数字环境下的信息资源编目

[作者] 喻爽爽
[单位] CALIS 联机编目中心
[摘要] 探讨了数字环境下图书馆编目工作的相应变化及其未来的发展方向。
[关键词] 编目工作，元数据，都柏林核心，可扩展标记语言

1 引言

数字化在信息技术处理上是一次划时代的革命。随着数字化信息资源的迅速膨胀，以及人们对信息情报在数量、层次上要求的迅速增长，如何将庞大的、处于混乱无序状态的网上信息资源进行有序化处理，以满足人们对信息情报多层次的检索需求，已成为全球性的重要课题。目前，由图书馆界或信息情报服务机构采用相关著录标准对数字化信息资源进行描述，供用户检索、浏览或下载，已逐渐成为解决这一问题的主流方案。

这一主流方案的实质，就是使用原有的图书馆编目工作原理，对新出现的数字化信息进行著录。由于传统图书馆编目工作的主旨是将无序的文献资料进行有序化组织，因而其工作原理完全适用于对数字信息资源进行著录，并且这种著录对数字信息的揭示要更科学、更深层。但在数字化环境中，人类知识的保存已经从以纸介质为主的时代进入多媒体信息载体的时代。与此同时，人们对信息的要求也向多元化、多样化方向发展，这些变化对图书馆编目提出了新的挑战，为了有效地对数字化信息进行多元化采集、海量化存储、有序化组织和有效化提供，编目工作已进行了一些相应的调整，这些调整主要表现在编目格式的变化，以及编目用数字语言的应用及发展上。

2 编目格式的变化

目前，在这方面有两种有益的探索：其中之一是用 MARC 格式整合所有类型的资料，它的依据是通过著录规则和规范控制实现记述的一致性及标准化。另一种是将不同类型的资料用特定格式的元数据来表现，在核心元数据的基础上进一步根据各种文献类型的具体情况增加及定义特殊的字段和子字段。

2.1 数字环境下的 MARC 格式——856 字段的设置

（1）856 字段产生的背景

20 世纪 90 年代初期，OCLC 因特网编目计划应运而生，其目的就是研究 USMARC 和 AACR2 对网上资源编目的适用性。该计划分为两个阶段：

第一阶段 1991－1992 年计划：尽管当时缺乏计算机文档编目的经验以及技术上的问题，编目人员在进行该试验计划时存在着一些困难，但是最终得出结论：MARC/AACR2 能够运用于网上资源编目，有可能找出连接书目记录和所描述的网上资源的方法，并且要为编目员提供网上资源的学习资料。该计划还提出了三条建议：一是为远程电子信息检索提供完备的机读目录格式；二是拓展编目规则和格式，以包含具有交互性的资源；三是检验所编制的

书目记录的作用和效益。

1993 年 4 月，OCLC 与国会图书馆联合发起了一项建议，即修改 USMARC 书目格式以适应联机信息资源编目，建议在 MARC 格式中设立新字段——856 字段，即“ 电子资源定位和检索 ”。它可用于一项资源的书目记录或馆藏记录中，前提是该资源或其子集有电子版可以获得。它还可以用于定位和检索一项在书目记录中描述的非电子资源的电子版、该资源的一部分或相关的电子资源。在以下两种情况下可以启用 856 字段：一是编目的文献资源如可经由电子方式获取时；二是记录中所著录的非电子资源具有相应的电子版本或相关的电子资源时。

第二阶段 1994－1996 年计划 :为了更好地解决在 1991－1992 年计划中发现的书目记录与其所描述的远程资源的直接连接问题，OCLC 再次组织了 1994－1996 年编目计划。大约有 500 多个 OCLC 成员馆联合编制了 18，000 条网上资源，以进一步证实 MARC/AACR2 可适用于网上资源的编目。该计划将 856 字段作为一个正式的 MARC 字段用于网络资源的编目，自此 856 字段的利用逐渐被全世界的图书馆界所接受。利用 MARC 格式进行网上资源的编目就使图书馆的书目数据库不再是一个“ 静态 ”的仓库，从而拓展了书目记录的功能，开创了编目的新时代。

（2）856 字段的内容

856 字段是 MARC 记录中专门用于揭示电子资源的字段，它包括电子资源的定位与检索所需要的各种信息，如：电子资源的地址、登录方式、读取方式、传输方法甚至口令等重要信息。需要指出的是，中文文献编目采用的是 CNMARC，目前的做法是套用 USMARC 的 856 字段，所有的定义及使用方法均参见 USMARC 中的 856 字段。

856 字段记述数字化信息源的所在及连接方法，配合必要的网络环境，它可扩展机读目录的功能——为读者检索远程电子信息和网络资源提供更为快捷的有效途径。因此，在 OCLC（Online Computer Library Center，美国联机计算机图书馆中心）的 Inter Cat 计划所进行的一项调查中发现，与其他类型的元数据相比，在描述电子情报资源时，MARC 的概括性、检索的依赖性、数据构造的详细性等方面依然表现优秀。

尽管如此，MARC 在追加新数据要素和对原有形式加以修正等方面缺乏融通性，同时，与图书馆以外的出版社、情报中心等关联机关具有互换性弱等缺点，在书目记录间连接机能等方面也存在着不足。在此情况下，出现了 DC、CDWA、EAD、FGDC、GILS、TEI、VRA 等其他类型的元数据。

2.2 元数据

所谓元数据（Metadata），即关于数据的数据，或称为描述数据的数据。数字化时代之前传统图书馆中的卡片式目录、书本式目录及 MARC 目录等都属于元数据。在数字图书馆系统中，常用的元数据可分为：描述型元数据（Descriptive Metadata）、管理型元数据（Administrative Metadata）和应用型元数据（Application Metadata）等几种类型。其功能如下：

（1） 描述功能。对所组织信息对象的内容、属性等进行描述。

（2） 检索功能。为用户提供多层次、多途径的快捷检索体系。

（3） 管理功能。为所保存的信息资源加工、存档，并具备使用权限管理（版权、所有权、使用权等）和防伪措施。

（4） 还有选择、定位、评估和交互功能。

目前在互联网上存在的多种格式的元数据中，DC（Dublin Core，都柏林核心）以其简便、灵活、易于理解及可扩展性等性能，倍受人们推崇。1998 年 9 月 IETF（Internet

Engineering Task Force，因特网工程任务组）将其作为一个正式标准（RFC2413）予以发布，成为国际通用的元数据标准。目前已有德、日、葡、西语等 10 余种不同语种版本。

DC 元数据为网上各种信息资源的著录、编目提供了一种容易掌握和使用的著录格式，必然会推动人们对网络信息资源的利用。它结构简单，只有 15 个核心元素（element），其中 7 个是有关内容的元素：题名（Title）、主题词和关键词（Subject）、内容描述（Description）、资源类型（Type）、来源和出处（Source）、关系（Relation）、范围（Coverage）；4 个是有关知识产权的元素：著者或创建者（Creator）、出版者（Publisher）、其他责任者（Contributor）、权限管理（Rights）；4 个是有形化元素：日期（Date）、格式（Format）、资源标识（Identifier）、语种（Language）。正是由于 DC 结构紧凑，便于对元数据记录数据库进行管理，而且检索效率也较高。同时，由于 DC 是一种结构化的元数据，可以实现各种元数据间的转换，并在此基础上建立搜索引擎，这不仅使大量存在的机读书目数据能转换为都柏林核心的元素集，从而实现网络存取，有效地对网络资源进行组织和利用。同时对于某些有质量保证的重要的网络资源，也可采用抽取以 DC 元数据格式著录的简编数据，将其转换为 CNMARC 数据格式，并根据各馆的需要进行修改，增添满足特殊描述、检索需求的字段或子字段。对于 DC 与 MARC 的相互转换，是基于二者各自的特点。如前所述，MARC 格式由于其格式的完整性、著录的详尽性，历经四十年的发展，已成为一种成熟的信息著录格式。在对那些相对稳定的、重要的网络资源进行著录时，采用 MARC 格式更能详细准确地表达其内涵。而 DC 能较全面地概括网络信息资源的重要检索点以及有价值的说明性信息。

3　编目用数字语言的发展

在编目格式随着数字化的到来而进行补充、调整的同时，数字语言也开始应用于编目，并且对它的开发也在飞快地进行着，特别是作为 SGML（Standard Generalized Markup Language，标准通用置标语言）子集的 XML（Extensible Markup Language，可扩展标记语言）的推出，解决了其他数字语言在扩展、结构及数据确认上的缺陷。

SGML 是表现结构化文献的信息语言，也是 1986 年国际标准化组织（ISO）发布的信息处理标准。作为大规模系统内信息交换的文档格式，它在欧、美国家被广泛应用于文献管理方面的历史已有十几年。我国于 1994 年将其定为国家标准，主要应用在新闻出版信息处理领域。由于 SGML 的设计早于 Web 的出现，不能作为 Web 的对应语言应用。这样，作为 SGML 在网上应用的置标语言 HTML 应运而生。它将信息简单地加工处理后通过 Web 发送到世界各地，是一种在 SGML 标准上开发的作为信息发布和阅览所用的最成功的 Web 应用语言。但随着网络信息化的不断发展，其应用范围的局限性逐渐显露——有限的标签种类不能精确地描述信息，不能充分表现数据的含义、层次以及结构过于简单等等。XML 的诞生解决了 HTML 存在的许多问题。

XML 是万维网联盟（W3C－the World Wide Web Consortium）开发的用于网络环境下网页设计和交换、管理的新技术，被认为是第二代 Internet 信息组织的标准。它是一种元数据语言，用于定义不限定数量的特殊标记语言。XML 最大的特点就是其可扩展性，可以将数据的存贮和数据的显示分离，同时可以轻易地完成不同元数据格式间的相互转换，具有联接各种元数据格式的重要作用，因而逐渐为各种元数据格式所采用。对于电子资源著录的结果就可采用 XML 建立中央数据库，数据库中每条记录就是一个网页的元数据。既采用了可供人阅读的文件形式，又采用了可供程序理解的数据形态，具有记述文件和数据的两面性，实现了人机共享的目标。

4 结语

随着国际互联网的快速发展，以网页、网站的形式在互联网上传播的数字化信息呈指数上升，对其它载体形式的资源进行数字化处理的工作也十分普遍。它涉及数字信息资源的加工、分类、剔除、存储、检索、传递、保护和利用等诸多方面。如何将庞大的数字信息资源进行有序化处理已成为全球性的一大课题。如果正确的目录系统没有被构筑起来，就不可能准确、快捷地进行检索，在这一背景下，只有应用图书馆编目的科学原理，才能为海量的数字化信息构筑正确的目录系统，以准确、快捷地进行检索。相信随着信息技术的飞速发展，MARC 格式的不断完善以及其他元数据和数字语言的不断开发应用，对虚拟性的网络资源和各种电子资源的编目将日臻成熟。

参考文献

1、吴建中．DC 元数据．上海：上海科学技术文献出版社，2000
2、肖珑，陈凌等．中文元数据标准框架及其应用．大学图书馆学报，2001（5）
3、赵慧勤．数字图书馆的信息组织－元数据描述技术．图书情报工作，2001（7）
4、石英弘，李颖．未来网络的基磐技术－XML 的理论与应用．北京：华艺出版社，2002
5、李慧．可扩展标记语言 XML 及其在数字图书馆中的应用．图书情报工作，2001（12）
6、Alan Hopkinson. Traditional Communication Formats：MARC is far from dead. ICBC（International Cataloging and Biblio graphic Control）. 1999（1）