

# 基于语义关联的高校图书情报档案数字资源整合研究<sup>\*</sup>

谭 静

(天津城建大学图书馆, 天津 300384)

**[摘 要]**分析了资源整合的内涵及作用,介绍了基于语义关联的海量数字资源整合方法。详述了高校图书、情报、档案资源整合的前提条件,主要包括:高校图书、情报、档案知识资源的数字化存贮,工作流程的革新化改造和高校组织之间基于协议的知识联盟。最后,重点论述了基于语义关联的图书、情报、档案数字资源整合模式。

**[关键词]**语义关联 数字资源 资源整合模式

**[分类号]**G250.73

信息社会的到来使得数字图书馆的服务模式发生了巨大转变,更加专注于深度挖掘知识的内容和关联性。随着互联网技术的不断发展,数字资源的增长速度也在不断加快,数字图书馆的知识资源也逐渐虚拟化,由此改变了用户的使用习惯和交流方式<sup>[1]</sup>。为了将分散的数字资源进行整合,数字图书馆尝试构建基于语义关联的数字资源整合模型,从而进一步提高数字图书馆的知识服务效果。具体来讲,数字资源整合技术就是利用相关技术将异构的数字资源进行整合,并通过后期的加工和排序处理将数字资源形成一个统一的整体,并将数字资源的规律性和知识性表现出来。笔者所研究的基于语义关联的数字资源知识整合技术,就是利用语义处理相关技术,分析隐藏在数字资源中的知识,挖掘出语义关联性,然后利用聚类和重构的方式将基于语义关联的数字资源整合为一个有机整体,从而为用户展现出丰富的知识关联性。

## 1 资源整合的内涵及作用

资源整合技术就是根据特定需要将具有一定关联性的数字对象、个体以及相应的功能进行重组、融合以及聚类处理,从而形成同一种类的数字资源体系<sup>[2]</sup>。一般将描述资源的规范称为元数据规范,而用于描述数字资源整合的规范又称为数字资源整合的元数据规范。元数据规范已经成为基于语义关联的数字资源管理系统的重要规范,甚至可以直接将该规范认定为各种数字资源的统一规定。宽泛地讲,数字资源的整合描述在分布式资源检索、资源定位以及基于异构系统的相互操作等方面有着积极的作用。

### 1.1 分布式信息检索

在分布式数字资源环境下,为了更好地满足用户关于信息系统的资源检索需求,本研究根据数字资源开发原则为不

同领域的数字资源以及不同网络检索技术提供数字整合描述,这也是在分布式数字资源中知识发现的重要方式。与此同时,针对数字资源进行有效描述不仅可以为异构数字资源的联合范围扩大,从而形成一种有机整体,而且还可以为用户提供资源整合工具和相互操作机制,以此扩大影响范围,并进一步提高信息检索水平。

### 1.2 异构信息系统间的互操作

与关于资源对象的描述不同,关于资源整合的描述具有显著的分布式特点,而且层次性特征也十分明显,描述的对象可以存储在不同的物理空间,甚至可以存储在不同的信息系统之中,因此关于这些描述对象的数据存储结构、存在方式以及检索模式都是异构的。此外,基于资源对象的动态描述规范还能够帮助实现异构资源对象的无缝链接,从而为整个异构信息系统的相互联系提供支持。

## 2 基于语义关联的海量数字资源整合方法

### 2.1 海量数字资源采集、描述与整合机制

海量数字资源的来源方式多种多样,存在的类型也繁多复杂,已经呈现出分布式的特点。针对数字资源的采集方式应该结合自身的分布特点进行,利用合适的数字资源采集工具,并且制定相应的数字资源采集方案。具体来讲,可以在数据库资源中利用转化和抽取技术以及分档分析软件,来提取异构数字资源的特征向量,并进行必要的语义标引<sup>[3]</sup>。数字资源的采集方式可以按照由近到远的方式,也就是首先采集日期较近的数字资源,然后对日期较晚的数字资源采用回溯的方式,从而保证较新的数字资源优先被采集和整理。当然,在具体描述数字资源之前,还要根据元数据模块进行语义描述,从而让异构的数字资源更容易被计算机识别。在语

<sup>\*</sup>本文系2014年天津市教育委员会高等学校人文社会科学研究项目“高校图书、情报、档案一体化管理模式研究”(项目编号:20142155)成果。

义整合阶段,还可以利用已经存在的先验数字资源进行语义分类处理,接着分析数字资源的内容差异和语义相似度,然后将较为相似的数字资源整合为一个整体;或者参照基于领域本体的映射关系进行数字资源重组,进一步揭示数字资源的内在语义关联,从而将隐藏的知识资源进行深度整合。

## 2.2 以引证与概念为基础的知识整合方式

引证关系反映出数字资源的流通情况,而且是一种单向流通。通过引证关系可以挖掘出数字资源的语义关联性,比较常用的方法主要以引证耦合以及引证路径为基础构建通用模型,从而实现关于数字资源的语义分级和整理处理<sup>[4]</sup>。如果数字资源之间存在直接的引证关系,就可以直接在引证联系网中搜索相似的数字资源,并进行多维度分析和评估。如果数字资源之间存在间接的引证关系,就需要利用分析引证方法来分析关联的强度,进一步确定数字资源之间的关联性,为资源的深度整合提供支持。此外,还可以利用数字资源的概念联系进行资源整合,可以借助语义表达方式的不同来区分数字资源,利用不同的语义单元探讨其关联性;还能够根据不同类型的数字资源关系和映射方式,并结合用户的资源需求特征,利用语义关联和概念关系来挖掘数字资源中的内在规律性。针对知识组织系统,可以根据知识粒度来分析语义关联,并在应用过程中挖掘不同知识粒度的语义关联性,从而得到在不同知识粒度条件下的知识整合效果。

# 3 高校图书情报档案资源整合的前提

## 3.1 高校图书情报档案知识资源的数字化存贮

一体化方式的知识组织、知识编码以及知识挖掘的主要对象不仅包括了在线网络信息,而且还包括数字化图书、档案和情报信息等<sup>[5]</sup>。数字资源的来源主要分为两个部分,其一为文档文献形式的数字资源,该类型在整个资源中占有的比例较大;其二为视听形式的数字资源。比如文档文献形式的数字资源主要包括以档案、情报以及图书数字化形式进行存储,还包括经过数字化存储的纸质文献。基于文献文档的数字化存储方式主要有两种方式:第一种为构建基于图像的存储方式,第二种为构建基于文本的存储方式。值得注意的是,第一种存储方式占用的物理空间较大,成本较高,不利于长远的存储规划;第二种存储方式需要人工进行数字化,也就是将文献的资料,主要是图像、文字以及数据表格等以 Word、PDF、TXT 或者 GIF 格式进行存储。如果需要数字化的图像、文字或者数据无法用人工输入的方式完成,就可以利用复制或者扫描的方式将数字资源进行存储;如果需要处理视听类数字化资源,就需要利用相关设备和技术进行转录、降噪以及压缩处理,并最终实现以 MP3 和 AIV 等格式进行存

储。

## 3.2 高校图书情报档案工作流程的革新化改造

一体化过程本身就是一种科技创新活动,是将图书类、档案类以及情报类资源进行结构调整和整合处理,而且要求这些工作流程与创新机制相一致。高校工作业务再造思想是进行一体化科学充足的重要参考依据,也就是强调整体性工作与业务分工的相互连接,减少多余重复的构建过程,重视基于决策的业务建设,从而实现从职能管理到业务流程管理的转变,坚持效能最优以及性能最强的宗旨,进一步保证每个工作环节的增值最大化<sup>[6]</sup>。根据上述的高校工作业务员再造思想倡导的宗旨,并结合一体化结构组织原则,设计了一体化业务流程图,如图1所示。

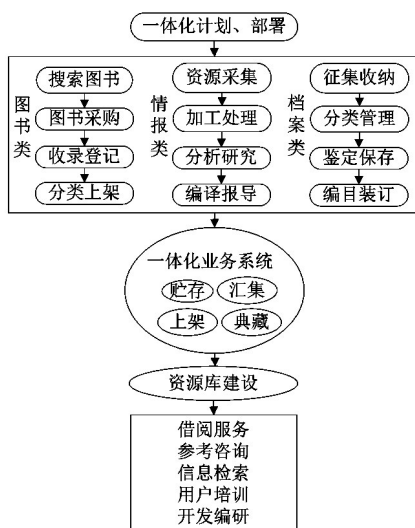


图1 一体化业务流程

## 3.3 高校组织之间基于协议的知识联盟

知识联盟主要指社会各界的组织机构以及单位系统为了更好地分享知识资源,加快知识交流以及实现知识创新,而利用各种协议和契约链接为一个团体,从而达到知识优势互补的目的。高校的知识联盟的主要链接方式为知识共享协议或者知识交换契约。通过知识联盟渠道,一体化机构可以直接从外界的高校知识环境中交换获得本校所需的文献资源,比如图书资料、档案资料以及情报资料等,从而更好地充实本校的知识体系。一体化机构还可以利用知识联盟的知识集约性,改进机构的知识吸收能力,加强资源结构调整能力以及追踪能力等,从而为知识整合提供便利条件。

# 4 基于语义关联的图书情报档案数字资源整合模式

若要实现基于语义关联的图书、档案以及情报等数字资源的整合,必须要处理好数据共享问题以及知识互操作问题

等。为了保证向用户提供一个统一、高效的知识发现机制,笔者尝试将OAI-PMH协议应用到以图书、档案以及情报等数字资源为主要内容的系统中,并将数字资源中的元数据进行集成,从而构建一个基于语义关联并且为用户提供统一、高效的知识服务整合模式<sup>[7]</sup>。具体来讲,本研究将语义关联的应用程序接口定义为HTTP格式,经过格式扩展后还能以Slash或者Hash格式转发。利用语义关联技术可以针对数字资源访问模式进行统一的标准化,也就是用户或者代理机构无需了解语义关联发布网站的运行模式、体系架构以及存储方式等内容,只需利用SPARQL技术根据Web服务器的IP地址进行访问即可。需要注意的是,基于图书、档案以及情报等资源的元数据存在两个方面的问题:(1) OAI-PMH协议只为用户提供基于Identifier等参数规定范围的收集服务,目前还不允许用户自行设定收集参数,比如用户不能按照作者或者资料语种收集,但这恰好是用户所熟悉的收集方式。(2) 在基于OAI的数据库中,每个元数据条目都有唯一的标识符,但是这个标识符并不能直接被HTTP识别,因而无法利用元数据条目直接收集数字资源。

因此,如果要利用语义关联技术实现关于图书、情报以及档案等数字资源的整合,首先要针对OAI-PMH元数据进行语义关联处理,也就是将OAI的数据库的元数据转化为具有语义关联性,从而帮助解决用户在收集元数据时遇到的技术障碍。

#### 4.1 OAI-PMH元数据的语义关联化

如果要实现基于OAI-PMH元数据的语义关联性处理,就需要参照语义关联的原则进行,以此来确定关于URL的分配方式、关联规则以及相应的关联信息生成方案等。接着根据OAI的存储特点利用基于URL收集技术来收集元数据,并把收集结果存储在本地的元数据库中,然后借助D2R等工具根据收集结果生成相应的映射文件,从而让存储在本地元数据库中的数据具有语义关联性。根据上述的原理,图书、情报以及档案等资源都可以将OAI-PMH类型元数据转化为具有语义关联性,然后利用URL就可以直接访问元数据资源,当然用户还可以参照SPARQL协议设定数据查询条件,从而实现针对元数据的有效检索。值得注意的是,一定要针对图书、情报以及档案等资源构建专门的OAI数据库,资源需要按照元数据的标准统一进行存储,只有这样才能够真正实现关于图书、情报以及档案等资源的语义关联处理。

#### 4.2 基于语义关联的图书情报档案数字资源整合模式

在关联开放数据项目的推进下,现在已经有超过140亿的传统数据转换为具有语义关联的数据。语义关联技术让图书、情报以及档案等数字资源的相互交流和共享更加便

利。现阶段,大部分的政府机构和组织机构已经认识到数字资源整合的重要性,并有意识地利用语义关联技术来实现数字资源整合的目的。当然,图书、情报以及档案等资源需要将资源之间的语义关联性数据存储起来,因此数据存储服务机构(比如云端服务提供商)将在数字资源整合服务占有更加重要的地位。

随着图书、情报以及档案等数字资源的数据量不断增长,急需针对这些数字资源进行整合,而利用语义关联技术以及相应的Web应用框架机制,并通过URL将不同类型数据资源进行语义关联,笔者根据用户和系统功能需要尝试设计一种基于语义关联的图书、情报以及档案等数字资源整合模式,具体如图2所示。

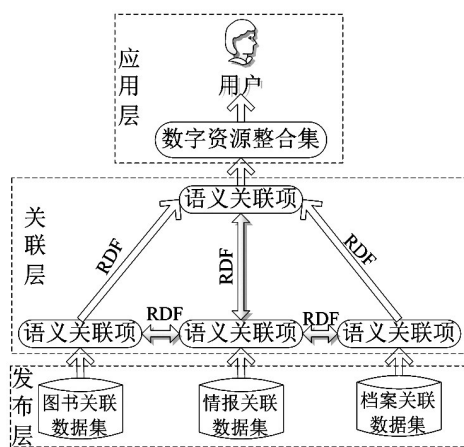


图2 基于语义关联的图书、情报、档案数字资源整合模式

从图2中不难看出,基于语义关联的图书、情报以及档案等数字资源的整合模式主要分为数据应用层、数据关联层以及数据发布层等。具体来讲,数据应用层的服务对象为SPARQL以及语义关联等相似的网络应用。例如国外数字图书馆已经尝试开发资源的语义检索服务,该服务可以根据检索词关联到更多的信息,如关联出题名、摘要以及作者等信息,从而为用户提供更有价值的检索信息;数据关联层的服务对象为图书、情报以及档案等数字资源内部存在关联性,比如一个作者可以有多部著作,一部电影可以有多个制作人等,可以利用RDF技术将这些联系进行关联,形成一个关联数据网络,不同的资源利用关联数据进行链接;数据发布层的服务对象为图书、情报以及档案等数字资源,主要以图像、音频、视频以及档案类资源为代表,设计人员可以将这些数据按照语义关联的原则进行处理并上传到网络中,从而让用户可以方便地进行学习和交流。需要注意的是,与传统的图书、情报以及档案等数字资源不同,在基于语义关联的整合模式下的资源都是按照语义关联原则进行描述的。

(下转第45页)



## 参考文献:

- [1] 唐淑香,李利.国内文献传递研究概述[J].图书馆工作与研究,2010(10):8-11.
- [2] 邹晓蕾.网络环境下高校图书馆文献传递服务发展对策[J].江西图书馆学刊,2012(2):66-69.
- [3] 黄丹.文献传递服务中存在的问题及对策——以武汉大学图书馆为例[J].图书情报知识,2006(2):53-55.
- [4] 杨雪萍,牛爱菊,刘兰.全媒体环境下高校图书馆馆际互借与文献传递服务宣传与推广路径——以北京师范大学图书馆为例[J].图书馆工作与研究,2016(6):100-103.
- [5] 肖景.提高文献传递满足率的对策[J].四川图书馆学报,2015(2):74-75.
- [6] 承欢.高校图书馆的文献资源建设怎样走出困境——兼论文献传递服务[J].大学图书馆学报,1994(4):4-7.
- [7] 赵立杰.论网络文献传递对信息资源建设的特殊意义[J].图书馆工作与研究,2008(6):49-51.
- [8] 杨坚红.CALIS CASHL NSTL 系统文献传递服务比较[J].情报科学,2009(1):83-88.
- [9] 李瑞芬,张晓青.国外网上文献传递服务系统的发展现状及特点[J].情报理论与实践,2007(5):714-717.
- [10] 祁卓麟,李其圣.百链云图书馆与高校文献传递服务对比分析[J].图书馆论坛,2013(1):43-45.
- [11] 张元莹.近五年国内馆际互借和文献传递研究的文献统计分析[C].国中小型公共图书馆联合会2014年研讨会,2014.4.
- [12] 卢纯昕.图书馆馆际互借与文献传递版权例外的立法构建[J].图书馆杂志,2016(5):26-32.
- [13] 魏艳霞.“读秀”图书搜索引擎资源及利用方式[J].中国科技信息,2007(6):153-154.
- [14] 吴云珊.读秀学术搜索与 Medalink 述评[J].农业图书情报学刊,2010(6):70-73.
- [15] 洪跃.读秀学术搜索系统述评[J].新世纪图书馆,2010(3):76-78.
- [16] 伍清霞.读秀知识库评述[J].图书馆论坛,2007(4):58-60.
- [17] 李幸.读秀学术搜索的信息组织探析[J].图书馆杂志,2012(4):29-32.
- [18] 杨军.读秀学术搜索与资源整合[J].科技情报开发与经济,2009(33):42-45.

石光莲 女,1989年生。硕士,助理馆员。研究方向:信息组织与检索。

侯艳 女,1983年生。硕士,馆员。研究方向:图书情报分析。

容军凤 女,1988年生。硕士,助理馆员。研究方向:现代图书情报技术。

(收稿日期:2016-09-29;责编:杨新宽。)

(上接第40页)

## 5 结语

笔者尝试利用语义关联技术来处理图书、情报以及档案等数字资源,借助于语义关联的链接性针对数字资源进行有效的组织和整合,还能够进行数字资源的深层联系挖掘和展示,并在互联网平台上为用户提供统一、高效的数字资源整合服务。如今人们对于信息的需求愈加强烈,借助于先进的信息处理技术进行数字资源整合符合其发展要求,而语义关联技术则为数字资源整合起到了非常重要的推动作用。

## 参考文献:

- [1] 丁楠,潘有能.基于关联数据的图书馆信息聚合研究[J].图书与情报,2011(6):50-53.
- [2] 李琳.关联数据在图书馆界的应用与挑战[J].图书与情报,2011(4):58-61.
- [3] 刘瑜.当代图书馆信息资源整合的若干模式[J].图书馆杂志,2010(3):8-41.
- [4] 楼白宇.公共图书馆图书档案情报一体化数字网站的实践效应[J].兰台世界,2013(4):99-100.
- [5] 贺德方,曾建勋.基于语义的馆藏资源深度聚合研究[J].中国图书馆学报,2012(7):79-86.
- [6] 肖希明,田蓉.国外公共数字文化资源整合的现状与发展趋势[J].国家图书馆学刊,2014(5):48-56.
- [7] 徐翠艳.网络环境下图书、情报、档案一体化建设研究[D].郑州:郑州大学,2013.

谭静 女,1978年生。学士,馆员。研究方向:高校图书情报、档案管理。

(收稿日期:2016-09-29;责编:姚雪梅。)