

理论园地 ·

我国图书情报档案界 XML 研究现状综述

李 琨 (琼州大学图书馆 海南五指山 572200)

【摘要】XML 自 1996 年推出至今已经引起了各个领域的关注,图书情报档案界也有不少专家和学者对 XML 进行了研究和探讨。本文主要借助中国期刊网对图书情报档案领域中关于 XML 研究的现状进行了综述。

【关键词】XML 数字图书馆 元数据 数据库

【中图分类号】G250 **【文献标识码】**A

【文章编号】1003 - 6938(2004)05 - 0019 - 03

XML 全称为 Extensible Markup Language,中文可译为“可扩展标识语言”,它是由 SGML(标准通用标记语言)发展而来的一种简单灵活的文本形式的标记语言,可提供跨平台、跨网络、跨程序语言的数据描述方式。自 1996 年推出至今,得到了如网络服务、电子商务等多方面的关注,造成了很大的影响。各领域对这种技术迅速做出反应,图书情报档案领域也有不少学者开始投入了对 XML 的研究。本文将主要对图书情报档案界对 XML 的研究现状进行综述。

笔者检索中国期刊网全文数据库,以“XML”为检索词,统计“电子技术及信息科学辑”中研究 XML 问题论文的数量,得出下表。

查询范围	检索字段			
	篇名	关键词	中文摘要	引文
图书情报学、档案学及博物馆学	60	79	67	97
其它	578	916	743	830

备注:1. 单位为“篇”。

2. 检索日期为 2003 年 9 月 17 日。

从上述统计数据我们可以看出,与信息技术领域相比,图书情报档案界目前的研究还处于起步阶段,研究论文的数量比较少。另外我们在检索中还发现目前档案界几乎还没有人进行过该问题的研究。

通过浏览以“篇名”为字段检索到相关论文的目录和内容,得出与数字图书馆有关的论文 22 篇,与信息检索有关的论文 23 篇,与原数据有关的论文 6 篇,与数据库有关的论文 2 篇,其它 7 篇。

这说明目前图书情报档案界对 XML 的研究主要集中在该技术在数字图书馆、信息检索、元数据方面的关系及应用,研究的视角不多,研究的深度较浅,其具体研究内容详细阐述如下:

1. XML 的起源及特征

虽然几乎没有人专门撰文对 XML 的起源和特征进行叙述,但在这些研究论文中都或多或少地对 XML 的起源和特征进行了阐述,观点基本一致。随着 Web 的发展和 Java 的出现,从 1996 年 W3C 就开始派专门的工作组致力于设计一个超越 HTML 能力范围的新语言,该语言后来就被命名为 XML。人们一致认为,XML 是可扩展置标语言,是一种基于 SGML 的简单灵活的语言。确切地讲,其本身并不是一个单独的置标语言,而是一种元语言,是一种能够创建标记语言的语言。XML 也源自 SGML(标准通用标记语言,1986 年成为 ISO 标准),具有 SGML 的 80% 的能力,但是其复杂性只有 SGML 的 20%。

XML 将 SGML 的灵活性和强大功能与已被广泛采用的 HTML 结合起来。它略去了 SGML 中使用频率低的一些内容,重新定义了 SGML 的内部值和参数,并删去大量繁复的、不常用的特征,从而使编程简化,易于在 Web 上传输和交互。同时 XML 保留了 HTML 所具有的简洁性和适合网上传输和浏览的优点,克服了 HTML 的种种不足,将网络上传输的文档规范化,并赋予标记一定的含义。^[1]众位学者虽然在论述其特征时,存在诸多表达方式,但基本上都围绕简单性、自描述性、可扩展性、互操作性和开放性等几个方面来总结 XML 的特征。XML 文档是建立在一个基本嵌套结构的核心集基础上的,这些基本结构可以被用来代表复杂

的信息集合,而不需要改变结构自身;XML 允许创建用户自己的 DTD,又可以根据自己的需要任意地扩展 XML,由于这种扩展机制是标准的,所以可以自动地将扩展信息传递给任何读取我们数据的人;XML 可以在多种平台上使用,而且可以用多种工具进行解释;XML 支持用于字符编码的许多标准,可在全世界许多不同计算机环境中使用;在 XML 中,由于数据和显示样式是分离的,数据信息更易于访问,可重用性好,并且已建立好的 XML 标识可以为其它 XML 程序使用,数据的提取利用更加灵活和方便。

2 XML 与 HTML 和 SGML 的联系与区别

XML 的诸多技术优势在一定程度上是人们通过将其与 HTML 和 SGML 进行比较分析总结出来的。这三者之间具有非常密切的联系,但也存在较大的区别。施燕斌、刘春红介绍了三者的区别和联系,“从本质上看,XML 和 HTML 都是由 SGML 派生来的,但是 XML 是一种元标记语言,而 HTML 是一种特殊化的标记语言。XML 将 SGML 的丰富功能与 HTML 的易用性结合到 Web 的应用中,并保留了 SGML 的可扩展功能,这使 XML 从根本上有别于 HTML。”^[2]张咏则认为:“和 HTML 相比,XML 的元素不是事先定义的,允许客户定义他们自己的文件元素集合,同时也可以指示这些元素在屏幕上如何按指定的要求表现。”“XML 也不同于 SGML,SGML 太过复杂、软件支持也较昂贵,XML 保留了 SGML 的可扩展性和适用性,但作了简化。XML 可被看作是 SGML 一个简化的子集,XML 规则的容量仅不到 SGML 的 1/10,这大为方便了 XML 的推广和使用。XML 与 SGML 有着密切的关系,一个证明是有效的 XML 文档同样是一个有效的 SGML 文档。”^[3]王伟专门撰文从介绍标记语言的发展历史入手,对 HTML 和 XML 进行了比较分析。林甫则以三者都是可应用于数字图书馆中数据交换与处理的数据格式的标识语言为切入点来分析了它们的异同性。

3 XML 对数字图书馆的影响及其应用

通过上面的分析我们也发现目前图书情报界对 XML 的研究论文中与数字图书馆有关的就占了 37%,可见目前人们对 XML 对数字图书馆的影响以及应用已经引起了人们的关注。总体来说,XML 在数字图书馆中的应用都是基于其优势性特征的,主要体现在 XML 强大的语法功能、互操作性和开放性、可扩展性、数据存取与显现可分离性、在它基础上建立的资源描述框架等几方面。虽然针对该

问题进行研究的论文较多,但综合归纳起来主要体现在以下几个方面:(1)专业导航库建设。利用 XML 设计资源描述框架不仅可以保持和 HTML 一样的浏览效果,而且可以很好地识别各专业库不同字段的内涵;(2)在信息资源查询检索中应用。XML 具有强大的语法功能,能够自定义自己的标记系统或字段集,可以真正实现图书资源数字化和共享,有利于信息交换与发布,增强检索的查全率。孙晓菲认为“用 XML 编写的网页克服了 HTML 不能区分信息和元信息且不支持信息嵌套体系结构的缺陷,使全文检索功能增强,检索的针对性更强。”^[4](3)在数字图书馆管理上的应用。XML 在很大程度上为数字图书馆的管理、维护和应用提供了方便。主要表现在数据更新、读物分级、规范馆内管理、开展读者培训、建设学区和社区图书馆(室)几个方面;(4)有序地组织信息资源。MARC 和都柏林核心集都是元数据。而元数据与 XML 发展是密不可分的。为解决网络信息资源有序组织的效率与质量之间的矛盾而产生了元数据,可是任何一种元数据格式都无法对所有信息资源进行著录。但如果使用有语义的标识标记数据或文档,就可以使数字图书馆信息著录和组织走出困境,XML 正好适应了这样的需要;(5)知识产权保护。知识产权问题是数字图书馆目前面临的棘手性问题之一。数字图书馆兼具公益性和商业性,针对不同机构,使用权利范围将有所不同,数字图书馆有必要分门别类地做出有区别的授权使用规定。申飞驹等在文中论述到:“数字图书馆知识产权的保护对象是对象数据库的集合,利用 XML 可划分知识产权保护的级别或权重,可注明 Web 页的知识产权要素,如权利、出版者、创造者、贡献者等;建立以各地的特色馆藏为中心的对象数据系统,并能有效地根据调度系统与知识产权、用户协议等因素将对象数据发向所需的用户。这些将对数字图书馆知识产权的保护起到积极作用。”^[5]

4 XML 与元数据

前面的叙述中我们已经明确提出元数据与 XML 的发展密切相关。目前在图书情报档案领域已经有一些人对二者的关系进行了研究和探讨,可以说是已经取得了一定的进展。张惠文认为,所谓元数据(Metadata)是关于数据的结构化数据,是用来描述和规定数据的编码数据。元数据能为各种形态的数字化信息单元和资源集合提供规范、普遍的描述基准和办法,在网络信息资源的组织和整理利用中发挥越来越大的作用,并成为各界研究的热

点论题。单一媒体的元数据研究开始趋于成熟,为单一应用领域的语义互操作提供了可能,但由于各类资源之间的差异性,使得各类元数据标准彼此不能兼容,符合某种标准规范的元数据不能被其他规范接受,给元数据的发展带来了不利的影响。XML 的出现使各类元数据标准之间的互通成为可能。1999 年 W3C 发布了一种基于 XML 语法的元数据规范 RDF (Resource Description Framework),即资源描述框架,目的是为元数据在网络上的各种应用提供一个基础结构,使应用程序之间能够通过网络实现数据的交换和处理。RDF 在利用 XML 严谨结构的基础上,采用了避免语义二义性的结构,从而为标准元数据的编码、交换及机器自动处理提供了保证,是一个开放的元数据框架。段荣婷在文中根据文件连续体理论,把元数据运转的整个流程中元数据功能的实现过程粗分为电子文件元数据动态获取和提供利用两大阶段,并分别探讨了两阶段中元数据是怎样与 XML 技术结合的,从而实现了对电子文件科学管理的。还有几位学者讨论了图书情报档案界 MARC、DC 等现有元数据进行了基于 XML 的元数据描述技术方面的研究,为元数据的深入研究奠定了一定基础。档案界更应该进一步加强基于 XML 的元数据描述技术方面进行研究,以增强对电子文件管理,解决著录、鉴定、质量控制、数据交换方面的困惑。

5 XML 与数据库

徐建平、王光明对 XML 与数据库的交互问题作了初步探讨,并列举了多个应用实例进行描述,指出现在可以有多种方案解决该问题,IBM 和 Sybase 等大公司都已经开发了专门工具,可以对数据库进行检索,产生 XML 格式的数据以及保存 XML 格式的数据到数据库表中。徐健则基于 XML 的三层 C/S 模型,探讨了利用 XML 实现图书馆 Web 数据库动态发布的设计思想,并通过示例展示了基于 XML 的图书馆 Web 数据库动态发布的具体实现。

结束语

有人已经把 XML 称为“世界语”,其应用前景已经得到了诸领域的认可和关注。如何在数字化、网络化时代保持图书情报档案事业可持续性发展是图书情报档案界的学者和实际工作者共同担忧的重要问题。XML 这一先进性技术应该也必须引起我们的重视。希望图书情报档案界的仁人志士能够在今后的研究中拓宽研究视角、拓展研究深度

和广度,充分挖掘 XML 的技术优势,为我们图书情报档案事业的发展服务。

【参考文献】

- [1] <http://www.oasis-open.org/cover/xml.html>
- [2] 施燕斌,刘春红. XML 简介及其应用浅析[J]. 高校图书馆工作,2002,(2):55-59,68.
- [3] 张咏. XML 及其在图书馆和情报检索中的应用[J]. 现代图书情报技术,2001,(2):30-34.
- [4] 孙晓菲. XML 与数字图书馆[J]. 现代图书情报技术,2000,(4):14-15.
- [5] 申飞驹,袁红,董建成. XML 在数字图书馆中的应用[J]. 中华医学图书情报杂志,2002,(6):40-42.
- [6] 邓凯,吴家春,王洪伟. 基于 XML 的移动数字图书服务体系结构研究[J]. 情报学报,2002,(5):559-562.
- [7] 段荣婷. XML 在电子文件元数据管理中的应用[J]. 图书情报知识,2002,(6):53-54.
- [8] 郭少友. XML 及基于 XML 的广播式检索[J]. 情报学报,2002,(5):568-572.
- [9] 黄文. XML 技术及其在数字图书馆中的应用[J]. 情报理论与实践,2003,(1):69-71.
- [10] 徐健. 利用 XML 实现图书馆 Web 数据库的动态发布[J]. 现代图书情报技术,2003,(1):54-56.
- [11] 林甫. 试析常用于数字图书馆中数据交换与处理的三种数据格式的标识语言(SGML、HTML、XML)的异同性[J]. 现代情报,2002(9):118-119.
- [12] 徐建平,王光明. XML 与数据库交互技术和方法[J]. 情报理论与实践,2003,(1):67-68,58.
- [13] 徐英,刘申学,毕强. XML 的技术特征及其对超文本导航的影响[J]. 情报学报,2002,(4):437-440.
- [14] 徐仲. XML 技术及在数字图书馆建设中的应用[J]. 图书馆理论与实践,2002,(2):59-60,86.
- [15] 王伟. 标记语言及 HTML 和 XML 的比较分析[J]. 现代图书情报技术,2000,(5):22-24.
- [16] 张惠文. 基于 XML 的元数据架构[J]. 情报科学,2002,(10):1072-1074.

【收稿日期:2004-03-26;责任编辑:陈 军】