The Implementation of Router Service Engine iSwitch for Open Access Papers

Authors: Zhang Xiaolin, Qian Li, Shi Hongbo, Liang Na, Zhang Xiaolin

Date: 2016-04-12T00:00:00+00:00

Abstract

Open Access academic papers have become an important instrument for the world' s leading countries to promote knowledge sharing, collaborative open innovation, economic growth, and inclusive development. Moreover, the open sharing of academic papers resulting from publicly funded projects has become a consensus among the world' s leading countries, with an important requirement being the deposition of those papers in open access institutional repositories (IRs) affiliated with funders and authors' institutions. However, the situation regarding institutional repositories in China faces significant challenges, including low awareness of deposition, incomplete or incorrect submitted data, and other issues. This affects the development and improvement of China' s Open Access and open sharing mechanisms. Although RJ-Broker can assist in addressing these problems, it is primarily related to European PMC data and is located abroad.

To better address these issues, the National Science Library of the Chinese Academy of Sciences (CAS), as the leading institution for this study in China, with hundreds of research institutes of the Chinese Academy of Sciences serving as demonstration sites, and with reference to the OA-RJ model, developed a push and routing service for Chinese academic papers, namely iSwitch, that implements automatic deposition. It can help institutions and funders construct their IRs effectively and promote the open utilization of academic papers by others.

Since its public release, the iSwitch service has routed more than 360,000 paper metadata records pushed by Web of Science and some experimental Open Access paper data from other publishers to CAS IRs. It is now a stable service for exchanging updated WOS data and other publisher data. Additionally, with the help of iSwitch, Web of Science has provided full-text links to CAS IR papers since July 27, 2015.

Full Text

Preamble

Zhang Xiaolin, Qian Li, Shi Hongbo, Liang Na
National Science Library, Chinese Academy of Sciences, Beijing 100190, China

Abstract

Open Access to academic papers has become an important measure for the world' s leading countries to promote knowledge sharing, collaborative open innovation, economic growth, and inclusive development. Moreover, the open sharing of academic papers funded by public research projects has become a global consensus, with a key requirement being the deposition of these papers in open access institutional repositories (IRs) attributed to funders and authors' institutions. However, the situation regarding institutional repositories in China is concerning, characterized by low awareness of deposit requirements, incomplete or incorrect submitted data, and other issues. These challenges affect the development and improvement of China' s Open Access and open sharing mechanisms.

While the RJ-Broker can assist in addressing these problems, it is primarily focused on European PMC data and is located abroad. To better solve these issues, the National Science Library of the Chinese Academy of Sciences (CAS), as the leading institute for this study in China, has constructed a pushing and routing service for Chinese academic papers—namely iSwitch—implementing automatic deposit, with hundreds of CAS research institutes serving as demonstration sites. This service references the OA-RJ model and can help institutes and funders construct their IRs effectively while promoting the open utilization of academic papers by others.

Since its public release, iSwitch has routed more than 360,000 paper metadata records pushed by Web of Science and some experimental Open Access paper data from other publishers to CAS IRs. It now provides a stable service for exchanging WOS update data and other publisher data. Additionally, with the help of iSwitch, Web of Science has linked to full-text versions of CAS IR papers since July 27, 2015.

Keywords: iSwitch; Open Access; Automatic Routing; Pushing and Routing; Resolving Authors and Institutes

1. Introduction

Open Access to academic papers has become an important measure for the world' s leading countries to promote knowledge sharing, collaborative open innovation, economic growth, and inclusive development. The Global Research Council, founded in 2012, released an Open Access action plan in 2013 that promotes the open sharing of academic papers funded by research projects. Both the UK

Research Council and the European Union' s Horizon 2020 program mandate open sharing of research papers funded by their projects.

The Chinese Academy of Sciences (CAS) issued its Open Access Policy on May 15, 2014. This policy requires researchers and graduate students to deposit an electronic version of the final, peer-reviewed manuscript of any research paper resulting from publicly funded research projects into their respective institute' s open access repository at the time of publication, with the paper to be made publicly available within 12 months of the official publication date. If the article is published as an open access paper, its PDF or HTML/XML version should be deposited and made openly accessible immediately. If the publisher agrees, the deposited copy should be made openly accessible before the 12-month embargo period. The National Natural Science Foundation of China (NSFC) issued a similar Open Access Policy with the same requirements for papers arising from its funding.

Beyond Open Access itself, IRs also need external data to check or validate their records. Repository managers cannot be certain whether authors have deposited all required articles or whether archived records contain correct information, yet they lack data sources to validate and check their records.

Therefore, depositing articles and their metadata into institutional repositories is an essential requirement of Open Access and IRs. While the ideal approach would be for publishers to push data directly to IRs, several challenges arise in this process:

For authors: (1) Authors are often unfamiliar with open access policy details and IR deposit procedures, and they are generally unwilling to devote time and effort to the process. (2) Authors may not retain or may find it difficult to confirm the correct version for deposit into the IR. (3) A majority of papers have multiple authors, and a significant number are written by authors from multiple organizations. This situation requires each record to be entered separately into each institutional repository, leading to massive duplication of effort.

For publishers: (4) If publishers directly push research papers to IRs, they encounter problems such as accurately locating the author's IR, mapping journal metadata with IR metadata, establishing and controlling a reliable pushing process, and providing reports of push activities to authors.

Other challenges: (5) Interoperability between IRs remains a weakness.

To address these problems and ease the burden on authors and publishers, the authors propose the iSwitch router service, which effectively transforms the relationship from "Many to Many" to "Many to One to Many," thereby greatly reducing complexity.

2. Related Work

The OA-RJ (Open Access Repository Junction) project, sponsored by JISC (Joint Information Systems Committee), started in 2009 and ran until 2011. The aim of OA-RJ was to assist deposit into multiple existing repository services by developing middleware that, for a given paper, would aid both the discovery of repository targets and the delivery of content to appropriate locations. The project would benefit multiple stakeholder groups, as shown in Table 1 .

To achieve this aim, the project designed the RJ Broker model, with discovery via the Junction and delivery via the Broker. The Junction functions as an API service that can be queried to discover repository targets at any time, while the Broker serves as the JISC Publications Router service that automates the delivery of research publications from multiple data suppliers (such as publishers and subject repositories) to multiple repositories (such as institutional repositories). To date, there are two data suppliers: European PMC and ELIFE, with ELIFE beginning to support the JISC Publications Router in March 2015.

Additionally, the American library community proposed the SHARE Service Architecture, which also provides similar services. In March 2014, the Institute of Museum and Library Services (IMLS) announced a $500,000 out-of-cycle National Leadership Grant for Libraries to the Association of Research Libraries (ARL) to develop and launch the Shared Access Research Ecosystem (SHARE) Notification Service. The SHARE notification service gathers information about research release events through both a direct push protocol and a harvest strategy, then notifies consumers of these events through free subscriptions to predefined channels of notices and by allowing searches of its digest of research release events. Research release events describe the release of publications, datasets, and other scholarly research results.

However, given the current state of IR construction in China—characterized by low deposit awareness, incomplete or incorrect submitted data, and the fact that many research institutes still do not have their own IR—this situation seriously affects the development and improvement of China' s Open Access and open sharing mechanisms. While RJ-Broker can assist in addressing these problems, it is primarily focused on European PMC data and is located abroad. Therefore, to better solve these problems, the National Science Library of CAS, as the leading institute for this study in China, has constructed the iSwitch pushing and routing service for Chinese academic papers, implementing automatic deposit of authors' academic work in Chinese research institutes, with hundreds of CAS research institutes serving as demonstration sites and referencing the OA-RJ model.

3. Main Ideas of iSwitch

The main idea behind iSwitch is to design and develop a data exchange service that receives academic papers pushed by publishers and routes them to IRs. As a data exchange service, iSwitch should effectively receive data, accurately parse

and route it, fully audit the process, and provide a stable and trustworthy data exchange service for publishers, authors, and IR managers.

3.1 Overall Architecture

To design and implement iSwitch, an overall architecture of the service is proposed. In this architecture, an automatic data Receiver is designed to receive data objects pushed by publishers using FTP or SWORD protocol. To locate the data target, an automatic Data Parser is designed to parse institute names, which are the target objects for routing. Subsequently, an automatic data Router is designed to route data objects packaged as original files to IRs using FTP protocol, as illustrated in Fig. 1 [Figure 1: see original paper].

3.2 Technical Workflows and Standards

To enable stable and effective pushing and routing between multiple publishers (Many) and institutes (Many), iSwitch (One) employs general standard specifications and transforms the complicated M:M relationship into an M:1:M relationship. Moreover, iSwitch needs to establish cooperation with publishers and target institutes respectively, and create pushing and routing service agreements, service running process specifications, and common technical standards.

The stakeholders of iSwitch are primarily pushers (publishers), authors, the router (iSwitch), and receivers (research institutes). Key components include:

(1) Pushing and Routing Protocol: Publishers push data objects to iSwitch using FTP or SWORD protocol, while iSwitch routes data objects using FTP protocol. Main features are as follows: FTP supports fast loading and downloading of large data packages, while SWORD supports flexible handling of single-paper data objects or smaller data objects.

(2) Data Exchange Format Standard: The metadata format is JATS, and the full-text format is WORD, PDF, or XML. Additionally, iSwitch requires the general standard of metadata description for authors, funders, journals, and affiliations as far as possible, such as ORCID.

(3) Authors and Institute Name Parsing: iSwitch requires publishers to follow specific schema for submitted data to ensure accurate parsing of authors' names and institute names. First, the mapping relationship between authors and institutes must be clearly marked. Second, authors' names should be composed using common rules. Finally, institute names should be structured, including abbreviation, full name, and complete affiliation information such as address, sub-institute, or laboratory.

(4) Routing and Package Mechanism: Using an agent, iSwitch automatically routes data objects packaged with all original files pushed by publishers and a new paper metadata file built by iSwitch to IRs. This mechanism can effectively and stably handle many complicated cases, such as one paper with multiple authors, different authors belonging to different institutes, one author

affiliated with multiple institutes, one paper funded by multiple funders, and so on.

4. Method of iSwitch Implementation

Based on the ideas above, the authors developed the iSwitch service, which consists of four main components: Reception, Parsing, Routing, and Audit.

4.1 Papers Reception

Through FTP and SWORD protocols, iSwitch receives paper data from providers and then resolves and preserves them in a local repository. The following sections describe the main processes of data reception.

4.1.1 Implementation of FTP and SWORD Protocol Modules For the FTP protocol, the authors encapsulated a common FTP operation class based on Apache Commons Net and use it to harvest paper data provided by publishers.

For the SWORD protocol, there are two main series: 1.x and 2.0. The highest version of 1.x is 1.3, which is the most commonly supported by publishers (such as BMC), so the authors chose to implement SWORD 1.3 for iSwitch. The 2.0 version will be developed in the future if needed. The authors developed the SWORD protocol primarily based on SWORD-Common1.1 open source software and referenced the implementations of Fedora and DSpace.

4.1.2 Implementation of Data Receiving and Preservation Module **(1) Batch Information Parsing:** iSwitch receives data from publishers in units called batches. Every batch contains a file named Batchinfo.txt with information about the batch, such as the batch ID provided by the publisher, the total number of papers, the paper IDs, and other details. The system ingests this information and compares it with the actual ingested papers in the batch, which is used for audit purposes.

(2) Paper Metadata Parsing: As agreed with publishers, they provide papers encoded in JATS format. However, JATS has three sub-formats and different versions. iSwitch provides specific metadata parsers for each publisher and ingests them into the repository.

For each paper provided by a publisher, iSwitch assigns a unique iSwitch identifier (iSwitch ID) and version number. If different papers have the same publisher paper ID, iSwitch assigns the same iSwitch ID but with different version numbers.

The original data for each paper provided by the publisher is preserved in a ZIP file with the iSwitch ID.

4.2 Data Parsing Module

The Data Parsing Module consists of two main parts: Pushing Target Resolving and Data Checking.

4.2.1 Pushing Target Resolving Accurately and completely identifying a paper' s pushing target is crucial for iSwitch.

(1) Author Institute Recognition: iSwitch resolves the author' s affiliation information to identify the paper' s target IR. Methods include: Requesting publishers to provide institute items in their paper metadata. JATS provides institute items, and most publishers have detailed affiliation information for papers. If provided, resolution accuracy is high. Configuring an alias table for institute names. Each institute name is configured with different aliases, such as Chinese name, English name, English short name, and others, each with a type value for distinction. Through the alias table, iSwitch can accurately identify the institute from the author affiliation. Handling the special comma character (","). The comma is a separator in affiliation text. For institute names containing commas, iSwitch compares them with the affiliation text first, thereby reducing the effect of commas.

(2) Fund Resolving: The target for a paper includes its funder' s IR, so iSwitch resolves fund information from the paper metadata. If the paper contains fund items, they are resolved directly. If fund information appears in the acknowledgement text or other text, iSwitch resolves it using regular expression matching, as the target funder IRs are limited.

4.2.2 Data Checking To ensure that received data is complete and correct, iSwitch checks the data from publishers in two aspects.

(1) Necessary Item Checking: According to the agreement between iSwitch and publishers, publishers should provide paper metadata with necessary items such as title, abstract, publication date, author affiliation, fund information, and others. iSwitch checks the received data for these items to ensure completeness.

(2) Batch Package Checking: For each batch pushed by a publisher, iSwitch checks whether the package contains the "Batchinfo.txt" file and whether the received data matches the "Batchinfo.txt" . iSwitch verifies the received paper count and paper IDs to ensure it has received the correct data package from the publisher.

4.3 Papers Routing and Pushing

After identifying the target institute and funder, iSwitch pushes the data to their IRs.

4.3.1 Creation of Pushing Task (1) Auto Task Scheduling Agent: For each IR, iSwitch creates a pushing task to distribute paper data. To enhance

iSwitch' s automatic distribution capability, the authors developed an "Auto Task Scheduling Agent." The agent creates pushing tasks for each IR based on the IR' s reception frequency.

(2) Create Pushing Task: Each paper has one or more institutes, but the institute may not have a direct IR. Two main situations exist: one where the institute is an older one and the paper should be pushed to a newer institute' s IR, and another where different institutes use a combined IR. In these cases, the authors built a relation table for these institutes and their relationships to ensure each paper finds the correct IR.

If an IR can be pushed at a given time based on its frequency and there are papers for it in iSwitch, the agent creates the pushing task, assigns a TaskID, and assigns the TaskID to each paper that needs to be pushed to the IR.

4.3.2 Pushing Data After the pushing task is built, the pushing data module sends paper data to the target FTP folder.

(1) Pushing Data to FTP: After task creation, paper data is pushed to its target FTP folder immediately (iSwitch's pushing data module scans for pushing tasks every 5 seconds). The folder is named based on the IR name, publisher, and TaskID. The IR then harvests data from the designated FTP folder.

(2) Web Interface for Pushing Result: Since pushing data is based on FTP, iSwitch cannot directly obtain the harvest results from IRs. To address this, the authors developed a web interface to receive the results of each pushing task, allowing iSwitch to track the status of each pushing task.

4.4 Audit of Received and Routed Papers

iSwitch provides an audit function for received data from two dimensions—publisher and batch—and offers a clear view of the reception process results for data managers with time filtering capabilities.

Regarding routed data, the system also provides four dimensions for auditing routed paper data with time filtering: publisher, institute (IR), funder (IR), and batch. In the audit results, managers can easily identify the expected, successful, and failed paper counts. In the future, the system will provide more detailed and specific audit results for publishers, IR managers, and other stakeholders.

5. Application of iSwitch

Currently, iSwitch has established cooperation with many publishers, including Web of Science (WOS), Chinese Science Citation Index (CSCD), IOPP, PLoS, and others, and has pushed paper data to Chinese Academy of Sciences (CAS) IRs.

5.1 Cooperation with WOS

iSwitch accepts both paper metadata (to help check IRs) and Open Access paper data (to facilitate deposit into IRs) from publishers. Through cooperation with WOS, iSwitch not only helps check IRs but also improves the utilization of IR papers.

WOS maintains the well-known Science Citation Index (SCI) and has collected metadata for SCI papers, which are considered the most influential in the natural sciences. iSwitch accepts SCI paper metadata and distributes it to target CAS IRs to help them verify whether these papers have been deposited, whether authors have submitted correct metadata to the IR, and whether full-text documents have been submitted. IRs also assign the corresponding WOS paper ID (WOSID) to archived papers and provide full-text links to the WOS website if full-text documents are available. WOS displays the full-text link to the IR if one exists. In this way, iSwitch improves the integrity, accuracy, and utilization of scientific research output in IRs. The cooperation relationship among iSwitch, WOS, and IRs is illustrated in Fig. 3 [Figure 3: see original paper].

WOS has pushed 3,625,512 CAS paper metadata records (1949-present, SCI-EXPANDED) to iSwitch, which have been distributed to 99 IRs in 484,088 batches. iSwitch has established a stable router service between IRs and WOS to accept updated data and distribute it, and WOS has displayed CAS IR full-text links on its website since July 26, 2015.

5.2 Cooperation with Other Publishers

CSCD, a citation index for Chinese scientific articles corresponding to SCI, has also pushed data to iSwitch for IR checking. iSwitch has built a similar service for CSCD and IRs, which can help IRs archive valuable papers published in Chinese journals.

In addition to metadata, PLoS and IOPP have reached preliminary agreements with iSwitch for automatic deposit of Open Access papers. PLoS has sent some test papers, and IOPP is in close contact with iSwitch regarding future work.

6. Conclusion

In this paper, the authors developed the iSwitch router service system to facilitate automatic deposit of Open Access papers and IR checking. iSwitch accepts metadata and full-text papers from publishers, resolves their target IRs, and pushes this data to the IRs. iSwitch uses FTP and SWORD protocols to receive paper data encoded in JATS (recommended) and other standard formats. Through alias and relation tables and other methods, iSwitch identifies the target IRs for papers and pushes them to the appropriate IRs via FTP. Currently, iSwitch has cooperated with WOS, CSCD, PLoS, IOPP, and CAS IRs, establishing a stable service for them. As an important infrastructure demonstration for libraries, iSwitch will facilitate content deposit and distribution in IRs, bridging

publishers and repositories more effectively. As a public service, iSwitch will be of great help to all stakeholders involved in IRs, including authors, funders, publishers, and institutions.

Acknowledgments

This article is supported by the project "Developing International Open Access Paper National Switch Center Demonstration System" (Grant No. Y14008), funded by the Chinese Academy of Sciences.

References

1. Zhang Xiaolin, Liang Na, Qian Li, Shi Hongbo. Router Service Engine iSwitch for Open Access Articles: The Concept, Strategy, and Framework. *New Technology of Library and Information Service*, 2014, 30(10): 4-8.
2. RCUK Policy on Open Access [EB/OL]. [2015-07-10]. http://www.rcuk.ac.uk/research/openaccess/policy
3. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 [OL]. [2015-07-12]. http://ec.europa.eu/research/participants/data/ref/h2020/grants_man hi-oa-pilot-guide_en.pdf
4. Chinese Academy of Sciences Policy Statement on Open Access to Articles from Publicly Funded Scientific Research Projects [2015-07-15]. http://www.cas.cn/xw/yxdt/201405/P020140516559414259606.pdf
5. The National Natural Science Foundation of China Policy Statement on Open Access to Research Publications from its Funded Projects [2015-07-15]. http://www.nsfc.gov.cn/publish/portal0/tab38/info44471.htm
6. OA-RJ Open Access Repository Junction - Overview. [2015-07-26]. http://edina.ac.uk/projects/oa-rj/index.html
7. Welcome to Jisc Publications Router-Publications Routers. [2015-07-26]. http://broker.edina.ac.uk/
8. eLife supports the Jisc Publications Router | eLife. [2015-07-26]. http://elifesciences.org/eLife-news/eLife-supports-the-Jisc-Publications-Router
9. OA-RJ Open Access Repository Junction-About-Stakeholders. [2015-07-26]. http://edina.ac.uk/projects/oa-rj/stakeholders.html
10. Institute of Museum and Library Services. IMLS and Sloan Foundation Award $1 Million to ARL for SHARE Notification Service. [2015-07-26]. http://www.imls.gov/imls_and_sloan_foundation_award_1_million_to_arl_for_share_notification_
11. Serv-U File Server Administrator Guide [EB/OL]. [2015-07-25]. http://www.serv-u.com/Serv-U-Administrator-Guide.pdf

12. A Brief History of SWORD [EB/OL]. [2015-07-25]. http://swordapp.org/about/a-brief-history/

13. Liang Na, Zhang Xiaolin, Qian Li, Shi Hongbo. Router Service Engine iSwitch for Open Access Articles: Technical Workflows and Standards. *New Technology of Library and Information Service*, 2014, 30(10): 9-13.

14. SWORD (protocol) - Wikipedia, the free encyclopedia. [2015-07-25]. https://en.wikipedia.org/wiki/SWORD_(protocol)

15. JATS, Journal Article Tag Suite [EB/OL]. [2015-07-25]. http://jats.nlm.nih.gov/

16. Shi Hongbo, Qian Li, Zhang Xiaolin, Liang Na. Router Service Engine iSwitch for Open Access Articles: Articles Reception and Resolving. *New Technology of Library and Information Service*, 2015, 31(6): 1-6

17. Qian Li, Shi Hongbo, Zhang Xiaolin, Liang Na. Router Service Engine iSwitch for Open Access Articles: Pushing and Routing. *New Technology of Library and Information Service*, 2015, 31(6): 7-12

18. Apache Commons Net [EB/OL]. [2015-07-25]. http://commons.apache.org/proper/commons-net/index.html

19. SWORD-Common 1.1 [EB/OL]. [Reference details not fully provided in original]

20. DSpace SWORD [EB/OL]. [2015-07-25]. https://github.com/DSpace/DSpace/blob/master/dspace-sword/src/main/java/org/dspace/sword/DSpaceSWORDServer.java

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