信息资源整合的建模与实现方法研究

章成志 苏新宁

(南京大学信息管理系 南京 210093)

【摘要】 信息资源整合是数字图书馆建设的重要基础工程,当前对信息资源整合理论与实现方法缺乏系统化研究。作者针对这一现况进行了初步研究,首先给出信息资源整合的基本框架,信息资源整合的三维模型及各个维度的含义和功能,然后给出信息资源整合实现的四层体系结构,分别说明了各层次的整合方法。接着在分析比较各层整合方法的基础上,给出信息资源整合的实施原则和方法

【关键词】 信息资源整合 整合模型 体系结构 数字图书馆 【分类号】 G250.73

Modeling and Method of Information Resources Integration

Zhang Chengzhi Su Xinning

(*Department of Information Management, Nanjing University, Nanjing 210093, China*)

【**Abstract**】 Resources integration plays an important role in the digital library development. The foundational frame, model of information integration, concept and function of dimensions are put forward definitely for the first time. The integration architecture and method are analysed from three levels. Based on comparing the integration method, the principle and method of implement are provided.

【**Keywords**】 Information resources integration Integration model Integration architecture Digital library

1 引 言

信息资源整合,即按照信息资源之间的知识关联进行优化重组,形成系统化的、智能化的数字资源体系,是解决这些问题的根本方法。

信息资源整合研究主要来自于计算机界的数据库研究者和人工智能研究者以及图情界。近年来,国外对信息资源整合的理论基础,信息资源整合方法和技术进行了相应的研究[1-4]。国内主要是图情界,对信息资源整合进行了探索,范围主要包括:相关资源整合工具、产品、系统的介绍、比较[5-7];数字资源整合理论研究[8-10];数字资源整合技术研究[11-13]。从目前的研究现状来看,国内外对信息资源整合的理论框架、基本模型与实现方法缺乏系统化研究。作者针对这一现况进行了初步研究,主要研究包括:给出信息资源整合的基本框架,信息资源整合的三维模型及各个维度的含义和功能;给出信息资源整合实现的四层体系结构,分别说明了各层次的整合方法;在分析比较各层整合方法的基础上,给出信息资源整合的实施原则和方法。

收稿日期:2005-05-19

2 信息资源整合的框架和模型

2.1 信息资源整合的框架

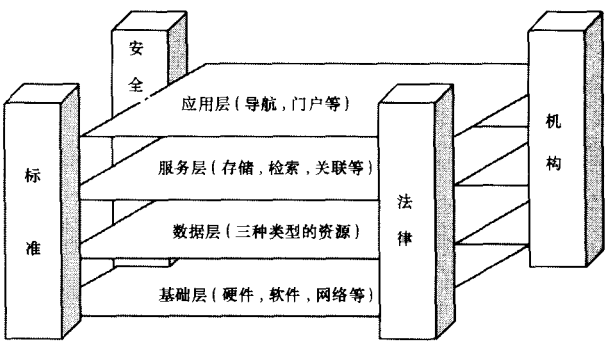

图1 信息资源整合的基本框架

信息资源整合的对象包括信息内容、信息系统、信息基础设施,其中,信息技术设施为信息资源整合的基础。为了保证信息资源整合的"资源共知、共建、共享"的目标,需要"机构、安全、标准,法律"为支柱。图1描述了信息资源整合的基本框架,它由四个支柱支撑和四个层次叠加组成。其中,两个支柱属于社会人文环境的组织、法律保障支柱,另两个支柱属于自然科技环境的安全和标准化支柱。四个层次是自下而上的,即:从基础层(包括硬件、软件及网络)、数据层(包括结构化数据、半结构化

数据、非结构化数据)、服务层(包括数据存储、信息检索、信息关联等)到应用层(包括学科导航、知识门户等)。本文根据该框架,对信息资源整合进行了建模。

2.2 信息资源整合模型的建立

信息资源整合从直观上来说,是指在统一的用户查询界面和检索结果的要求下,共享异构信息资源,为用户提供不同层次的知识服务。从系统角度来看,信息资源整合是将分散的异构系统中的异构信息资源,进行优化或重组,生成一个更加有序化、智能化、综合化的系统。生成的系统是一个逻辑上虚拟的系统或者一个实际的物理整合实体。

在以上整合思想的指引下,作者提出了信息资源整合的三维模型。图 2 给出信息资源整合的三个维度,分别为:资源维、服务维及应用维。图 3 给出信息资源整合的三维立方体视图。其中资源维显示了信息资源的类型,从资源整合范围来看,整合从结构化资源到半结构化资源,再到非结构化资源,整合的范围不断扩大;服务维显示了信息资源整合的资源利用效率,在"信息存储—信息检索—信息关联—知识发现"的信息服务过程中,资源处理的智能化不断增强,服务的效率不断提高;应用维显示了信息资源整合的资源应用层次,在"数据库—专题数据库—学科导航库—知识门户"的应用过程中,资源整合的层次不断提高。

2.3 信息资源整合模型的数学描述

设 Ω 为信息资源整合前的信息分布空间,Π 为信息资源整合前的信息分布空间,则信息资源整合的过程可表示为:从空间 Ω 经过 R,S,A 三个维度整合后,转换为空间 Π。记为:$f(R,S,A):\Omega\rightarrow\Pi$

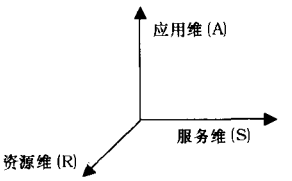

图 2 信息资源整合的三个维度

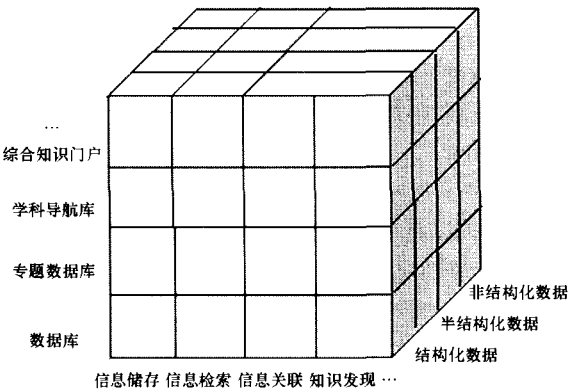

图 3 信息资源整合的三维视图

资源整合的程度表示为:$F = f(R,S,A)$ (1)

其中:

R:资源维变量,表明了信息资源整合的资源对象范围,亦即信息资源整合的广度。

S:服务维变量,表明了信息资源整合的资源利用效率,亦

表 1 信息资源整合的三个维度及状态分析

维度/状态 \ 项目		含义/功能	范围	关键技术	实例
R	结构化数据(R1)	具有属性-值格式的数据,提供描述结构化资源的格式	数据库,结构化标识语言等	数据库技术,信息组织技术	Oracle,XML,中国分类主题词表
	半结构化数据(R2)	部分附加属性-值格式的数据,附件部分提供描述信息样式	带标记的文档,经标引的多媒体信息等	数据标识,格式转换技术	TXT,HTML,PD,DOC 等格式文档
	非结构化数据(R3)	不具有属性-值格式的数据,描述动态的,非常规的信息内容	无标记的文档,未经标引的多媒体信息	信息采集,内容分析,特征提取,标引,摘要	音频,视频,流媒体
S	信息储存(S1)	按照特定格式进行存数据放	信息资源 R	数据组织,数据压缩	MARC,MPEG 格式
	信息检索(S2)	按照特定需求进行信息获取	信息资源 R	索引技术,语义匹配	图像,视频检索批
	信息关联(S3)	按照特定联系将信息互相链接	信息资源 R,目前主要为 R1,R2	连接分析,开放链接协议,知识元关联	Metalib & SFX, PageRank 算法
	知识发现(S4)	挖掘信息内部含义,发现新知识	信息资源 R,目前主要为 R1,R2	知识推理,智能 Agent,知识服务	文本挖掘,Web 挖掘
A	数据库(A1)	遵循某种范式的数据存储技术	信息资源 R 的子集	数据库优化,访问,控制	Oracle, DB2
	专题数据库(A2)	遵循某种标准的数据存储技术	信息资源 R,目前主要为 R1,R2	数据库选择,信息转换,信息去重	CALIS 专题数据库①
	学科导航库(A3)	遵循某种学科体系的信息储存和组织,信息表示技术	信息资源 R 的子集,	资源评估与选择,信息组织与分类	中科院化学学科信息门户②
	综合知识门户(A4)	遵循某种知识逻辑体系的知识表示,知识组织技术	信息资源 R	知识组织,语义网技术,知识服务	CNKI③(雏形)

(注:①http://www.calis.edu.cn;②http://chin.csdl.ac.cn/SPT-Home.php;③http://www.cnki.net)

即信息资源整合的深度。

A:应用维变量,表明了信息资源整合的资源应用层次,亦即信息资源整合的高度。

由此可得,信息资源在 R,S,A 三个维度上的整合分别体现了信息资源整合的三个目标,即:信息资源组织的有序化(或结构化)、信息资源处理的智能化、信息资源应用的集成化。因此,信息资源整合所要达到的最理想的结果应该是:在一定约束条件下(通常表示为用户的实际需求),信息资源整合的广度、深度及高度在"Ω→Π"转换中,达到一个最优点,即信息资源整合的程度达最优值,用数学模型表示为:

$$\text{Max} \quad f(R,S,A)$$

$$\text{s.t.} \quad R>0, S>0, A>0$$

由图 3 可以看出,信息资源的整合程度的最低点和最高点。信息资源整合程度的最低点,即最低层次整合的状态为:结构化数据存储在数据库中。信息资源整合程度的最高点,即最高层次整合的状态为:将非结构化、半结构化、结构化等三种数据,以专题数据库、学科导航库、知识门户等应用方式,为用户提供信息资源的知识关联、知识发现、知识推荐等服务。

3 信息资源整合的实现

3.1 信息资源整合实现的体系结构

根据信息资源整合的基本框架和三维模型,作者给出信息资源整合实现的体系结构,如图 4 所示。由图 4 可以看出,信息资源整合可以从三个层次,即:从数据层、中间层、表现层上进行整合。

(1)数据层整合

数据层,亦即基础层。数据层整合是一种物理整合方法,是在"数据大集中"思想指导下进行的信息资源整合方法。数据层整合是对现有的信息资源重新组织、深度加工和知识服务的过程。它需要建立一个新的存储仓库,将收集到的各种资源装入其中,不同结构的信息资源被组织为相同的数据格式,用统一的检索平台很容易的检索所有的信息资源。数据层整合的优势在于,经过该层整合后,便于进行数据的统一存储或迁移,便于在其上进行数据挖掘,进行分析和决策等。

(2)中间层整合

中间层整合是一种逻辑(或称虚拟)整合方法,可将它细分为"检索入口层"和"元数据层"两个子层次。其中,"检索入口层"整合针对同构系统(如不同高校使用的相同版本的汇文 OPAC 系统)而进行的一种逻辑整合方法,"元数据层"整合是针对异构系统(如汇文 OPAC 系统与 CNKI)而进行的一种逻辑整合方法。这两个子层次的整合方法的共同点是:整合本身不建立资源库,而是以代理的角色接受用户的请求,通过中间件技术把查询请求转换成相应信息系统的查询语言和检索方法,分别对各个检索系统发出检索请求,然后将各个系统返回的命中结果经过处理后在同一界面上呈现给用户。其中,元数据层整合方法通过开放链接技术(OpenURL),将不同信息资源之间建立元数据级的链接,这种信息关联极大地方便了需获取信息的用户。现阶段,元数据层整合只是提供信息的集成,不是本质上的整合。

(3)表现层整合

表现层,也为最高层。表现层整合也是一种逻辑整合方法,它是将各种应用进行集成,建立学科知识导航和知识门户,为用户提供整合服务。表现层整合实现相对简单,整合的结果比较清晰,便于鸟瞰学科和知识体系。

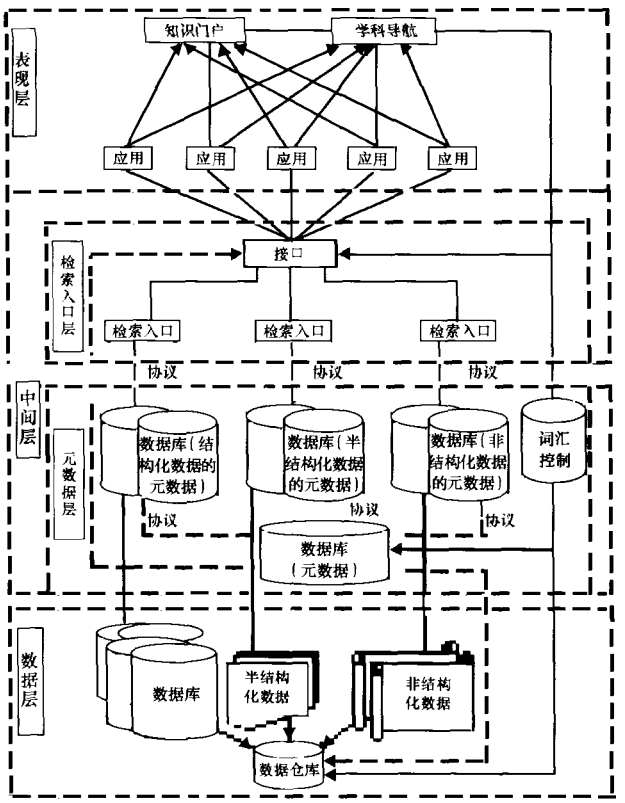

图 4 信息资源整合的体系结构图

3.2 各层次整合实现方法的比较

表 2 给出了各个层次上的整合的整合对象、整合结果、整合方法和主要技术。从表 2 可以看出,不同层次的信息资源整合方法都具有各自的优势和缺点。现阶段,三个层次整合方法都面临的问题是如何有效地将词汇知识(如同义词、上下位类词、词语翻译等)用于信息资源整合的控制中。在数据层,利用对词汇的控制可以建立信息资源在内容上的关联,在元数据层,通过本体的概念映射,可以得到元数据之间的语义关联,在检索入口层,通过对检索词汇的控制,可进行扩展检索或语义检索,最后,在表现层,词汇控制可以作为资源组织和关联的依据。

表 2 信息资源整合实现方法比较

项目 / 层次		整合对象	整合结果	整合方法	主要技术	实 例	优 势	缺点
表现层		各种应用,或服务	知识门户或学科导航界面	筛选资源,建立学科导航库,知识门户	资源评估与选择,可视化技术,个性化技术等	中科院化学学科信息门户①, SNTL 认知科学门户②	实现简单,便于鸟瞰学科体系和知识体系	表层的资源整合,不能挖掘深层的知识关联
中间层	检索入口层	同构系统中查询式,检索入口	元搜索(一站式检索)	中间件技术或者异构资源访问协议	Z39.50 协议,检索式的语义扩展,检索结果的数据归并排序等	CALIS 联合目录服务系统③, TPI USP④, TRS IIP⑤	建设周期短,更新容易	数据库规模增大时,检索效率明显下降
	元数据层	主要为异构系统中信息资源的元数据	信息的关联,相关资源排序(通过链接学习得到)	建立元数据数据库,建立元数据形式和语义上的集成	开放链接(如 Metalib&SFX),元数据语义映射,连接学习等	Web of Knowledge⑥(元数据关联), TPI⑦(元数据关联)	对知识元进行关联,进行知识评价,消除信息孤岛	实现方式复杂,现阶段只是提供信息的集成,不是本质上的整合
数据层		各种系统中存在的数据	有序化的数据存储于数据仓库中	将不同类型数据整合到统一的数据仓库中	数据采集,转换,去重,自动标引,分类,摘要,数据挖掘等	Web of Knowledge⑥(内部资源整合), DIPS⑧	数据的存储,知识发现,分析,决策等	实时更新困难,占用比较大的额外存储空间

(注:①http://chin.csdl.ac.cn/SPT-Home.php;②http://cogsci.nstl.gov.cn/SPT-Home.php;③http://opac.calis.edu.cn/;④http://202.204.32.151/USP/main/main.asp;⑤http://www.trs.com.cn/products/dls/trsiip/index.jsp;⑥http://www.isiknowledge.com/;⑦http://www.cnki.net;⑧http://www.gotodigit.com/home.htm)

在进行信息资源整合的实践中,应根据整合的对象,整合所要达到的效果和现有技术条件等多方面来考虑,选择一个可行的信息资源整合方法。现有的系统大多从一个层次上进行整合,已经不能满足大部分的实际需求,因此,综合几种整合方法,对信息资源进行多层次的、智能化的整合是今后信息资源整合的发展趋势。

3.3 信息资源整合的若干原则

(1)整体性原则 又称完整性原则,是指要保持数字资源对象学科的完整性。整合后的资源系统应涵盖各子系统内部功能,反映数据对象间的内在关系。

(2)针对性原则 又可称之为个性化原则,是指数字资源整合的目的性。整合后的数字资源应满足特定用户需求。

(3)层次性原则 是指数字资源整合的结构性。数字资源本身和用户需求的层次性,要求按多种类型、多种层次、多种方式进行多维整合。

(4)科学性原则 是指对数字资源的整合对象、内容、方式、要进行科学论证,切忌随意拼合。

(5)最优化原则 又称优化性原则,是指运用一定的技术手段和方法,使数字资源得到合理组合,取得最好的组织结构和组织功能。

(6)动态性原则 又称开放性原则,是指整合系统是个开发系统,它并不是永恒不变的,而是与数字资源及用户需求等环境有着密切的联系,并且随着外界环境的变化而不断变化,从而显示出系统整体功能的开放性和进化性。

4 结 语

信息资源整合是数字图书馆建设的重要基础工程。信息资源整合的目的是为了解决数字图书馆建设中存在的信息重复和冗余、信息孤岛、知识关联程度低等现象。在进行信息资源整合的实践中,应在信息资源整合的原则下,根据整合的对象,整合所要达到的效果和现有技术条件等多方面来考虑,选择一个可行的信息资源整合方法。对信息资源进行多层次的、智能化的整合是今后信息资源整合的发展趋势。

参考文献:

1 Pearce, Judith, Warwick Cathro and Tony Boston. 2000 "The Challenge of Integrated Access: the Hybrid Library System of the Future". Books and Bytes: Technologies for the Hybrid Library. Melbourne, Victorian Association of Library Automation Inc., 2000

2 Peter Haddad. Integrating Digital Resources into the Library Information Infrastructure. 14th National Cataloguing Conference, Geelong Waterfront Campus, Deakin University, 2001(11):7-9

3 Grahne, GÃÃsta; Kiricenko, Victoria. Towards an algebraic theory of information integration. Information and Computation. 2004, 194(2):79-100

4 Rousset, Marie-Christine; Reynaud, Chantal. Knowledge representation for information integration. Information Systems. 2004, 29(1):3-22

5 范爱红,姜爱蓉. 基于知识管理的学术信息资源整合体系——对 ISI Web of Knowledge 的评介. 现代图书情报技术. 2001(6):43-46

6 李富玲,卢振波. SFX——信息资源整合新工具. 现代图书情报技术. 2002(6):69-71

7 王平,姜爱蓉. 国内外数字信息资源整合管理系统的对比研究与思考. 上海交通大学学报, 2003, 37(S):164-170

8 马文峰. 数字资源整合研究. 中国图书馆学报, 2002, 28(4):64-67

9 黄晓斌,夏明春. 论图书馆数字资源的整合. 图书情报工作, 2005,(1):50-53

10 黄晓斌,夏明春. 数字资源整合研究的现状及发展方向. 图书情报工作, 2005,(1):75-77

11 金更达. 网络资源界面整合和 Agnet 界面实现探讨. 大学图书馆学报, 2002, 20(1):30-34

12 崔宇红,刘涛. 图书馆数字资源与 OPAC 系统的整合. 图书馆杂志, 2003,(1):55-56

13 张文德,戴晓翔. 信息资源整合系统与技术研究. 现代图书情报技术, 2003,(6):72—73,71

(作者 E-mail:zcz51@citiz.net)