中国图书馆学报(双月刊) 1999 年第 5 期
ZHONGGUO TUSHUGUANXUEBAO

●方 曙

SGML 及其在图书馆中的应用

摘 要 SGML (Standard Generalized Markup Language)是一种标准结构化电子文献格式或描述文献的计算机语言,是文献信息生产、管理的标准和技术。它在图书馆被应用于文献资源的开发、存取和共享,处理购来的全文电子文献,向用户提供全文电子文献的查询、检索、发送、咨询。参考文献 14。

关键词 SGML 电子文献格式 MARC 数字图书馆
分类号 G258.94

ABSTRACT SGML (Standard Generalised Markup Language) is a standard structured electronic format or a computer language describing documents. It is applied in the development, access, sharing of document resources, processing of purchased full-text electronic documents, etc. 14 refs.

KEY WORDS SGML. Electronic format. MARC. Digital library.
CLASS NUMBER G258.94

1 引言

近年来信息技术飞速发展,因特网先进的信息传递、交流功能及丰富的信息资源使人类社会正在发生根本性的变化。从网络获取信息、发布信息,成为当今信息社会的重要特征。向网络提供电子或数字化文献信息资源,为终端用户提供优质信息服务,是图书馆的主要任务之一。传统上,图书馆使用MARC 格式来处理、加工各种印刷型文本文献,对它们著录、标引,创建书目数据库,向读者或用户提供联机查询、检索等服务。然而对多媒体文献,MARC 格式则无能为力。随着因特网的进一步发展,下一代因特网(NGI)将为多媒体信息的传递、交流和使用等创造良好的网络环境,网上文本文献占统治地位的格局将发生改变。因此,突破MARC 局限,采用新的标准和技术,实现各种类型文献的电子化或数字化,将是今后网络文献资源建设和数字图书馆研制的关键。现在,发达国家信息产业界、出版界和图书馆界选择的解决方案之一就是采用 SGML 标准和技术。

2 SGML 简介

SGML (Standard Generalized Markup Language)即所谓的标准通用标引语言,它是国际标准化组织 1986 年颁布采用的一个文献信息生产、管理的国际标准,即 ISO 8879。SGML 定义了在文献中插入描述性标记的标准格式,规定了描述文献结构的标准方法。它不仅是一种标准,而且更重要的是它是一种强大的技术。在基于文献信息的开放系统中,SGML 得到了广泛的应用。

虽然人们在 90 年代以来才日益关注 SGML,但它的历史可以追溯到 60 年代。当时 IBM 公司的 Charles Goldfarb 和他的小组在进行一个关于法律办公室集成信息系统研究时,设计开发出一种叫做 GML (Generalized Markup Language)的通用标记语言的方法,以在文本编辑、格式化和信息子系统中实现文献的共享。经过 20 多年有关人员和组织的努力,GML 演变为当今的国际标准的 SGML,并得到信息产业界、出版界和图书情报界等的广泛重视和应用,成为当今文献信息处理的主流技术和数字图书馆的关键技术之一。

SGML 是一种标准结构化电子文献格式或描述文献的计算机语言。它是一个复杂的系统,简单说它包括 3 部分: (1) 前言: 系统特殊信息如字符使用等; (2) 文献类型定义(DTD): 目的、区别标记符、文献结构标记规则、文献所附的其他东西; (3) 文献实体,包括标记符在内的实际文献。SGML 的应用核心是文献类型定义,其规则通常具有以下

形式: 〈! 关键词 参数 1 个或多个相关参数〉。1 篇文献由一系列元素组成, 如标题、作者、段、段标题等, DTD 规定了这些组成元素之间的关系。SGML 可在 DTD 中建立一组规则, 定义自己的标记语言, 因此它是一种元语言(metalanguage)。插入电子文献中的简短符号或字符串叫做标签(tags), 使用 DTD 中声明的描述性标签对文献进行标记, 就生成 1 个文献实体。文献实体中标记出来的元素反映了该文献的逻辑结构。

SGML 是一个极端精密复杂的系统, 它规模庞大, 功能丰富, 充满各种选项。它不依赖特定的软硬件, 因而具有许多突出的优点, 即长久有效性、可互操作性、共享性、不同目的反复使用性、高效性等。用 SGML 建立的文献能在文字处理或桌面出版系统等中转换, 能在不同的机器及不同的程序中转入转出, 可以节省大量开支和费用。正因 SGML 有上述优点, 所以它成为当前文献信息加工、处理、储存和发布的主流技术之一。

目前, SGML 的研究应用非常活跃, HTML 就是其成功应用的成果之一。Web 的流行当归功于 HTML。然而 HTML 仅是 SGML 的 DTD 的简单应用实现, 它丧失了 SGML 的许多功能, 而且不具扩展性。在处理多媒体及超媒体文献时, 它面临严峻的挑战。解决的方案之一就是利用其父系 SGML 的可扩展性。1996 年底, 可扩展标记语言 XML (Extensible Markup Language) 诞生, 它是 SGML 的简化版, 它继承了 SGML 的可扩展性, 但又避免了 SGML 的一些复杂性, 将信息处理技术向前推进了一步。XML 代表了 SGML 应用发展的一个新的方向。

在 SGML 应用领域的另一个典型是 TEI (Text Encoding Initiative) 编码方案。TEI 是由计算机与人文协会(ACH)、文学与语言计算协会(ALLC)及计算语言学会(ACL)3 个专业团体发起的一个国际合作研究项目。其目标是为各种不同的复杂文本产生一个普通的编码标记方案, 以实现人文科学及语言学等领域文献资源数据的交换、共享和标准化。目前, 国际上已有数十个大型项目采用 TEI 来生产 SGML 文献。如英国国家大全项目、美国密歇根大学的人文科学文本计划、欧洲大全计划等。在档案界也开发出 SGML 的应用子集 EAD 等。

另一方面, SGML 的相关标准也得到发展。最典型的是 DSSSL (Document Style Semantics and Specification Language) 和 HyTime (Hypermedia/Time-based Document Structuring Language)。DSSSL 是 1996 年 1 月才完成认定的国际标准, 即 ISO/IEC 10179, 它的主要用途是实现 SGML 文献之间以及向其他文献格式的转换, 这对 SGML 的广泛应用极其重要。因为 SGML 没有标准的强制性标记符号, 相同文体和结构的文献, 由不同的作者和出版社出版, 便会有不同的 DTD。实现 SGML 文献之间以及与其他格式之间的交换, 才能实现文献的共享与交换。HyTime 是关于超媒体标记方面的一个国际标准, 即 ISO/IEC 10744, 1992。它使用 SGML 作为它管理数据的基础编码语言, 它是 SGML 的应用与扩展。DSSSL 和 HyTime 正在开发一种对树型结构和查询语言共享的方案。其成果将发展一套在所有基于 SGML 标准中共享的一般工具, 从而使文献处理更容易, 信息发布更强大。

目前, 已有多种有关 SGML 的软件产品面世。它们大致可分为将词处理文件转化为 SGML 文献的软件, 如 Novell 公司的 Wordperfect SGML 版等; 用于出版光盘文献或电子图书等的 SGML 软件包, 如 Electronic Book Technologies 公司的 Dynatext、SoftQuad 公司的 Explorer 等; 以及处理 SGML 文献的数据库管理系统软件, 如 Information Dimensions 公司的 SGML server 和 Fulcrum 技术公司的 Ful/Text 等。这些 SGML 软件使 SGML 的应用迅速发展。

3 SGML 在图书馆中的应用

SGML 是文献信息生产、管理的标准和技术。信息产业界、出版界早就意识到广阔的应用前景及潜在巨大的优势和效益。目前, 出版界的电子出版几乎都瞄准 SGML, 越来越多的电子文献在以 SGML 格式出现。SGML 在图书馆界的应用是近几年的事。主要体现在图书馆文献资源的开发、存取和共享, 处理出版界、数据生产者、信息服务公司等发售的全文电子文献, 向读者或终端用户提供全文电子文献的查询、检索、发送、咨询的中介服务等。前者主要涉及书目数据库如联合目录、联

机目录的生产、馆藏文献资源的数字化。万维网信息服务、电子文献传递、共享服务等; 后者涉及存取、加工和处理电子期刊、图书、全文数据库、多媒体文献和全文期刊文献的订购与传递等。它们几乎都与数字图书馆的建设紧密相关。下面我们将对它们进行一些讨论。

3.1 书目数据库的生产

传统上, 图书馆采用各种MARC 格式如USMARC、UNIMARC、CNMARC 等格式以及非MARC 格式等来编制、生产书目数据, 产生联合目录、联机目录等。但由于不同的图书馆使用的软件、硬件不同, 自动化系统不同, 采用的分类规则不同等, 数据质量差异较大, 因此数据的兼容性、交换、共享等方面存在一定的问题。另外, 随着电子出版物的增加, 尤其是多媒体文献的增加, 以及数字图书馆发展的要求,MARC 格式已无法处理包括多媒体文献在内的各种文献的统一著录和标引。因此, 一些发达国家的图书馆界等开始转向SGML, 以期利用SGML 来统一书目数据的标准, 实现图书馆之间数据广泛的交流与共享。由于书目数据的内容具有严格的逻辑结构, 它完全适合用SGML 的DTD 来描述。因此采用SGML 生产书目数据库是非常容易和行之有效的。SGML 文献的众多优点, 为今后书目数据的交换与共享和长久使用打下了良好的基础。实际上, 一些发达国家从90 年代以来, 已成功地用SGML 技术和标准创建了SGML 书目数据库、期刊联合目录数据库等, 实现了数据广泛地交换与共享, 从中受益匪浅。例如: 美国加大伯克利大学图书馆等开发了USMARC 记录与SGML 格式之间相互转换的程序, 比利时图书馆界采用SGML, 逐步实现了全国性书目数据格式的统一和标准化, 建立了包括40 多所大学及专业图书馆在内的国家联合目录以及国家期刊联合目录等。结束了以前比利时图书馆界书目数据格式多样、图书馆之间难以进行数据交换共享的复杂混乱局面, 同时避免了重复劳动, 节省大量开支, 使图书馆的文献信息服务出现新的局面。

3.2 文献资源的订购、存取与共享

随着信息技术的发展, 电子文献资源大量涌现。因特网使图书馆的文献订购、收藏职能等发生重大变化。通过网络来存取而不是"拥有"文献资源是当今图书馆文献订购、共享的新动向。图书馆的文献订购方式逐渐转向网络进行。若干图书馆以集体用户名义订购版权文献, 各馆通过因特网等来存取和使用, 并为读者提供服务。传统意义上的"馆藏"(拥有) 变为存取和使用。同时在资源共享方面, 利用先进的网络技术, 当地图书馆不再拥有或收藏的文献, 可及时从其他地方的图书馆获取, 即一些馆根据他馆的要求, 通过计算机网络向异地馆发送当地馆没有但用户和读者需要的文献(类似于传统意义上的馆际互借)。这就要求图书馆建立文献资源订购、传递和共享系统。SGML 文献能方便地以数字形式传输和检索。它可在不同的机器中转入转出, 所以目前大多数出版商均在以SGML 格式出版发送各种电子或数字化文献, 例如著名的Elsevier 科学出版社等单位发起的电子图书馆SGML 应用项目(ELSA)。该项目旨在向图书馆及终端用户提供、发送基于SGML 的文献, 即全文期刊文献直接从出版社以SGML 格式发送, 图书馆收到后, 选择在DTD 中定义的元素产生索引, 文献本身则转换成不同用途形式使用, 如联机阅读、超文本导航等。用户还可建立SDI, 每周可收到有关主题的最新文献信息。图书馆要存取使用SGML 文献, 就必须了解掌握SGML 技术并建立相应的图书馆的文献订购、传递和共享系统。比利时图书馆界的Impala 就是采用SGML 技术建立的联机文献订购、共享系统。其他类似的还有DECOMATE 系统(Delivery of Copyright Materials to End-Users), 这是英国、荷兰和西班牙几所大学的一个合作项目, 它的目标是通过图书馆向终端用户提供、传送来自出版商的拥有版权的电子文献。

3.3 馆藏文献资源的数字化

数字图书馆是网络环境下图书馆文献资源的组织形式, 也是图书馆自动化发展的高级阶段。数字图书馆的主要内容之一是数字化资源建设。它涉及两个方面: 一是增加电子文献资源的订购和收藏; 二是加快馆藏特色资源的数字化。前者是外来电子资源的存取和收藏, 在上面已作讨论, 后者则是图书馆资源的转化。现有图书馆因条块分割、馆藏重复率高, 所以, 馆藏资源数字化时要合理规划, 统一部署, 避免重复劳动和浪费, 在技术上, 要有超前意识。鉴于今后超媒体信息的广泛使用和

信息长久性等因素, 采用 SGML 标准和技术是当今的首选方案之一。目前, 大多数数字图书馆项目均采用 SGML 技术来数字化特色馆藏资源。如美国康奈尔大学与美国化学会等单位合作进行的 CORE 化学文献数字图书馆项目。他们采用 SGML 将美国化学会的 20 种化学刊物数字化。他们发现采用 SGML 后, 大大地增强了数据使用的灵活性及效率。SGML 技术在馆藏资源数字化中最突出的特点就是一旦数字化, 就可重复、永久使用, 即一馆将某一文献数字化后, 其他馆根据不同目的和用途加以利用, 使信息增值, 从而加速图书馆文献资源的数字化进程。

SGML 在图书馆的其他工作方面, 如编制工作人员手册、介绍图书馆服务内容的读者指南、在万维网上发布新消息、读者培训资料的准备等都大有用武之地。同时 SGML 也能帮助图书馆适应迅速变化的学术需要, 创造新的终端用户服务。如许多图书馆正在建立新的系统, 这些系统不仅提供目录服务, 而且能为当地图书馆用户提供全文期刊数据库文献服务, 即将馆藏目录与馆藏全文期刊文献联系起来。总之, 新的信息技术正在改变传统图书馆的许多职能, 使图书馆工作进入一个新的阶段。

4 结论

SGML 将使图书馆的电子文献资源逐步走向规范化和标准化, 并拓展图书馆的服务方式和范围。然而, 图书馆应用 SGML 技术还存在许多障碍。首先是经费上的问题。由于 SGML 较为复杂, 它需要一定的软硬设备。而当前的 SGML 应用软件、工具等较为昂贵。要设计、实施一个 SGML 文献处理系统, 必须对这个标准相当全面地熟悉了解, 弄清它的原理和应用细节, 需要对有关人员进行培训, 这些都需要费用。其次, SGML 并非是专门为图书馆使用而设计的。一般图书馆员对它都很陌生, 使用它技术要求高、难度大, 人员培训任务重。另外, 还有思想认识、观念转变等问题。然而, 由于 SGML 客观上的巨大优势, 从长远的观点看, 它代表了今后文献信息生产管理的发展方向, 它能给我们带来巨大的效益。虽然 SGML 在图书馆中的应用前期投入大, 但是从它获得的收益却是无法估量的。因此, 我们应密切注视它的研究应用动向, 并逐步在我国图书馆中推广使用, 以跟上世界发展潮流, 从而加速我国图书馆文献信息资源的建设和数字图书馆的发展。

参考文献

1 J. Corthouts and R Philips. SGML: a librarian's perception. The Electronic Library, 1996, 14(2)
2 A. Dorward. SGML in Publishing-Why Use The Standard? The Electronic Library, 1995, 13(1)
3 Y. Marcoux, M. Sevigny. Why SGML? Why Now? J. of Am. Soc for Info. Sci, 1997, 48(7)
4 许绥文. 漫笔之二: 数字式图书馆在美国. 北京图书馆馆刊, 1998(1)
5 刘霞. 关于数字图书馆建设的几个问题. 图书情报知识, 1998(3)
6 XML. Web 标记的第二次机会. 个人电脑, 1998(1)
7 刘树森. SGML ——信息时代的生力军. 中国计算机用户, 1997, 9(上)
8 J. D. Mason. SGML and Related Standards: New Directions as the Second Decade Begins, J. Am. Soc for Info. Sci, 1997, 48(7)
9 S. C. Adler. The "ABCs" of DSSSL, J. Am. Soc for Info. Sci, 1997, 48(7): 597～602
10 D. T. Barnard, N. M. Ide. The Text Encoding Initiative: Flexible and Extensible Document Encoding, J. Am. Soc for Info. Sci, 1997, 48(7)
11 J. Fausey, K. Shafer. All My Data Is in SGML. Now What? J. Am. Soc for Info. Sci, 1997, 48(7)
12 黄晓斌. SGML 与数字图书馆. 图书馆学刊, 1996(6)
13 N. M. Ide, et al. The Text Encoding Initiative: Its History, Goals and Future Development. Computers and the Humanities, 1995, 29(1): 5～15
14 T. W. Cole, M. M. Kazmer. SGML as a component of the digital library. Library Hi Tech, 1995, 13(4)

方 曙 中科院成都文献情报中心编辑出版部主任、副研究员。通讯地址: 成都市人民南路四段九号。邮编 610041。

(来稿时间: 1999-01-26。编发者: 刘喜申)