Postprint Institutional Repository Content Construction Based on the iSwitch Open Access Paper Push and Forward Service System

Authors: Zhang Wangqiang, Zhu Zhongming, Yao Xiaona, Liu Wei

Date: 2017-10-11T00:00:00+00:00

Abstract

【Purpose】To automatically synchronize and deposit the institutional knowledge output data distributed by the open access paper push forwarding service system iSwitch into the institutional repository. 【Method】Using scheduled task scheduling and FTP protocol for data synchronization, preloading data into the database through file packages and file parsing, while providing functions such as import management, imported data management, and auditing. 【Results】Achieved automatic synchronization and semi-automated import of data. Completed the reception and deposit of over 60,000 Web of Science data entries. 【Limitations】The accuracy and timeliness of iSwitch push data need improvement, and the IR needs to further optimize its data import functionality to increase automation. 【Conclusion】The content construction of institutional repositories based on iSwitch greatly reduces the burden on researchers and institutional repository managers while ensuring data quality. This model has certain promotional value.

Full Text

Preamble

ChinaXiv Partner Journal, Issue 269, 2016, No. 4

Building Institutional Repository Content Based on the Open Access Paper Push and Forward Service System iSwitch*

Zhang Wangqiang, Zhu Zhongming, Yao Xiaona, Liu Wei
(Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000, China)

Abstract

[**Objective**] To automatically synchronize and deposit open access papers distributed by the iSwitch service into institutional repositories. [**Methods**] We employed scheduled task synchronization via FTP protocol, preloaded data into the database through file package parsing, and provided functions for import management, deposited data management, and auditing. [**Results**] The system achieved automatic synchronization and semi-automated import, successfully processing over 60,000 records from Web of Science. [**Limitations**] The accuracy and timeliness of iSwitch data distribution need improvement, and IR systems must optimize import functions to increase automation. [**Conclusions**] iSwitch-based IR content construction significantly reduces the burden on researchers and repository managers while ensuring data quality, representing a model worthy of broader adoption.

Keywords: Open Access; Institutional Repository; iSwitch; Content Recruitment
Classification Number: G250.7

Institutional Repository (IR) content development has long faced the challenge of low researcher participation. According to the European PEER project survey, even when publishers invite authors to upload their final peer-reviewed manuscripts, the actual deposit rate is only 2% [1]. Experience from Chinese Academy of Sciences IR construction reveals that low participation stems primarily from researchers' unfamiliarity with IR submission procedures, concerns about time investment, and insufficient IR user appeal [2].

Acquiring institutional scholarly output from external systems through standard web protocols (such as OAI, REST API, and SWORD) has become a popular IR content development approach. This method reduces redundant resource construction and prevents metadata loss that may occur during manual submission.

With the development of open access, an increasing number of publishers support pushing paper metadata (and even full text) to authors' institutional repositories through machine interfaces. For example, MIT Library collaborates with BioMed Central, which automatically pushes article metadata and full text to MIT' s IR [3]. The Italian National Institute of Geophysics and Volcanology (INGV) [4] has an agreement with the open access journal *Annals of Geophysics*, which automatically submits papers to INGV' s Earth-prints repository upon publication [5]. The JISC Repository Junction Broker (RJB) project [6] aims to establish a paper exchange hub that receives data from multiple publishers and distributes it to each author' s institution.

The Chinese Academy of Sciences Library' s paper push and forward service system—iSwitch [7]—was conceived similarly to JISC' s RJB. Its primary function is to acquire and distribute papers published by Chinese Academy of Sciences corresponding authors from relevant publishers, providing standard interfaces

to support data sharing [8]. The Chinese Academy of Sciences IR system collaborates with iSwitch, with each institute' s IR automatically harvesting and importing its institutional output data distributed by iSwitch.

iSwitch-based IR content construction primarily employs scheduled task synchronization via FTP protocol. File packages are parsed to extract distribution batch information and scholarly output metadata, which is preloaded into the database. Administrators are provided with data import management functions to achieve semi-automated final import. Additionally, error management, batch updates for deposited data, and auditing functions are included.

iSwitch uses the FTP communication protocol for data sharing. When new data is received from publishers, it parses and identifies author institutions and funding organizations, then distributes the output data to the corresponding institutional directories. The main functions of iSwitch-based IR content construction include data synchronization, batch data browsing and import, deposited data management, work claim, and auditing. The overall system functional framework is shown in [Figure 1: see original paper].

[Figure 1: see original paper]

The iSwitch services directly used by IRs mainly include authorization management, FTP servers, and batch status interfaces.

1. **Authorization Management:** iSwitch employs IP-based access control. IR developers typically register relevant information by logging into the iSwitch authorization system.
2. **Data Automatic Synchronization:** Through scheduled task scheduling, IRs synchronize institutional output from iSwitch's FTP server. After downloading data to local cache, files are parsed and metadata is mapped and converted before being saved to the database, with import reminder emails sent to IR administrators.
3. **Batch Browsing and Import:** Supports IR administrators in browsing batch lists and detailed information about outputs within each batch. Due to potential network transmission interruptions and administrators' needs for real-time synchronization, the system supports manual synchronization of selected or all batches. Administrators ultimately complete the import execution, determining import methods, target collections, and import fields. After a batch is fully imported, the import status is fed back to iSwitch' s batch status information interface.
4. **Deposited Data Management:** Supports administrators in browsing deposited data by import method classification, managing status records, and re-importing erroneous or deleted data. Considering possible updates to iSwitch' s original data, an automatic metadata update function is provided for deposited output.
5. **Claim and Auditing:** iSwitch-imported data generally includes only

metadata without full text. Through work claim reminders and full-text solicitation, along with storage auditing functions, IRs audit and remind users to complete work claims and full-text uploads in a timely manner.

3. Key Function Design and Implementation

3.1 Data Synchronization

The organizational structure of iSwitch-distributed scholarly output data on FTP servers follows "Publisher → Institution → Batch → Article." One batch corresponds to one file distribution, which generally contains multiple articles. Each article is stored as a ZIP file package, including the publisher' s original metadata description document (typically in XML format) and a document re-encoded by iSwitch using the JATS (Journal Article Tag Suite) standard [9].

iSwitch provides a separate service interface for batch and output list descriptions, but the FTP directory structure itself contains this information, and original data acquisition also requires downloading from the FTP server. Therefore, IRs obtain batch information by directly reading the FTP directory structure. To facilitate file synchronization, IRs establish a local cache directory with an organizational structure consistent with iSwitch' s FTP batch file storage directory.

IRs created scheduled tasks based on the Quartz framework [10] to automatically harvest iSwitch data. The current default harvest frequency is once per week. FTP file download functionality is implemented primarily through the FTP Client module of the Apache Commons-Net package public service interface. JDOM components are used to parse and read metadata from XML description documents. The batch data automatic synchronization process flow is shown in [Figure 2: see original paper].

[Figure 2: see original paper]

After synchronization tasks launch, IRs retrieve the institution' s official name from parameter configuration and construct a URL in the format "iSwitch domain/institution name." Accessing this address obtains all batch directories distributed to the current institution, then iteratively reads the list of all output files under each batch, checking whether each output exists in local cache. If not, it is downloaded. Each downloaded output is a ZIP file whose name contains distribution year-month and article unique ID information, such as "201410.00024." IRs use the ZIP filename as each article' s unique identifier to query the database and determine whether the article has been loaded. If not, IRs decompress and extract the metadata description document for parsing. Considering potential metadata loss from re-encoding, IRs choose to directly read the publisher' s version. Since IR underlying data description is based on the DC metadata framework, parsed metadata requires mapping before being stored in the IR database. Different publishers generally use different encoding formats, requiring creation of corresponding data parsers for each publisher' s

data source. During synchronization, IRs also detect batches and output data deleted from iSwitch but already loaded in the local database, removing these entries (this operation does not affect deposited data). After automatic synchronization completes, if new data was downloaded, reminder emails are sent to IR administrators.

In addition to automatic synchronization, IRs provide manual synchronization functionality. Besides implementing similar functions to automatic synchronization, manual synchronization supports optional batch selection and whether to re-parse metadata from description documents and update the database.

3.2 Data Import

The storage organization of synchronized but not yet imported output data is essentially consistent with formally deposited and publicly accessible data in the IR database, except it is marked as not imported and not publicly accessible.

IRs provide data import management functions for administrators to import iSwitch data. Import operations are not fully automated because IR knowledge output is organized by collection, requiring administrators to determine which collection each article belongs to. Additionally, handling duplicate data requires administrators to determine import methods and metadata fields.

The import function supports three optional deduplication methods: title only; content type + title; content type + (publisher article ID or DOI or title). Three import methods are supported: add new, supplement, and skip. "Add new" creates a new record regardless of whether it already exists in the IR. "Supplement" updates existing output metadata with iSwitch data. "Skip" can be selected for outputs not belonging to the institution.

To reduce manual operations during import, the system performs some data preprocessing when loading pending import data, primarily detecting existing data and determining target collections. Using the default deduplication method, the system detects whether each pending import entry already exists and displays deduplication results in the import management interface. For existing entries, the target collection defaults to match the existing data' s collection, with "supplement" as the import method. For non-existing entries, "add new" is the default method. The system preliminarily determines potential owning users through fuzzy matching between author names and the system' s user alias database, using the collection of that user' s published journal papers as the default collection for pending import data. IR administrators can modify default collections and import methods during execution.

The import management interface is shown in [Figure 3: see original paper].

[Figure 3: see original paper]

The upper portion of the page contains public parameter settings for the current import, including target batch, content type, deduplication method, items per

page, and fields to import. The lower portion displays the list of non-imported output in the current batch. The left side shows non-imported entries parsed from iSwitch original data and saved in the database, with Chinese Academy of Sciences authors and address information highlighted in yellow background to facilitate administrator verification of institutional affiliation. The right side shows existing entries detected through the selected deduplication method. Below each entry are options for target collection and import method. The current function supports importing multiple entries at once. The side-by-side layout clearly displays non-imported and corresponding existing data, maximizing simplification of administrator operations and improving work efficiency.

3.3 Fault Tolerance

In iSwitch-based IR content construction, errors may occur at every stage from iSwitch distribution to IR data download, parsing, import, and even post-import processes. Main error phenomena and IR solutions are as follows:

1. **iSwitch Distribution Errors:** Due to inconsistent author affiliation information across different journals, iSwitch may distribute some non-institutional data. For such cases, IR administrators manually identify and select "skip" during import.
2. **Synchronization Task Errors:** Synchronization errors mainly manifest as network transmission interruptions and parsing failures for special encoding formats or characters in metadata description documents. These errors cause inconsistency between synchronized IR data and iSwitch original data. IRs provide manual synchronization functionality, with options to reload iSwitch original data. When parsing errors are discovered, upgrading program code and re-executing manual synchronization resolves the issue. For data already imported into the system, IRs also provide batch update functionality, allowing administrators to select specific metadata fields for updating.
3. **Re-importing Skipped or Deleted Data:** Some iSwitch data imported with "skip" may later be identified as institutional output. Additionally, some imported data may be intentionally or unintentionally deleted by administrators but later need recovery. To address this need, IRs do not delete iSwitch original data entries after successful import but save imported data as separate entries while preserving associations between them. This design records whether deposited data has been deleted and enables associated updates when iSwitch original data is updated. Based on this underlying design, IRs support re-marking skipped or deleted data as not imported for re-import.

IR iSwitch data auditing functions support system administrators in real-time understanding of total batch count, import status per batch, deposited work claim status, and full-text upload status. For iSwitch, it is necessary to know whether each institution' s distributed data has been downloaded and whether

certain batches have been fully imported into IRs. IRs promptly return batch download and import status to iSwitch during synchronization and import processes to support iSwitch auditing requirements.

1. **Institute IR iSwitch Data Auditing:** Auditing at the institute IR level is implemented within the IR. The IR records import status and method for each batch' s output, supporting batch import progress tracking and classification statistics of deposited data by import method. IR work claim reminders and full-text solicitation functions maintain real-time records of user-work associations, work claim status, and full-text storage status, enabling administrators to audit unclaimed works and claimed works without full-text submission, and to send batch notification emails to relevant users.

2. **Support for iSwitch Auditing:** iSwitch auditing of IR data usage, which requires IR system collaboration, currently focuses primarily on obtaining batch import success status. When administrators execute batch import operations, the IR checks whether the current batch has been fully imported. If import is successful, batch unique identifiers and status information are returned to iSwitch' s interface.

4. Application Effectiveness Evaluation

Currently, iSwitch has completed historical backfilling of Chinese Academy of Sciences author output articles indexed in WoS (Web of Science) and supports automatic reception and distribution of new data. As of December 8, 2015, 83 Chinese Academy of Sciences institutes have deployed iSwitch data monitoring and import functionality in their IRs. According to statistics from the Chinese Academy of Sciences Grid System (IR Grid) [11] on institute iSwitch data download and import status, 67,024 iSwitch-sourced output entries have been successfully imported, with 49,885 containing full text, achieving a full-text storage rate of over 74%. As iSwitch continuously receives and distributes new output, these statistics continue to grow. The top 10 institutes by iSwitch data import volume are shown in .

**** iSwitch Data Import Top 10 Institutes

Institute Name	Data Import Volume
Dalian Institute of Chemical Physics, Chinese Academy of Sciences	
Institute of Process Engineering, Chinese Academy of Sciences	
Institute of Oceanology, Chinese Academy of Sciences	
Kunming Institute of Botany, Chinese Academy of Sciences	

Institute Name	Data Import Volume
Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences	
Institute of Psychology, Chinese Academy of Sciences	
Xi' an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences	
Institute of Hydrobiology, Chinese Academy of Sciences	
Institute of Chemistry, Chinese Academy of Sciences	
Institute of Mountain Hazards and Environment, Chinese Academy of Sciences & Ministry of Water Resources	

iSwitch-based IR content construction represents an ideal model in the context of low researcher participation. This approach not only reduces the burden on researchers and IR administrators but also prevents metadata errors or loss that may occur during manual operations.

System usage has revealed some limitations. For instance, some institutes report inability to obtain the latest WoS-indexed data from iSwitch, with delays in new data distribution. There are also cases where WoS-indexed data is not distributed by iSwitch. These issues relate to data completeness and push frequency from publishers themselves, as well as iSwitch' s own data distribution mechanisms. Additionally, due to non-standard author affiliation entries, inconsistent affiliation information across journals from different sources, and institutional name changes over time, some articles cannot be correctly distributed to their true institutional affiliations. IR data import automation needs further improvement. For example, currently administrators must confirm import of existing data, but future enhancements could allow administrators to predefine handling rules for existing output. During synchronization, if iSwitch data matches IR deposited data by title and first author, import could proceed automatically according to predefined rules.

iSwitch currently only supports distribution of WoS-sourced data; support for automatic acquisition and forwarding of open access papers from more publishers is desired. Both systems continue to improve and optimize during actual operation, making inter-system interaction smoother, functions more complete, and automation higher, thereby truly solving deposit challenges through iSwitch-based IR content construction.

References

[1] How to Increase Content in OA Repositories –What Can Be Learnt from the Special Case of the Research Project PEER-Publishing and the Ecology of European Research [EB/OL]. [2015-12-10]. http://www.peerproject.eu/fileadmin/media/ppt_about_peer/PEER-How_to_increase_content_in_repositories_April2012.pdf.

[2] Zhang Xiaolin, Liang Na, Qian Li, et al. Router Service Engine iSwitch for Open Access Articles: The Concept, Strategy, and Framework [J]. New Technology of Library and Information Service, 2014(10): 4-8.

[3] Duranceau E F, Rodgers R. Automated IR Deposit via the SWORD Protocol: An MIT/BioMed. Central Experiment [J/OL]. UKSG Series, 2010, 23(3): 212-214. http://dx.doi.org/10.1629/23212.

[4] Italian National Institute of Geophysics and Volcanology [EB/OL]. [2015-12-10]. http://www.ingv.it/en/.

[5] Lewis S, De Castro P, Jones R. SWORD: Facilitating Deposit Scenarios [J/OL]. D-Lib Magazine, 2012, 18(1-2). http://www.dlib.org/dlib/january12/lewis/01lewis.html.

[6] Jisc Publications Router [EB/OL]. [2015-12-10]. http://broker.edina.ac.uk/information.html.

[7] Router Service Engine iSwitch for Open Access Articles [EB/OL]. [2015-12-10]. http://iswitch.las.ac.cn/.

[8] Shi Hongbo, Qian Li, Zhang Xiaolin, et al. Router Service Engine iSwitch for Open Access Articles: Articles Reception and Resolving [J]. New Technology of Library and Information Service, 2015(6): 1-6.

[9] Journal Article Tag Suite [EB/OL]. [2015-12-10]. http://jats.nlm.nih.gov/.

[10] Quartz Scheduler [EB/OL]. [2015-12-10]. https://quartz-scheduler.org/.

[11] IR Grid [EB/OL]. [2015-08-14]. http://www.irgrid.ac.cn.

[12] Article Match Retrieval [EB/OL]. [2015-08-14]. http://wokinfo.com/products_tools/products/related/amr

Author Contributions

Zhang Wangqiang: Drafted the manuscript, detailed system function design and implementation.
Zhu Zhongming: Proposed research ideas, revised final manuscript version.
Yao Xiaona: Statistics on iSwitch distributed paper data downloaded and imported by various institute IRs.
Liu Wei: System function upgrade and deployment.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the online version of the journal at http://www.infotech.ac.cn.

[1] Shi Hongbo. 201410.00016.zip. Sample data of single article distributed by iSwitch.

Received: 2015-12-14
Revised: 2015-12-29

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