关联数据集中开放资源的自动获取研究*

王思丽[1] 马建玲[1] 李慧佳[1] 王楠[1] 张秀秀[1]
（1.中国科学院兰州文献情报中心，甘肃，兰州，730000）

Study on Automatic Acquisition of Open Resources in Linked Data Sets

Wang Sili[1] Ma Jianling[1] Li Huijia[1] Wang Nan[1] Zhang Xiuxiu[1]
(1.Lanzhou literature and information Center of Chinese Academy of Sciences, Lan Zhou ,Gan Su,China,730000)

[摘要]关联数据集中的开放资源是当前数字图书馆知识服务系统的重要登记对象和利用对象之一。通过调研分析关联数据集的资源内容类型和应用接口等特点，设计 4 个数据集遴选指标，归纳提出 5 种从关联数据集中获取开放资源的自动获取策略。同时，通过实验研究，对 5 种策略的优劣性进行了对比分析并给出了应用建议。

[关键词]关联数据 开放资源 自动获取 语义搜索

Abstract : The open resources in linked data sets are one of important registered object and used object to digital library's knowledge service system at present. This paper mainly designs four datasets' select index and puts forward five strategies for automatic acquisition of open resources in linked data sets by surveying and analyzing linked data sets' resource content type and application interface. At the same time, the paper has done a comparative analysis of the good or bad of the five strategies and gives some recommendations for use in the future.

Keywords: Linked Data；Open Resource；Automatic acquisition；Semantic search
分类号：G250

0 引言

关联数据，从 2006 年被英国计算机科学家蒂姆•伯纳斯•李（Tim Berners-Lee）提出，到 2008 年被首次引进中国图书情报学界，再到 2009 年被称为“语义网的最佳实践”，逐渐掀起了国内外专家及学者的研究热潮，至今已发展了将近 10 年。目前，关联数据的基础原则、理论、方法、技术、标准规范、工具体系已得到了重大支持与广泛共识，越来越多的个人、政府机构、公司、学术团体如研究所、大学、图书馆等长期关注并积极参与关联数据的应用研究，关联数据相关的发布平台和消费工具正处于良好的开发和持续的完善中。

关联数据的研究进展使得网络上产生了大量的关联数据集，其中包含着大量的非传统的综合性学术性资源，可能涉及科研项目、科研机构、科学数据、软件工具等各类知识资源，始终贯穿和支撑面向用户的个性化知识服务过程，应当引起关注。过去，传统的图书馆资源库主要依靠人工进行搜集、获取并登记组织资源，工作效率不高。而关联数据集自身的开放关联、可免费获取、可复用、学术资源丰富等性质决定了它应该是当前新型数字图书馆知识服务系统的重要登记对象、利用对象和关联对象，对有效支持第三方系统对各类学术资源的开放关联和可计算化利用，持续提升和增强数字图书馆的支撑服务能力意义重大。

本文在前人的研究理论和技术基础上，通过调研分析关联数据集的资源内容类型、应用接口等，对关联数据集中开放资源的自动获取展开了研究，最终总结出了数据集遴选的 4 个指标和 5 种自动获取策略。

1 理论基础

2009 年 7 月，DERI（Digital Enterprise Research Institute，目前国际上比较大的语义网和关联数据研究机构之一）的首席数据工程师 Michael Hausenblas 在其发表的《Linked Data Application》[1]中阐述了关联数据消费的最佳实践，其中将关联数据的获

*本文系中国科学院国家科学图书馆青年人才领域前沿项目“开放知识资源登记系统集成关联数据的方法及应用研究”（项目编号：Y300231001）和中国科学院文献情报能力专项基金项目“开放知识资源登记系统（二期）”（项目编号：Y300051001）的研究成果之一。

取划分为三个技术层次：发现、存取和处理，并列举了一些实现方法，包括借助通用的发现机制、特定的查找服务如语义索引引擎、利用 RDF 的 HTTP URIs 的内在发现机制等。随后应用和研究比较广泛的主要有 Silk、Drupal 的 RDF 模块、LRDD、语义搜索引擎等。

其中 Silk[2]是一个关联发现框架，在 2009 年由柏林自由大学的 Christian Bizer 等和 Google 的工程师 Julius Volz 共同开发。Silk 的核心机制是一个链接发现引擎，主要功能是：通过 SPARQL 协议访问关联数据源，发现其他 RDF 链接；同时可以使用外部本体去计算链接的相似性，利用相似性聚合功能对相似结果进行合并，产生新的共指链接。Drupal[3]是一个模块化的开源的内容管理系统，2000 年诞生，从 2009 年开始增加 RDF 模块，支持关联数据的发布和消费，到 2011 年已形成一套完整的支持关联数据的机制体系。其中的 SPARQL Endpoint 模块为 Drupal 站点中的 RDF 数据提供检索支持，RDF Proxy 模块负责从其他含有关联数据集的站点获取资源数据。LRDD[4]提供一种标准的用于从具有 URI 标示的资源中发现和获取信息的方法体系。语义搜索引擎如 Falcons、SWSE、Sindice、Swoogle 等，提供获取关联数据的各种 API，支持人工和机器对关联数据的自动检索、结果分析和再利用。下文就是基于以上技术理论进行数据分析和后续研究。

2 数据分析

根据大量的调研和征求项目组意见，选取的初步指标有：（1）目标资源必须全部或部分包含关联数据集；（2）目标网站必须能够公开访问；（3）目标资源必须全部或部分包含学术性资源，且开放获取或免费获取。最终共选取了 12 个大型的关联数据集进行样本分析，主要从关联数据集的资源内容类型和应用接口 2 个方面进行重点分析。

首先以熟知的 DBpedia[5]为例，它包含的主要是从维基百科中抽取的结构化数据，涉及的资源内容类型比较复杂，有学术性的也有非学术性的资源，主要有人、地点、作品（音乐、电影、游戏）、组织（公司和教育机构）、物种、疾病等。其应用接口主要有 4 种：（1）批量下载：以 CSV 格式提供核心数据集的下载功能；（2）SPARQL 端点：以公共 SPARQL 端点提供在线查询服务，并支持将结果集导出为 HTML、Spreadsheet、XML、JSON、Javascript、Turtle、RDF/XML、N-Triples、CSV、TSV 格式；（3）分类 Web Service 接口：提供公共分类浏览器“search and find”用户接口和相应的基于 DBpedia 数据集的分类 web service。再如，ANDS[6]网站发布的澳大利亚研究数据，截止 2015 年 7 月 1 日，ANDS 共包含资源集 97375 个，主要是研究数据集和研究材料资源；机构 25392 个，主要是指创建或支持研究数据或资源的组织；活动 40706 个，主要是指创建研究数据或资源的项目或计划；服务 157 个，主要是指支持创建或使用研究数据或资源的服务。共提供了 6 种应用接口。其中 getRIFCS API 提供在注册系统里检索一个或多个对象的 RIFCS XML 文件。getMetadata API 用于在注册系统搜索或检索对象的结构化元数据，以 XML 或 JSON 格式返回。Vocabulary Service API 用于提供 ANDS 受控词表查询服务，允许科研机构查询“受控词表”。getExtRif API 允许用户检索在系统中因生成页面而产生的丰富的和被注释的记录复制信息。OAI-PMH Provider 基于 OAI 协议提供元数据收割服务。Research Grants API 提供获取有关澳大利亚研究经费和科研人员的数据信息。整体分析结果见表 1。

表 1 关联数据集分析

关联数据集	学科类型	内容类型	应用接口
DBpedia	多学科	人、地点、组织机构、物种、疾病等	批量下载：RDF、CSV、XML、JSON；SPARQL 端点；分类 WebService 接口
Freebase	多学科	书籍、人、机构、生物学、教育、项目等	批量下载：RDF、CSV 各种 API：RDF API、Search API
Nature.com	自然科学	期刊论文	快照下载：RDF；SPARQL 端点； OAI-PMH、OpenSearch、OpenURL、Z39.50
Data.gov	多学科：以农	科学数据	单个免费下载：RDF、CSV、JSON、XML、XLS；

	业、气候、能源为主		SPARQL 端点；OAI-PMH、OpenURL
Data.Gov.UK	多学科	科学数据	单个免费下载:RDF、CSV、XLS、TXT、DOC;RESTful API
ROAR	多学科	开放仓储	批量下载：RDF、XML、CSV、JSON；OAI-PMH、SearchSRU
OpenDOAR	多学科	开放仓储	OAI-PMH
Databib	多学科	科学数据仓储	批量下载：RDF/XML、spreadsheet；OpenSearch
NARCIS(荷兰国家学术研究系统)[7]	多学科	科学数据、科研项目、科研机构、科研人员	批量下载：HTML；OAI-PMH
NSF（美国国家科学基金会资助项目集）	多学科	科研项目	批量下载：XML
CORDIS(欧盟科研项目集)	多学科	科研项目	OpenSearch
ANDS （澳大利亚国家研究数据服务系统）	多学科：以地球科学、环境科学、生物学、海洋学为主	科学数据	getRIFCS API; Vocabulary Service API; getMetadata API; getExtRif API; OAI-PMH Research Grants API

3 确立遴选指标和自动获取策略

3.1 遴选指标

根据上文的数据分析结果，结合数字图书馆用户对学术性资源的需求，最终确立了关联数据集的 4 个遴选指标：学科指标、内容指标、接口指标和访问权限指标，见表 2。

（1）学科指标：学科指标主要用于保证所获取资源的专业性、学术性。主要包括自然科学及社会科学，不包括文艺、休闲、娱乐等与科技无关的资源。根据目前能够获取到的资源现状，从中图分类法和 DDC 分类法中已选择的主要学科领域有 10 种，见表 2。

（2）内容指标：参考 RSLP、DCMI 以及 NSTI 的划分规范，结合调研的用户需求及关联数据源目前的数据开放现状，着重设定以下内容类型为自遴选关联数据源中的内容指标：开放科研项目、开放科学数据（包括数据集、数据仓储和科学数据库）、开放仓储（包括学科仓储和机构仓储）、开放会议资源、开放期刊、开放课件。为了控制和保证所遴选数据集的质量，一般主要由来源判断资源内容的可信度，包括创建者的权威性，是否是该领域的权威机构、权威作者、权威企业或具有影响力的国际组织等。其中开放科研项目的质量控制,则重点考虑科研项目的级别,项目的负责单位，主要选择国内外比较重大的具有代表性的科研项目类型作为遴选目标和范围。如中国的国家自然科学基金项目、国家社会科学基金项目，美国的国家科学基金会 NSF 资助的项目，欧盟的研究项目等。其中开放科学数据着重选择比较著名的科学数据门户所登记的数据，如澳大利亚研究数据门户 Research Data Australia、美国开放政府数据门户 Data.gov，英国开放政府数据门户 Data.gov.uk 等。其中开放仓储主要有一些开放获取仓储目录、注册系统或世界排名如 OpenDOAR、ROAR、世界顶尖机构仓储网络排名等。其他类型的开放资源也都根据自身特点制定不同的遴选指标和质量控制方法。此外，资源的更新状况，如资源是否更新及时，是否具有良好的维护机制也作为内容指标的一个附加因素进行衡量。

（3）接口指标：接口指标，不仅关系着对学术资源的获取，也关系着对资源的定期维护和更新。为了能够方便快捷地获取到大量资源，一般选择公开提供了 SPARQL 端点的关联数据集或提供了传统 Web Service 接口如 OAI-PMH、OpenURL、RESTful 接口的关联数据源。同时兼顾关联数据集提供的在线免费批量下载功能。

（4）访问权限指标：为了不引起版权纠纷，一方面需要保证学术资源创建者或拥有者的合法权益不受侵犯；二要保证合法自动获取学术资源内容，并能够最大限度开放给用户浏览和查询。这需要在遴选数据集时注意数据集自身所规定的访问权限。原则上遴选访问方式为免费、可开放获取的的关联数据集，对于有限制的关联数据集可只获取其公开部分的元数据

内容，对其全文暂时不获取。

表 2　关联数据源的遴选指标

主要遴选指标	指标作用	指标内容
学科指标	保证遴选资源的专业性、学术性。主要是自然科学及社会科学。	Mathematics(数学) Physics(物理学) Chemistry and Chemical Engineering(化学与化学工程) Astronomy(天文学) Earth and Environmental Sciences(地球与环境科学) Biology and Life Sciences(生物学与生命科学) Medicine, Biomedical and Health Sciences(医药与健康科学) Materials Science(材料科学) Engineering and Technology(工程技术) Social Sciences(社会科学)
内容指标	控制遴选资源的质量和来源，保证权威性。	开放科研项目 开放科学数据（数据集、数据仓储、科学数据库） 开放仓储（学科仓储、机构仓储） 开放会议资源 开放期刊 开放课件
接口指标	保证资源可快速、持续获取，保证已获取资源的长期监测、维护和更新。	是否公开提供 SPARQL 查询端点 是否公开提供 Web Service 接口，如 OAI-PMH、OpenURL、RESTful、Z390.5、RSS 等 是否提供免费的批量下载或导出功能
访问权限指标	保证资源的创建者或拥有者的合法权益;保证合法获取到的资源能够最大限度的开放给用户浏览和查询。	需遵循 CC、CC BY、CC BY-NC、CC BY-NC SA、CC BY-NC ND 等开放获取使用许可等; 免费（free） 开放获取（OA） 有限制（元数据可开放获取）

3.2 自动获取策略

除了对已知关联数据集自身描述性信息的获取和存储，我们一般更希望发现并采集到更多的未知的相关联的学术性数据集。因此还需要针对各个具有不同特点的关联数据集的内容元数据进行解析和抽取。根据上文研究，主要制定出以下 5 种自动获取策略：

3.2.1 基于关联数据集 SPARQL 端点的获取策略

该策略主要针对那些提供了公开 SPARQL 查询端点的关联数据源。SPARQL 端点能够支持远程客户端的查询请求，通过 HTTP 协议实现查询传输，匹配相应的关联数据源，返回查询结果。一般返回格式为 XML、JSON、HTML、CSV 等。一般情况下一个 SPARQL 端点只针对一个关联数据集，是一对一的精确关系。但有些 SPARQL 端点也提供对多个关联数据集的联合查询，如 FactForge 支持对 DBpedia，Geonames，WordNet，Freebase 等 8 个数据集的联合查询。使用 SPARQL 端点对关联数据集进行查询与资源采集获取，一般需要通过语义 web 框架或工具,进行语义编程实现。目前已经有多种优秀的语义 web 程序设计框架，如 Jena、Sesame 等。这种策略不需要下载数据到本地，能快速、准确的获取指定数据集内的相关数据，但需要了解 SPARQL 的语法规则，并且查询结果的呈现形式不够直观。

3.2.2 基于关联数据集 Web Service API 的获取策略

该策略主要针对那些提供了各种 Web Service 接口的关联数据集。主要有 OAI-PMH、OpenSearch、OpenURL、SRU/SRW、Z39.50、RESTful、RSS feeds 等，此外还有各数据集自定义的专业 Web Service API。当前的 Web Service 技术已经非常成熟，基于 HTTP Client 远程调用起来也相对简单。并且能够持续、快速地、全面地获取到数据源提供的全部元数据。其中最常见的是基于 OAI 协议的 OAI-PMH 元数据收割接口，该接口通常支持 oai_dc、mets、marc 等通用的标准元数据格式，便于机器自动采集和解析获取数据。OCLC 发布了标准的 Harvest2 组件，包含一系列的 jar 包和类库，用于支持 OAI 收割。一般这些接口的返回格式都为 XML 或 JSON。同时 OAI-PMH 接口支持通过自定义时间戳（对应的是资源集

合被创建时间的上下限）来获取资源，非常有利于大规模的、持续的、长期的、稳定地定期监测和更新获取资源。

3.2.3 基于关联数据集批量下载功能的获取策略

该策略主要针对那些提供了各种批量下载、导出、镜像下载、快照下载、RDF dump 功能的关联数据集。这些数据提供者将关联数据源中的资源数据序列化或转换为一些常规数据格式，提供免费下载功能。常见的数据格式有 RDF、XML、CSV、XLS、TXT、JSON、Spreadsheet 等格式。该策略比较简单，其关联数据集中的资源一般被预先索引并保存在本地，获取的准确率高，速度较快。且获取到的资源由于格式固定，便于被解析或直接导入到关系型数据库进行存储。但由于受数据更新方面的限制，该策略不适合规模较大且经常更新的关联数据集。

3.2.4 基于关联数据集动态网页抽取的获取策略[8]

该策略适用于互联网上绝大部分的那些既不提供 SPARQL 端点，也不提供 Web Service API 源接口，更不提供在线免费下载或导出功能的关联数据源。这一类的数据集一般是直接发布在网站上，通过 RDFa 将 RDF 三元组嵌入 XHTML 文档，使得符合标准的第三方用户或机器用户可以从 RDFa 文件中提取 RDF 三元组，获得资源元数据和语义信息。这类数据集的样式多变，结构不固定，不能完全通过编程进行自动获取，可能需要借助人工的预处理，对网页深度，网页主体块进行划分，形成可动态更改配置的参数，借助网页内容抽取工具分批分层次完成采集。目前开源的网页内容抽取工具或组件有很多，如 HtmlParser、MetaSeeker、GooSeeker 等。

3.2.5 基于语义搜索引擎或关联数据爬虫的获取策略

前 4 种策略的共同点是需要事先遴选好数据源，针对特定的数据源根据该数据源的特性去选择相应的采集策略。该策略不同于前 4 种策略，这种情况下不能确定应用相关的关联数据源，需要在整个语义网空间中进行数据发现和采集获取，主要有两种实现方式：语义搜索引擎和关联数据爬虫。该策略的主要目的是为发现更多的可能存在的未知关联数据集，即潜在的资源对象。目前，有多种语义搜索引擎提供人机交互接口，可以通过跟踪 RDF 链接获取关联数据。如 Falcons、Swoogle、Sindice 等。图书馆资源建设人员不需要自己实现底层的数据抓取和索引，通过应用接口就能发现包含特定关键词或 URI 的 RDF 资源。关联数据爬虫能够从一系列种子 URIs 跟踪 RDF 连接，找到相关的关联数据，并将检索到的数据存在 RDF 存储库或本地文件中。如 LDspider[9]，是一个开源的关联数据爬虫，核心是采用宽度优先的采集策略，同时以负载均衡方法对应用状态进行监控从而动态控制爬取进程。它的优点是可以随机发现更多的新资源。缺点是由于目标不明确，也不存在比对判断，可能重复抓取数据，后期需要人工进行查重处理。

3.3 实验与分析

选取第 3 章节中的部分数据集，对上述的 5 种获取策略进行试验研究，结果见表 3

表 3 自动获取结果

数据集	获取内容	获取策略	获取数据量（条）
ROAR	开放仓储	批量下载：RDF/XML	3830
OpenDOAR	开放仓储	基于 OAI-PMH	2780
Databib	开放仓储	基于 OpenSearch	1015
EU Research Project - FP7	科研项目	基于 OpenSearch	23048
NSF-Project	科研项目	批量下载：XML	119508
NSFC-Project	科研项目	基于动态网页抽取	9569
ANDS	科学数据	基于 OAI-PMH	37856
Global Think Tank Directory	科研机构	基于动态网页抽取	4357
Data.gov	科学数据	基于 SPARQL 端点	104815
DBpedia	组织机构	关联数据爬虫：LDSpider	49000

通过实验，对上述 5 种策略进行对比分析发现，每种策略都有最适用自己的情况和相应的优缺点，见表 4。在实际应用中，需根据实际情况如获取内容或其他需求等，选择一种或

多种策略相结合的方式进行数据采集。实验中也发现，对于那些既提供了 SPARQL 端点又提供了传统 Web Service API 的关联数据源，优先考虑使用基于 Web Service API 的采集策略，因为获取到的资源属性会比较全面，更有助于资源分类标注和做进一步的处理。

表 4 几种获取策略的对比分析

获取策略	适用情况	优点	缺点
SPARQL 端点	提供公开 SPARQL 端点	查询准确率高 支持联合查询	需熟练掌握 SPARQL 语法；获取到的资源属性比较少
Web Service API	提供 Web Service 接口之一	可长期、稳定、批量地监测和获取到最新数据。资源属性比较全。	需熟悉 API 背后的各种协议。
批量下载	提供批量下载功能，数据规模不大，更新频率不高	简单、快速、直接	不适用于大规模、更新频率高的数据源
动态网页抽取	不提供任何接口或下载功能	有针对性，可直接定位到需求的模块，摒除垃圾信息	不适用于页面构造复杂，如纯 js 生成或客户端加密过的数据源；需随着数据源网站的改版而更改采集抽取策略。
语义搜索引擎或关联数据爬虫	发现更多新的未知的关联数据源	能够同时获取多个数据源的资源	数据可能存在重复；存在较多与主题无关的垃圾信息；获取结果受爬取策略影响较大

4 结语

关联数据技术的快速发展使得互联网上关联数据集中的开放性学术资源也越来越丰富，为数字图书馆充盈资源库提供了潜在的可用资源。同时，作为图书馆的资源建设和信息服务人员，为了加速资源库建设，获取更多更全面的资源，从而满足用户日益增长的对新型数据资源的需求，也必须时刻关注新技术新数据的发展，并及时总结出有效的资源获取方法，才能更大的发挥图书馆员的作用。本文在前人的研究基础上，最终总结提出了 5 种自动获取策略，并进行了实验分析与研究，结果表明一定程度上提高了对关联数据集中的开放性学术资源的获取速度和效率，希望能够给其他研究人员提供一些借鉴和帮助。 当然，也肯定会存在着不足，笔者会在后续的研究中积极探索更多的资源获取方法，为图书馆扩大资源规模，提高服务质量而继续努力。

参考文献

[1]Michael Hausenblas.Linked Data Applications[C].DERI Technical Report,2009:1-27.

[2]Julius Volz;Christian Bizer;Martin Gaedke; Georgi Kobilarov.Silk – A Link Discovery Framework for the Web of Data[C].Linked data on the Web workshop at www2009,2009:1-6.

[3]Drupal[EB/OL].[2015-04-02].https://www.drupal.org/.

[4] Hammer-Lahav E.LRDD:Link-based Resource Descriptor Discovery[EB/OL].[2015-05-30]. https://tools.ietf.org/html/draft-hammer-discovery-04.

[5]Dbpedia[EB/OL].[2015-04-20].http://wiki.dbpedia.org/About.

[6]ANDS[EB/OL].[2015-03-01].http://researchdata.ands.org.au

[7]NARCIS[EB/OL].[2015-06-01].http://www.narcis.nl/?Language=en

[8]王思丽,马建玲等.开放知识资源的元数据自动采集策略研究[J]. 图书馆学研究,2013(12):47-51.

[9]Robert Isele; Jürgen Umbrich; Christian Bizer; Andreas Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data[C]. Proceedings of the ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, 2010:1-4.

作者姓名及单位

王思丽 马建玲 李慧佳 王楠 张秀秀

单位均为：中国科学院兰州文献情报中心

作者简介：

王思丽（1985-），女，馆员，中国科学院兰州文献情报中心

马建玲（1969-），女，研究馆员，中国科学院兰州文献情报中心

李慧佳（1984-），女，馆员，中国科学院兰州文献情报中心

王楠（1979-），女，副研究馆员，中国科学院兰州文献情报中心

张秀秀（1981-），女，馆员，中国科学院兰州文献情报中心

邮编：730030

通信地址：甘肃省兰州市城关区天水中路 8 号中国科学院国家科学图书馆兰州分馆 2#601

通信作者：王思丽

联系电话：13919149873 / 0931-8270076

电子邮箱：wangsl@llas.ac.cn

定稿日期：2015-07-15