

开放获取论文推送转发服务系统 iSwitch: 技术流程与标准*

梁娜 张晓林 钱力 师洪波

(中国科学院文献情报中心 北京 100190)

摘要:【目的】为开放获取论文推送转发服务系统 iSwitch 提出具体的技术要求。【方法】根据推送方、转发方和接收方的需求进行 workflow 分析、重点环节技术需求分析和可参考标准规范分析。【结果】提出 iSwitch 技术流程及其技术功能要求, 提出从出版社到 iSwitch 的必备推送信息、从 iSwitch 到接收方的必备转发信息以及应采用的标准规范。在此基础上, 提出对出版社的推送服务要求和对系统开发的具体技术要求。【结论】iSwitch 推送转发流程中论文及其元数据的描述、封装、传输等应遵循相应的标准规范。

关键词: 开放获取 推送转发 作者与机构解析 传输管理 标准规范 iSwitch

分类号: G250.7

1 iSwitch 背景

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, 笔者在《开放获取论文推送转发服务系统 iSwitch: 概念、功能与基本框架》中提出开放论文推送转发服务系统 iSwitch, 支持将公共资助项目发表的论文及时完全地存储到论文作者所属机构的知识库和资助机构的知识库中。iSwitch 接收相关出版社推送的论文及其元数据, 解析和识别相应的作者、作者机构和资助机构, 映射目标机构(包括作者机构和资助机构)知识库接口, 向目标机构知识库推送相应的论文包, 进行推送审计。

为高效可靠地在多个出版社与多个机构之间进行推送转发, iSwitch 将采取业界通用的标准规范, 把复杂的 N:M 关系转化为 N:1:M 关系, 并通过部署标准化插件便于目标机构接受推送转发服务。同时, iSwitch 需要与出版社和目标机构建立合作, 建立推送转发服务协议、运行流程规范和共同接受的技术标准规范。本文在此基础上分析 iSwitch 的技术流程、关键技术要

求和可应用的标准规范。

2 iSwitch 逻辑技术流程

根据中国科学院文献情报中心的实践, 本文提出以下逻辑技术流程要求。

2.1 iSwitch 流程的对象定义

- (1) 推送方: 出版科技期刊的国内外出版社;
- (2) 作者: 符合相关资助机构或科研教育机构开放获取政策定义的、在推送方期刊发表论文的人员;
- (3) 论文: 符合相关资助机构或科研教育机构开放获取政策定义的论文及其版本, 下文简称“论文”。如果论文在订购期刊发表, 提交存储的是同行评议后录用的最终审定稿 (Author Accepted Manuscript, AAM)。如果论文本身是开放出版的, 存储的论文应是出版的 PDF 或 XML 文件;
- (4) 论文包: 由推送方推送给 iSwitch 或由 iSwitch 推送给接受方的、由一定机构范围的作者在一定时间范围内出版的论文的集合, 一般包括说明推送论文数量等信息的包和实际的论文数据集合;

收稿日期: 2014-08-13

收修改稿日期: 2014-08-22

*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。

(5) 转发方: 负责接收出版方推送的论文、并负责向作者所属机构的知识库推送作者论文的公共服务者。在本文中, 转发方为 iSwitch;

(6) 接收方: 负责从转发方接收本单位作者论文的具体机构, 包括作者机构和资助机构, 往往由其机构知识库管理部门负责管理。

2.2 iSwitch 推送转发流程

推送方与 iSwitch 系统协商确定推送服务所覆盖的期刊、论文和接收方范围后, 推送方和 iSwitch 按照以下流程进行推送服务:

(1) 推送方按照约定频率和约定的传输协议、封装格式、论文元数据格式、全文格式等, 向 iSwitch 推送指定范围的作者论文包;

(2) iSwitch 对推送方推送的论文包进行解析, 检验每次推送包的论文范围和数量, 并按照约定的差错信息编码提供必要的差错信息;

(3) iSwitch 对推送方推送的论文元数据进行解析, 检验论文元数据(必备数据项)的完整性, 解析论文的作者、作者单位和论文资助单位;

(4) iSwitch 按照规定时限向推送方提供推送审计报告。如果推送中出现差错, 推送方需重新推送存在差错的纪录;

(5) iSwitch 根据解析出的论文作者、作者机构和资助机构, 按照约定频率和约定传输协议、封装格式、论文元数据格式、全文格式等, 按符合接收方开放获取政策的作者范围分装论文包, 向接收方的机构知识库推送相应的论文包;

(6) 接收方对 iSwitch 推送的论文包进行解析, 检验每次推送包的论文范围和数量, 并按照约定的差错信息编码提供必要的差错信息;

(7) 接收方对 iSwitch 推送的论文元数据进行解析, 检验论文元数据(必备数据项)的完整性, 解析确认论文作者、作者机构和论文资助机构;

(8) 接收方按照规定时限向 iSwitch 提供推送审计报告。如果推送中出现差错, iSwitch 需重新推送存在差错的纪录;

(9) iSwitch 汇集各个接收方定期提供的接收审计报告, 并向相应的原始推送方提供汇总的接收审计信息。

2.3 推送转发中的技术要求

(1) 传输过程管理: 各个传输方(包括推送方、

iSwitch 和接收方)均需提供或接受可靠的数据包封装, 提供标准的 FTP 和 SWORD 协议传输接口, 保证可靠的传输流程控制, 提供技术流程管理支持;

(2) 数据格式支持: 各个传输方应支持 JATS 元数据格式, 支持论文 DOI, 还可支持业界通用的元数据格式, 尽可能支持规范的作者、期刊、机构唯一标识符, 尽可能支持业界通用的开放获取版本状态、开放获取状态、统计审计等元数据规范;

(3) 数据包解析检验支持: iSwitch 和接收方应对每个推送数据包进行准确解析, 检验是否收到指定的全部论文, 是否包含正确的元数据内容和全文格式, 是否能够准确和完全地解析作者、作者机构和论文资助机构, 并针对每批推送数据准确提供推送审计数据;

(4) 作者与机构的准确解析: 应保证对作者姓名的准确解析, 有效区分作者的姓和名。如有多个作者, 应保证每个作者可解析, 应尽可能提供和保留完整的作者姓名信息; 应保证对作者机构的准确解析, 如果有分别隶属于不同机构的多个作者, 或者一个作者有多个不同隶属机构, 应保证对每个作者对应的隶属机构准确解析; 应提供和保留完整的作者隶属机构名称, 并有效区分不同层次的机构名称; 应保证对资助机构的准确解析, 如果有多个资助机构, 应保证对每个作者对应的资助机构准确解析, 提供和保留完整的资助机构名称。推送方应尽可能支持 iSwitch 对作者机构和资助机构的解析, 包括但不限于提供机构名称缩略规则、机构缩略名称与全称对照等;

(5) 转发封装支持: iSwitch 应能把对推送方推送的数据包中不同机构的不同论文准确分装到相应机构的论文包中。要有效和可靠地处理同一论文有多个作者、不同作者可能隶属不同机构、同一作者可能隶属多个机构、同一论文或不同作者得到不同资助机构资助等复杂情况。

3 iSwitch 推送转发中的必备信息

为了支持对推送转发过程的管理和对论文作者、作者机构和资助机构的解析, 需要对推送转发各个阶段的必备信息做出规定。

3.1 从出版社向 iSwitch 推送阶段的信息要求

(1) 推送方论文包必备封装信息: 出版社名称、出版社标识符、论文包编号、推送时间、联系信息; iSwitch

名称、iSwitchID、iSwitch 联系信息; 论文发表时间范围、论文数量、起始论文编号、结束论文编号; 论文元数据格式和全文格式; 推送审计信息编号、出错编码、审计详细信息、发送时间;

(2) 推送论文包必备论文元数据: 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。如果允许, 可提供论文文摘、关键词, 作者 ORCID、资助机构和作者机构标识符等元数据;

(3) 推送审计必备信息: 出版社名称、标识符、论文包编号、推送时间、联系信息; iSwitch 名称、iSwitchID、iSwitch 联系信息; 推送审计信息编号、出错编码、发送时间、审计详细信息(出错数量、出错论文编号集)。

3.2 从 iSwitch 向接收方转发阶段的信息要求

(1) 转发论文包必备封装信息: iSwitch 名称、iSwitchID、iSwitch 联系信息、论文包编号、推送时间; 接收方名称、接收方 ID、接收方联系信息; 论文发表时间范围、论文数量、起始论文编号、结束论文编号; 论文元数据格式和全文格式; 推送审计信息编号、出错编码、审计详细信息、发送时间;

(2) 转发论文包必备论文元数据: 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。在出版社提供信息的情况下, 应提供论文文摘、关键词、作者 ORCID、资助机构和作者机构标识符;

(3) 转发审计必备信息: iSwitch 名称、iSwitchID、iSwitch 联系信息、论文包编号、推送时间; 接收方名称、接收方 ID、接收方联系信息; 推送审计信息编号、出错编码、审计详细信息、发送时间。

4 iSwitch 技术流程中的标准规范

在前述技术流程和要求中, 需要以不同方式采用多种标准规范。

4.1 SWORD 传输协议

SWORD(Simple Web-service Offering Repository Deposit)协议^[1]是英国 JISC 开发的支持向机构知识库批量存缴论文的传输协议, 可用于作者或出版社向机构知识库以及机构知识库之间推送批量内容对象。SWORD 在 Atom Publishing Protocol 基础上, 建立能够向指定的一个或多个机构知识库或机构知识库群推送指定的包含多个文件的资源集, 能够接受直接存缴或代理存缴的资源集。

SWORD 协议可通过一致方式接收不同的数字对象类型, 例如学术出版物、学习对象、数据、图像等, 可以用复杂内容对象或内容包的形式接收数据, 支持多种工作流的存缴。SWORD 可以完整地接收、传送完整的论文内容, 包括元数据、全文以及其他相关文档(比如论文的图表、图片)等。目前多种机构知识库平台已支持 SWORD 协议, 比如 DSpace、EPrints、Fedora 与 IntraLibrary 等。

4.2 JATS 期刊论文元数据格式

美国国家信息标准化组织 NISO 发布期刊文章标注元素集 JATS^[2], 通过一系列 XML Tags 描述期刊和论文。总体上, JATS 包括: 论文基本元数据(论文头标信息, Article Header Information, 必备), 包括论文标题、期刊名、发表日期、版权声明等, 由论文级(Article-level)和期刊卷期级(Issue-level)元数据构成; 论文内容元数据(Body, 可选), 包括关于论文章节结构的描述信息, 每章、节可能包含图、表、旁述栏等。论文附录元数据(Back Matter, 可选), 包括对参考文献、附录数据等信息的描述; 漂浮资料(Floating Material, 可选), 出版者可以把论文的图、表、附录等放到一个专门的漂浮资料包中。JATS 定义了三种标注集合: 期刊存档与交换标注集(Journal Archiving and Interchange Tag Set)、期刊出版标注集(Journal Publishing Tag Set)和论文编辑标注集(Article Authoring Tag Set)。笔者要求出版社提供期刊存档与交换标注集中的必备元数据内容。

4.3 其他期刊论文元数据元素

英国 RIOXX 项目提出开放获取知识库元数据指南^[3], 在 Dublin Core 核心元数据上, 增加了 rioxxterms.projectid 和 rioxxterms.funder 两个核心元素, 前者使用资助机构资助项目号, 后者从 RIOXX 项目资助机构

名称表中选用规范机构名称,支持资助机构发现和调用受资助项目的开放获取论文。同时 RIOXX 增加了强制元素 `rioxterms: author`,表示文章的作者必须包含作者的唯一标识符;可选元素 `rioxterms: contributor`,表示文章的贡献者必须包含贡献者的唯一标识;强制元素 `rioxterms: version`,表示论文版本,沿用 NISO/ALPSP 的期刊论文版本(JAV)推荐标准^[4],例如 AO(作者原稿)、SMUR(评审中稿件)、AM(被录用稿件)、VoR(出版稿)等。建议在元数据中吸纳以上元素。

4.4 唯一标识符

论文唯一标识符采用 DOI(Digital Object Identifier)^[5]。作者唯一标识符包括 ORCID^[6]和 ResearcherID^[7]等,是业界解决学术交流中作者姓名歧义问题的举措,作者可通过 ORCID 开放注册平台获得 16 位的 ORCID 标识码,记录个人研究记录;ResearcherID 现也与 ORCID 兼容。超过 300 家出版社、资助机构等都加入了 ORCID。资助机构唯一标识符 FundRef^[8]是学术成果资助机构和资助项目的规范化信息,出版社可把 FundRef 中规范信息整合到论文元数据中的资助出版投稿系统,作者在投稿时可选择资助机构名称、项目和编号。笔者要求出版社在论文元数据中嵌入论文的 DOI 和 URL;建议出版社在论文元数据中增加 ORCID 或 ResearcherID,增加资助机构的 FundRef 规范描述。另外,笔者拟积极探索用国内组织机构代码来标识国内研究机构,这是对国内境内依法注册与登记的机关、企事业单位、社会团体和民办非企业单位等颁发的唯一标识,按照《全国组织机构代码编制规则》(GB11714)编制。

4.5 开放获取状态规范元素

英国 RIOXX 项目已建议在 `rights` 元素中使用 Creative Commons 许可名称描述论文的使用许可,其 RIOXX 2.0 beta1 版通过`<license_ref>`强制元素对论文开放时滞期进行描述,该元素通过 `start_date` 属性表示论文的开放许可生效日期。英国 V4OA^[9]项目致力于建立一套规范的元数据语言,描述开放获取状态、开放时滞期、使用授权等信息,支持机构知识库和第三方系统准确了解机构知识库内容的可获取状态。美国 NISO 的开放获取元数据和指标集项目(NISO OA Metadata and Indicators)^[10]也试图提供关于开放状态与

使用许可的规范描述,指明一个特定的作品是否是开放获取的,并且哪些复用权利是可用的。此推荐实践草案提议采用元数据和相关标签的两个元素:`<free_to_read>`和`<license_ref>`,前者指明在指定时间范围内(如果适用的话)作品是免费获取的,后者包含一个 HTTP URI,指向许可条款。建议尽可能在论文元数据中使用这些元素。

4.6 推送审计与使用统计规范

使用统计(Usage Statistics)是衡量开放机构知识库的使用状况和影响力的重要手段。比较有代表性的规范包括 COUNTER、SUSHI、PIRUS、OA Statistics 等。COUNTER (Counting Online Usage of NeTworked Electronic Resources)^[11]是由图书馆、出版商等共同建立的记录在线资源使用情况的规范,包括对文章级别的使用统计和对期刊与数据库的使用统计。SUSHI (Standardized Usage Statistics Harvesting Initiative)^[12]标准化电子资源使用统计获取标准,是为获取 COUNTER 格式的电子资源使用统计数据而制订的标准,目前是 NISO 的正式标准(ANSI/NISO Z39.93-2013),定义了系统间传输使用数据须基于 XML 格式、遵循 SOAP 协议,为收集不同出版商提供的数据进行处理和分析提供了支持。另外,英国支持 Publisher & Institutional Repository Usage Statistics-IRUS 项目^[13],研究利用 COUNTER 指标统计机构知识库内容的使用情况,支持机构知识库内容使用和影响的可靠评价。欧盟 Knowledge Exchange 项目也提出使用统计指南(Guidelines to Facilitate the Exchange of Usage Statistics)^[14]。笔者建议出版社、iSwitch 和目标知识库尝试使用其中的规范进行未来的使用统计信息交换。

参考文献:

- [1] About SWORD (Simple Web-service Offering Repository Deposit) [EB/OL]. [2014-08-12]. <http://swordapp.org/about/>.
- [2] ANSI/NISO Z39.96-2012 JATS: Journal Article Tag Suite [EB/OL]. [2014-08-12]. http://www.niso.org/apps/group_public/project/details.php?project_id=93.
- [3] The RIOXX Metadata Profile and Guidelines: Application Profile Version 2.0 beta 1 [EB/OL]. [2014-08-12]. <http://docs.riox.net/v2-0-beta-1/>.
- [4] Journal Article Versions (JAV): Recommendations of the

- NISO/ALPSP JAV [EB/OL]. [2014-08-12]. <http://docs.rioxnet.net/v2-0-beta-1/>.
- [5] The DOI(Digital Object Identifier) System[EB/OL]. [2014-08-12]. <http://www.doi.org/>.
- [6] ORCID[EB/OL]. [2014-08-12]. <http://orcid.org/>.
- [7] ResearcherID [EB/OL]. [2014-08-12]. <http://www.researcherid.com/>.
- [8] FundRef [EB/OL]. [2014-08-12]. <http://www.crossref.org/fundref/>.
- [9] V4OA[EB/OL]. [2014-08-12]. <http://v4oa.net/about/>.
- [10] A Proposed NISO Work Item: Specification for Open Access Metadata and Indicators [EB/OL]. [2014-08-12]. http://www.niso.org/apps/group_public/download.php/9845/Open%20Access%20Metadata%20-%20Work%20Item%20for%20ballot.pdf.
- [11] COUNTER: Counting Online Usage of Networked Electronic Resources [EB/OL]. [2014-08-12]. <http://www.projectcounter.org/about.html>.
- [12] The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol [EB/OL]. [2014-08-12]. http://www.niso.org/apps/group_public/download.php/10253/Z39-93-2013_SUSHI.pdf.
- [13] Shepherd P, Needham P. Publisher and Institutional Repository Usage Statistics: The PIRUS2 Project—Final Report. October 2011[EB/OL]. [2014-08-12]. <http://www.jisc.ac.uk/publications/reports/2009/pirusfinalreport.aspx>.
- [14] Guidelines to Facilitate the Exchange of Usage Statistics [EB/OL]. [2014-08-12]. <http://knowledge-exchange.info/Default.aspx?ID=395>.

作者贡献声明:

梁娜: 提出技术流程和技术要求, 提出参考技术标准, 撰写论文;
张晓林: 参与提出和梳理技术流程及技术要求, 撰写和审核论文;
钱力, 师洪波: 参与梳理技术流程及其技术要求, 承担 iSwitch 系统技术开发。

(通讯作者: 梁娜 E-mail: liangna@mail.las.ac.cn)

Router Service Engine iSwitch for Open Access Articles: Technical Workflows and Standards

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

Abstract: [Objective] Provide workflow and standards description for router service engine iSwitch of open access articles. [Methods] Analyze the workflows, key technical needs and applicable standards according to the requirements of publishers, the router and the receiving organizations during the push and routing process. [Results] Describe the iSwitch tech workflows, required information items from the publisher to iSwitch and from iSwitch to the recipient, and suggest the standards to be used. [Conclusions] Open access articles and its metadata description, packaging, transmission should follow the corresponding standards in iSwitch workflow.

Keywords: Open access Push and routing Resolving authors and institutions Transmission management Standards iSwitch