, sharing, and interaction. Applying the IIIF protocol enables functions such as manipulation, commentary, citation, sharing, and authenticated access to images [33].

(5) Resource Transfer. Distributed, networked, cloud storage models are the core architecture of next-generation IRs, requiring distributed deployment of resource content. Cloud storage and computing technologies have matured and can support required applications.

IPFS is a peer-to-peer hypermedia protocol designed to make the web faster, safer, and more open. Applying the IPFS protocol can meet the needs for sharing large data collections among multiple parties [34]. ResourceSync is a sitemap-based specification that repository managers can use to provide information allowing third-party systems to continuously synchronize with resources in their repositories, including creation, updates, and deletion. Sitemaps enable public repository content and metadata needed by search engines. ResourceSync can use Sitemaps XML format to implement content and metadata discovery and synchronization [35]. SWORD (Simple Web-service Offering Repository Deposit) is a lightweight protocol for depositing content from one location to another [36].

(6) Batch Discovery. As IRs develop, users need unified, cross-platform repository resource discovery services and full-text content search. Implementing global repository academic search functionality is an important goal of next-generation IRs. Using ResourceSync, Signposting, and Sitemaps protocols enables batch search and enhances repository resource value. Sitemaps provide an easy method for search engines to crawl website content. In its simplest form, a Sitemap is an XML file listing each available resource URL along with optional metadata (e.g., modification date, change frequency) to help crawlers obtain data timely and accurately [37].

(7) Collecting and Exposing Activities. IRs need to actively and in real-time collect and expose activities (including any modifications, additions, comments, annotations, peer reviews, accesses, downloads, etc.), and send real-time notifications to relevant users, providing various value-added services needed by users to make IRs academic communication communities. Implementing notification mechanisms requires not only unique identifiers for resource objects and user identity authentication but also application of multiple standard protocols and technologies.

ActivityStreams 2.0 provides semantic definition specifications for resource activity information, offering structured description methods for activities through JSON format and vocabulary specifications. Linked Data Notifications is a generic notification protocol describing how servers (receivers) handle messages pushed by applications (senders) and how other applications (consumers) retrieve these messages. Any resource can notify a receiving endpoint (inbox). Messages are defined in RDF format and can contain any data. Any resource can notify an inbox that publishes notifications related to that resource, such

as annotation, commentary, or review information, notifying about interactions with the resource, interaction content, and participants. Notifications are expressed as JSON-LD using the ActivityStreams 2.0 vocabulary [38]. ResourceSync Change Notifications is a publish/subscribe protocol based on WebSub that sends subscribers notifications about repository resource modifications (creation/update/deletion). ResourceSync notifications can be used for content and metadata discovery and synchronization, using the Sitemaps XML format [39]. Webmention is a peer-to-peer trackback/pingback method designed to notify resources of link changes, supporting bidirectional links [40]. WebSub is a publish/subscribe protocol where publishers notify subscribers of resource updates. IRs can interact with publishers through WebSub to timely obtain paper citation, commentary, and review data [41]. Other messaging protocols (e.g., AMQP, Kafka) provide common communication mechanisms for all web content publishers and subscribers.

(8) Identification of Users. All value-added functions and services such as resource interaction and activity exposure require users to have unique identity identifiers and need to identify relationships between resources and users. Identity identifiers can use ORCID, Social Network Identities, WebID, etc. ORCID (Open Researcher and Contributor Identifier) provides researchers with a permanent digital identifier and automatically links researchers with scholarly activities by integrating with major research workflows (e.g., manuscripts and publications) to identify researcher outputs [42]. Social Network Identities are provided by multiple social networking platforms. WebID is a proxy HTTP(S) URI, typically created by an agent (person, organization, device, etc.) in its domain. WebID uses RDF-based machine-readable profiles, typically combined with WebID/TLS authentication and Web Access Control authentication methods [43].

(9) Authentication of Users. Providing user interaction and personalized value-added services requires user identity identification and authentication, including academic identities (e.g., ORCID) and social network identities (e.g., Twitter, Google, Facebook, Weibo, Mastodon).

HTTP Signatures provide an authentication method similar to WebID/TLS. In addition to authentication, Sign HTTP Messages allows verification that communication between client and server has not been tampered with. This method is currently being standardized by the IETF and deserves further attention [44]. OpenID Connect 1.0 is a simple identity layer on top of the OAuth 2.0 protocol for distributed identity verification. OpenID Connect allows client applications (e.g., IRs and browsers) to authenticate through user identity providers. After successful authentication, basic user information can be returned to client applications. The protocol is extensible, allowing developers to use optional features such as identity data encryption, OpenID provider information, and session management. Major social media platforms already support OpenID Connect, and ORCID is currently in the testing phase [45]. WebID/TLS is a protocol for secure user authentication based on Transport Layer Security (TLS), X.509

certificates, and WebID. It enables users to authenticate by simply selecting the required certificate from those provided by the browser, solving the problem of servers obtaining user private key information and user WebID. Through WebID, personal information containing the user's private key is obtained and verified. Although WebID/TLS is a fully distributed and efficient method, it has not been widely adopted due to difficulties in certificate generation and user interface issues [46].

(10) Exposing Standardized Usage Metrics. By sharing user interaction data, IRs can develop and provide more value-added services needed by users. Collecting, managing, and providing standardized usage metric data is an important service that allows authors and all users to understand the value of IRs. To ensure data accuracy, reliability, and trustworthiness, common standard protocols, methods, and interoperability are needed to provide users with complete metric data. The significance would be even more profound if a global IR metric system could be established based on data, providing an evaluation system independent of commercial journals. By combining quantitative data with qualitative data from user interactions (annotations, comments, reviews), IRs could potentially achieve this goal.

The COUNTER standard enables users to obtain usage statistics for electronic resources. This "code of practice" ensures that vendors and publishers can provide consistent, reliable, and comparable usage data to users [47]. SUSHI is an ANSI/NISO standard defining an automated request and response model for harvesting electronic resource usage data, used together with COUNTER. ETag or entity tag is part of HTTP and one of several mechanisms HTTP provides for web cache validation, allowing clients to make conditional requests. This enables more efficient caching and saves bandwidth, as web servers need not send full responses if content has not changed. ETags can also be used for concurrency control to prevent simultaneous updates from overwriting resources, helping systems obtain only new metric data [48].

(11) Preserving Resources. The significance of open access lies not only in providing open access to current scholarly resources but also in permanent access and long-term preservation. Long-term preservation does not require each repository to operate independently but should establish a network for long-term preservation of institutional and global scholarly resources through standards, protocols, and interoperability. Preservation requires maintaining complex interconnections among resources, metadata, and structural information, as well as implementing real-time data acquisition and preservation through new technologies. Data formats should attempt to use reusable formats (e.g., LaTeX and TEI, rather than PDF). Long-term preservation is an extremely complex activity involving policies, standards, practices, and technologies that require attention, research, and application.

3.2 Development Trends of Next-Generation Institutional Repositories

3.2.1 From Institutional Academic Archives to Institutional Information Infrastructure The goal of next-generation IR construction is to become critical infrastructure for the new global scholarly communication ecosystem, capable of managing diverse types of scholarly resources including papers, books, reports, data, software, and tools, and serving as an academic resource center providing services for scholarly communication worldwide. Next-generation IRs are network repositories enabling data interoperability, user-friendly and machine-friendly repositories, and global scholarly communication infrastructure providing facilities, data, and services for every scholar and institution. Their design is based on complete scholarly resource data collection to build academic resource management and service platforms, achieving academic resource management and service functions. This enables: improving global scholarly communication information infrastructure and environment; promoting interdisciplinary exchange and collaboration; providing complete, rich, diverse, and innovative academic information for researchers worldwide; providing an open scholarly information environment for researchers as part of world-class academic exchange; and participating in the reconstruction of global scholarly communication to become part of new scholarly communication rules.

Next-generation IRs have the following functional characteristics: Cloud platform architecture based on open frameworks providing multiple open service interfaces for data interoperability and collaboration with global scholarly communication information infrastructure; Support for management and services of multiple resource types including scholarly outputs, archival materials, research data, and software tools; Compliance with open identifier standards ensuring all data can interoperate with global IRs based on identifiers; Provision of long-term preservation, management, and services for scholarly resources following the OAIS framework, supporting exposed identifiers, and providing data format identification and migration services; Provision of unified, complete academic search services. Through open service frameworks and interfaces, next-generation IRs can either provide new discovery services superior to existing academic discovery tools or collaborate with better academic discovery services to provide users with unified, cross-platform repository resource discovery and global repository academic search functionality.

Taking Peking University's IR as an example, by December 2018, the repository had collected 540,000 metadata records and 300,000 full-text documents from Peking University since 1949, gradually establishing a complete institutional academic paper database. Based on this data construction, a research management system output subsystem was built to become an institutional scholarly output repository. Future improvements will enhance functional services to achieve the goal of building institutional scholarly output information infrastructure.

3.2.2 From Self-Archiving to Automated Submission Next-generation IRs transform from self-archiving to automated submission workflows through interoperability technologies and collaboration with publishers, providing multiple deposit methods including self-archiving, automated data harvesting, proxy submission, and cross-system collaboration. Multiple submission functions are implemented: Through interoperability and collaboration with database providers to achieve automated data submission, or negotiating with publishers for direct content provision; Providing data tools and interfaces embedded in user workflows to reduce self-archiving steps. For example, implementing direct submission from Word based on the SWORD interoperability protocol. Data tools can automatically extract or generate metadata to improve metadata quality; Data interaction between global IRs. Taking the U.S. National Science Foundation's IR (NSF-PAR) as an example, NSF-PAR establishes interoperability with the Department of Energy (DOE) and the Office of Scientific and Technical Information (OSTI) repositories. Since spring 2018, authors of publications funded by both agencies can deposit the final version of their manuscripts once. Authors who successfully deposit eligible publications in the DOE/OSTI system can now seamlessly disseminate their publications through NSF-PAR, which exchanges data with PubMed [49].

3.2.3 From Independent Platforms to Integration and Development with Research Management Systems Practice has proven that integration with research management systems makes IRs more sustainable, better supports institutional academic and research management activities, and achieves the goal of next-generation IRs becoming information infrastructure. Through integration, IRs provide collection, preservation, management, and services for scholarly resource data, offering automated submission services for researchers and services for research management. Research management processes and researcher confirmation of scholarly output data support IRs in completing data verification workflows, obtaining complete and accurate output data, and establishing relationships between outputs and scholars to build high-quality scholarly output repositories, laying the foundation for providing more value-added services and implementing academic evaluation.

Integration can take multiple forms: Integration based on IRs, i.e., extending IRs to add research management functions. The University of Hong Kong currently uses this approach by adding a CRIS module to DSpace to implement a CERIF-compatible DSpace-CRIS system. Integration based on research management systems, i.e., adding open-access-enabled IR modules to research management systems. King's College London and Queen's University currently use this approach by adding open-access front-end interfaces to Pure, allowing users to download research output data. Integration based on system interoperability, i.e., IRs and research management systems operate independently and achieve interoperability through APIs. The University of St Andrews currently uses this approach, using Pure to collect and organize relevant output data and pushing open-access data to the IR to continuously inject fresh outputs.

Peking University's research management output subsystem, managed by the library, extends IR functions to achieve unified management and services for research management and institutional repositories. The output subsystem includes seven modules (as shown in Figure 2 [Figure 2: see original paper]): login authentication, permission management, output submission, output claiming, output awards, data statistics, and API interfaces. The original DSpace system could not meet current requirements, so extensive secondary development was conducted on DSpace. All these functional modules are implemented in the IR, while to better meet user habits, modules such as output claiming, output submission, and data statistics are also implemented in the research management system. The development and application of Peking University's research management output subsystem expands IR functions into research management processes and, through system interoperability, ultimately achieves integration between Peking University's IR and research management system, injecting stronger vitality into the IR.

Figure 2. Modules of Peking University's Research Management Output Subsystem

3.2.4 From Scholarly Output Management Platforms to Academic Resource Service Centers With the rapid development of big data and artificial intelligence, data and software tools have become increasingly important. Research data, as important outcomes of scientific research, has gained growing attention from the international academic community and publishing sector. Supporting the management of diverse types of scholarly resources (including data and software tools) and research data services to transform into academic resource service centers has become one of the important functions of next-generation IRs. Currently, multiple IRs provide data services in various ways: some build data service platforms to collect research data and provide services (used by Peking University, Fudan University, Harvard University, and others—Peking University also links IR outputs with data used through persistent identifiers such as Handle and DOI); others directly extend their IR platforms to collect research data and provide services.

D. J. Lee interviewed 15 IR managers from 13 large U.S. research universities about research data management services to study what research data services IRs can provide [50]. The University of Bristol extends research data services on its IR platform [51]. Next-generation IRs need to support massive data management and services in information architecture and provide research data services offering complete research lifecycle data services for users.

3.2.5 From Massive Academic Output Data Retrieval to Big Data and Semantic Research Support The rapid development of big data and artificial intelligence technologies, combined with semantic search and machine learning, provides more comprehensive and accurate search results. Commercial databases have begun applying and releasing corresponding services. For

example, the patent database InnovationQPlus (innovationqplus.ieee.org), a collaboration between IEEE (ieee.org) and ip.com, uses semantic search to enable searching by concepts rather than keywords. By constructing semantic relationships, it searches for equivalent words and phrases when executing queries and uses machine learning to improve the accuracy of its conceptual search. AI companies Luminance (luminance.com) and iManage (imanage.com) use machine learning and pattern recognition technologies to scan massive legal documents, analyze data, and assist lawyers in analyzing legal contracts [52]. Next-generation IRs need to research and apply artificial intelligence and machine learning technologies to build semantic search and provide data mining and text mining functions to serve users' research needs.

3.2.6 From Output Archiving to New Academic Communication Communities In the new scholarly communication ecosystem, online community exchange has become an important scenario for academic communication. Next-generation IRs prioritize value-added services that interact with users (annotation, commentary, review, subscription, proactive push, etc.), aiming to become user academic communication communities through these services. Social media platforms such as Facebook, Twitter, Weibo, and WeChat have accustomed people to obtaining and exchanging information through social media and communities. Based on cloud service architecture and complete academic resource centers, next-generation IRs provide value-added services needed for academic communication, building IRs into new academic communication communities where people exchange scholarly information.

Changes in technology, education, and the academic communication environment are transforming IRs' roles and values. Finding new roles in the evolving scholarly communication lifecycle has become the goal of next-generation IR construction.

3.2.7 From Providing Metrics to Establishing a New Academic Evaluation System Extensive practice and data have proven that providing metrics such as access statistics and citation frequency can effectively enhance the academic impact and visibility of IR outputs, with numerous IRs already providing various statistics. Next-generation IRs need to provide richer, real-time, and more accurate metric data. By generating knowledge directories at various levels through multi-angle statistics (overall, institutional, researcher, temporal, etc.), and establishing knowledge graphs through multi-dimensional logical semantic relationship analysis and association of scholarly resources, next-generation IRs can analyze and evaluate institutional knowledge capabilities, knowledge relationships, knowledge asset applications, and requirements. By analyzing user behavior, user profiles can be established. Based on metrics, statistical data, knowledge graphs, and user profiles, an entirely new academic evaluation system independent of existing commercial publishing evaluation data can be established to reconstruct the academic evaluation ecosystem.

In the process of rapid scholarly communication transformation, IRs need to reconstruct their functions and services to become critical infrastructure for the new scholarly communication ecosystem. Researching the objectives, functions, services, and technologies of next-generation IRs and exploring new applications have become important content for current IR construction. China's IR construction needs to seize this development and transformation opportunity to participate in building the global scholarly communication ecosystem and become an important component leading worldwide scientific research progress.

References

[1] OpenDOAR statistics—an overview of the data held in OpenDOAR [EB/OL]. [2019-03-11]. http://v2.sherpa.ac.uk/view/repository_{visualisations}/1.html.

[2] Environmental Scan 2017 [EB/OL]. [2017-07-01]. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/p

[3] Ma J X. Trends in institutional repository content construction and service design [J]. Information Theory and Practice, 2010(9): 23-27.

[4] Zhang X L. Development trends and challenges of institutional repositories [J]. Data Analysis and Knowledge Discovery, 2014, 30(2): 1-7.

[5] Liu W, Zhu Z M, Zhang W Q, et al. Knowledge analysis and visualization based on institutional repositories [J]. Library and Information Service, 2016, 36(3): 125-131, 137.

[6] Wu Z Q, Zhu Z M, Liu W, et al. Image retrieval in institutional repositories based on LireSolr [J]. Library Science Research, 2016, 385(14): 58-63, 39.

[7] Wu Z Q, Zhu Z M, Yao X N, et al. Research and practice on extending audio-visual resource support capabilities in CSpace institutional repositories [J]. Data Analysis and Knowledge Discovery, 2017(9): 90-96.

[8] Wu Z Q, Zhu Z M, Liu W, et al. Research and practice on 3D model retrieval and display technology in institutional repositories [J]. Data Analysis and Knowledge Discovery, 2017(1): 73-80.

[9] The inheritance and change of TAIR [EB/OL]. [2018-01-01]. http://ntur.lib.ntu.edu.tw/handle/246246/2705

[10] Cui H Y, Nie H, Luo P C, et al. Research and construction of funder open access repositories—taking the NSFC Basic Research Repository as an example [J]. Library and Information Service, 2017, 61(11): 45-54.

[11] HKUST Institutional Repository [EB/OL]. [2018-11-01]. http://repository.ust.hk/ir/.

[12] The University of Hong Kong Scholars Hub [EB/OL]. [2018-11-01]. http://hub.hku.hk/.

[13] The PolyU Institutional Research Archive (PIRA) [EB/OL]. [2018-11-01]. http://ira.lib.polyu.edu.hk/.

[14] Sterman L, Borda S. Making visualization work for institutional repositories: Information visualization as a means to browse electronic theses and dissertations [J]. Journal of Librarianship and Scholarly Communication, 2017, 5(1): 1-17.

[15] Cocciolo A. Can Web 2.0 enhance community participation in an institutional repository? The case of PocketKnowledge at Teachers College, Columbia University [J]. The Journal of Academic Librarianship, 2010, 36(4): 304-312.

[16] Jones R. Giving birth to next generation repositories [J]. International Journal of Information Management, 2007, 27(3): 154-158.

[17] From academic repository (IR) to current research information system (CRIS)—how and why [EB/OL]. [2017-11-01]. https://core.ac.uk/download/pdf/38034688.pdf.

[18] Reshaping library support for research at King's College London [EB/OL]. [2017-11-01]. http://www.unica-network.eu/sites/default/files/GB%20and%20NW%20UNICA%20Brussels%2

[19] DeCastro P, Shearer K, Summann F. The gradual merging of repository and CRIS solutions to meet institutional research information management requirements [J]. Procedia Computer Science, 2014, 33(2014): 39-46.

[20] Finaf, Proven J. Using a CRIS to support communication of research: Mapping the publication cycle to deposit workflows for data and publications [J]. Procedia Computer Science, 2017, 106(2017): 232-238.

[21] Technical recommendations for next generation repositories [EB/OL]. [2018-12-20]. https://www.coar-repositories.org/news-media/technical-recommendations-for-next-generation-repositories/.

[22] Next generation repositories [EB/OL]. [2018-11-15]. https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/.

[23] Science Europe [EB/OL]. [2018-11-01]. https://www.scienceeurope.org/coalition-s/.

[24] Wellcome is updating its open access policy [EB/OL]. [2018-11-08]. https://wellcome.ac.uk/news/wellcome-updating-its-open-access-policy.

[25] Zhang X L. Making all research papers free to read: Chinese institutions explicitly support open access [EB/OL]. [2018-12-20]. http://zhishifenzi.com/depth/depth/4778.html.

[26] Guidance on the implementation of Plan S [EB/OL]. [2018-12-20]. https://www.coalition-s.org/feedback/.

[27] COAR's response to draft implementation requirements in Plan S [EB/OL]. [2018-12-20]. https://www.coar-repositories.org/news-media/coar's-response-to-draft-implementation-requirements-in-plan-s/.

[28] Impact of Plan S implementation guidelines on DSpace repositories [EB/OL]. [2018-12-20]. https://www.atmire.com/articles/detail/impact-of-plan-s-implementation-guidelines-on-dspace-repositories.

[29] Crow R. The case for institutional repositories: A SPARC position paper, 2002 [EB/OL]. [2018-12-21]. https://uta-ir.tdl.org/uta-ir/bitstream/handle/10106/24350/Case%20for%20IRs_{SPARC}.pdf?sequence=1.

[30] Signposting the scholarly web [EB/OL]. [2018-11-29]. http://signposting.org/.

[31] ActivityStreams 2.0 [EB/OL]. [2018-11-28]. https://www.w3.org/TR/activitystreams-core/.

[32] Open annotation community group [EB/OL]. [2018-11-28]. https://www.w3.org/community/openannotati

[33] International image interoperability framework [EB/OL]. [2018-11-08]. https://iiif.io/.

[34] IPFS is the distributed web [EB/OL]. [2018-11-01]. https://ipfs.io/.

[35] ResourceSync framework specification [EB/OL]. [2018-11-01]. http://www.openarchives.org/rs/toc.

[36] About SWORD [EB/OL]. [2018-11-01]. http://swordapp.org/about/.

[37] What are sitemaps? [EB/OL]. [2018-11-01]. https://www.sitemaps.org/.

[38] Linked data notifications [EB/OL]. [2018-11-01]. https://www.w3.org/TR/ldn/.

[39] ResourceSync framework specification—change notification [EB/OL]. [2018-11-01]. http://www.openarchives.org/rs/notifications/1.0.1/notification.

[40] Webmention [EB/OL]. [2018-11-01]. https://www.w3.org/TR/webmention/.

[41] WebSub [EB/OL]. [2018-11-01]. https://www.w3.org/TR/websub/.

[42] ORCID [EB/OL]. [2018-11-01]. https://orcid.org/.

[43] WebID 1.0 [EB/OL]. [2018-11-01]. https://www.w3.org/2005/Incubator/webid/spec/identity/.

[44] Signing HTTP messages [EB/OL]. [2018-11-01]. https://datatracker.ietf.org/doc/draft-cavage-http-signatures/.

[45] OpenID Connect [EB/OL]. [2018-11-01]. https://openid.net/connect/.

[46] WebID authentication over TLS [EB/OL]. [2018-11-01]. https://www.w3.org/2005/Incubator/webid/spec/

[47] COUNTER [EB/OL]. [2018-11-01]. https://www.project-counter.org/.

[48] HTTP ETag [EB/OL]. [2018-11-01]. https://en.wikipedia.org/wiki/HTTP_{ETag}.

[49] Public access plan: Today's data, tomorrow's discoveries: Increasing access to the results of research funded by the National Science Foundation [EB/OL]. [2018-10-01]. https://www.nsf.gov/news/special_{reports}/public_{access}/#.

[50] Lee D J, Stvilia B. Practices of research data curation in institutional repositories: A qualitative view from repository staff [J]. PLoS One, 2017, 12(3): e0173987.

[51] Tang F. Linking research information systems and institutional repositories for research data management [J]. Information Theory and Practice, 2018, 41(2): 73-76.

[52] Ojala M. Big data and AI: Technology, transparency, and trust [EB/OL]. [2019-01-01]. http://www.infotoday.com/cilmag/dec18/Ojala–Big-Data-and-AI–Technology–Transparency–and-Trust.shtml.

Author Contributions:
Cui Haiyuan: Designed the paper framework, wrote and revised the content;
Sun Chao: Conducted key technology research and revised the paper;
Luo Pengcheng: Conducted key technology research and revised the paper.

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