The Implementation of Router Service Engine iSwitch[1] for Open Access Papers

Zhang Xiaolin, Qian Li, Shi Hongbo, Liang na

National Science Library, Chinese Academy of Sciences, Beijing 100190,China

Abstract

Open Access academic paper has become important measures of the world's leading countries which promote knowledge sharing, collaborative open innovation, economic growth and inclusive development. Moreover, Open Sharing of academic papers founded by public funding projects has become the consensus of the world's leading countries, and an important requirement is to deposit those papers in the open access IRs which are attribute to funder and authors' institute. But the situation of institutional repository in China is serious, such as poorer deposit awareness, incomplete or incorrect submitted data and so on. It affects the development and perfection of China Open Access and open sharing mechanism. Though RJ-Broker can assist and promote to solve problems above, it is mainly related to European PMC data and located abroad.

In order to better solve those problems, National Science Library of CAS, as a leader institute of this study in China, with hundreds of Research institutes of Chinese Academy of Sciences as demonstration, and reference to the OA-RJ model, constructed a pushing and routing services of Chinese academic papers, namely iSwitch, implementing automatic deposit. It can help institutes and funders to construct their IRs effectively and promote academic papers utilization by others openly.

After iSwitch service is released publicly, it has routed more than 360,000 paper metadata pushed by Web of Science and some experimental Open Access paper data from other publishers to CAS IRs. Now it is a stable service to exchange WOS update data and other publisher data. Besides, with the help of iSwitch, Web of Science has linked full-text link of CAS IR papers since 27, July, 2015.

Keywords: iSwitch; Open Access; Automatic Routing; Pushing and Routing; Resolving authors and institutes

1. Introduction

Open Access academic paper has become an important measure of the world's leading countries which promotes knowledge sharing, collaborative open innovation, economic growth and inclusive development[1]. Global Research Council, founded in 2012, which released an Open Access action plan in 2013, promoting the open sharing of academic papers founded by project. UK Research Council[2] and European Union Horizon 2020[3] all force Research Papers Open Sharing funded by projects.

Chinese Academy of Sciences[4] (CAS) issued its Open Access Policy in May 15, 2014. The policy requires researchers and graduate students to deposit an electronic version of the final, peer-reviewed manuscripts of their research papers that submitted and consequently published in academic journals, resulted from any public funded research projects, into the open access repositories of their respective institutes at the time of publication, and be made publicly available within 12 months of the official data of publication. 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. National Natural Science Foundation of China[5] (NSFC) issued its Open Access Policy with the same requirement for papers arising from its funding.

1 http://iswitch.las.ac.cn

Except the Open Access, the IRs also need outside data to check or validate their records. They are not sure whether the authors have deposited their all needed articles or not and whether archived records have owned correct items or not. But there are no data for IRs to validate and check their records.

So, deposit article and its metadata to its Institutional Repository (IR) is an essential requirement of Open Access and IRs. The best way is the publishers push data to IRs directly, but there are some challenges and problems in the process, including:

For authors:

(1) Authors are not familiar with open access policy details and IR deposit operation process. Meanwhile, they don't incline to devote time and energy.

(2) Authors may not retain or be difficult to confirm the correct version of depositing into IR.

(3) A majority of papers are attributed to more than one author, and a significant number of those papers are written by authors from multiple organizations. This situation requires each record be entered separately into each Individual Repository, leading to a massive duplication of effort.

For publishers:

(4) If publishers directly push research papers to IR, they also encounter some problems, such as: locate the author IR accurately; map journal metadata with IR metadata; establish and control a reliable pushing process and provide reports of push activities to authors and so on.

Others:

(5) Interoperability between IRs is still a weakness.

In order to solve problems above and ease the burden of authors and publishers, the authors proposed the method of iSwitch router service that it could effectively build the relationship from "Many to Many" to "Many To One To Many", and greatly reduce the complexity between them.

2. Related work

OA-RJ [6](Open Access Repository Junction) project sponsored by JISC (Joint Information Systems Committee) started in 2009 and run in 2011. The aim of OA-RJ is to assist deposit into multiple existing repository services by developing middleware that, for a given paper, will aid both discovery of repository targets and delivery of the content to the appropriate locations. The project will be benefit more than one group, as follow Table 1. So in order to implement this aim, they designed RJ Broker model, discovery via the Junction and delivery via Broker. Junction as an API Service is queried to discover repository targets at any time and Broker as JISC Publications Router service[7] automates the delivery of research publications from multiple data suppliers (such as publishers and subject repositories) to multiple repositories (such as institutional repositories). Until now there are two data suppliers, including European PMC and ELIFE, and ELIFE started to support the JISC Publications Router in March, 2015[8].

Table 1 Related information of Stakeholder[9]

Stakeholder	Interest / stake
Principal Investigator/Researcher	i. ability to deposit (directly or by notification) into multiple repository locations. ii. compliance with research funders' open access policies.
Journal Publishers	i. supply of service to authors via controlled OA deposit of the correct version (and metadata)
Research Funders	i. easier for researcher to follow OA mandates ii. Rise in mandate compliance as a clear success indicator
Repository Managers	i. additional content, via RJ broker, from publishers into institutional repositories ii. interoperability with subject/funders repositories iii. adoption of SWORD client
Repository Developers community (Eprints, DSpace, SWORD, etc)	i. promotion of use of standards, eg SWORD protocol, across platforms

Besides, American library community proposed the SHARE Service Architecture which is also available to provide some services above. In March 2014, The Institute of Museum and Library Services (IMLS) announced the award of 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[10]. The SHARE notification service will gather information about research release events through both a direct push protocol and a harvest strategy. The service will then notify 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 results of scholarly research.

But in the face of the current construction situation of IR in China, such as, poorer deposit awareness, incomplete or incorrect submitted data, especially many research institutes still do not own their own IR. So this situation seriously affects the development and perfection of China Open Access and open sharing mechanism. Though RJ-Broker can assist and promote to solve problems above, it is mainly related to European PMC data and located abroad. So in order to better solve those problems, National Science Library of CAS as a leader institute of this study in China, with hundreds of Research institutes of CAS as demonstration, with reference to the OA-RJ model, constructed the pushing and routing services of china academic papers, namely iSwitch, implementing automatic deposit of authors' academic of in the research institutes of China.

3. Main ideas of iSwitch

The main idea behind the iSwitch is trying to design and develop a data exchange service that receives academic papers pushed by publishers and routes them to IRs. To be a data exchange service, iSwitch should effectively receive data, accurately parse and route data, fully audit data and provide a stable and trustworthy data exchange service for publishers, authors and IR managers.

3.1 Overall Architecture

To design and implement the iSwitch, an overall architecture of the service is proposed. In this architecture, an automatic data Receiver is designed to receive the data object pushed by publishers by using FTP[11] or SWORD[12] protocol. In order to locate the data target, an automatic Data Parser is designed to parse Institute Name which is the target object to route. And then an automatic data Router is designed to route data object packaged by original files to IR by using FTP protocol. As illustrated in Fig.1.

Fig. 1 Overall Architecture of iSwitch

3.2 Technical Workflows and Standards[13]

In order to push and route between multiple publishers (Many) and institutes (Many) stably and effectively, iSwitch (One) used general standard specification and transferred complicated M:M relation to M:1:M relation. Moreover, iSwitch needs to establish cooperation with publishers and target institutes respectively and establish pushing and routing service agreement, service running process specification and common technical standards.

Besides, stakeholders of iSwitch are mainly pushers (publishers), authors, router (iSwitch) and receivers (Research Institute).

(1) Pushing and Routing Protocol. Publishers push data object to iSwitch by using FTP or SWORD[14] protocol. iSwitch routes data object by using FTP protocol. Main features as follows: FTP supports big data packages to load and download fast, SWORD supports one paper data object or smaller data object flexibly.

(2) Data Exchange Format Standard. Metadata format is JATS[15]; full-text format is WORD, PDF or XML. Besides, iSwitch requires the general standard of metadata description of the author, funder, journal and affiliation as far as possible, such as ORCID.

(3) Authors and Institute Name Parsing. iSwitch should require some specifications schema of submitted data by publishers and ensure authors' name and institute' name to parse accurately as far as possible. First, mapping relationship between the author and the institute must be clearly marked. Second, authors' name is composed by using some common rules. Finally, institutes' name should be structured, abbreviation, full name and full affiliation information including address, sub-institute or laboratory and so on.

(4) Routing and Package Mechanism. By using agent, iSwitch will automatically route data objects which are packaged by all original files pushed by publishers and a new paper metadata file built by iSwitch to IRs. Meanwhile, this mechanism may support to deal with many complicated cases validly and stably, such as one paper being multiple authors, different authors belonging to different institutes, one author being attached to multiple institutes, one paper being funded by multiple funders and so on.

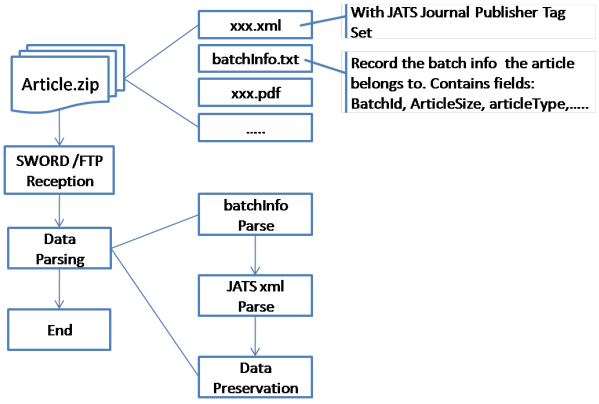
4. Method of iSwitch Implementation

Based on the idea above, the authors developed the iSwitch service contained following parts: Reception, Parsing[16], Routing[17] and Audit.

4.1 Papers Reception

Through the FTP and SWORD protocol, we get paper data from the providers, and then resolve and preserve them to our local repository. The following figure shows the main process of data reception.

4.1.1 Implement the FTP and SWORD Protocol Module



For the FTP protocol, the authors encapsulate a common FTP operation class based on the Apache Commons Net[18], and use it to harvest paper data provided by publisher.

For the SWORD protocol, there are two main series, 1.x and 2.0. The highest version of 1.x is 1.3 and it is the most common supported by the publisher (such as BMC), so the authors choose to implement the 1.3 version SWORD for iSwitch. As to the 2.0 version, we will develop it in future if needed. The authors developed the SWORD protocol generally based on SWORD-Common1.1[19] open source software and had referenced the implementation of Fedora[16] and DSpace[20].

4.1.2 Implement the Date Receiving and Preservation Module

(1) Batch Information Parsing

We receive data from publisher in a unit and we call it batch. Every batch contains a file Batchinfo.txt which contains the information about this batch, such as the ID of this batch provided by the provider, the total numbers of the paper, the IDs of the papers and others. The system ingests the information and compares it with real ingest papers of this batch, which will be used as the audit information.

(2) Paper Metadata Parsing

As the agreement with the publisher, they provide paper encoded by JATS format. However, the JATS has three sub formats and different versions. iSwitch provides specific metadata parser for each publisher and ingests them into our repository.

Each paper provided by publisher, iSwitch will assign a unique iSwitch identifier (iSwitch ID) and version number. If different paper with same publisher paper ID, iSwitch will assign same iSwitch ID too, and with different version number.

The original data of each paper provided by publisher will be preserved in a ZIP file with the iSwitch ID.

4.2 Data Parsing Module

There are two main parts in the Data Parsing Module: Pushing Target Resolving and Data Checking.

4.2 .1 Pushing Target Resolving

To find out the paper's pushing target precisely and completely is the crucial point of iSwitch.

(1)Author Institute Recognition

ISwitch resolves the author's affiliation information to find out the paper's push Target IR. Methods as follows:

① Request the publisher to provide institute item in their paper metadata. The JATS has provided institute item and most publishers have detailed affiliation information for paper. If they provide it, the resolving will have a high precise.

② Configure an Alias table for institute name. Each name of one institute will be configured with different aliases, such as Chinese name, English Name, English Short Name and others, and each of them has a type value to distinguish with each other. Through the alias table, iSwitch will find out paper's the accurate institute from the author affiliation.

③ Solve the special character comma (","). Comma is a special character in the affiliation text, where it is a separate sign. For institute names with comma, iSwitch will compare them with the affiliation text first. In this way, iSwitch decreases the effect of the comma.

(2) Fund Resolving

The target of paper contains its fund IR, so iSwitch will resolve the fund info from the paper metadata.

① If the paper has the fund item, resolves it directly.

② If the fund info in the acknowledgement text or other text, iSwitch resolves them by regular expression match, because the target fund IR is limited.

4.2.2 Data Checking

In order to ensure that received data is completed and correct. ISwitch will check the receiving data from publisher in two aspects.

(1) Necessary Item Checking

According to the agreement between iSwitch and publisher, the publisher should provide iSwitch paper metadata with necessary items, such as title, abstract, publishing date, author affiliation, fund and others. ISwitch will check the receiving data for the items to ensure them are complete.

(2)Batch Package Checking

For each batch pushed by publisher, iSwitch will check the package whether contains the "Batchinfo.txt" file and the receiving data is matching to the "Batchinfo.txt". ISwitch checks the receiving paper number, and receiving paper ID to ensure it has received the correct data package from the publisher.

4.3 Papers Routing and Pushing

After finding the target institute and fund, iSwitch pushes these data to their IRs.

4.3.1Creation of Pushing Task

(1) Auto Task Scheduling Agent

For each IR, iSwitch will create a pushing task to distribute paper data. To increase the auto distribution ability of iSwitch, the authors developed an "Auto Task Scheduling Agent". The agent will create pushing task for each IR based on the IR reception frequency.

(2) Create Pushing Task

Each paper has one or more institutes, but the institute may not have a direct IR. There are two situations mainly. One is that the institute is an old one and the paper should push to a new institute's IR. The other is that different institute use a combined IR. In these cases, the authors build a relation table for these institutes and their relations to ensure each paper find one and correct IR.

If an IR could be pushed this time based on its frequency and there are papers for it in iSwitch, the agent will create the pushing task and assign a TaskID for this task and assign the TaskID to each paper which needed to be pushed to the IR.

4.3.2 Pushing Data

After the pushing task built, the pushing data module will push paper data to their target FTP folder.

(1) Pushing Data to FTP

After the creation of pushing task, the paper data will be pushed to its target FTP folder immediately (iSwitch pushing data module scans the pushing task in every 5 seconds). The folder is named based on the IR name, publisher and the TaskID. And then, the IR will harvest data from the desired FTP folder.

(2) Web Interface for Pushing Result

The pushing data is based on FTP, so iSwitch system cannot get the harvest result of IR. In this case, the authors developed a web interface to receive the results of each pushing task. In this way, iSwitch can know each pushing task status.

4.4 Audit of Received and Routed Papers

iSwitch provides a audit function for received data from two dimensions of publisher and batch, and provides a clear view of the reception process result for data manager with time filter.

Regard to the routed data, the system also provides 4 dimensions to audit the routed paper data with time filter: publisher, institute (IR), fund (IR) and batch. In the audit results, the manager could easily find out the expected, successful and failing paper number.

In future, the system will provide more detailed and specific audit results for publishers, IR managers and other people.

5. Application of iSwitch

Now, iSwitch has cooperated with many publishers, such as Web of Science[21] (WOS), Chinese Science Citation Index (CSCD)[22], IOPP, PLoS, and others, and has pushed paper data for Chinese Academy of Sciences (CAS) IR.

5.1 Cooperation with WOS

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

WOS has a well-known Science Citation Index (SCI) and has collected the metadata of the paper of SCI. The papers in SCI are considered the most influence in nature science field. So iSwitch accepts SCI papers metadata and distributes them to target CAS IRs to help them check whether these papers had been deposited or not, the author had submitted the correct metadata in IR and full-text documents of these papers had been submitted. IR also assigns the corresponded WOS paper ID (WOSID) to the archived paper and provides a full-text link to the WOS web site if the paper has a full-text document. WOS will show the full-text link of IR if exists. In this way, iSwitch improved IR's integrity, accuracy and the utilization of science study output. The following figure shows the cooperation of iSwitch, WOS, and IR.

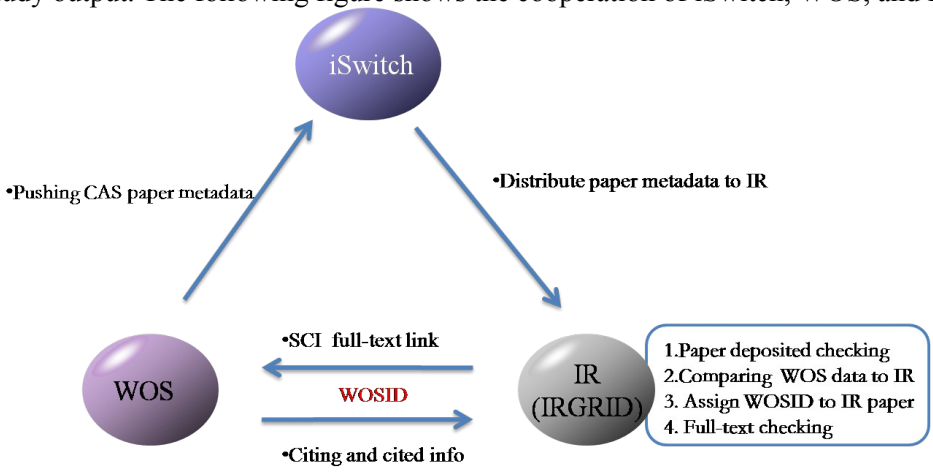


Fig. 3 Cooperation Relation of iSwitch, WOS and IRs

WOS has pushed iSwitch 362551[2] CAS papers (1949-now, SCI-EXPANDED) metadata and distribute them to 99 IRs with 484088 times. Now, iSwitch has built a stable router service

2 Until 2015-07-28

between IR and WOS to accept updated data and distribute them, and WOS has released CAS IR full-text link to its web site too since 26 July, 2015.

5.2 Cooperation with other Publisher

CSCD, a citation index for Chinese science article, built correspond to SCI, has pushed data to iSwitch for IR checking too. iSwitch has built a similar service as WOS for CSCD and IR, which can help IR archive valuable papers written in Chinese journal.

Besides the metadata, PLoS and IOPP have reached primary agreement with iSwitch for Open Access paper auto deposit. PLoS has sent some test papers and IOPP has closely contact with iSwitch for future work.

6. Conclusion

In this paper, the authors developed an iSwitch router service system to help Open access paper author auto deposit and IR checking.

iSwitch accepts metadata and full-text papers from publishers, resolves their target IR and pushes these data to IR. iSwitch uses FTP and SWORD protocol to receive paper data, which are encode with JATS(recommended) and other standard format. Through alias and relation tables, and other methods, iSwitch finds out the target IR of the papers, and push them to desired IR by FTP. Now iSwitch has cooperated with WOS, CSCD, PLoS, IOPP, and CAS IRs, and built a stable service for them.

iSwitch is an important demonstration for libraries, as an infrastructure, it will facilitate the content deposit and distribution of the IRs. It bridges publishers and repositories in a more effective way. To be a public service, iSwitch will be of great help to all those involved in IRs, 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), and funded by the Chinese Academy of Sciences.

Reference

1 Zhang Xiaolin, Liang Na, Qian Li, Shi Hongbo. Router Service Engine iSwitch for Open Access Articles: The Concept, Strategy, and Framework. , 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_manual/hi/oa_pilot/h2020-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 http://www.imls.gov/imls_and_sloan_foundation_award_1_million_to_arl_for_share_notification_service.aspx
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. , 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 [EB/OL]. [2015-07-25]. http://sourceforge.net/projects/sword-app/.
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.
21 Web of Science.[EB/OL].[2015-07-25]. http://apps.webofknowledge.com/
22 Chinese Science Citation Index. [EB/OL].[2015-07-25]. http://sciencechina.cn/