

资源同步框架ResourceSync及其在机构知识库中的应用场景*

王超

(辽宁工业大学图书馆, 锦州 121000)

摘要: 资源同步框架规范ResourceSync可实现互联网上不同系统之间的资源同步。本研究通过对ResourceSync的框架结构、同步过程、资源发现机制的分析, 阐述其用于资源同步的基本原理, 描述其在机构知识库中的应用场景及应注意的问题及对策。该框架规范在机构知识库中的应用可实现资源转移、批量发现、长期保存等新功能。机构知识库可通过ResourceSync框架规范执行与镜像库、用户、资源提供者之间的资源同步操作。

关键词: 机构知识库; 资源同步; 资源描述; ResourceSync

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2021.06.006

引文格式: 王超. 资源同步框架ResourceSync及其在机构知识库中的应用场景[J]. 数字图书馆论坛, 2021(6): 36-41.

2017年11月, 开放获取知识库联盟(COAR)经过长期调研发布报告^[1], 描述了新一代机构知识库(IR)应具有11项新功能, 并称新一代机构知识库的资源转移(resource transfer)、批量发现(batch discovery)、收集和公开活动(collecting and exposing activities)、长期保存(preserving resources)5项新功能主要依托于资源同步框架规范ResourceSync实现。

ResourceSync是由美国国家标准协会(ANSI)指导研发, 并于2014年颁布。2017年更新的资源同步框架规范^[2], 可用于实现互联网上不同系统之间的资源同步。郭少友等^[3]和曹迪^[4]对于ResourceSync在数字图书馆中的应用场景进行了探索, 其中部分应用场景同样适应于机构知识库的资源同步, 但ResourceSync在数字图书馆中的应用侧重于资源的收集和保存, 而机构知识库更侧重于资源的开放和共享。笔者通过相关调研并结合现有研究, 分析ResourceSync资源同步规范的基本原理, 并探讨其在机构知识库中应用场景及应注意的问题及对策, 使机构知识库功能更强大, 最大程度地实现资源开放与共享。

1 数字资源同步现行方法分析及ResourceSync的应用

笔者基于国内外学者郭少友等^[3]、曹迪^[4]、刘树等^[5]、Zhen^[6]、Haslhofer等^[7]、Klein等^[8]的研究并结合相关调研, 认为目前数字资源常用的同步方法可归纳为软件同步、中间件同步、发布变化通知同步3类。

1.1 软件同步

该方法的核心是将同步软件分别安装于源数据库和目标数据库, 资源变化时使用软件实现资源同步。此方法要求源数据库和目标数据库是完全相同的网络系统, 而且同步软件需要根据数据库情况自主研发。其优势在于源数据库和目标数据库不分主从、互相同步, 但软件对环境要求较高, 而且需要自主研发或修改, 通用性不强。国内此类同步软件较多, 比较有代表性的有刘树等^[5]研发的基于触发器的变化捕获器。

* 本研究得到2018年辽宁省社会科学规划基金图书馆、情报与文献学青年项目“高校智库助力辽西两地创建国家级高新区研究”(编号: L18CTQ001)资助。

1.2 中间件同步

中间件是一种安装于数据库系统的应用程序或者软件,用于数据中转站管理、传输计算机资源、实现网络通信。源数据库和目标数据库可通过中间件共享资源,当源数据库发生资源变化时,向中间件发送变化信息,目标数据库获取变化信息后进行解析和转储。Zhen^[6]使用JMS中间件程序,用4个步骤来实现不同数据库之间的数据同步:①通过触发器获取源端的增量数据;②增量数据转换为XML文件;③以消息格式发送XML文件到目标端;④使用XML解析器解析数据。利用中间件同步的优点是可以在不同网络系统之间共享资源,且中间件程序较为简单、开发难度低,还能实现一对多的资源同步。其缺点在于使用中转站传输数据,步骤较多、过程复杂,同时中间件需要使用专有的API和协议,不同中间件互操作困难。

1.3 发布变化通知同步

此类方法把资源提供者称为源端,目的地数据库称为目标端,源端通过Sitemap协议发布变化通知(XML格式文件),目标端获知变化信息,再通过某种规范执行资源同步操作。DSNotify的资源同步方法和ResourceSync资源同步规范均可归为发布变化通知同步的方法。维也纳大学Haslhofer等^[7]提出一种基于DSNotify的资源同步方法,也称资源断链检测与修复。从源端所有资源中抽取描述性特征,生成资源特征向量,通过DSNotify附加组件提取并导出资源的特征项(资源的URI及其特征向量),再由监视器通过反复比对监控本地资源变化情况,监视器将检测到的变化情况写入事件日志,并发送通知到目标端,目标端据此修复资源断链。DSNotify方法应用的可行性已经被网络同步仿真实验证明,但DSNotify应用仍有一定局限性,其适合处理关联数据,主要用于链接的同步,无法同步非文本资源。

1.4 ResourceSync的优越性及应用

ANSI颁布的ResourceSync资源同步框架规范,同时提供了两款开源软件——resync-simulator和resync,前者是资源同步信息发布工具,后者是资源同步工具。ResourceSync把元数据和元数据描述的对象

数据视为资源,赋予它们不同的URI,使用Sitemap协议语法对元数据和对象数据进行描述。ResourceSync把内容提供者称为源端,把打算复制源内容的站点称为目标端。源端对资源进行描述并对外发布能力列表,其中包括4种同步能力:资源列表、变化列表、资源转储库、变化转储库,目标端获取能力列表后按规范步骤完成资源同步。相较于现行的数字资源同步方法,ResourceSync具有如下优越性:与软件同步方法相比,它不但省略了软件研发,而且具有4种同步能力,即环境适用性和通用性更强;与中间件同步方法相比,它同样具有在不同网络系统之间实现一对多的同步能力,除此之外其源端和目标端使用Sitemap协议直接建立联系,不需要中间件,过程简单、不易出错,还可减少延迟;与DSNotify相比,ResourceSync同步规范具有更强的通用性,DSNotify无法同步非文本资源,而ResourceSync适用于任何具有URI的资源之间的同步。

自2014年颁布以来,ResourceSync便因其优越性受到广泛重视。ResourceSync已经在康奈尔大学图书馆和电子预印本数据库arXiv中实现部分功能的成功应用,并受到欧洲数字人文项目(Europeana)、美国数字公共图书馆(DPLA)等的关注。截至2020年底,DPLA的部分成员库之间已经采用ResourceSync实现数字资源互操作^[9]。COAR工作组也将ResourceSync作为新一代机构知识库的核心技术、标准、协议之一。

现代机构知识库不仅收录了各种类型的海量数字资源,而且其资源是高度动态化的,不断地收录新资源,更新或删除旧的数据。ResourceSync适用于任何具有URI的资源之间的同步,而且采用模块化设计,灵活性强,易于操作,具有广泛的适用性。根据COAR报告的描述,ResourceSync主要应用于3个方面。①资源长期保存。采用镜像同步备份等措施,为数据的长期保存提供保障。②批量发现和获取资源。在ResourceSync框架下,用户可实现资源的批量发现和打包下载,提升机构知识库资源价值。③资源转移。“ResourceSync是一种基于站点地图(Sitemap)的规范,存储库管理器可以使用该规范提供信息,允许第三方系统持续与其存储库中的资源保持同步,即创建、更新和删除。站点地图允许公开知识库内容和搜索引擎所需的元数据。ResourceSync可使用Sitemaps XML格式实现内容和元数据的发现和同步”^[10]。

2 ResourceSync资源同步的基本原理

2.1 ResourceSync框架结构

ResourceSync资源同步框架如图1所示。源描述是目标端了解数据源情况,实现同步的接入点。源描述枚举数据源的能力列表,每个资源集都指向一个能力列表。如只有单一的源集合,则同步描述只包含一个指针;对于多个数据源(多个源使用索引的方法)的情况,其描述包括源描述和能力列表两个层次。能力列表列举了一组资源支持的4种能力:资源列表、变化列表、资源转储库、变化转储库的URI。

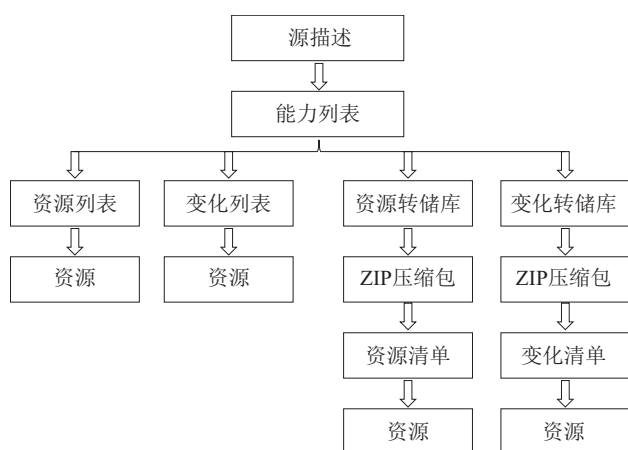


图1 ResourceSync资源同步框架

ResourceSync框架规范把数字资源同步定义为4种情况:资源列表、变化列表、资源转储库、变化转储库,分别与4种同步能力相对应。

(1) 资源列表。ResourceSync框架规范将源端某一时刻(列表中时间戳属性)允许同步的资源以列表呈现,使用<url>标签将其URI逐一列出。对于每个资源,资源列表最低限度地提供资源的URI,再添加<lastmod>和<rs:md hash="标签,显示最新修改时间和哈希(散列)及其长度信息,用于校验同步是否成功。

(2) 变化列表。当源数据发生变化时,Resource Sync发布变化列表,它与资源列表不同,变化列表添加两个时间戳以指明变化资源覆盖的时间间隔。例如,一个“变化列表”可以描述一天、一个小时或某个规定时间段的变化。变化资源的URI使用<url>标签逐一标出,与资源列表类似,可以传达3种资源更改(更新、删除、创建),每种资源变化类型使用一个<url>文档列出。目标端获取变化列表,通过其URI获取已创建/更新的资源,

删除已删除的资源。如果变化资源较多,则按时间升序排列。此外,ResourceSync还提供变化列表存档功能,方便目标端查看变化列表的历史版本。

(3) 资源转储库。如果目标端根据资源列表URI逐条获取资源,少量可行,但不适用于大型资源列表。为此,ResourceSync引入了打包同步的功能,为使用者提供资源转储库。资源转储库是资源包(URI)的列表,格式为一个或者多个Sitemap文件ZIP压缩包。目标端获取资源转储库,通过URI获取资源包,解包获得同步资源的比特流信息以及资源清单,清单描述了每个资源的URI、解压路径、最后修改时间等常用信息。

(4) 变化转储库。ResourceSync引入了变化转储库功能,为使用者打包同步变化资源。和资源转储库类似,变化转储库同样是一个或者多个Sitemap文件ZIP压缩包(包含指向ZIP压缩包的链接),其中包含变化资源的比特流信息以及描述变化资源的清单。此外,Sitemap文件容纳资源数量最多为5万个,如果资源清单或者变化清单超过规定数量,可通过ResourceSync的索引机制分别组织,在每个资源清单和变化清单的URI索引中予以指明。同理,源端也可建立相应的资源转储库索引和变化转储库索引。

2.2 ResourceSync资源同步过程

ResourceSync资源同步过程(见图2)是由目标端通过“拉”或“推”获取源端资源及其变化信息,执行基准同步、增量同步、审校,实现源端与目标端的同步。

	基准同步	增量同步	审校
•URI •元数据 常用信息 链接	资源列表	变化列表	资源列表 变化列表 常用信息
•URI •比特流 •元数据 常用信息 链接	资源转储库	变化转储库	资源列表 变化列表 常用信息

图2 ResourceSync资源同步过程

(1) 基准同步。基准同步是指目标端与源端首次同步,目标端复制源端基础数据。按照ResourceSync规范,可通过资源列表和资源转储库获取源端基础资源,

实现基准同步。第一，目标端通过资源列表中资源的URIs，逐一发出HTTP请求，完成基准同步。第二，目标端还可以获取源端发布的资源转储库ZIP文件中的比特流信息，在资源清单的指导下解引内容包。

(2) 增量同步。增量同步是指在基准同步的基础上，通过重复执行增量同步，目标端可以与源端保持动态同步。源端发布变化列表传达关于资源变化的信息，目标端通过变化列表中列出的新创建或更新资源的URI获取最新的内容，同时同步删除已删除资源。目标端还可以获取源端的变化转储库实现增量同步，和基准同步类似，目标端通过变化资源的URI获取数据包，然后在变化清单的指导下解包。

(3) 审校。为了验证同步效果，目标端必须能够检查它获得的内容是否与源端当前资源在覆盖率和准确性方面相匹配。审校时，目标端将获取资源（数据包）包含的元数据与源端提供资源的元数据进行匹配验证，这些元数据描述了资源的最新状态，如上次修改时间、长度和基于内容的散列及其长度。

2.3 ResourceSync资源发现机制

在实际操作中，目标端如何发现源端发布的能力列表是实现资源同步的前提。在ResourceSync框架规范下，资源发现主要有：well-known URI、Link链接、robots协议3种方式（见图3）。

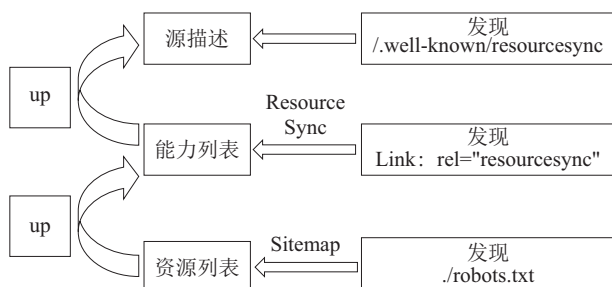


图3 ResourceSync资源发现机制

(1) well-known URI方法。一个数据源要发布源描述，源描述的URI: <http://example.com.well-known/resourcesync>。well-known URI资源发现方法就是在源描述URI添加前缀“/well-known/”，添加前缀的URI: <http://www.example.com/well-known/resourcesync>。此特殊URI是目标端获取源端能力列表的适当入口点。

(2) Link链接方法。目标端对于web（网页）类数字资源的同步，一般采用Link方法。在HTML文档顶部

(`<head>`)添加`<link>`元素，此`<链接>`必有一个`rel`属性值为`resourcesync`，这表明可以通过`<href>`的值URI可以发现源端的能力列表。由于Sitemap协议不提供目标端从具有根元素文档直接导航到其“父”文档的方法，所以可通过`<rs:ln>`标签`up`（向上）获取源描述。

(3) robots协议方法。该方法是首先向一个源端的`robots.txt`文件中添加一个Sitemap指令（Sitemap: <http://example.com/dataset1/resourcelist.xml>），“通过<http://www.example.com/robots.txt>，可获取源数据端的`robots.txt`文件，该文件的值是资源列表的URI，目标端据此发现源端的资源列表，进而获取能力列表。”^[4]如果一个源端支持多个资源集，那么可以添加多个指令，每个指令对应一个与特定资源集相关联的资源列表。

3 基于ResourceSync规范机构知识库资源同步场景

国内现有的机构知识库大多采用的是DSpace开源软件作为系统平台，利用OAI进行数据收割，然后导入中间元数据池，再通过数据接口或数据关联等数据处理工具实现不同合集或不同区域资源的整合和分享，完成源数据的更新或同步。这种资源同步由于使用中转站传输数据，不同中间件互操作较困难，而且操作步骤较多、过程复杂，容易出现数据迟滞或混乱。

ResourceSync框架规范由于环境适用性和通用性较强，既可批量“收割”元数据，也可实现对象资源的批量互操作，且过程简单，将成为下一代机构知识库资源同步的技术支撑。ResourceSync框架规范在机构知识库中实现资源同步主要有以下3个应用场景。

3.1 机构知识库与镜像库的同步方法

学术机构可以根据自身情况为机构知识库设立一个或多个镜像站点（备份库），机构知识库与其镜像库进行备份同步。这既有利于机构知识库资源的长期保存，也具有一定的分流作用，可减轻主服务器压力。ResourceSync框架规范在资源发现机制上，由于双方的隶属关系，作为目标端的镜像站点无论采用何种发现方法都可轻易获取作为源端的机构知识库资源。在同步方法上，镜像站点可略过资源列表，直接获取机构知识库资源转储库数据包，完成基准同步。增量同步和

基准同步类似,当机构知识库资源发生变化(更新、删除、创建)时,直接获取变化转储库,在资源清单的指导下完成增量同步。由于镜像数据作为机构知识库备份,同步数据准确性和完整性至关重要,同步操作完成后,再根据清单常用信息对所有同步资源进行审校。

3.2 机构知识库与用户之间的同步方法

ResourceSync框架规范在机构知识库中的第二个应用场景是通过机构知识库与用户之间的同步,帮助用户精准发现、批量获取机构知识库资源,提升用户的使用体验。机构知识库用户又分为个人用户和机构用户两类。首先,机构知识库按照ResourceSync框架规范对自建资源进行源描述,为了兼顾不同的用户,机构知识库为用户提供全部同步能力,用户根据需要选择同步能力。其次,用户可根据需求选择资源发现方法。ResourceSync提供了3种资源发现机制,用户可自主选择。如用户只需获取机构知识库资源列表,那么选用robots协议方法即可实现;若需要4种同步能力,则选用well-known URI或Link链接方法。在同步过程中,个人用户一般对于资源的覆盖率和准确率要求较低,只需执行基准同步和增量同步,审校可略过,而机构用户对于同步资源的准确性和完整性要求较高,因而要严格按照步骤执行同步,审校环节不可省略。

3.3 机构知识库与资源提供者之间的同步方法

近年来机构知识库从机构下辖科研院(所)或团队数据库获取资源正成为扩大机构知识库资源持有量的主要途径。纵观国内外知名机构知识库(国外如麻省理工图书馆机构知识库等,国内如中国科学院系统机构知识库等)都采用ResourceSync框架规范与下辖的研究院(所)数据库保持资源同步,这样既可扩大机构知识库资源持有量,也能保障资源实时更新。

尽管机构知识库与机构下辖的研究院(所)数据库具有隶属或者合作关系,但与镜像站点数据备份不同,机构知识库从子数据库中获取资源要遵循既定目标选择性地同步。因此,在资源发现机制上,双方可采用约定的发现方法。由于是选择性同步,同步过程也较为简单,作为目标端的机构知识库可略过子库的资源转储库和变化转储库,只需获取子库的资源列表和变化列

表,根据资源的URI下载元数据或全文数据,即可完成基准同步。当子库资源发生变化时,与基准同步方法相同,通过获取变化列表,完成增量同步。最后再根据获取资源的常用信息(最新修改时间和散列及其长度信息)验证同步效果。

4 机构知识库资源同步中应注意的问题及对策

机构知识库运行是复杂的“生态系统”,ResourceSync资源同步框架规范在实际应用中可能出现以下若干问题。

4.1 资源同步的实时性问题及对策

在机构知识库资源同步实践中如何减少延迟,保障同步的实时性,是无法回避的现实问题。在ResourceSync同步框架规范中,为保障同步的实时性,在实际操作中应采用“推”“拉”结合的方式。当机构知识库资源发生变化时,应及时发布(推送)变化通知,并提供变化列表或变化转储库,以便目标端及时同步更新。同时,目标端应动态监测源端资源变化情况,及时发送同步请求,根据比特流信息判断资源变化情况,量少则通过变化列表完成同步,量大则打包下载变化转储库。

4.2 资源同步过程中数据格式转换问题及对策

当两个网络系统之间要实现大量同步资源,在数据传输的过程中极易出现格式不一致的数据异构性问题。导致数据格式异构的原因很多,主要源于计算机系统本身的复杂性。如何有效解决此类问题是实现资源同步的关键。在机构知识库与镜像库同步中,“IR数据格式应尝试应用可重复使用的Latex或TEI格式”^[10],镜像备份过程中尽量避免数据格式转换,直接传输数据保存备份。在机构知识库与用户或资源提供者同步中,IR与资源使用者或提供者之间资源同步时,目标端获取源端资源后,先使用ETL(抽取-转换-加载)工具进行格式转换,数据转换成功之后,再执行基准同步、增量同步以及审校。

4.3 元数据与对象数据的协调同步问题及对策

现代机构知识库不仅收录海量的数字资源,而且资源类型十分丰富。ResourceSync同步框架规范将元数据和对象数据分别赋予不同URI,两者之间是描述与被描述的关系,如果元数据发生变化,其描述的对象数据也随之变化,反之亦然。因此,在同步过程中可能出现两类资源不协同的问题,为了避免此类问题,“应在继续保持二者之间描述与被描述关系的基础上,分别对元数据资源及其所描述的对象资源间实施同步”^[3]。

5 结语

机构知识库的研究和应用已受到广泛的重视和获得长足的发展,而ResourceSync资源同步框架规范在机构知识库中的应用研究才初见端倪,利用ResourceSync同步原理,研究其在机构知识库中具体应用场景,提出应用过程中需要注意的问题并给出相应对策,才能更好地实现机构知识库的资源转移、批量发现、长期保存等新功能,使机构知识库功能更强大,最大程度地实现资源开放和共享。

参考文献

[1] Next generation repositories [EB/OL]. [2021-01-15]. [https://www.](https://www.coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/)

coar-repositories.org/activities/advocacy-leadership/working-group-next-generation-repositories/.

- [2] ResourceSync framework specification (ANSI/NISOZ39.99-2014) [EB/OL]. [2021-02-11]. <http://www.openarchives.org/rs/1.0/resourcesync>.
- [3] 郭少友,曹迪,简丹,等.资源同步框架ResourceSync及其在数字图书馆中的应用场景[J].图书情报工作,2016,60(22):109-115.
- [4] 曹迪.基于ResourceSync的数字图书馆资源同步研究[D].郑州:郑州大学,2018.
- [5] 刘树,徐硕,黄其泉,等.一种基于触发器的数据同步系统的研究及关键技术实现[J].计算机应用与软件,2012(12):189-191.
- [6] ZHEN Y. A Method of Data Synchronous Updating Based on Message Oriented Middleware [J]. Ordnance Industry Automation, 2013, 32(7): 93-96.
- [7] HASLHOFER B, POPITSCH N. DSNotify-Detecting and Fixing Broken Links in Linked Data Sets [C] //International Workshop on Database & Expert Systems Application. IEEE, 2009.
- [8] MKLEIN M, VAN DE S. Extending sitemaps for ResourceSync [P]. 2013.
- [9] ISMAIL L. Digital Public Library of America (DPLA) [J]. The Charleston Advisor, 2017, 19(2): 25-30.
- [10] 崔海媛,孙超,罗鹏程.新一代机构知识库的关键技术和发展趋势研究[J].图书情报工作,2019(19):96-104.

作者简介

王超,男,1983年生,硕士,馆员,研究方向:图书馆数字资源建设,E-mail: 309729169@qq.com。

Resource Synchronization Framework ResourceSync and its Application Scenarios in the Institutional Repository

WANG Chao

(Library of Liaoning University of Technology, Jinzhou 121000, China)

Abstract: Resource synchronization framework ResourceSync can synchronize resources between different systems on the Internet. This study analyzes the framework of ResourceSync, the synchronization process, and the resource discovery mechanism, describes the basic principles of ResourceSync for resource synchronization, and describes the application scenarios of ResourceSync in the institutional repository, the problems that should be noticed, and the countermeasures. The application of the framework specification in institutional repository can realize new functions such as resource transfer, batch discovery and long-term storage. An institutional repository can synchronize resources with mirror libraries, users, and resource providers using the ResourceSync framework.

Keywords: Institutional Repository; Resources Synchronization; Resource Description; ResourceSync

(收稿日期: 2021-05-03)