开放获取论文推送转发服务系统 iSwitch:论文接收与解析*

师洪波[1] 钱力[1,2] 张晓林[1] 梁娜[1]
1（中国科学院文献情报中心 北京 100190）
2（中国科学院大学 北京 100049）

摘要：
[目的]对开放获取论文推送转发服务系统 iSwitch 接收与解析等模块介绍。**[方法]**根据系统前期技术、标准等调研、需求分析及关键问题解决方案，设计实现系统的论文接收及解析模块。**[结果]**实现 iSwitch 系统论文接收及解析模块，并对 Web of Science 中 34 332 条文章数据进行测试接收及解析。**[局限]**主要针对实验数据进行测试，对于系统实际运行可能遇到的更多问题考虑不够全面。**[结论]**论文作者机构的解析是很多研究中面临的共同问题，本文的解决方案对相似系统功能的设计开发有借鉴参考价值。
关键词：推送转发　论文接收　机构解析　iSwitch
分类号：G250.7

Router Service Engine iSwitch for Open Access Articles: Articles Reception and Resolving

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

Abstract:
[Objective] This paper provides description of the implementation of iSwitch system's Reception and Resolving. **[Methods]** Based on the investigation and analysis of technology, standards, and the key problems, this paper shows the design and implementation of the Reception and Resolving of iSwitch. **[Results]** Implement the Reception and Resolving portions of iSwitch System and make a test based on Web of Science article data(34 332). **[Limitations]** The problems and difficulties may encountered in real service are not considered enough. **[Conclusions]** Articles' affiliation resolving is a common problem for library and information study, and the solution of this article has reference value for other similar system's design and implementation.
Keywords: Push and routing　Article reception　Resolving institutions　iSwitch

1　iSwitch 背景及技术框架

为支持开放获取论文从多个出版社向多个研究机构或资助机构的自动推送服务，文献[1]提出了开放论文推送转发服务系统 iSwitch。其主要功能即通过及时接收相关出版社推送的论文及其元数据，解析并识别相应的作者、作者机构和资助机构，并通过标准接口转发论文到对应机构知识库（包括作者机构和资助机

* 本文系中国科学院文献情报能力建设专项"国际开放论文国家交换服务中心示范系统"(项目编号:Y14008)的研究成果之一。

构）中。文献[2]提出了 iSwitch 系统相关角色，包括推送方、转发方和接收方，并分析了工作流程、重点环节及可参考的标准和规范。

以上述研究背景为基础，笔者对 iSwitch 系统进行了实现。依据 iSwitch 工作流程，将其划分为三个主要模块，即数据接收、数据解析、数据分发。数据接收模块使用标准协议接收来自推送方的论文数据并进行统一格式保存；数据解析模块对论文的作者数据、资助机构数据进行解析分析；数据分发模块将接收的论文数据转发到其对应的目标知识库中。在这个过程中需要遵守约定的标准及协议。图 1 对 iSwitch 系统进行了技术框架描述。本文主要对接收及解析模块的实现进行介绍，分发推送模块在《开放获取论文推送转发服务系统 iSwitch：论文分发推送》一文中详细讨论。

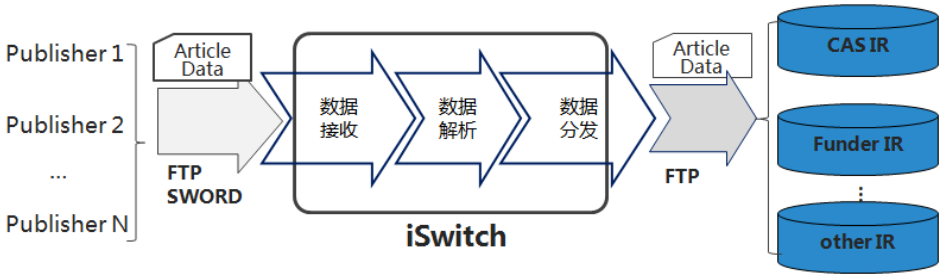

图 1 iSwitch 技术框架

2 iSwitch 论文接收模块设计

2.1 数据接收模块规范要求

该模块完成从推送方接收或收割（简称为接收）符合系统要求的开放论文数据，同时存储为 iSwitch 系统数据，以备解析分发使用。主要要求如下：

（1）使用标准协议传输方式

FTP 协议在传送大体量数据方面有先天优势，是一种传统、应用广泛的数据传输方式，多数出版商及相关机构支持 FTP 协议，如 PMC 使用 FTP 协议接收各出版商提交数据[3]。FTP 的优点有：大体量数据安全传输；得到广泛使用及支持。缺点主要是：其并非一种标准的仓储库互操作协议，需双方约定好传输数据的格式。

SWORD 协议[4]具有以下几方面优势：是一种专门的仓储数据传输协议；得到众多仓储软件支持[5]，使用 DSpace[6]、Fedora[7]、EPrints[8]等软件作为基础仓储软件的知识库，可以快速实现对 SWORD 协议的支持；有比较长的实践历史（始于 2008 年），得到部分出版商的支持（例如 BMC[9]）；拥有开源类库[10-11]，可以比较灵活的个性化本地实现。但是 SWORD 协议也有局限性：其实现基于 HTTP 协议，在一次性传输大体量数据时可能会出现中断、超时等问题[12]；不是所有出版商都支持 SWORD 协议，需要额外的开发成本。

综上所述，两种协议各有利弊，所以本系统接收模块同时支持两种协议。

（2）使用个性化配置数据接收方式

由于推送方直接提供的元数据格式（包括版本）、传输方式等各不相同，同时也很难要求出版商按照完全一致的方式向 iSwitch 提供论文数据。所以系统设计过程中，建立了多种的通用模块，可以通过灵活配置，为出版商提供个性化数据推送方式及数据解析功能。

（3）使用归一化数据保存

对于数据推送方提供的各种异构数据，需要转换为 iSwitch 系统数据统一存储管理。主要提供的功能包括：对文章元数据，特别是文章的作者及其对应机构、对应 E-mail 进行准确解析保存；对文章的基金资助信息进行保存；对文章的多个版本进行保存、区分；对接收的原始数据进行安全保存及备份等功能。

2.2 数据接收协议实现

（1）FTP 接收模块实现

FTP 接收模块是基于开源软件 Apache Commons 的 Net[13]模块实现，对 FTP 操作的公共方法类进一步封装，结合个性化元数据数据解析模块提供 FTP 数据接收。主要流程如下：

①下载 Apache Commons 的 Net Jar 包，实现 FTP 操作的公共方法封装，如登录、列出目录、下载等操作。

②编写数据收割模块，对指定目录进行下载。

③调用数据解析保存模块，完成数据接收。

（2）SWORD 服务模块实现

SWORD 服务模块，主要实现 SWORD 协议的服务端（也就是 iSwitch 面向出版商数据的接收端）部分。目前 SWORD 主要支持两个版本协议，分别是 1.x 和 2.0 系列。1.x 系列最高版本为 1.3，1.3 系列被更多的出版商支持，所以项目目前对 1.3 版本的服务端进行开发。后续根据对 2.0 版本的需求进行后续开发。

具体实现使用 Java 语言，基于 SWORD-Common1.1 开源工具包[14]，实现 SWORD 1.3 服务端的开发。服务端的实现参考了 DSpace SWORD 服务端[14]和 Fedora SWORD 服务端[14]的实现源码。实现流程如下：

①在 http://sourceforge.net/projects/sword-app/下载并编译 SWORD Java Common Library 1.1（具体指令参见包内说明文件）。

②在 iSwitch 工程 web.xml 中插入 servicedocument、deposit、media-link 三个 Servlet 配置项及相应 servlet-mapping 部分。

③在 web.xml 中配置 sword-server-class 参数，对应 Class 实现 SWORD 服务端三个功能模块；配置 authentication-method 参数，指定使用 Basic 模式登录验证。

④在 LasSWORDServer 类继承 SWORDServer 类，并实现其中 doServiceDocument、doDeposit、doAtomDocument 三个方法，分别对应 SWORD 规定的三个请求，即请求服务文档(servicedocument)、请求存储资源(deposit)、请求资源 Atom 文档(media-link)。可参考 DSpace 的 SWORD 客户端实现方式[15]。

2.3 数据接收保存

通过实现 FTP 和 SWORD 两种数据接收协议，完成了论文数据从推送方到 iSwitch 的传输过程，但是数据还是以推送方提供的方式存在，需要对这些数据进行解析、结构化存储等操作才能够进行后续推送操作利用。

（1）数据接收流程

图 2 展示了数据解析的主要功能及流程。其中比较重要的部分是批次解析和元数据解析两个部分。

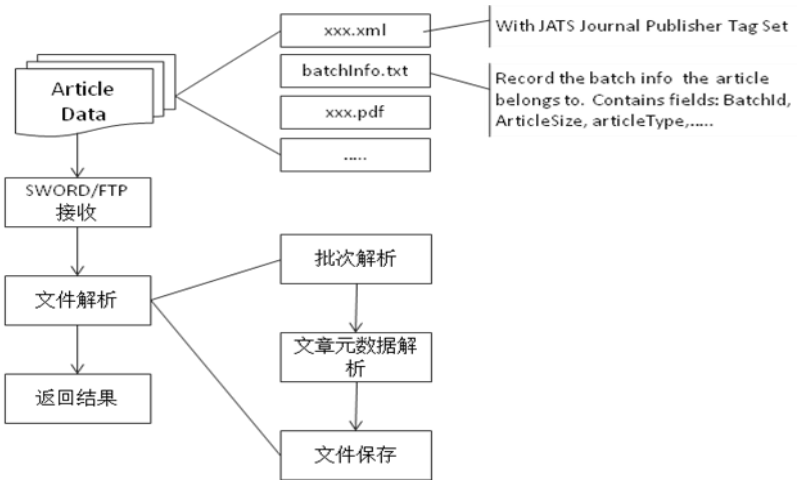

图 2 接收解析流程图

（2）批次解析

推送方通过 FTP/SWORD 等方式提交的论文数据包，以一个批次向 iSwitch 推送。该批次以双方约定的推送频次为基础，如每周、每月等推送一次。批次信息应记录该批次中论文数量、卷期数、论文编号等信息，这些信息在后续的数据检查中会得到利用。

（3）文章元数据解析

根据与出版商约定，文章元数据应使用标准的元数据描述方式，例如 JATS[16] 标准。但是 JATS 本身包含三个子标准，同时各个子标准也有不同的版本。这就需要与出版商详细约定标准，以开发特定的解析工具，并在配置部分配置到相应的出版商。

值得注意的是，除了解析文章常规的字段如标题、摘要、关键词、作者、地址、出版商、期刊、卷期页等常规信息外，还需要解析记录开放获取类型、开放时限、作者与地址对应关系、资助情况等信息，这些信息既是文章对外开放的依据，也是分发过程依赖的重要信息。

（4）数据保存

首先将论文的元数据按照一定的结构保存，为文章分配 iSwitch ID。对于推送方同一 ID 的文章，在 iSwitch 中也应该分配同一个 iSwitch ID，并且保存相应版本信息。

对于推送方的不同数据格式，统一转换为 iSwitch 数据存储规范格式。例如出版商对于地址信息的分隔字符可能使用“,”，也可能使用“;”，在存储的时候统一转换为使用“,”分隔符存储。

对于出版商提交的论文原始文件也需妥善保存，以保证数据的完整性。在系统中，以 iSwitch ID 为基础，以 ZIP 文件包的形式对出版商提交的数据按篇进行打包保存。

3 iSwitch 数据解析模块设计

3.1 数据解析模块主要功能

（1）地址解析

准确全面地解析文章作者所在机构以及文章资助基金和资助号码，才能够保证文章正确地转发到对应机构知识库。地址解析模块主要功能包括：

①作者机构地址解析。作者所在机构信息是确定文章所属接收方，也就是所属 IR 的直接对应信息。在解析过程中，对于明确列出机构信息数据项的数据，处理方便且准确度高；但很多数据没有提供机构信息的数据项，而只提供了作者机构地址数据项，这些数据则需要按一定的规则匹配来实现作者机构的正确解析。

②资助机构解析。开放论文的接收方也包括资助机构，主要通过论文资助基金信息获取。基金信息在论文元数据中往往不单独列出，而是放在 Acknowledgement 当中，需要对其中的中国主要资助机构及资助号码进行匹配解析。

（2）数据检查

对于推送方提供的数据，需对这些数据进行多方面的检查校验，包括：必备字段检查、推送数据包完整性检查等，详见 3.3 节。

3.2 机构解析地址实现

（1）解析难点

作者所属的机构一般在作者地址信息中，而地址信息并不是完全结构化的信息；文章的资助基金信息往往在文章的致谢部分，往往是一段描述性的文字。具体难点如下：

①地址（Affiliation）字段中，国家、城市、机构等写在一起，运用规则区分有时会出现匹配出错。以地址信息为例：“State Key Laboratory of Luminescence and Applications, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, People's Republic of China”，该信息中的准确机构为 Changchun Institute of Optics, Fine Mechanics and Physics，而该信息中既混杂了其上层机构“Chinese Academy of Sciences”，又有下层机构“State Key Laboratory of Luminescence and Applications”，而且机构名称中包含分隔符（“,”），直接匹配出准确的机构比较困难。

②机构名称变化。同一个机构，可能由于历史原因，存在多种名称，这些名称可能同时使用，也可能有些是旧名称；有些机构有分拆、合并等情况，名称可能也会有变化。如中国科学院文献情报中心，又称国家科学图书馆。也有一些使用缩写的情况，如在 Web of Science 中的二级机构名称。这些在解析过程中也需要考虑。

③资助基金及资助号码没有独立的描述字段。如在“致谢（Acknowledgements）”字段中描述资助信息。下面展示了这种情况——“This work is supported by the National Basic Research Program of China (973 Program) under Grant No. 2011CB302004 and by the National Natural Science Foundation of China under Grant No. 60506014 and 11004187.”这里需要单独的规则抽取或者比对。

④一个科室属于多个机构情况。一个实验室可能由不同机构共同组成，如边缘海地质重点实验室由中国科学院广州地球化学研究所及中国科学院南海海洋研究所合作组成，对于该实验室的文章需要解析出上述两个机构，这就要求在设计数据结构及查找机构的过程中要加以考虑。

⑤中国科学院机构关系复杂。中国科学院与大学等科研机构不同，其有众多研究机构共同组成，在查找属于中国科学院的文章时，既要确定该文章是中国科院的，又要确定该机构的具体名称。

（2）应对策略

①尽可能让出版商提供详细的地址，直接提供机构及基金名称。如在 JATS 的描述标准中，可以将作者的机构、国家以及资助机构、资助号码等单独列出，而且大部分出版商的数据库中保存了作者及对应机构信息，比如 Springer。如果出版商能够直接提供机构、基金等信息，能够大幅度提高目标机构知识库匹配准确度。

②对于机构名称中存在分隔符“,”的情况，在匹配过程中使用全字匹配的方式优先匹配含逗号的机构名称，消除机构名称使用逗号分割匹配产生的错误。例如上述机构“Changchun Institute of Optics, Fine Mechanics and Physics”，在匹配过程中首先从数据库中找出包含逗号的机构名称，并与机构信息逐个匹配，如果没有找到机构，再使用逗号分割机构信息，进行后续的匹配查找。

③在 iSwitch 机构配置方面，将机构存在的多种名称形式配置成机构别称，保证机构名称匹配的全面性。如中国科学院文献情报中心的英文标准名称是“National Science Library”，但也可能有人使用“Library of Chinese Academy of Sciences”，在机构别称表中加入该名称，则可以匹配出该机构。这种方式需要比较全面地掌握机构名称变化，需要结合接收的数据更新配置信息。

④对于基金，由于目标基金较少，同时基金编号规则一般固定而且彼此区分，则可以使用正则表达式等方式对于基金信息逐个匹配。

⑤增加手工分配功能。对于一些找不到具体机构但能确定其为中国科学院的文章，或者有重名机构的文章，需要人工检查到对应的机构知识库，对其进行手工分配。同时生成自动配置规则，保证后续匹配的自动运行。

3.3 数据完整性检查实现

完整性检查主要包括以下两个方面：

（1）必备字段检查

按照与推送方的约定，推送方提供的数据应该包含必备的一些字段，包括论文编号、DOI、标题、URL、期刊名称、出版卷期年、出版页码、作者、作者机构（多个作者和机构时应注明作者和作者机构对应关系）、论文资助机构、资助项目名称与编号（多个资助机构和项目时应注明作者、资助机构和资助项目对应关系）、论文版本、开放获取状态、开放时滞期标志、出版时间等。对于接收完毕的数据，对各字段逐一检查。对于字段有缺失的数据需要进行记录，与推送方协同重新获取数据。同时该批次数据在数据完全通过检查之前，不能转发给数据接收方。

（2）推送数据包完整性检查

检查本批次信息与推送方提供的批次信息是否一致。检查内容包括：论文数量、论文编号与推送包封装批次信息是否一致，如果不一致，需要联系推送方，核对批次信息及数据报信息。该批次数据在通过数据包完整性检查之前，不能转发给数据接收方。

4 iSwitch 论文接收与解析实验

依据接收模块及解析流程设计，项目组创建完成了 iSwitch 系统。系统利用 Java 语言，实现了 FTP/SWORD 数据接收、基于出版商个性化数据接收配置、机构解析、机构及基金配置等功能并进行了数据接收解析实验。

4.1 实验说明

（1）实验数据

通过合作，由汤森路透公司提供了 Web of Science 引文数据库（仅 Science Citation Index Expanded）收录的中国科学院 2013 年发表论文作为实验数据，共 34 332 条。

（2）硬件条件

CPU：Intel i3-3210M，内存：4GB，操作系统：Windows7 专业版 32 位，硬盘：500GB。

（3）推送方式

由于 FTP 是数据推送方使用最广的方式，本实验模拟 FTP 推送数据。

4.2 实验结果及分析

（1）实验结果

实验中，模拟推送方以 FTP 方式提供数据，系统完成了对提交数据的接收、解析及保存工作，实现了对文章作者机构及资助基金数据项信息的正确解析。34 332 篇文章共解析出对应的 99 个机构（含同一机构包含多篇文章情况），目标基金 2 个，分别包含 16 735 篇（国家自然科学基金）和 4002 篇（973 计划）文章。接收及解析情况如下表 1 所示：

表 1 接收及解析情况统计表

接收批次数	接收论文篇数	解析覆盖机构数/篇数	解析目标基金数量/篇数
1	34 332	99/35 562	2/20 737

（2）总体实验效率分析

实验中使用内网 FTP 模拟推送数据，所以数据处理起来比实际处理时间可能稍快。接收数据的处理时间如表 2 所示：

表 2 总体时间效率分析

总体接收及解析时间（分）	平均每篇消耗时间(秒/篇次)
60.75	0.106

通过以上实验过程及结果可以看出，目前 iSwitch 系统可以满足论文数据接收及转发的功能需求。

5 iSwitch 运行系统配置

iSwitch 系统的正常运行，依赖于对系统的准确、详细配置，系统的配置功能主要包括以下两个方面：

5.1 推送方配置

（1）推送方信息注册

接收推送数据之前，要对推送方进行系统注册，对推送方如某个出版商的名称、联系方式等进行注册，对每个推送方的数据予以区分。

（2）推送方式注册

对于不同的推送方式，需要不同的信息支持。对于 FTP，主要配置 FTP 用户名、密码、对应的目录等，同时在系统中配置 FTP 收割任务。对于 SWORD，需

要配置 SWORD 用户名、密码、SWORD 提交数据集以及对于该出版商的文章解析工具等。

5.2 接收方配置

（1）机构名称配置

对文章分发机构主要配置其中英文名称。替代名称包括现有的中英文名称、各种简称、曾用名等。对于不同的出版商，使用的简称形式可能不同，需要结合推送方名称对机构名称进行配置并予以区分。

（2）基金配置

对于资助基金配置，需要配置其中英文名称、替代名称，同时还需要配置基金号正则表达式。在匹配过程中，对于没有将基金信息单独列出的文章，可以使用该表达式直接匹配出基金号供后续使用。

6 结论

为建设支持从多个出版机构接收论文并转发到多个机构的 iSwitch 系统，结合系统技术流程和标准需求，笔者将 iSwitch 系统划分为三个功能模块并进行实现，包括数据接收、数据解析、数据分发。本文主要对数据接收、数据解析两个模块的实现进行了详细介绍。

数据接收的主要难点在于使用合适的标准协议对异构数据统一接收保存。FTP 协议是一种比较推荐的数据接收和推送协议，其在双方数据传输过程中安全性比较高。SWORD 虽然设计为一种标准的知识库互操作协议，但是其存在交互过程没有批次校验信息、大尺寸文件传输效率低等问题，应用在本系统中需要重新设计封装包的信息，二次开发的工作也比较多。在数据解析模块中克服了地址格式多样、名称形式多样等问题，这也可为其他机构识别解析工作提供借鉴。

参考文献:

[1] 张晓林, 梁娜, 钱力, 等. 开放获取论文推送转发服务系统 iSwitch：概念、功能与基本框架[J].现代图书情报技术, 2014(10): 4-8. (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.)

[2] 梁娜, 张晓林, 钱力, 等. 开放获取论文推送转发服务系统 iSwitch：技术流程与标准[J]. 现代图书情报技术, 2014(10): 9-13. (Liang Na, Zhang Xiaolin, Qian Li, et al. Router Service Engine iSwitch for Open Access Articles: Technical Workflows and Standards[J]. New Technology of Library and Information Service, 2014(10): 9-13.)

[3] File Submission Specifications: Delivery[EB/OL]. [2014-08-25]. https://www.ncbi.nlm.nih.gov/pmc/pub/filespec-delivery/.

[4] A Brief History of SWORD[EB/OL]. [2014-08-25]. http://swordapp.org/about/a-brief-history/.

[5] SWORD V1 Clients and Demonstrator Repositories[EB/OL]. [2014-08-25]. http://swordapp.org/sword-v1/sword-v1-clients-and-demonstrator-repositories/.

[6] DSpace[EB/OL]. [2014-08-25]. http://www.dspace.org/.

[7] Fedora Repository[EB/OL]. [2014-08-25]. http://www.fedora-commons.org/.
[8] EPrints[EB/OL]. [2014-08-25]. http://www.eprints.org/.
[9] Automated Article-Deposit[EB/OL]. [2014-08-25]. http://www.biomedcentral.com/libraries/aad.
[10] SWORD V1 Downloads[EB/OL]. [2014-08-25]. http://swordapp.org/sword-v1/sword-v1-downloads/.
[11] SWORD V2 Implementations[EB/OL]. [2014-08-25]. http://swordapp.org/sword-v2/sword-v2-implementations/.
[12] Simeon W, Thorsten S. SWORD V1 Case Study--arXiv[EB/OL]. [2014-08-25]. http://swordapp.org/sword-v1/sword-v1-case-studies/sword-v1-case-study-arxiv/.
[13] Apache Commons Net[EB/OL]. [2014-08-25]. http://commons.apache.org/proper/commons-net/index.html.
[14] SWORD[EB/OL]. [2014-08-25]. http://sourceforge.net/projects/sword-app/.
[15] DSpace SWORD[EB/OL]. [2014-08-25]. https://github.com/DSpace/DSpace/blob/master/dspace-sword/src/main/java/org/dspace/sword/DSpaceSWORDServer.java.
[16] JATS, Journal Article Tag Suite [EB/OL]. [2014-08-25]. http://jats.nlm.nih.gov/.

作者贡献：

师洪波：iSwitch 试验系统需求调研及论文接收及解析部分的开发，撰写论文；
钱力：iSwitch 试验系统需求调研及论文分发部分的开发，撰写论文；
张晓林：提出和梳理技术流程及技术要求，撰写和审核论文；
梁娜：梳理提出技术流程和技术要求，提出参考技术标准。

（通讯作者：师洪波，ORCID：0000-0002-6450-4115，E-mail:shihb@mail.las.ac.cn。）