• 医学信息组织与利用 •

面向 OAI - PMH 协议的西太平洋地区医学索引数据服务设计与实践*

范云满 方 安 王 蕾

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 介绍西太平洋地区医学索引（WPRIM）系统在实现 OAI - PMH 协议，提供数据服务过程中面临且必须解决的 3 个问题，制定 WPRIM 元数据规范、WPIRM 数据分组策略，以及基于网络带宽自适应的 resumptionToken 生成策略的基本方法，目前已在 OAI - PMH 官方网站上通过协议验证并实现数据服务注册。

〔关键词〕 西太平洋地区医学索引；OAI - PMH 协议；元数据规范；数据分组策略

〔中图分类号〕R - 056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673 - 6036.2019.08.016

Design and Practice of Data Services of Western Pacific Region Index Medicus（WPRIM）with OAI - PMH Protocol *FAN Yunman, FANG An, WANG Lei, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper introduces three problems to be solved when WPRIM system achieves OAI - PMH protocol and provides data services, that is, the formulation of WPRIM meta data specification, WPRIM data grouping strategy and basic approaches of adaptive resumptionToken generation strategy based on network bandwidth. Nowadays WPIRM has been verified on OAI - PMH official website and achieved the registration of data services.

〔**Keywords**〕 Western Pacific Region Index Medicus（WPRIM）; OAI - PMH protocol; metadata specification; data set strategy

1 引言

数字资源的不断增加，数字信息的极大丰富，导致对不同资源与组织形式的元数据统一加工、保存、跨库检索存在困难。在这种背景下，产生了基于开放文献预研（Open Archives Initiative，OAI）的元数据互操作协议（Protocol for Metadata Harvesting）（OAI - PMH 协议），这是一种独立于应用、能够提高 Web 上资源共享范围和能力的互操作协议标准，主要通过指定的命令集合，利用 Internet 和元数据技术，提供数字资源的元数据信息的互操作。随着加入开放获取（Open Access，OA）的机

〔收稿日期〕 2019 - 01 - 17

〔作者简介〕 范云满，助理研究员，发表论文 10 余篇；通讯作者：方安，副研究馆员。

〔基金项目〕 中国医学科学院医学与健康科技创新工程服务“一带一路”战略先导科研专项“卫生信息服务研究”（项目编号：2017 - I2M - B&R - 10）；国家重点研发计划 精准医学研究专项“精准医学大数据管理和共享技术平台”子课题“重大疾病精准医学数据库群”（项目编号：2016YFC0901602）；中国医学科学院医学与健康科技创新工程（协同创新团队项目）；“医学科技创新评价与卫生服务体系构建研究”（项目编号：2016 - I2M - 3 - 018）。

构不断增多，已经在 OAI 官方机构登记注册数据提供者从 2005 年 139 家[1]发展到 2018 年 12 月超过 3 680家[2]。

西太平洋地区医学索引（Western Pacific Region Index Medicus，WPRIM）面向西太平洋地区的医学期刊用户，提供收录、索引服务，依托于 WHO 西太平洋区域办事处与其成员国若干机构合作的项目。同时 WPRIM 收录索引期刊文献后，需要向全球卫生图书馆（Global Health Library，GHL）提供收录文献的元数据。WPRIM 已经为 GHL 提供两种数据提交的方式，定期按照指定格式批量生成文献 XML，通过 FTP 的方式上传到 GHL。WPRIM 以应用程序接口（Application Interface，API）的方式提供两个函数，根据期刊名称等信息获取文献 ID 列表，根据文献 ID 获取相应的元数据。虽然能根据检索条件为有检索需求的用户提供文献服务，但存在元数据信息不能自解释的问题；同时由于该 API 服务是一个 WPRIM 系统定制的服务，不是一个通用的协议标准，存在元数据的语义不能自我表达，导致与其他仓储系统的互操作性不够完全兼容。

本文主要介绍 WPRIM 系统实现 OAI 协议的过程中解决自我制定元数据标准、制定数据分组策略、Resumptiontoken 的实现机制 3 个问题，通过 OAI 协议数据提供者验证注册工具实现 WPRIM 服务的注册，验证 WPRIM 系统 OAI 数据服务工作的完成。

2 背景技术

2.1 OAI 协议

OAI 协议包含两种角色的参与者，数据提供者（Data Provider）和服务提供者（Service Provider）。数据提供者以 OAI－PMH 方式发布元数据，服务提供者（用户）以 OAI－PMH 为基础获取元数据来建立增值服务。本文工作目前实现了前者，即数据提供者的角色。

表 1 文中用到的 OAI 协议术语

术语名称	术语含义	术语名称	术语含义
仓储（repository）	提供数据服务的服务器	resumptionToken	用于对返回列表数据分页的链接字段
命令动词（verb）	OAI 协议中用户通过命令动词向数据提供者发送请求	pageSize	列表数据分页每一页数据量的大小
分组（set）	仓储中按照不同的 set 对数据进行组织，用户可以选择不同的分组的数据	metadataPrefix	OAI 协议中仓储提供元数据规范的名称

仓储依靠 6 个命令动词（verb）对外提供数据服务，见表 2。需要特别说明的是 ListMetadataFormats 列出仓储提供的元数据规范，ListRecords 和 LisIdentifier 都是列出仓储中的数据列表，但是前者列出数据的详细信息（按照某一种元数据规范），后者则只是列出数据的标识。

表 2 数据服务者提供的 6 个动词

动词	意义	动词	意义
Identify	标识仓储的有关信息	ListRecords	列出仓储中的数据列表
ListMetadataFormats	列出仓储提供的元数据规范	LisIdentifier	列出仓储中的数据标识列表
ListSets	列出仓储提供的数据分类体系	GetRecord	根据标识给出数据的详细元数据

由于 OAI－PMH 是一个互操作协议标准，不是可以直接利用的工具，具体应用可以根据协议内容各自实现，例如开放文献预研社区的成员提供了很多的版本。WPRIM 系统在实现 OAI 协议的数据服务时，主要解决自定义元数据规范并且按照自定义的元数据规范列出数据信息，以及实现分页列出数据，其中最为关键的是 resumptionToken 的实现策略。

2.2 自定义元数据规范

OAI 协议中返回一个条目的数据 xml，其中第 2 部分元数据为返回的主体部分，通常仓储选择 DC（Dublin Core）的元数据规范描述该主体，但是也可以按照仓储自定义的并且已经发布的元数据规范予以描述。DC 元数据规范以 15 个元素为核心，包括 7 个资源内容描述类元数据项、4 个知识产权描述类元数据项和 4 个外部属性描述类元数据项。DCMI 成立了多个工作组，针对不同领域的需求进行扩展，国内的很多机构也以 DC 为基础，制定了各自的元数据方案。但是 DC 元数据仍然存在元数据项受限、已有元数据项缺少能够表达复杂关联关系的能力，另外各个机构制定的元数据方案之间并不能互相通用，缺少互操作机制。

2.3 数据分组策略制定

仓储需要通过分组对内部数据组织才能更好为用户提供服务，因此需要根据 WPRIM 的数据特点设置分组。

2.4 resumptionToken 实现

2.4.1 resumptionToken 如前所述 OAI 协议通过 6 个命令动词，由数据提供者向服务提供者提供数据，其中 ListIdentifier 和 ListRecords 都有一个独特的参数 resumptionToken，用来处理不完整列表。数据提供者由于受到网络带宽、硬件资源等的限制，在提供批量元数据时每次提供的数据列表为固定大小，如 100。当服务提供者向数据提供者查询到的列表数量大于 100 时，数据提供者每次只提供 100 条数据，该服务提供者继续申请获取后面的数据时，数据提供者再提供后面的 100 条。由于数据服务者和服务提供者之间没有采取握手机制，即数据服务者并不能记住每次请求是由哪一个服务提供者发出的，更不知道该服务提供者上次提出请求时的查询条件是什么，因此需要在两者之间有一个联系的标识，称为 resumptionToken。该标识由数据提供者发出，随数据列表一起发给服务提供者，服务提供者收到该标志的时候，将该标志保存，当需要获取下一批次的数据时，将该标志发给数据提供者。如果服务提供者没有收到该标志，说明收到的数据是查询到的全部数据。

2.4.2 DSpace Hewlett Packard 和麻省理工学院于 2002 年开展的联合开发的一个开源的数字资源管理软件平台，使机构能够捕获和描述数据内容[2]。它可以在各种硬件平台上运行，支持 OAI - PMH2.0 版。由于 DSpace 作为世界上很多机构知识库建设的中间件，很多机构知识库也利用其中实现的 OAI - PMH 功能。DSpace 中的 resumptionToken 的生成机制主要有两个关键点：一是需要每次仓储中的所有数据条目加载到服务器中，二是采用 Base64 算法对每一次的查询条件及返回的数据加密生成 resumptionToken，但是前者在中大型规模的仓储中容易产生内存溢出的问题。

2.4.3 开放源码软件 ARC Old Dominion University 数字图书馆小组开发的基于 OAI 的搜索引擎，提供了 OAI 协议中服务提供者完整的实现，是一个通用的服务提供者平台[3]。ARC 的 resumptionToken 的生成机制相比 DSpace 有改进，采用 URLencoder 生成 resumptionToken。可取得根据查询条件得到的所有数据。这样可在很大程度上缓解内存溢出的问题，但是当查询条件是查询所有数据时仍然容易导致内存溢出。

2.4.4 Archimede 拉瓦尔大学图书馆开发的一个用于机构知识库的开源软件，具有全文搜索、Web 用户界面等功能，完全支持 OAI - PMH 协议 2.0 版[4]。Archimede 的 resumptionToken 生成机制进一步优化：根据查询条件查询（0 - pagesize + 1），pageSize 为每次分页数据量的大小。同样，Archimede 采用 URLencoder 生成 resumptionToken，利用 urldecoder 对 resumptionToken 解密。利用该方法已经解决了内存溢出的问题，但是 pageSize 的选择需要根据经验加以设定，不能依据数据提供者、数据收割者两者之间互通的带宽动态设定。

3 WPRIM 系统 OAI 协议实现

3.1 WPRIM 元数据规范

为了能够满足 WPRIM 系统完整描述自身的元

数据，让数据服务提供者理解 WPRIM 的元数据项，WPRIM 系统借鉴了 DC 元数据[5]、PUBMED[6]、Web of Science 的元数据方案[7]，发布了自身的元数据方案，命名空间是 http://wprim.whocc.org.cn/WPRIM_2，所有的元数据项（Term）在该空间下定义。图 1 是 WPRIM 制定元数据规范的主要片段，略去了 Title，Abstact 等简单的元数据项。本元数据规范包含的 Journal、AuthorList、KeywordList 和 Meshlist 都是组合元数据，是 DC 元数据规范无法表达的。

```
<schema
    xmlns=' http://www.w3.org/2000/10/XMLSchema'
    targetNamespace=' http://wprim.whocc.org.cn/WPRIM_2'
    xmlns:wprim=' http://wprim.whocc.org.cn/WPRIM_2' >
  <element name=' Article' >
    <complexType>
      <sequence>
        <element ref=' wprim:Journal' />
        …
        <element ref=' wprim:AuthorList' minOccurs=' 0' maxOccurs=' 1' />
        …
        <element ref=' wprim:KeywordsList' minOccurs=' 0' maxOccurs=' 1' />
        <element ref=' wprim:MeSHList' minOccurs=' 0' maxOccurs=' 1' />
      </sequence>
    </complexType>
  </element>
</schema>
```

图 1 WPRIM 元数据规范主要片段

3.2 WPRIM 分组策略

WPRIM 系统中收录的论文都是期刊论文，其元数据都是期刊论文元数据，同时 WPRIM 需要定期按月向 GHL 提供元数据。基于以上考虑采用两种分组策略：第 1 种按照期刊名称、ISSN 分组，用户可以按照期刊名称、ISSN 获得对应的论文；第 2 种按照更新时间分组，将数据的更新时间划分到每个月，用户可以按照每个月取得该月的更新数据（包含新收录的数据）。随着下一步的工作继续开展，考虑加入更多的分组类别。

3.3 WPRIM 系统 resumptionToken 的生成算法

WPRIM 以 DSpace 中的 OAI 协议模块为基础实现数据提供者的功能，其生成逻辑如下所述：第 1 次根据请求条件查询（0 - pageSize + 1）条数据，判断得到的数据集大小是否大于 pageSize，如果大于 pageSize 返回值中包含 resumptionToken，否则不包含。同时利用算法侦测数据收割者的网络带宽动态决定 pageSize 的大小，具体的生成机制，见图 2。本生成机制基于一个假设，数据收割者访问 WPRIM 的 OAI 仓储时，第 1 次先发出 Identify 请求，而 ListRecords 请求因为需要指定数据分组、metadataPrefix 等条件，因此不会在第 1 次请求时就发出。探测网络带宽的算法如下：记录下用户 Identify 请求时的来源 IP。在向该用户返回结果 XML 数据时，记录下该响应所需要的时间。用户的网络带宽反比于响应时间，时间越长，则带宽越小，反之越大。通过大量的实验制定网络带宽基数，以及对应响应时间的网络带宽阶梯差值。当接收到该用户的 ListRecords 请求时，通过用户的 IP 和网络带宽的关联关系查到对应的带宽，进而决定 pageSize 的大小，从而实现用户的带宽不同，其批量返回数据集大小的不同。

3.4 WPRIM 系统 OAI 协议的验证

OAI 官方组织提供了一个能够提供数据服务的仓储注册的功能，目前已经登记注册 3860 个。注册时需要经过两个步骤的验证，第 1 步为验证 Identify 响应，检验其是否满足 OAI - PMH2.0 版本的协议要求。第 2 步的验证包括 6 个动词请求分别验证、错误请求的处理、OAI 2.0 特有异常的处理、POST 请求的处理以及 resumptionToken 能够正常使用的验证。验证成功后的 WPRIM 系统的数据服务者链接已经出现在官方列表中[4]。

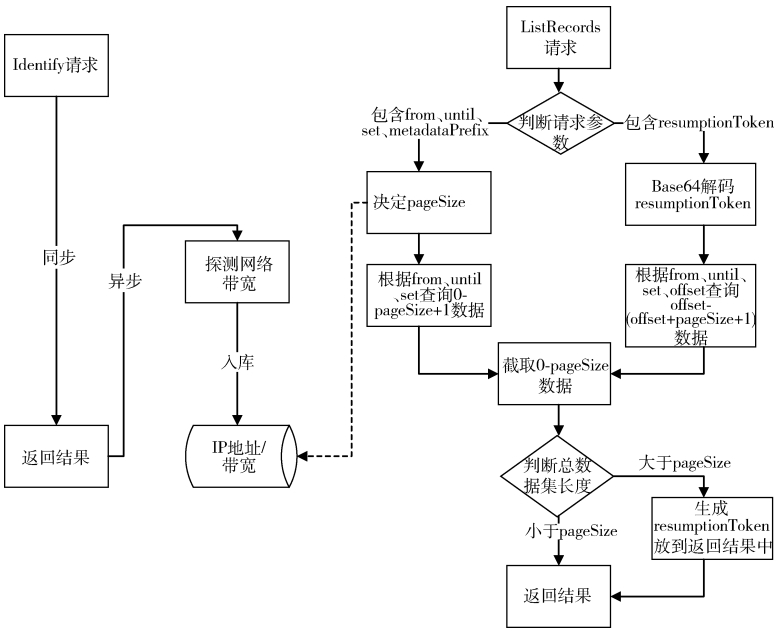

图 2 WPRIM 系统 resumptionToken 生成机制

4 结语

WPRIM 收录、索引西太平洋地区的医学期刊，对用户提供检索服务，同时为了提高 WPRIM 与其他的期刊仓储的互操作性，除了在原有已经具有的 API 接口的基础上，需要加入对 OAI－PMH 协议的支持。本文介绍 WPRIM 在实现 OAI－PMH 协议时面临的 3 个问题提供的解决策略：（1）制定 WRIM 元数据规范。利用该元数据规范实现 WPRIM 的元数据能够被其他的数据仓储理解、支持。（2）制定 WPIRM 数据的分组策略。根据 WPRIM 的数据特点以及 WPIRM 需要向 GHL 定期提交数据的要求，采用期刊名称、ISSN 和数据更新年月作为分组的策略，实现了对数据的有效组织，同时能够实现数据对 GHL 数据定期同步更新。（3）resumptionToken 生成策略的制定。WPRIM 在分析其他知名系统实现策略的基础上，提出了一种基于网络带宽自适应的高效生成策略。通过解决上述问题，WPRIM 系统实现了对收录数据的索引并提供数据服务，也已通过 OAI 组织网站上的验证注册服务，证明本文的实现策略是可行的。下一步的工作拟在通过为其他的仓储提供数据服务的基础上，优化 WPRIM 的 OAI－PMH 协议的服务。

参考文献

1 齐华伟，王军. 元数据收割协议 OAI－PMH［J］. 情报科学，2005，23（3）：414－419，425.

2 Anon. Dspace － a Turnkey Instutitional Repository Application［EB/OL］.［2019 － 01 － 01］. https：//duraspace. org /dspace/.

3 Anon. Arc Harvester and Search Engine Download | Sourcefo-rge. net［EB/OL］.［2019－01－01］. https：//sourceforge. net/projects/oaiarc/.

4 Anon. Archimède － Archimede：a Canadian software solution for institutional repositories［EB/OL］.［2019－01－01］. https：//www. bibl. ulaval. ca/archimede/index. en. html.

5 Anon. Dcmi：Dcmi Metadata Terms［EB/OL］.［2018－01－01］. http：//dublincore. org/documents/dcmi－terms/.

6 Anon. Medline/pubmed Data Element（field）Descriptions［EB/OL］.［2018 － 01 － 01］. https：//www. nlm. nih. gov/bsd/mms/medlineelements. html.

7 Anon. Web of Science Core Collection Schema［EB/OL］.［2019 － 01 － 01］. http：//help. prod － incites. com/wosWebServicesExpanded/wosSchemaWoSCCGroup/wosSchema. html.