Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

科技文献搜索引擎元数据仓储建设实践*

甘大广 苏学 张正峰 / 北京万方数据股份有限公司 北京 100038

摘要：文章分析了用户查询行为、行业资源出版模式等变化对数字资源整合的机遇与挑战，重点结合实际工作介绍了科技文献搜索引擎底层元数据仓储的建设过程，包括元数据采集、元数据规范、元数据整合等环节。

关键词：元数据仓储，元数据整合，数字资源
DOI：10.3772/j.issn.1673—2286.2013.06.007

1 引言

伴随着数字图书馆建设的进程，以网络数据库、数字期刊以及电子图书为代表的数字资源数量越来越多，搜索引擎的流行及普及，行业资源出版格局及模式的变化，以及用户信息查询行为的变化，对元数据整合、元数据仓储建设既是机遇又是挑战。

（1）行业资源出版格局及模式的变化

随着电子期刊以及独家代理的出现，出版模式发生重大变化：期刊由原来的单纯由出版社出版，演变为出版社出版、代理商出版和跨学科的网络出版系统等出版模式并存[1]，尤其是在出版社转制后，出版模式变化显现得更为明显。

（2）用户信息查询行为的变化

根据CNNIC发布的《第22次中国互联网络发展状况统计报告》，搜索引擎是用户在互联网中获取信息的重要工具，2007年12月的使用率为72.4%，规模达到15204万人；2008年6月比重虽然有所下降，但仍然达到69.2%，用户群人数也增长到17508万人[2]。在另一份OCLC的报告《大学生对图书馆和信息资源的认知》中称：89%的大学生从搜索引擎开始信息检索[3]。在用户获取信息的途径中，搜索引擎已经成为了第一选择，其次才是数字图书馆或是数据库商提供的联邦检索。

一方面行业出版格局及模式的变化导致资源更“条块化”，更“分散化”；另一方面用户期望信息查询更“简单化”，更“快速化”。如何面对行业资源出版格局及模式的变化、用户信息查询行为变化已成为数字图书馆领域重要的研究课题。目前解决此问题的方法主要集中在两个层面[4]：一是系统层面，二是数据层面。从清华大学图书馆、北京大学图书馆、上海交通大学图书馆等各大学图书馆纷纷新上的基于中央索引的资源发现系统来看，元数据层面的整合是一种趋势，关于资源发现系统的比较[5,6]、资源发现系统的应用[7,8]研究较多，然而关于资源发现系统底层元数据仓储建设过程研究较少，因此本文重点结合实际工作介绍科技文献搜索引擎底层元数据仓储构建。

2 基于元数据仓储的科技文献搜索系统逻辑结构

从用户角度看，科技文献搜索系统作为与用户交互的前端，为用户提供单一入口的检索与获取；从资源服务角度看，科技文献搜索系统是将分布式的海量学术数字资源从异构到有序的一个过程。图1从资源服务角度描绘了科技文献搜索系统的逻辑结构，数据集合层相当于数据采集，有序数据层相当于数据加工、整合，元数据仓储建设就是实现科技文献搜索系统中的数据集合层以及有序数据层。

3 元数据仓储建设实践

元数据仓储建设旨在通过出版商、大学公开的网站收集学术文献信息，规范整理、整合，将异构、分布和海量的学术文献信息得以汇聚，形成无重的元数据仓储，通过

* 本文系国家高科技发展计划（863计划）“云计算一期”重大专项课题“以科技文献为主的搜索引擎研制”子课题（编号：2011AA01A206）成果之一。

预索引的方式，为用户提供简单、快捷的数字资源发现服务[7]。元数据仓储建设包括元数据采集、元数据加工、元数据整合等三个模块，图2为元数据仓储整体建设总体流程图。

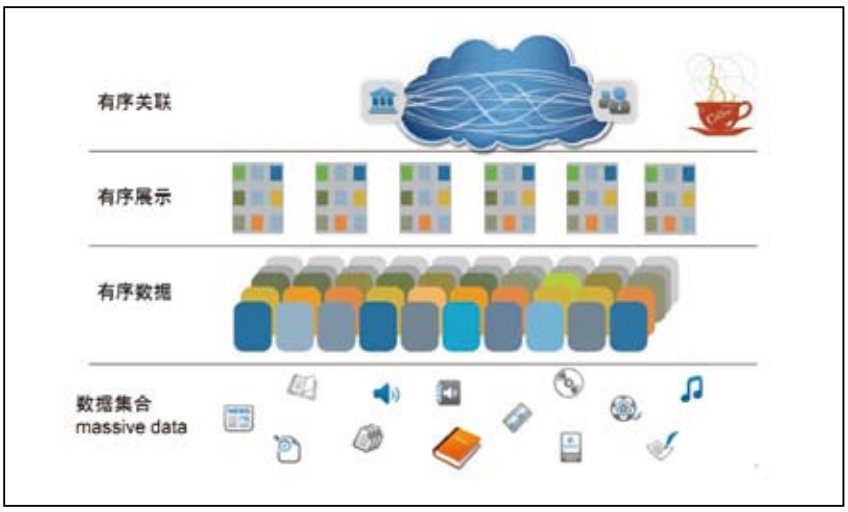

图1 科技文献搜索引擎逻辑结构图

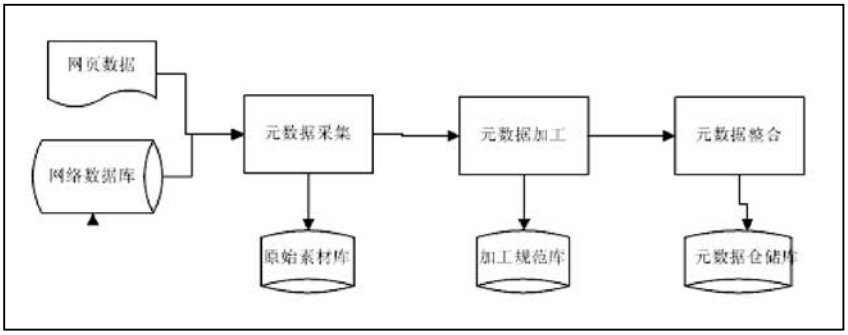

图2 元数据仓储建设流程图

3.1 元数据采集

元数据采集模块是元数据仓储建设的基础，负责元数据记录的收割采集。采集模块允许系统通过多种采集方式获取元数据记录，采集模块将采集到的元数据记录直接存入系统中的原始素材库，不改变元数据的原格式，以供后期元数据加工、元数据整合模块使用。元数据采集主要有两种方式：一种是资源提供商提供；另一种是收割资源提供商系统。资源提供商提供的方式无论在数据全面性还是数据质量方面都是通过收割资源提供商系统所无法比拟的。然而碍于商业利益，多数资源提供商不愿提供元数据。目前已与部分资源提供商进行了合作，通过FTP方式收割数据，另有部分需要通过互联网采集来实现数据采集，面对海量的元数据资源，如何快速采集和及时更新是采集的难点，这需要通过增量元数据收割方式。针对不同的网络数据库主要采用以下几种方式：按刊年期[9]、资源厂商资源唯一号、数据更新时间，以及RSS订阅等方式。考虑到增量的采集方式多少会有所遗漏，定期通过全量的方式收割资源唯一号，通过与已收割的数据比对后再补充。上述元数据收割方式多少都存在些先天不足，如按刊年期的方式前提是网络数据库出版每次整期出版，随着优先出版方式的出现，网络出版出现了非整期出版，在实际采集过程中使用以上多种方式组合来保证数据采集的全面性、及时性。

3.2 元数据加工

元数据加工模块是元数据仓储质量的保障，按照功能细分为元数据转换、元数据清洗、元数据深度规范以及元数据质量检查。元数据转换又称元数据映射：利用转换技术将不同结构的数字资源规范化，消除异构资源间的不一致性，为来自不同资源厂商的元数据根据元数据仓储著录标准进行字段映射，为资源的整合及统一存储奠定基础。元数据清洗包括规范大小写；全角转半角；字段拆分（主要集中在由于网页模版不统一，元数据采集时未拆开）；格式化处理，如期刊中卷期的描述有的来源是固定位数，不足位前面补0，有的来源不是固定位数；非学术论文记录剔除，如征稿启事等；关联补充字段项，如ISSN；以及规范用于整合查重的字段。元数据深度加工是对“知识获取五要素”进行了深度加工标引，“知识获取五要素”[10]是指学者、科研机构、主题、学科、基金项目。如对作者单位进行分级处理，一级单位、二级单位、邮编等；论文基金资助信息规范出基金名称、项目名称、项目编号等，深度加工后的元数据便于用于对科研实体、科研主

题、学科、项目进行分析。元数据质量检查是对元数据的质量进行控制，主要通过必备性检查，如标题、期刊、年等；字典检查，如学位级别、授予学位单位等；以及正则检查，如ISSN、年、DOI等。经过上述加工后的元数据，进入资源厂商镜像库。

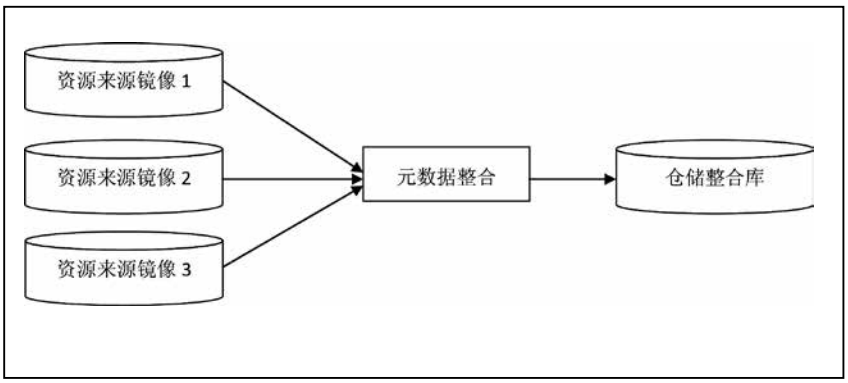

图3 元数据管理架构

3.3 元数据整合

3.3.1 元数据整合程度及模式

元数据整合程度简单讲，就是各种来源不同的元数据，是在系统内简单地堆积，还是在系统内部转换为统一格式、合并重复记录[11]。前者，保持了原始数据的原貌，输入数据处理简单，但是存在数据量大、元数据判别问题，元数据质量、编制索引、提供显示、按指定格式输出都非常复杂，并且需要读者自己判断记录是否相同等问题；后者，输入数据处理复杂，甚至需要工作人员辅助去重，但是系统检索服务简单快捷，读者使用数据集中明了。笔者赞同后一种整合程度，并由此引申出元数据管理模式及去重等问题。鉴于既要去除重复数据，又要保留原始数据的原貌，笔者在实际处理中，借鉴FRBR模型的理念，基于“版本”思想。将同一条记录来自不同资源供应商分别作为一个“版本”。图3为元数据整合阶段的数据管理模式，经过元数据采集、元数据加工后，每个资源厂商的数据都进入到相应的镜像库，在元数据整合阶段，根据各资源来源的数据质量优先顺序进入仓储整合库，对于同一条记录有多个来源时，仓储整合库中只存一条质量较优的记录，同时标记数据来源，以及能关联到各来源镜像库的ID。

3.3.2 元数据整合去重

元数据整合去重规则关系到整合后数据的质量，应尽量减少漏查和误查。针对不同的文献类型及不同来源的数据分别针对不同的查重规则。元数据仓储的数据渠道多样，来源复杂，加上当前学术期刊发布平台各不相同，造成一本期刊/一篇文章被多个平台收录和揭示，且不同平台的元数据描述规范不统一，给海量的元数据仓储数据的去重带来很大困难。另外，在来源数据内部也涉及去重问题，但单一来源内部查重相对简单，一种是按照来源数据库内部唯一号，如万方数据期刊论文每篇论文都有内部唯一号；另一种按照来源URL，前提是来源数据库的URL未发生变化。对于多来源元数据的整合去重规则相对复杂，主要是由于来源渠道多样、各学术资源平台的元数据描述标准不统一等原因导致。以中文期刊论文为例，严格意义的中文期刊论文查重规则：标题、第一作者、刊名、年、期、页码相同，但由于各资源平台著录标准不同，主要体现在以下几个方面：

（1）期刊合并拆分

出版社由于CN刊号有限，常使用同一CN号出版不同的版本，如绍兴文理学院学报存在教育版、社科版、自科版等三个版本，对于这种出版方式，各资源来源的著录方式又有所不同，详见表1。对于有CN/ISSN号的期刊可通过CN/ISSN合并，对于没有CN、ISSN的通过比较来源不同信息源论文相同的比例来判断是否是一本刊，为此建立一个不同资源来源刊名间对照关系表。

（2）跳转页

对于某期内一篇文章存在跳转页的现象，各资源来源的著录标准有所不同，如表2所列有两家资源来源把跳转页论文作为多篇论文对待，这点万方数据做的相对较好，仅作为一篇。因此，在整合多家来源之前需要先对单个资源来源内部进行查重清理。

（3）期号

学术类的期刊的刊号相对较规范统一，各不同资源来源期刊期号著录不同主要集中在增刊：有些资源来源增刊用“Z1”，有些资源来源增刊用“S1”；合期出版，以3、4期合刊为例，有些资源来源出2期，期号分别为3、4，内容完全相同，有些资源来源则出1期，期号为：3-4；以及由于期刊合并分拆导致出周期不同，从而期号受到影响，如表2所示。

表1 期刊合并拆分

资源来源	刊　　名	ISSN	CN	期数/年
信息源1	绍兴文理学院学报	1008-293X	33-1209/C	12
信息源2	绍兴文理学院学报（教育版）	1008-293X	33-1209/C	2
信息源2	绍兴文理学院学报（社科版）	1008-293X	33-1209/C	6
信息源2	绍兴文理学院学报（自科版）	1008-293X	33-1209/C	4
信息源3	绍兴文理学院学报：教育教学版	1008-293X	33-1209/G	2
信息源3	绍兴文理学院学报：自然科学版	1008-293X	33-1209/C	12
信息源3	绍兴文理学院学报：哲学社会科学版	1008-293X	33-1209/C	6
信息源3	绍兴文理学院学报	1008-293X	33-1209/C	12

表2 跳转页著录样例

资源来源	论文标题	作　者	出　　处
信息源1	水稻稻曲病与纹枯病的发生规律	徐立佳；刘学亮；范洪玉	《农民致富之友》2012年第10期 103-103页，共1页
信息源1	水稻稻曲病与纹枯病的发生规律	徐立佳；刘学亮；范洪玉	《农民致富之友》2012年第10期 107-107页，共1页
信息源2	某深基坑支护及降水方案设计研究分析	张桂林；徐顺泉	《中华民居（下旬刊）》2012年第6期 136页
信息源2	某深基坑支护及降水方案设计研究分析	张桂林；徐顺泉	《中华民居（下旬刊）》2012年第6期 198页

鉴于以上分析，笔者所用到的查重字段见表3，并对各个字段内容进行尽量规范统一，减少因不同资源来源著录标准不同所导致的数据合并不干净的情况。根据表3中字段进行多种组配方式查重，如标题、作者、刊、年；标题、刊、年、期等。另外考虑同一家资源来源是否把所用的查重组配方式作为多条记录。通过对上述影响去重数据的处理，数据合并量相比未处理前增加了450万。

表3 中文期刊查重字段

字段名	备　　注
标题	去除掉标点符号的标题
第一作者	
刊名/ISSN/CN	刊名去除标点符号
期刊ID	元数据仓储期刊ID
年	
期	期格式统一
页码	
来源数据库唯一号	

4 结语及展望

面对用户查询行为、行业资源出版模式等变化，构建科技文献元数据仓储具有重要意义。基于构建的元数据仓储一方面可服务于科技文献搜索系统，另一方面基于海量数据的分析，可揭示出隐含的、有潜在价值的信息和知识，如基于文献计量概念的学科或主题研究趋势及热点分析，学者、科研机构的科研产出分析及评价。

元数据仓储建设仍面临诸多难题：元数据采集作为元数据仓储建设的起始阶段，元数据采集情况直接影响元数据仓储建设的数据量

以及数据质量，由于商业利益等原因，目前仅有少数资源提供商愿意提供元数据。另外，笔者在实际建设过程中虽然对元数据进行了多方位的规范处理，但仍存在少量因不同来源数据不规范导致的数据重复现象，要解决元数据仓储中各种来源数据的规范性问题，仍然是任重道远。

参考文献

[1] 刘树新,王庆良,杨春.中美发行代理商服务方式之比较[J].出版发行研究,2006(7):48-50.
[2] CNNIC.第22次中国互联网络发展状况统计报告[EB/OL]. [2008-08-06]. http://www.cnnic.cn/uploadfiles/pdf/2008/7/23/170516.pdf.
[3] OCLC. College Students' Perceptions of Libraries and information Resourccs - A Report to the OCLC Membership [EB/OL]. [2008-08-15]. http://www.oclc.org/asiapacific/zhcn/reports/pdfs/studengperceptions.pdf.
[4] 赵悦,富平.数字资源与传统文献元数据整合[J].国家图书馆学刊,2007(2):63-66.
[5] 刘颉颃,陈定权,郭婵.用户对图书馆资源发现系统功能的期望:基于广州大学城高校图书馆学生用户的调研[J].图书情报工作,2012(7):27-31.
[6] 包凌,蒋颖.图书馆统一资源发现系统的比较研究[J].情报资料工作,2012(5):67-72.
[7] 窦天芳,姜爱蓉.资源发现系统功能分析及应用前景[J].图书情报工作,2012(7):41-43.
[8] 徐荣华.基于元数据仓储的资源整合应用[J].图书馆杂志,2012(4):67-73.
[9] 吴广印,苏学,甘大广.基于HTTP协议的OA期刊元数据动态收割研究[J].数字图书馆论坛,2011(9):43-47.
[10] 甘大广.学术论文导航系统知识库的构建与实现[D].北京:中国科学技术信息研究所,2008.
[11] 国家图书馆.国家图书馆同人文选[M].北京:国家图书馆出版社,2009:193.

作者简介

甘大广（1984-），男，研究方向：情报检索与知识组织。E-mail: gandg@wanfangdata.com.cn
苏学（1986-），男，研究方向：情报检索与知识组织。
张正峰（1977-），男，研究方向：情报检索与知识组织。

Construction of Metadata Repository in Scientific Literature Search Engine

Gan Daguang, Su Xue, Zhang Zhengfeng / Wanfang Data Co., Ltd, Beijing, 100038

Abstract: This paper analyzes the opportunities and challenges of a user query behavior, industry resources publishing model change on the integration of digital resources, focusing on the actual work of scientific literature search engine underlying metadata repository construction, including metadata harvesting, metadata specification, and metadata integration.

Keywords: Metadata repository, Metadata integration, Metadata repository

（收稿日期：2013-05-10）