

●黄晓斌 钱国富

## 数字图书馆开发平台及其发展趋势\*

**摘要** 数字图书馆开发平台具有创建、获取、存储和管理数字资源、管理用户权限和资源查询、以及信息发布与服务等功能。目前国内已开发了支持非结构化文件系统、支持全文检索系统、支持元数据管理、支持异构数据库检索及出版发行业务等功能的开发平台。数字图书馆开发平台的发展趋势,是逐步走向标准化、功能整合化,以及开放源码软件与商业软件并行发展,建立完善的用户权限控制机制和不断增强信息服务能力等等。参考文献 17。

**关键词** 数字图书馆 开发平台 发展现状 发展趋势

**分类号** G250.76

**ABSTRACT** The authors summarize the functions of digital library development platforms, and the current status of such kind of platforms in China. Then, they discuss the future trends of digital library development platforms, such as standardization and function integration. 17 refs.

**KEY WORDS** Digital library. Development platform. Development status. Development trends.

**CLASS NUMBER** G250.76

数字图书馆平台是开发和利用数字图书馆的管理系统。数字图书馆建设的基本要素包括数字化资源、分布式管理以及智能化服务等方面,数字图书馆开发平台需要解决的就是如何通过计算机、网络以及软件系统来构建这些基本的要素,从而开发出一个合理高效的数字图书馆<sup>[1]</sup>。

### 1 数字图书馆开发平台的基本功能

一般而言,数字图书馆开发平台需要具有以下五大功能:数字资源的创建和获取、数字资源的存储和管理、用户的权限管理、数字资源的访问和查询以及信息发布和服务。

#### 1.1 数字资源的创建和获取

数字图书馆开发平台应当能够支持通过计算机来获取已经数字化的文章、图片、录音、录像等多种来源的信息,支持通过扫描、识别、压缩和转化等多种技术来创建数字信息。数字图书馆开发平台应该支持通过开放的内容创建应用程序接口和其他厂商的相关技术产品来完成上述不同种类信息的数字化及内容的提取。此外还应提供多种工具和软件包来支持内容创建图像校正和图像转置、协调、编辑和色彩校正,以及 MPEG 压缩等特殊功能。可以采集网上的信息资源和建立特色数据库等功能。

#### 1.2 数字资源的存储与管理

数字资源的存储与管理包括自动索引、建档、特性抽取和翻译功能,利用先进的组织和筛选工具,用户能够用中文语言进行特定信息的查询。使用全文数据库来存储数字资源,并进行数字资源的索引和查询。数字图书馆开发平台应当能够综合利用包括全文数据库技术、面向对象的技术和多媒体等技术来为数字图书馆用户提供实用性强、完整性较好和安全性高的数字图书馆解决方案。此外数字图书馆还应提供独立于内容之外的数据管理工具,使其能够具有对多种系统的操作能力。

#### 1.3 用户权限管理

数字图书馆开发平台应该能够根据完整的用户权限管理方案来提供一系列全面的管理工具,包括对数字资源建设过程中的用户权限管理,对用户访问和使用数字资源进行许可、控制和监督,并保护资源拥有者和最终用户相关利益等功能。如在开发平台中集成数字水印技术等版权保护技术,来实现附加版权信息、注册商标的水印嵌入等操作,通过用户管理中的计费模块来实现对数字资源利用的计费、结账等功能<sup>[2]</sup>。此外,由于数字图书馆是基于互联网来进行信息发布和利用的,因此开发平台一般需要有加密信封等加密技术,来保证数字信息在网上传递过程中的安全性、完整性。

#### 1.4 数字资源的访问与查询

数字图书馆开发平台一般都具有强大的访问控

\* 本文为教育部规划项目(01JA870013)的研究论文之一。

制以及信息查询功能,包括文本和图像分析工具以及数字化音频和视频信息的查询工具,提供全文检索、基于声音和图像的检索以及自然语言检索等多种检索方式。用户可以根据自己的喜好和用途来选择一种或多种检索手段,而且要求检索的响应要及时,结果要准确。

### 1.5 信息发布与服务

数字图书馆开发平台应当支持用户选择从现有的任何计算机网络系统上来进行信息发布与服务,支持包括传统的 C/S 到 B/S 在内的多种信息发布和服务途径,并且所发布的信息在任何具有图形化用户界面的计算机系统上都可以进行呈现和阅读,对于特殊用户还应当支持包括通过触摸屏、手写及语音识别等技术来提供信息服务,使得系统做到对用户透明,并具有良好的安全性、易用性和可扩展性。数字图书馆开发平台还应当能够通过多种多样的信息服务手段来进行个性化的信息服务,如各种的数字化参考咨询服务和电子文献的传递提供服务。

除上述之外,数字图书馆开发平台的功能应当具有扩展性,满足图书馆工作的不同需要,如支持用户教育、资源共享与业务合作、安全与系统管理等。

## 2 国内数字图书馆开发平台的研发现状

随着数字图书馆研究、建设的发展,越来越多的公司开始推出自己的数字图书馆开发平台软件,目前主要有:TPI—清华同方数字图书馆建设和管理平台、超星数字图书馆软件、麦达数字图书馆应用管理软件、方正 Apabi 数字图书馆软件、天宇数字图书馆解决方案、金鑫数字图书馆解决方案、联想大型综合类数字图书馆解决方案以及较早的 IBM 数字图书馆应用平台和 Horizon 数字图书馆系统等。开发平台的类型和数量增加,市场竞争的进一步加强,扩大了图书馆在进行系统选型时的选择空间,而且竞争也有利于激励企业进行技术创新,不断开发更好的产品。但是另一方面,如果竞争是无序的,缺乏一定的标准,各个厂家各自为政,那么也不利于数字图书馆的发展。

目前数字图书馆开发平台的基本情况大致如下:

### 2.1 支持非结构化文件,包括音频、视频等多媒体文件的管理

数字图书馆在资源上是多种多样的,不仅有文本,还有音频、视频等多媒体资源,如何对这些资源进行有效的整合和管理,是数字图书馆开发平台需要解决的一个重要问题。以目前国内用户最多的 TPI—

清华同方数字图书馆建设和管理平台为例,它是为大中型信息服务机构量身开发的基于非结构化文档管理的大型智能化内容管理系统。该系统以 FTS 全文检索数据库为核心,采用 B/S 浏览器检索方式和三层 C/S 架构,能够同时管理文字、图片、多媒体等信息,并提供全文检索,支持网页的动态发布<sup>[3]</sup>。

### 2.2 支持全文检索

检索功能是数字图书馆中使用频度最大,也是技术要求最高的功能之一。目前各个数字图书馆开发平台都宣称自己支持全文检索。以 TRS 数字图书馆软件平台为例,该公司由于较早开展了在信息检索、中文信息处理以及数据挖掘和知识管理(自动分类、自动聚类、自动摘要、相似性检索)等方面的研究和开发,并拥有了自己的专利技术,在数字图书馆开发平台中以领先的 TRS 内容管理平台为核心基础,因此对全文检索的支持做得相当不错,支持 PDF 等其他格式文件的全文检索等<sup>[4]</sup>。

### 2.3 支持元数据管理

目前各个开发平台均支持元数据管理,在资源标引和整合过程中采用元数据来进行封装,这样有利于不同系统之间的数据共享,可以减少重复建设和资源浪费。上海金鑫数字图书馆解决方案中就内置了元数据管理组件(Metadata Management Component),通过该组件可以灵活定义数字图书馆整体元数据方案,支持对数图知识体系、资源对象的添加删除等综合管理,并可基于关系型数据库,实现元数据封装文件的存储和查询<sup>[5]</sup>。此外还可单独作为元数据存储工具,实现任意结构的 XML 文件到元数据库的映射存储。

### 2.4 支持异构数据库检索

通过提供对异构数据库检索的支持,对多个不同结构、不同使用规范的数据库进行统一、并发的检索,可以免除用户在不同数据库之间来回穿梭的麻烦以及用户熟悉各个不同数据库操作和使用规范的压力,节约用户时间,更好地为用户服务。目前在这方面做得较好的是 TPI—清华同方数字图书馆建设和管理平台。该平台可以选配异构统一检索平台 USP 模块,该模块可以对多个异构数据库进行统一检索,检索结构在统一的界面中输出,检索方便、快捷,结果准确,并且能够方便地加入新数据库和修改数据库的配置以及访问权限控制等。该平台具有良好的通用性,它不要求被用来统一检索的数据库必须支持 Z39.50 或者 OAI 协议,也无须数据库厂商提供专门的接口,

它采用通用的 HTTP 协议进行网络数据库的统一检索,覆盖面广<sup>[6]</sup>。

## 2.5 对于出版发行业务的支持

目前各个开发平台均提供了对出版业务的支持,如清华同方就提供了专门的电子书制作工具 E-Book。北大方正的 Apabi 电子书/电子书发布平台,获得了完全版权,构筑了完整的电子书产业链,采用完全还原版式技术,可以做到“越大越清晰”,采用 DRM 技术进行版权保护,有效地保护知识产权<sup>[7]</sup>。

此外在用户管理、用户交流、个性化信息服务等方面,目前各个数字图书馆开发平台也都取得了一定的成果。由于各个开发平台的侧重点不同,所以在功能模块上也有差异,形成功能互补的状况。如 TPI—清华同方数字图书馆建设和管理平台共有七大功能模块,包括数据库制作管理与发布系统、异构统一检索平台 USP、数字参考咨询系统、互联网信息资源整合系统、辅助教学工具、个性化信息服务系统 PIS、以及其他的专项工具等。用户可以按需定制,具有良好的可拓展性<sup>[8]</sup>。TRS 数字图书馆的解决方案包括集数据加工标引、建库、发布和检索服务为一体的信息资源(元数据或全文数据)库建设解决方案;集用户管理、认证计费 and 个性化服务为一体的资源发布服务系统;实现异构、分布式、跨平台、多样化资源统一检索和参考链接服务的资源整合门户;基于国际标准 ILL10160/10161 的馆际互借系统、分布式虚拟参考咨询服务系统;以及面向高校论文库建设、教参库建设的应用系统等等<sup>[9]</sup>。

## 3 数字图书馆开发平台的发展趋势

未来图书馆的主要形式将是数字图书馆和传统图书馆有机结合的复合图书馆,数字图书馆开发平台也应当着眼于这一大的发展趋势,在立足于数字资源的收集、加工、管理、利用的同时,为利用现代技术提升传统图书馆的服务水平提供条件。

### 3.1 标准化

在国家标准中将标准化一词定义为在一定的范围内获得最佳秩序,对实际的或潜在的问题制定共同的和重复使用的规则的活动。标准化的范围和对象是在经济、技术、科学和管理社会实践中重复性事物和概念。数字图书馆开发平台的标准化是指开发平台在所采用的数字资源建设和服务规范的标准化,具体而言就是在数字资源加工、组织上的标准化,在元数据方案上的标准化,资源的标识符以及指示系统的

标准化,采用国际通行的资源标识符,在数字资源检索与应用上的标准化,开放接口的标准化,避免各自为政,便于不同系统之间的资源整合和共享,也为数字资源的长期利用创造条件,避免因系统所采用标准的时效性而导致资源的失效。目前由科技部科技基础性工作专项资金和科技基础条件平台专项资金重点支持的《数字图书馆标准与规范建设》项目已经取得初步成果。中国科学院文献情报中心、中国科学技术信息研究所、国家图书馆等 20 多家单位在国家科技图书文献中心统一协调和组织下,共同参与了研究,目前已完成《数字图书馆建设中标准规范应用指南》、《基本元数据规范》、《元数据扩展规则》等近 20 多份标准规范初步方案,将逐步进行各个标准规范的开放试验。为实现数字图书馆系统的可广泛使用性、可互操作性以及可持续性要求,项目围绕数字资源的生命周期,制定了数字图书馆核心标准规范体系,建立了数字图书馆标准规范开放应用机制等一系列任务<sup>[10]</sup>。当前各个开发平台都在关注异构数据库之间的统一检索问题,但是由于个别数据库在检索式的构造以及检索接口上没有采用通用的标准,导致系统的互操作性很差,影响了系统的二次开发和读者的使用。可以肯定,随着数字图书馆建设中标准规范建立,数字图书馆开发平台的兼容性和可互操作方面将会进一步加强。

### 3.2 功能的整合化

数字图书馆需要开放集成各种分布、异构和多样化的数字信息资源和数字信息资源服务系统,动态构建满足各种用户群和业务流程需要的信息服务系统。这包括对分布和多样化的资源和服务系统进行搜寻、调用和集成,可以搜寻、解析和转换数据资源,可以支持和其他信息系统的互操作和集成管理<sup>[11]</sup>。目前的图书馆网站都是于自动化系统之外独立开发的,与馆内的 OAPC 资源以及其他数字资源也没有完全融合到一起。用户在访问图书馆的网站时,一般需要访问多个入口,比如查询书目需要通过 OAPC,查询数据库和使用其他的 Web 服务也是如此。但是对于那些使用惯了类型 Google 和百度等搜索引擎的读者而言,他们已经习惯于通过统一的检索界面,介绍不同类型、不同载体形态的资源,同样地,他们也希望图书馆能够做到这一点,能够在在一个系统内部检索图书馆的全部馆藏资源,而不用去理会所需的是纸质印刷品、光盘、磁带还是新的数字资源。但是传统图书馆自动化系统和新的数字图书馆开发平台之间共享的

数据很少,即使是数据库之间也未能实现统一的、高效的跨库检索,未来图书馆要想提供门户式的一站式服务,要么需要对图书馆的自动化系统进行技术改造和升级,要么就是新的开发平台里对原有的馆藏资源进行重复的元数据建设,前者有技术门槛和壁垒,后者则是对图书馆原有工作的浪费。因此对于未来的复合图书馆而言,迫切需要的是一个能够把图书馆的用户、馆员、管理者联系在一起,并对图书馆的印刷资源、数字资源和其他类型的馆藏资源进行统一管理的系统,这个系统要求不仅能够便于图书馆员和读者使用,而且要能够对图书馆的各项业务流程和馆藏资源进行有效的组织与管理。许多开发平台都在考虑如何将图书馆集成系统和数字图书馆的功能整合起来,提供更加全面的服务。如深圳图书馆拟在原来的图书馆自动化集成系统(ILAS)基础上增加数字图书馆的其他功能;上海金鑫公司的“数字图书馆系统软件平台”将按照现代图书馆五大核心职能(馆藏、教育、科研、情报、娱乐)的建设导向,结合丰富的数字图书馆系统开发和实施经验,提供各类基于因特网技术、合理组织大规模数字化资源的配套解决方案<sup>[12]</sup>;TPI—清华同方数字图书馆建设和管理平台目标是能够同时管理文字、图片、多媒体等信息,并提供全文检索服务,支持网页的动态发布,成为一个面向内容管理的最佳应用、管理和信息发布的工具,提供从印刷品到电子文档、从静态资源处理到网络资源实时整合、从单一资源管理到集群检索服务,从单一数据库到异构数据库的统一检索<sup>[13]</sup>。

### 3.3 开放源码软件和商业软件并举

目前的数字图书馆开发平台大多是由商业性的企业团体开发的,图书馆需要通过支付一定的费用来购买,才能获得并使用该软件。商业软件本身并无可厚非,但是在图书馆经费预算不断萎缩,而信息资源价格逐年上涨的情况下,购买商业软件的费用就显得很宝贵了,尤其是对于经费有限的中小型图书馆而言。此外商业软件出于知识产权考虑,一般在使用许可上有很多限制,图书馆要想进行二次开发以及重复利用都需要和软件开发商协商甚至额外支付费用。开放源码软件(Open Source Software,以下简称开源软件)兴起,为图书馆在数字图书馆开发平台的选择问题上提供一条新的解决途径。

1983年,为了反对软件所有权的私有化,Richard M. Stallman 创立了自由软件基金会,通过 GNU 计划和 GPL(General Public License)来推广开源软件。开

源软件就是在开放源代码许可证下发布的软件,以保障软件用户自由使用及接触源代码的权利,它也保障了用户自行修改、复制以及再分发的权利,开源软件不仅开放软件的源代码,还包括了应用软件、开发工具、数据和文档的开放。开源软件的发展改变了商业软件一统天下的单一局面,通过竞争和协作,开源软件和服务的总体质量得以改善,并逐渐获得认可<sup>[14]</sup>。如联合国教科文组织于2001年正式接纳开源软件,并在发展中国家宣传推广开源软件的应用,OCLC也将它组织开发的 Site Search 作为开源软件进行发布。

目前在数字图书馆领域的开源软件已经很多,如由美国麻省理工学院(MIT)和惠普公司(HP)实验室联合开发的 DSpace 系统。VDC(Virtual Data Center)是由哈佛大学等多家研究机构联合开发的一个完全开放源代码的数字图书馆系统,该系统还提供了对分布式和协作馆藏资源的广泛支持,包括资源命名空间,分布式的认证和访问权限控制,元数据获取,远程知识库存储等<sup>[15]</sup>。

Fedora 是开放源软件,用 Java 编写和支持 Windows、Linux/Unix 等操作系统,具有多种数据库支持,由美国的弗吉利亚和科奈尔大学开发。其后由 Mellon 基金会支持 VTLS 提供软件包的进一步完善和技术支持。设计特点包括:提供 API 接口来添加和访问资源;采用数字对象框架(DOA)进行整体设计。它可以用在很多地方,包括数字图书馆、内容管理、数字资产管理、资源保存等方面。Fedora 实现了数字对象框架,能够统一地存储管理各种属性不同的数字资源;数字对象使用 METS 编码存储成 XML;实现了 OAI 协议;提供的管理和访问服务作为 Web Service 进行发布。清华大学数字化项目中的建筑史数字图书馆、数学史数字图书馆、机械史数字图书馆以及正在进行建设的科技史数字图书馆采用 Fedora 开放源软件,并且通过设计并实现了可视化的配置管理工具来完成 Web 服务;通过设计虚拟馆藏配置文件,实现了基于不同的资源提供不同的服务;设计并实现了外建索引自动创建和自动更新模块,为不同的数字资源建立了不同的索引;设计并实现了检索服务模块,实现了检索界面和检索结果界面的定制和自动生成,基于不同的数字资源提供了不同的检索服务;设计并实现了 METS 数据著录模块;从而提供了元数据著录、资源管理、Web 发布的完整解决方案<sup>[16]</sup>。

### 3.4 完善的用户权限控制机制

现有的数字图书馆开发平台在数字资源的用户

访问权限的控制上,一般都采取 IP 地址段过滤来作为身份认证和使用授权方式。这一认证访问方式在有效保护数字资源知识产权的同时,也带来了新的问题。随着研究者活动范围和交流空间的扩大,一旦用户离开了规定的 IP 地址段后就不能使用原本所能利用的资源。因此用户迫切希望能够有一个更加灵活完善的用户权限控制机制,使得他们可以随时随地都能够获取和利用所需的资源。

中国科学院数字图书馆(CSDL)新推出的随易通(Access Anywhere)服务为这一问题的解决提供了一个崭新的思路。随易通是中科院数字图书馆分布移动用户认证系统的简称,它为中国科学院的用户提供了通过用户名/密码的认证方式访问 CSDL 开通的网络文献数据库的途径,即其注册用户到外地出差、参加会议和在家时均可通过用户名/密码访问 CSDL 网络数据库。随易通提供 2 种认证方式,分别是:常规登录(用户名+密码)以及 E-Key 用户(E-Key+用户名+密码)。这里所指的 E-Key 是一种形状类似 U 盘的智能存储设备,用于存放识别随易通用户身份的数字证书,内有 CPU 芯片,可进行数字签名和签名验证的运算,外形小巧,可插在电脑的 USB 接口中使用。由于 E-Key 具有存储信息不可读取、导出的特征,安全性高,用于身份识别可有效防止用户账号被窃取、散发,从而保证了系统认证的可靠性和安全性,在保护了知识产权的同时,也方便了用户<sup>[17]</sup>。

### 3.5 更强的信息服务能力

信息服务能力的水平也是衡量一个数字图书馆系统功能是否完善和强大的重要指标。在数字图书馆开发平台的未来发展中,应当充分利用新技术、新方法来提高数字图书馆的信息服务能力,提供个性化服务,如提供文献传递、订阅和推送信息的功能;系统能够根据用户的需求过滤筛选和推荐有关信息,主动发送用户需要的信息;提供多种形式的数字参考咨询服务;提供基于网络实时交互技术的远程教学服务;提供功能齐全的个人数字图书馆服务等。又如可以在系统中扩充内容聚合(RSS)功能,根据用户需求利用 RSS 方式动态聚合各类信息来实现信息主动推送服务。比如可以利用 RSS 来实现传统的定题检索功能,当用户检索数据库时,系统可以根据用户的检索词自动生成相应的 RSS 订阅地址,用户在客户端

订阅了该 RSS 链接之后,就可以通过定期更新来获取自己所需检索的最新信息。由于 RSS 将传递从数据库中检索到的新记录,原有的搜索结果也可以在客户端长期保留,一来可以把用户从数据库检索界面中解脱出来,二来也免除用户进行查重的问題,充分节约读者时间和精力。

### 参考文献

- 王大可. 数字图书馆应用平台的研制与开发. 艺术科技, 2002(4)
- [美] William Y. Arms 著;施伯乐等译. 数字图书馆概论. 北京:电子工业出版社,2001
- 孙书平. 清华同方数字图书馆管理与建设平台 TPI 试用报告. 图书馆学研究,2004(3)
- 4,9 TRS: 数字图书馆解决方案. <http://www.tris.com.cn/solutions/sztsg/> (Accessed Oct. 30. 2004)
- 5,12 上海金鑫计算机系统工程有公司. 数字图书馆解决方案. <http://www.finegold.com.cn/sztsg-1.htm> (Accessed Oct. 30. 2004)
- 6,8,13 TPI 数字图书馆建设与管理平台. <http://tpi.cnki.net/product/product/product01.htm> (Accessed Oct. 30. 2004)
- 7 北大方正电子有限公司. 方正 Apabi 网络出版(e-Book)系列解决方案. [http://www.apabi.com/project/pro\\_1/pro\\_a.htm](http://www.apabi.com/project/pro_1/pro_a.htm) (Accessed Oct. 30. 2004)
- 10 陶世安. 全国数字图书馆标准与规范建设项目成果显著. [http://news3.xinhuanet.com/st/2004-05/28/content\\_1495451.htm](http://news3.xinhuanet.com/st/2004-05/28/content_1495451.htm) (Accessed Oct. 30. 2004)
- 11 刘科等. 基于 Java 技术解决数字图书馆异构平台数据库访问. 现代图书情报技术,2002(3)
- 14 毛军. 开放源码软件在科学数字图书馆建设中的应用. 图书情报工作,2004(4)
- 15 DSpace. <http://www.dspace.org/> (Accessed Oct. 30. 2004)
- 16 郑晓惠. 开放源代码内容管理系统. <http://lib.utsz.edu.cn> (Accessed Oct. 30. 2004)
- 17 随易通常见问题问答. <http://q.csdl.ac.dn/doc/FAQ.htm> (Accessed Oct. 30. 2004)

黄晓斌 中山大学信息管理系副教授、博士。通信地址:广州市新港西路中山大学。邮编 510275。

钱国富 中山大学信息管理系硕士研究生。通信地址同上。(来稿时间:2004-12-18)