·信息工作·

元数据在汉语文古籍数字化中的应用

熊 静 (中山大学资讯管理系 广东广州 510275)

摘 要:文章首先对古籍和古籍元数据的定义和范围进行了界定,介绍了我国古籍元数据标准建设的现状,在比较MARC格式和基于DC的元数据格式的基础上,认为DC元数据更能适应网络环境,应当成为古籍数字化建设的首选。

关键词:古籍数字化 元数据 机读目录格式 都柏林核心元数据

中图分类号:G255.1 文献标识码:A 文章编号:1003-6938(2010)01-0092-04

The Application of Metadata in Digitizing Chinese Ancient Books

Xiong Jing (The School of Information Management of Sun Yat-sen University, Guangzhou, Guangdong, 510275)

Abstract: This article defines ancient books and their metadata. Then it briefly introduces the status quo of standardized building of ancient books metadata. Finally it suggests that DC metadata should be first choice for the construction of digitizing Chinese ancient books since it adapts well to the network environment easily.

Key words: digitization of ancient books; metadata; Machine-Readable Catalogue; Dublin Core metadata

CLC number: G255.1 Document code: A Article ID: 1003-6938(2010)01-0092-04

据调查,我国现存的古籍数量约有13万种,1000万册以上,[1] 为了更好地整理和利用这批数量巨大的珍贵历史文献,各大图书馆和收藏机构均立足馆藏,致力于古籍数字化建设。

1 古籍和古籍元数据

古籍是指1912年以前在中国书写或印刷、具有中国古典装订形式的书籍。[2]但在实际工作中,对民国年间乃至1949年以后产生的以反映中国传统文化为主的文献,图书馆多将其与古籍一起统称为线装书或古旧书,在收藏和编目时不作严格的区分。因此,在确定古籍元数据的著录范围时,只将1912年作为古籍历史分期的概念,[3]以装帧形式和内容特征作为主要分类标准,既包括产生于1912年以前的汉语文典籍,也包括民国年间甚至1949年以后书写或印刷的、具有中国古典装订形式并反映中国传统文化的书籍。

元数据通常被定义为关于数据的数据,其目的在于提供一个中间级别的描述,人们据此就可以确定孰为其想要浏览或检索的信息包,而无需检索大量不相关的全文文本。[4]古籍元数据可以简单定义为,描述的信息对象为古籍的元数据。

根据各自的不同作用,古籍元数据被分为三种类型:描述性元数据、管理性元数据、应用性元数据。[5]传统的MARC格式,其实也是一种描述性元数据;而DC元数据以及基于DC的其他数据格式,是网络环境下应用最为广泛的一种元数据标准。下面将从两方面,对我国古籍元数据标准的建设情况进行梳理。

2 我国古籍元数据标准建设现状

2.1 MARC机读目录格式

(1) 中国国家图书馆《汉语文古籍机读目录格式使用手册》

国家图书馆自1999年开始古籍书目数据库建设,经过论证,选择CNMARC格式作为描述信息对象的标准,编制了《古籍机读目录格式字段表》,该表规定了古籍机读目录所使用的字段、子字段及其记载各项古籍书目信息的格式。2001年10月,为了适应实际工作的需要,国家图书馆在对古籍机读目录格式深入研究的基础上,制定了《汉语文古籍机读目录格式使用手册》,将其作为规范文件,指导该馆的古籍书目数据库建设。迄今为止,国图共完成了30余万条的古籍书目数据,[6]并在2003年初完成了全部普通古籍回溯书目数据库建设

收稿日期:2009-06-05 责任编辑:宋 戈

工作，并进入该馆馆藏书目数据库系统，供用户通过网络访问。

国图古籍机读目录格式的著录规则完全按照国家标准《古籍著录规则》执行。考虑到字库原因，以及与馆内外其他文献类型书目数据库统一合库的问题，著录用文字采用了规范的简体字。[7]该格式的主要特点有：(1) 和普通图书机读目录尽量保持一致。(2) 遵循完整本著录原则，古籍完整本的特征著录在140或193、200、205、215、305、306、307等字段。没有完整本时（如没有全本），参照完整本的书目著录进行手头古籍复本的编目。(3) 区别复本的标准是古籍写刻成书时已有的版本特征，与成书后在流传收藏过程中形成的版本特征，如藏书章、圈点、残缺等无关。复本统一著录为一个数据，不同的题跋者在316字段说明。(4) 连接字段的处理，启用4字段，使用该字段的各个子字段反映丛书子目、合刻、合订等复杂的关系。

总体来说，国图格式是CNMARC在古籍编目上的一次尝试，最为突出的特点在于保持了与图书馆现有书目系统的一致性和兼容性，但是从另一方面而言，却忽视了对古籍著录特殊性的要求，[8]如在2字段的著录中部分规定就未能符合客观著录的要求。

(2) CALIS古籍联合目录——《CALIS古籍联机合作编目规则》

CALIS（中国高等教育文献保障系统）的古籍数字化建设始于2001年。2001年10月14~16日，在北京大学召开的“CALIS古籍联合目录数据库建设研讨会”，揭开了CALIS古籍联合目录系统建设的序幕。在元数据建设方面，CALIS选用了国家图书馆《汉语文古籍机读目录格式》（2001年6月版）为蓝本，组织部分高校古籍编目骨干编写《CALIS古籍联机合作编目规则》。经过一系列的完善和修改，《CALIS古籍联机合作编目规则》于2003年定稿。[9]同年12月中旬，CALIS古籍联合目录系统正式启动。经过多年的努力，目前，用户已经可以通过CALIS联合目录公共检索系统访问成员馆提供的数据。

CALIS古籍机读目录以国图格式为蓝本，相对于国图格式，CALIS的创新点在于：(1) 严格遵守客观著录原则，凡取自规定信息源以外的信息，或编目员自拟的著录信息均需置于“”内，并在附注项中加以说明。(2) 依据品种和版本立目。对同品种、同版本的书，仅收录一条书目记录。通常以先递交的记录为主，收藏复本的成员馆只需于该记录下添加馆藏信息。(3) 书影提交原则。[10]为了便于查重，要求编目员提供原本首卷卷端原大书影图像一页或其它能客观反映该文献版本信息的原大书影图像一页。

2.2　基于DC的元数据标准

(1) CDLS子项目——古籍元数据规范

《我国数字图书馆标准与规范建设》项目（CDLS）是我国科技基础性工作专项资金重点项目，立足于制定我国数字图书馆标准规范发展战略与标准规范框架，建立数字图书馆核心标准规范体系。2002年10月，该项目启动了《我国数字图书馆标准规范专门数字对象描述元数据规范》子项目，古籍同金石拓片也属其中的一种专门对象。子项目由北京大学图书馆牵头，联合CALIS管理中心、上海图书馆等8家单位共同完成。2002年~2004年间，经过了资源分析、标准草案、试验著录、开放应用及试验等阶段，最终推出了包括古籍元数据标准在内的推荐报告。[11]

迄今为止，该项目产生的有关古籍元数据标准的规范文件包括：《古籍描述元数据规范》（2004.06.07）；《古籍描述元数据著录规则》（2004.06.07）；《古籍元数据规范》（2006.11.22）；《古文献系列-资源分析报告（舆图、古籍、拓片）》（2006.11.22）与《古籍著录规则》（2007.1.19）。

古籍元数据基于DC构建，在吸收了DC核心元素的基础上加入了部分古籍专门元素而成，共有17个元素，由核心元素、古文献系列核心元素组成。每个元素由数量不等元素修饰词，元素编码体系构成。在实际工作中，如有特别需要，可遵循《专门元数据规范设计指南》（CDLS-S05-001）中的扩展规则添加本地元素，以满足收藏机构的特殊要求。[12][13][14]

13个核心元素为：资源类型、题名、主要责任者、其他责任者、日期、出版者、附注、相关资源、主题词、古籍语种、时空范围、标识符、权限管理；4个古文献类型核心元素为：版本类别、载体形态、收藏历史、馆藏信息。

古籍元数据的主要特点有：(1) 著录层级至版印一级，版印是指同一书版的不同印次。由于同一书版在流传过程中可能会发生删改、剜补等各种变化，且考虑到书目数据与图像、全文的对应关系，著录时应该以单个的本子为著录单位。[15]对于复本，如数据库中已有其它馆关于该书的记录，则编目馆必须利用现有记录改填本馆馆址和典藏号以及其他藏本信息，采用套录的方式提交。(2) 丛书与子目：丛书记录可以不著录子目，子目单独著录，但子目单独著录时必须在“相关文献”元素中著录所属丛书名。(3) 原抄、原刻、原印与影抄、影刻、翻刻、影印各本均单独著录，但影抄本、影刻本、翻刻本、影印本应在“相关文献附注”子元素中注明所依据的底本。(4) 合刻书、合印书、合函书、合装书等除

同一著者的合刻书可采取合并著录的方法外，其他一般均单独著录。但应在“相关文献”元素中著录相关古籍的题名。⑷ 附录、附刻：附录、附刻一般不为之另做记录。附录如无题名但有卷数，在题名说明文字中著录；附录如有题名，在“附注”元素位置著录其名。附刻题名一般均应著录在“相关文献”元素中。[16]

⑵ 中科院古籍数据库——DC元数据格式

中科院图书馆于2003年初，开始“中国科学院图书馆古籍目录网络数据库”建设，在比较了MARC和其他一些元数据标准后，直接选择了DC元数据格式作为古籍著录的标准。截至2004年9月，已完成了8.4万余条数据的录入，用户可通过中科院图书馆网页进行访问和检索。并在在2003年6月编制了《中国科学院图书馆古籍目录网络数据库著录条例》和《中国科学院图书馆古籍目录网络数据库各字段著录解释》。

由于直接吸收了DC的元素构成，中科院古籍元数据格式分为：题名区、著者区、相关文献区、版本区、统计区、说明区、丛书名、丛书子目、责任区等9个主要区段。[17]每个区段下又设置若干个子字段，达到充分描述古籍对象特征的要求。

3 MARC和基于DC的元数据标准比较

3.1 信息对象描述能力

MARC是一种信息描述能力非常强大的数据格式，9大字段块，以及数以千计的子字段，为描述信息对象提供了无限的空间。相对于MARC，以DC为代表的元数据，往往只有十几个核心元素构成，各元素下的子字段也屈指可数。从表面看来，DC等元数据格式的信息描述能力要远远弱于MARC格式。但据最新的研究统计，在MARC格式的书目数据中，80%的书目记录只使用了36个字段或子字段，国图数据的抽样中多于30个字段的记录只占0.09%，[18] 几乎可以忽略不计。这说明了MARC繁复的字段格式，且重复严重，而真正对读者有意义的字段（主要指与内容描述有关的字段）即可供用户检索利用的字段较少。DC等元数据格式，字段的数量较少，但实际上对信息对象的描述能力并未降低。

3.2 信息对象描述的完整性

MARC格式的一个突出特点就是分段著录，在描述中文古籍时，为了适应MARC字段的定义，传统的MARC格式通常会将一个完整的描述割裂成几个部分进行著录，[19] 表1是笔者比较了CDLS的古籍元数据与CN-MARC字段后得出的，从表中可以看出，一个基于DC元数据的元素可能对应不止一个MARC字段，同类型的信息在不同的子字段里被重复著录，不仅增加了编目人员的工作量，同时也使用户产生了疑惑，难以在浩繁的子字段中找到自己需要的信息。相对于MARC，基于DC的元数据采用了分块著录的思想，将同类型的信息全部集中在一个元素块中，方便了著录，同时也避免了割裂记录。对比之下可以看出，在信息对象描述的完整性上，DC元数据克服了MARC的一些痼疾。由于古籍著录的特殊性，在流传过程中形态发生的每一点变化，都有可能成为辨别古籍版本的重要信息，因此要求如实照录原本上所有细微差别。显然，分块著录的方法更有利用集中同类信息进行比较判断，更加适应古籍著录的要求。

3.3 对古籍数字化的适应性

目前，数字化是公认的解决古籍收藏与利用矛盾的最佳途径。古籍书目数据库和古籍全文数据库建设是古籍数字化的两个重要的方面。古籍数字化的最终目标在于，通过各种数字化手段，为用户提供便利的网络访问环境，满足其研究和利用古籍的需要。为了达到

表1 MARC格式与古籍元数据对照表

元数据名称	对应的CNMARC字段	元数据名称	对应的CNMARC字段
题名	200、510、512、514、515、517字段	相关资源	4××字段
主要责任者	200（$f）、701、711、721字段	主题	600-608、610、696
其他责任者	200（$g）、702、712、722	时空范围	660、661
日期	210（$d、$h）	语种	101
出版者	210（$a、$b、$e、$f、$g）	类型	099
版本类别	205	标识符	099
附注	3××字段	馆藏信息	920
载体形态	010（$b）、215	权限	920（$z）
收藏历史	317字段		

这个目标，古籍数字化必须满足两个条件：第一是适应网络环境；第二是保证与古籍全文数据库的无缝链接。在这两个方面，基于DC的元数据标准无疑更具吸引力。首先，DC元数据本来就是为了适应网络环境而设计的，基于先进的网络技术和最通用的XML网络传输语言，简单易学的数据格式更加便于编目人员的掌握，在简化编目工作的同时，不仅可以更全面细致地对文献进行描述，也使用户对中文古籍的检索变得更易于操作。其次，以版印为著录级别，单本书为著录单位的规定，更加利于书目数据库与全文数据库一一对应关系的建立。而MARC格式设计所依赖的是以磁带为主要存储介质的技术，在目前各种集成系统的技术实现中早已采用了关系数据库技术，乃至其它更为先进的全文索引、面向对象技术甚至XML技术的情况下，著录古籍总不免有“削足适履”之叹。

3.4　数据格式的可扩展性

在扩展性方面，DC和MARC都为编目员预留了扩展本地记录的空间，编目员可以根据本馆需要，按照规定增加相应的子字段。由于DC本身只提供了一些描述信息对象最为基本的属性，使得扩充变得非常简便，就是在DC元数据15个核心元素的标准框架下，根据古籍文献的特点而专门制订与之相适应的元数据标准，CDLS的古籍元数据采用的就是这一做法。经过扩充的元数据标准，既能和已有标准兼容，同时也满足了收藏馆的特殊要求，具有较好的可扩充性。而MARC格式虽然也预留了扩充的空间，但是由于这一格式本身的字段已经非常复杂了，在此基础上的任何添减，都可能导致系统实现上的混乱和不兼容。

综上所述，我们可以得出结论：MARC格式虽然在过去的图书馆工作中发挥了巨大的作用，但是，在网络技术日新月异，用户对获取数字对象内容及方式的渴求不断增强的背景下，MARC格式已经显得不堪重负。因此，在进行古籍数字化建设时，我们应该充分利用后发优势，借鉴信息产业领域业已成熟的理论和标准，选择更加适应网络环境的数字技术标准，如DC元数据等，加快文献信息数字化进程。

参考文献：

[1] 潘德利.中国古籍数字化进程和展望[J].图书情报工作，2002，(7)：117-120.

[2] 富平等.中国文献著录规则[M].北京：北京图书馆出版社，2005.

[3] [4] 刘嘉.元数据导论[M].北京：华艺出版社，2002：61-64.

[5] 姚伯岳等. 古籍元数据标准的设计及其系统实现[J].大学图书馆学报，2003，(1)：17-21.

[6] 董馥荣.国家图书馆普通古籍书目数据库建设工作综述[EB/OL].[2008-06-20].http://www.nlc.gov.cn/old/old/wjls/html/8_09.htm.

[7] 张磊.再论《汉语文古籍机读目录格式使用手册》使用中的问题[J].图书馆工作与研究，2005，(3)：27-29.

[8] 鲍国强. 古籍书目数据库整改工作构想.[EB/OL].[2007-06-20].http://www.nlc.gov.cn/old/old/wjls/html/8_08.htm.

[9] 杨健.CALIS中文古籍联机合作编目的缘起与进展[J].图书馆理论和实践，2006，(9)：54-56.

[10] 喻爽爽，谢琴芳.汉语文古籍文献目录资源的共建共享—CALIS古籍联合目录系统[J].大学图书馆学报，2005，(3)：22-26.

[11]《我国数字图书馆标准规范专门数字对象描述元数据规范》子项目[EB/OL].[2007-06-10].http://cdls.nstl.gov.cn/cdls2/w3c/2003/SpcMetadata/.

[12] 古籍描述元数据规范[EB/OL].[2007-06-10].http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/古籍描述元数据规范.pdf.

[13] 古籍元数据规范[EB/OL].[2007-06-10].http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/CDLS-S05-013.pdf.

[14] 古籍描述元数据著录规则[EB/OL].[2007-06-10].http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/古籍描述元数据著录规则.pdf.

[15] 沈芸芸等.古文献系列资源分析报告(舆图、古籍、拓片)[EB/OL].[2007-06-10].http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/古文献系列-资源分析报告(舆图、古籍、拓片).pdf.

[16] 古籍著录规则[EB/OL].[2007-06-10].http://cdls.nstl.gov.cn/mt/blogs/2nd/archives/docs/CDLS-S05-014古籍著录规则.pdf.

[17] 罗琳.“中国科学院图书馆古籍目录网络数据库”解读[J].中国索引，2004，(3)：8-13.

[18] 让MARC安乐死？[EB/OL].[2007-08-10].http://my.donews.com/keven/2007/03/18/post-070318-134027-225/.

[19] 程佳羽.古籍全文数据库的理想实现模式[J].图书馆建设，2006，(3)：54-56.

作者简介：熊静(1984-)，女，中山大学资讯管理系08级博士生。