·实践平台·

有关 CNMARC 与 DC 元数据之间的对应转换

刘圆圆 (西北工业大学图书馆 陕西西安 710072)
刘军华 (西安财经学院 陕西西安 710065)

摘 要: 文章概述了 CNMARC 的缺陷和 DC 的优势,提出 CNMARC 向 DC 转换的必要性,详细列出 DC 元素与 CNMARC 各字段的映射和匹配关系,同时介绍了 CNMARC 与 DC 的转换方式,分析了目前我国在 CNMARC 与 DC 转换中存在的问题。

关键词: DC CNMARC 元数据转换

中图分类号: G254 文献标识码: A 文章编号: 1003-6938(2007)03-0103-04

Thinking about Transforming of CNMARC to DC

Liu Yuanyuan (Northwest Polytechnic University Library, Xi'an, Shanxi, 710072)
Liu Junhua (Xi'an Finance and Economics College, Xi'an, Shanxi, 710065)

Abstrct: This article gives a brief account of the limitation of CNMARC and advantage of Dublin Core, and pointes out the necessity about transformation from CNMARC to DC. It also lists the matching relationship between the DC element and every field of CNMARC, and introduces transform mode from CNMARC to DC. In the end, it analyzes existent problems in the transformation from CNMARC to DC in China at present.

Key words: DC; CNMARC; metadata transform

CLC number: G254 Document code: A Article ID: 1003-6938(2007)03-0103-04

1 CNMARC 与 Dublin Core 简介

MARC 是用于描述、存储、交换、控制和检索的一套机读书目数据标准,它开始主要是针对印刷型文献的描述。其数据结构严密,能很好地描述文献信息,尤其是在检索点的选取原则上,能确保其数据元素组成具有统一性,有利于文献信息资源交换。我国于 1991 年编制了《中国机读目录通讯格式》,即 CNMARC,从而正式开始了我国机读目录的建设工作。现在,CNMARC 已成为我国图书出版行业、图书情报机构普遍使用的标准。

Dublin Core——都柏林核心元素数据集,简称 DC,它是由 OCLC 和 NCSA 在 1995 年联合建立的一个针对网络信息资源的著录格式,其目的是提供一种容易掌握和使用的网络资源著录格式,方便网上资源的著录、编目,从而促进网络资源的利用。DC 由 15 个核心元素组成,每一个元素都可以选用和重复,具有简练、可扩展、能与其它数据形式匹配等特点。[1]经过历届 Dublin Core 工作会议的探讨与改进,DC 已经发展成为一种比较成熟的元数据格式,得到了多个国家、不同学科专业领域、各种研究机构和组织的使用。

2 CNMARC 向 DC 转换的必要性

2.1 CNMARC 的缺陷

CNMARC 在规范印刷型文献资源的管理中,起到了很重要的作用。但是随着网络信息的日益丰富,在网络环境下如何对网络资源进行编目,就成为数字图书馆首先要思考的一个问题,这时 CNMARC 的缺陷和 DC 的优势就明显地显现出来。CNMARC 的缺陷主要表现在:

(1) 格式复杂,著录速度慢

收稿日期: 2006-09-12; 责任编辑: 宋 戈

CNMARC 的著录格式主要由三部分组成：记录头标区、地址目次区、数据字段区。其中数据字段区又分为 10 个功能块：标识块、编码信息块、著录信息块、附注块、连接款目块、相关题名块、主题分析块、知识责任块、国际使用块、本地使用块，一共 176 个字段，522 个子字段，而且字段重复之处较多。比如：200 字段的责任者与 7—责任者块的数据单元内容就是重复的；200 字段的题名与 5—题名块的内容也相互重复，这是因为传统的书目卡片是以检索点的提取为出发点，而 CNMARC 的设计理念正是以卡片目录的思想为主导。[2] 同时因其复杂的著录内容而使著录速度较慢。

（2）著录的专业性要求高

由于 CNMARC 格式复杂，著录规则繁多，所以只有接受过专业训练的编目人员才能建立记录，没有受过专业训练的普通作者或使用者一般无法建立。而且其数据处理过程复杂，致使数据完成滞后，时效性较差，无法满足网络资源快速增长的需求。

（3）灵活性、扩展性差

CNMARC 数据格式的标准有六个部分，用于分别描述图书、连续出版物、地图等不同类型的文献资料。但它只是在编码信息块用字段名的方式标出不同格式的文献，如果出现新的文献格式，个人或图书馆只有在经过专家认证和相关机构的认可后，才可能为新文献格式创造新的字段，不能随意增加新的字段，所以灵活性、扩展性比较差。[3]

（4）网络应用能力差

CNMARC 的制定是为了方便印刷型文献的著录，方便各个图书馆间书目数据的交换，本身缺乏与网络语言相结合的基础，在当今网络发展迅疾的时代，面对数量巨大，内容及形式繁杂的网络信息资源，CNMARC 则显得力不从心了。

2.2　DC 元数据的优点

（1）元素集结构简洁，著录速度快，适合网络资源的著录

DC 只有 15 个核心元素，结构简洁而且易于使用，它简化了许多著录规则，能同时为普通人和资源描述专家所用，大部分元素都具有能够被普遍理解的语意，而不像 CNMARC 格式那样复杂，大大减少了项目的著录成本和时间。它是针对网络资源设计的，所以更适合网络资源的著录。

（2）较强的可扩展性

DC 所有的元素都是可选择、可重复、可延伸的，而且所有的元素都是可以以任何顺序显示，它可以用于不同格式的文献资源描述。DC 元数据的创建和维护简单，任何制作者无需经过专门的培训就可以为自己的文件创建元数据。

（3）较强的适应性

为了适应网络文献资源的著录需求，CNMARC 中特别增加了 856 字段，可以对网络信息资源的主机名、URL、URN、路径、口令等进行著录和超文本链接。[4] 但 CNMARC 从整体上来说仍然无法很好地适应网络文献资源变动性强、更新快、类型多样等特点。而用 DC 著录的信息与网页的信息相当契合，动态地适应了网络信息不断发展变化的趋势，满足了网络文献资源著录内容及时更新的需要。

（4）较强的兼容性

这种兼容性体现在两个方面：一是它与不同浏览器和操作系统相兼容，DC 本身就是利用 XML 语言作为描述语言，而 XML 作为脚本语言适合于任何平台，这就显示出 DC 的兼容性；二是 DC 与目前其他类型的元数据相兼容，可以作为结构化元数据来进行相互间的编码和转换。

CNMARC 作为一种为存储、交换、处理和检索文献资源而设计精密的数据格式标准，其描述对象全面详细，应用广泛，DC 元数据在一定程度上参考了它的格式特点，并在吸取 CNMARC 大部分优点的基础上克服了其缺点，优化了其结构。由此可见，CNMARC 向 DC 元数据的转换既能满足传统文献资源数字化的基本要求，也解决了数字图书馆建设和发展中网络信息资源组织与管理的关键问题。DC 与传统印刷型文献组织的唯一标准 MARC 格式正在呈现出并驾齐驱的形势，大有成为网络环境下信息资源组织的统一标准的趋势。

3　DC 与 CNMARC 的映射关系

目前，关于元数据的匹配和转换，国外的研究比较多，并已经建立了多种元数据的匹配机制，例如：DC 与 USMARC，DC 与 EAD，MARC 与 EAD，GILS 与 DC 等。[5] DC 与各种元数据格式的匹配机制，成为数字图书馆建设中信息资源组织与管理的有效工具。

DC 由三个主要部分组成：资源类型描述类、知识产权描述类、外部属性描述类，包括 15 个基本著录项：Title（题名）、Subject（主题）、Description（说明）、Language（语种）、Source（来源）、Relation（关联）、Coverage（覆盖范围）、Creator（创建者）、Publisher（出版者）、Contributor（其他责任者）、Rights（权限）、Date（日期）、Type（类型）、Identifier（标识符）、Format（格式），这 15 个核心元素可重复使用或者有选择地使用，而且还可以延伸有子类型和子模式。CNMARC 是由 000-999 之间的上百个字段构成，通过对 DC 和 CNMARC 结构和内容的分析，找出 DC 的每一个核心元素及其子元素与 CNMARC 中每个字段的映射关系（见表 1）。

4　CNMARC 与 DC 的转换

有了 CNMARC 与 DC 详细的映射关系，就可以逐步实现两者之间的转换了。目前，CNMARC 与 DC 之间的相互转换有两种方式：一种是动态的 CNMARC 转换，就是根据 DC 元素与 CNMARC 字段映射表，把 DC 元数据转换成 CNMARC 格

表 1 DC 与 CNMARC 映射关系对照表

DC	限定词	CNMARC	DC	限定词	CNMARC
Title（题名）		200 1$a 正题名	Description（说明）	Version（版本说明）	205 $a 版本说明
	Translated（翻译题名）	510 1$a 并列正题名		Version Details（版本细节）	305 $a 版本附注
		500 1$a 统一题名	Date（日期）		210 $d 出版发行日期
	Subtitle Alternative（交替题名）	200 1$e 副题名及其他题名信息			210 $h 重印日期
		517 1$a 其他题名			306 $a 出版发行附注
		200 1$a 交替题名	Format（格式）		8564$q 电子格式类型
		517 1$a 其他题名			215 $a 文献数量
	Long	5321_$a 展开题名			215 $a 尺寸
Identifier（标识符）	Identifier	010 $a 国际标准书号（ISBN）			336 $a 计算机文件类型附注
		011 $a 国际标准连续出版物号			337 $a 计算机文件技术细节附注
		020 $b 国家书目号	Language（语种）		101$a 作品语种
		8564$u 统一资源定位地址			
	canceled	011 $y 注销的 ISSN 号	Rights（权限）		300 $a 一般性附注
	incorrect	010 $z 错误的 ISBN 号	Relation（关联）		430 1$a 继承
		011 $z 错误的 ISSN 号			442 1$a 由…继承
		020 $z 错误的国家书目号			8564$u 统一资源定位地址
Publisher（出版者）	Name	210 $c 出版发行者			451 1$a 同一载体的其他版本
	Name onference	71112$a 会议名称			452 1$a 不同载体的其他版本
	Name Corporate	71102$a 团体名称			488 1$a 其他相关作品
	Name Persional	7010$a 个人名称			321 $a 外部索引/摘要/参考附注
	Place	210 $a 出版发行地			337 $a 电子资源细节附注
Contributor（其他责任者）		2001$g 其他责任者	Source（来源）		300 $a 一般性附注
	Name onference	71212$a 会议名称	Subject（主题）		600 - $a 个人名称主题
		71202$a 团体名称			6010$a 团体名称主题
	Name Personal	7020$a 个人名称			6011$a 会议名称主题
Coverage（覆盖范围）	Temporal（时间范围）	1222$a 作品内容时间范围			606 $a 论题主题
		661 $a 年代范围代码			607 $a 地理名称主题
	Spatial（空间范围）	660 $a 地区代码			610 $a 非控制主题词
	Note	300 $a 一般性附注			675 $a 国际十进制分类号（UDC）
Creator（创建者）		2001$f 第一责任者	Subject（主题）		676 $a 杜威十进制分类号（DDC）
	Name Personal	7010$a 个人名称			680 $a 国会图书馆分类号（LCC）
	Name onference	71112$a 会议名称			686 $a 其他分类法分类号
	Name Corporate	71102$a 团体名称			690 $a 中图法分类号（CLC）
Description（说明）	Summary bstract	330 $a 提要或文摘			692 $a 科图法分类号（LCCAS）
	Note, Award	300 $a 一般性附注	&Type（类型）		105（4-7）内容类型代码
	Audience	333 $a 用户/使用对象附注			200 1$b 一般资料标识
		337 $a 电子资源细节附注			608 $a 形式、体裁、物理特征主题

式记录，再返给用户。另一种是静态的 CNMARC 转换，在数据库中另外设计 CNMARC 数据记录表，利用 DC 元素和 CNMARC 字段映射表，通过单独的转换程序把 3W 元数据表中的数据转换成 CNMARC 记录，也就是在数据库中，对同一 3W 文档的元数据描述保存 DC 元数据与 CNMARC 记录两种描述格式，这两者之间存在一一对应关系。[6] 当服务端接收到来自用户的查询请求时，查询所需的 DC 元数据，并直接从 CNMARC 表中返回到对应的 CNMARC 记录。

5 CNMARC 与 DC 转换过程中存在的问题

虽然有了 CNMARC 与 DC 的映射关系表，但是由于 CNMARC 的结构严密性比 DC 完备，CNMARC 的描述能力也远

远丰富于 DC，所以在两者实际转换过程中就存在一些亟待解决的问题，具体如下：

5.1　CNMARC 与 DC 元素的对应

在 CNMARC 中，正式定义的字段有 176 个，每个字段下面一般又含有几个子字段。DC 虽然只有 15 个元素，但它具有可扩展性，并且每个元素、子元素都可以选择和重复。因此，要把 CNMARC 中各个字段与 DC 中的元素准确地一一对应，显然是一个难题。而且，由于 CNMARC 与 DC 的结构差异较大，所以在 CNMARC 与 DC 的转换中，就存在许多一对多的问题。比如：DC 中的 Creator 元素就对应 CNMARC 中 200$f、200$g 和 7- - 字段（个人责任者以及团体责任者说明）。而且，并非所有的 DC 元素都是可以通过映射关系转换的，比如：Coverage（覆盖范围）、Rights（权限管理）这两个元素目前在 CNMARC 字段中就很难找到对应的部分。这是因为 DC 是为网络资源的著录而制定的，其对象是电子资源，而 CNMARC 主要针对的是传统信息资源，有印刷型出版物、本地存取的计算机文件等，所以两种资源的具体类型、使用的编排格式、编码标准等均不同，二者的设计主导思想也不完全相同。因此，上述两个元素就无法在 CNMARC 中找到对应的映射，所以只好暂时将这两个元素放入 300 字段（一般性附注）中。

5.2　CNMARC 与 DC 的记录转换

在字段对应的基础上，还存在着记录转换的问题。DC 的特点是简洁，但简洁也容易造成字段定义概念模糊，这种模糊说明容易造成使用者对字段的不同理解，很容易使 DC 与 CNMARC 在转换过程中产生歧义和不确定性。比如：Creator 和 Contributor 应该如何划分，DC 中并没有一个明确详细的说明，不同的使用者就可能有不同的理解。在 CNMARC 与 DC 记录转换中还存在一个实际问题，那就是国内图书情报机构在进行 CNMARC 的著录时，绝大多数都没有把 CNMARC 的所有字段进行著录，而只是著录一些基本的项目和字段。这样，CNMARC 中的一些与 DC 对应的字段就无法找到。[7]比如：与 DC 中 Format（格式）对应的 336 字段（计算机文件类型附注）、337 字段（计算机文件技术细节附注）字段基本没有著录，这在记录转换中是一个需要考虑的问题。

5.3　DC 的本土化

DC 是国外提出的一种元数据标准，应用到国内必然有一个本土化的问题。第一个面临的问题是：究竟是采用 DC 的英文元素名称进行著录比较好，还是用中文译名来进行著录比较好？目前，DC 的中文译名还没有统一，比如：Description 就有摘要、说明、内容描述 3 种译名；Creator 就有创作者、创建者、作者或创造者 4 种译名。至于 DC 的扩展集的译名就更多了，这种译名的不统一势必会影响 DC 的使用和发展，所以最好统一规范 DC 的元素译名。

5.4　转换过程中的数据缺失

从理论上讲，从 DC 到 CNMARC 的转换并不是一件很困难的事情，任何一种结构化的数据都可能转换成另一种数据结构，但是这种转换不可避免地会造成一定程度的数据缺失。国外有研究成果表明：在合适的条件下，一个 DC 记录有可能转换成一个比较全面的 MARC 记录，但该记录可能不是一个有效的 MARC 记录，因为它丢失了一些强制字段，如：记录标签；001 记录标记符；100 通用处理数据等，而构造“记录标签”和“通用处理数据”却是其中最难解决的问题。[8]如何避免这种转换中的数据缺失呢？美国国会图书馆和拥有全球最大的 MARC 编目资源数据库的 OCLC 正在积极致力于这一方面的研究，目前越来越多的图书馆和研究机构都加入到了这个课题的研究行列，取得显著成绩的有北欧 Metadata 计划小组和英国的 ROADS 计划小组等。

6　研究展望

综上所述，CNMARC 与 DC 元数据的转换在理论上具有必要性，在实践中具有可行性，在这种理论指导的前提下，最重要的就是根据我国图书馆的实际情况研制和开发出成熟的转换软件，并解决好转换过程中存在的问题。只有开发出完善、成熟的转换软件才能快速、批量地完成 CNMARC 与 DC 元数据间的数据转换，并放置到网络上供用户使用，从而为数字图书馆的建设奠定坚实的技术基础，更好地推动数字图书馆的建设和发展。

参考文献：

[1] 周建清. MARC 与 DC 元数据对比研究[J]. 中国科技信息，2006，(8).

[2] 庞孝梅.基于网络信息资源组织管理的 DC 与 MARC 比较研究[J].大学图书情报学刊，2006，(1).

[3] 杨小云.CNMARC 与 DC 的相互转换[J].科技情报开发与经济，2005，(11).

[4] 黄国忠.DC 元素集与 CNMARC 字段匹配研究[J].图书馆理论与实践，2005，(1).

[5] 李睿华.DC 与 CNMARC[J].科技情报开发与经济，2003，(12).

[6] 赵光林.MARC 与 Dulin Core 之比较研究[J].情报学报，2002，(4).

[7] 许四洋，柳晓春.Dulin Core 元素与 CNMARC 字段的匹配、对应[J].大学图书馆学报，2001，(5).

[8] 朱贵玲，焦素琴.从 DC 看 CNMARC[J].图书馆工作与研究，2002，(2).

作者简介：刘圆圆（1974-），女，西北工业大学图书馆馆员；刘军华（1970-），男，西安财经学院馆员。