

●真 漆

矛盾重重的元数据世界

摘 要 元数据世界的主要矛盾包括:由非专业人员还是由专业人员提供元数据;元数据格式应该是一个简单体系还是复杂体系;通用、专用和互操作问题。参考文献 13。

关键词 元数据 格式 标准化

分类号 G254.36

ABSTRACT In the world of metadata, there are controversies over the questions whether metadata should be provided by professionals or non-professionals, whether metadata format should be a simple system or a complex system, whether a metadata has a single standard or multiple standards, etc. 13 refs.

KEY WORDS Metadata. Format. Standardization.

CLASS NUMBER G254.36

1 元数据是什么

元数据是关于数据的数据。在图书馆信息界,一种较为正式的定义是:元数据是“结构化的编码数据,用于描述载有信息的实体的特征,以便标识、发现、评价和管理被描述的这些实体”^[1]。近年来被广泛讨论的 DC (Dublin Core; 都柏林核心) 是一种元数据,图书馆信息界长期使用的机读目录也是一种元数据。国外有学者将搜索引擎自动抽取的全文索引也归入元数据的范围,虽然它们的确包含在上文所述定义的范围之内,但为了更清楚地论述问题,下文所述的元数据不包括它们。

不管人们是否意识到,从开始利用元数据的第一天起,它们就在信息资源的查找、标识、挑选、获取和管理等方面发挥作用。对于传统的非电子型文献, MARC 和 CCF (Common Communication Format: 通用通讯格式) 等基于 ISO 2709 的元数据完全能够胜任。经过 30 多年的发展,图书馆信息界已经提出了一整套切实可行的方案,制订并使用较为完备的元数据标准及配套的编目规则。

但对于万维网上的信息,这些“老家伙”显得既笨拙又昂贵。Rachel Heery 指出网络信息资源具有以下特点:①文档位置和内容不稳定;②文档格式与版本多种多样;③存在大量冗余数据;④一件文档和一组文档界线模糊;⑤缺少统一的定位数据标准^[2]。笔者认为,最重要的一条是万维网上信息资源的规模和增长速度引起了“资源描述危机”,人们不得不对原有的图书馆信息界编目-标引体系做出调整。

第 1 种思路是,抛开元数据,完全依靠全文检索。这可以称为搜索引擎方案。衡量信息检索的两个传统指标是查全率和查准率,并且通常情况下人们更看重查准率。但仅

仅依靠搜索引擎很难得到满意的查准率。一般可以接受的查全率和查准率是 10%~20%。澳大利亚国家图书馆的研究人员发现,搜索引擎的查准率远远低于 1%,也就是说,搜索引擎提供的无用信息远远多于有用信息^[3]。第 2 种思路是,信息资料的创建者或发布者提供元数据,理想情况下,元数据和网络信息资源能够同步增长,再加上所提供的元数据质量较好,“资源描述危机”也就不复存在。这可以称为“人海”方案。第 3 种思路是,专业人员(即图书馆编目人员和信息机构标引人员)在简化原有编目过程的基础上提供元数据,编目效率得到提高,“资源描述危机”也就得到缓解。这可以称为“简化”方案。后两种思路是当前图书馆信息界讨论的重点,也是本文的重点。

2 元数据格式概况

2.1 基于 ISO 2709 的 MARC 和 CCF

MARC 格式目前已为大家熟悉,但人们现在很少提到另一种基于 ISO 2709 的元数据格式:CCF。和 MARC 格式相比,这是一种“年轻”的格式:联合国教科文组织 (UNESCO) 1984 年发布了第 1 版 CCF。制定 CCF 的一个目的是为了桥接已经存在的通讯格式,因此设计过程中参考了多种通讯格式(其中包括 MARC 格式)。CCF 的数据元素较少,但由于较多地考虑了连续出版物和会议文献的情况,它提供了一套备受争议的复杂的记录链接方法。1992 年发布的第 2 版 CCF 中,在 CCF/B (Common Communication Format for Bibliographic Information: 书目信息通用交换格式) 的基础上派生出了一个 CCF/F (Common Communication format for Factual Information: 事实信息通用交换格式)。这是目前已知的基于文献信息数据处理的机读目录格式标准中,惟一将非书目信息资源与书目文献信息资源融为一体的实例^[4]。

基于 ISO 2709 的元数据格式被认为是最严格的,需要配套使用编目规则。例如使用 MARC 时会使用 AACR2 或者 ISBD,使用 CCF 时也需要使用 ISBD。

2.2 基于 ISO 8879 (SGML) 的 TEI, EAD 和 CIMI

尽管 5 年前美国国会图书馆就提供了 MARC DTD (Document Type Definition: 文件类型定义),但习惯上,人们仍然认为 MARC 格式基于 ISO 2709 而不是 ISO 8879。无论 ISO 2709 还是 ISO 8879,都只是编码及传输语法层标准,而不是语义层标准。因此依据它们来归并元数据格式只是一种简便的方法,巧合的是传输语法的选择体现了元数据每个发展阶段的特点。

严格说,TEI (Text Encoding Initiative: 文本编码倡议) 格式并不是一种元数据格式。所有 TEI 文档都包括一个 TEI 头标(header)和文本的复本。TEI 头标是一种元数据格式。TEI 主要用于处理电子型文本,也可以用来处理各种非文本的电子型资源。其最初设定的目标是:电子型文本的哪些文本特性应该编码及描述;怎样才能实现无损的、独立于平台的及可交换的编码。因此 SGML (Standard Generalized Markup Language: 标准通用置标语言) 就成为必然的选择。

利用 SGML 的元素和属性,TEI 定义了 400 多个文本特性。这些元素构成了 4 组标签集(tag set):核心标签集、基本标签集、附加标签集和辅助标签集。TEI 头标是一个核心标签集。TEI 头标有 4 个组成部分:〈fileDesc〉记录电子文档本身及来源;〈encodingDesc〉记录采用的编码系统;〈profileDesc〉记录文档及其主题的描述性信息;〈revisionDesc〉记录文档的历次修改^[5,6]。

本文不介绍 EAD (Encoded Archival Description: 编码档案著录) 和 CIMI (Computer Interchange of Museum Information: 博物馆信息计算机交换) 的情况。基于 ISO 8879 的元数据格式也较为严格,通常认为它们与书目记录的要求最为接近,例如 TEI 头标字段的设计基于 ISBD。在 DESIRE I (Development of a European Services for Information on Research and Education I: 第 1 期欧洲研究与教育信息服务发展计划) 的一份报告中,它们和 MARC 同被列为具有复杂描述格式的第 3 类^[7]。但是和 MARC 不同的是,这些元数据格式没有指定配套使用的编目规则,当然不是因为疏忽,这一点下文还将论述。

2.3 独立于传输语法的 DC

DC 算得上是一种新的元数据格式。从 1995 年 3 月 OCLC 和 NCSA (National Center For Supercomputing Applications: 美国超级计算应用中心) 主持第 1 届元数据研讨会,到 2000 年 10 月第 8 届研讨会,历次研讨会都标志着 DC 的一个新阶段。DC 标准包括一个元素集 DCMES (都柏林核心元数据元素集) 和一个修饰词集。从第 3 届会议开始,DC 的 15 个元素就被确定下来,它们是 Title (资源名)、Creator (创建者)、Subject (主题和关键词)、Description

(说明)、Publisher (出版者)、Contributors (其他责任者)、Date (日期)、Type (类型)、Format (形式)、Identifier (资源标识符)、Source (来源)、Language (语种)、Relation (关联)、Coverage (覆盖范围) 和 Rights (权限管理)。DC 还包括两类修饰词:一类用于将元素的含义限定得更更为专指,另一类用于说明元素值的编码体系^[8]。

DC 基于这样一种假设:DLO (Document Like Object: 类文档对象) 是因特网上的主流资源,因此它主要用来描述网络环境中的 DLO 资源。和 TEI 的情况类似,利用 Warwick 框架,DC 也可以用于描述图像资源。第 3 届元数据研讨会专门讨论了网络环境下图像资源描述问题^[9]。

第 1 届元数据研讨会为 DC 确定了这样一些特性:内在性、可扩展性、语法独立性、可选择性、可重复性及可修改性^[10]。这些特性本身就是充满争议的、模糊的。仅以“内在性”为例,什么是信息资源的“内在”特性? 什么又是信息资源的“外在”特性? 电子型文档的版本是它的“内在”特性还是它的“外在”特性?

3 元数据世界的主要矛盾

3.1 非专业人员提供,还是专业人员提供

TEI 和 DC 都是前文所述第 2 种思路(即“人海”方案)的结果,最初的动机都是希望非专业人员提供元数据。TEI 希望对文本进行置标的人员提供相应的元数据,这些人员大多是人文学科的学者。尽管 TEI 头标的字段设计基于 ISBD,却有意识地回避了复杂的编目规则,这样不需要编目知识就能完成资源描述。DC 则更进一步,希望资源创建者提供元数据。置标人员或资源创建者在自己的领域里可能都是专家,但对编目规则很可能一无所知或者知之甚少。因此这一矛盾的实质就是,编目规则真的是必须的吗? 更进一步的问题是:专业人员编目的价值是什么?

编目主要涉及 3 类活动:组织过程、分析过程和判断过程。组织过程包括分类、归并及格式化等过程;分析过程则包括区分、验证、评估、比较、解释及综合过程^[11]。总而言之,编目是一个复杂的、信息增值的过程。利用编目规则及其他一些手段(例如规范记录),专业人员能够提供具有较高质量的元数据。那么这种方式究竟存在哪些问题呢? 主要是费用,一个是时间的花费,另一个是金钱的花费。仅以后者为例,1996 年 10 月至 1997 年 9 月,国会图书馆完成编目 289154 卷,平均每卷编目需要 87.05 美元^[12]。

非专业人员提供元数据的好处是:1) 费用降低;2) 元数据的专指性较好。但是同时应看到不利因素是:1) 如果是创建者提供元数据,客观性值得怀疑。由于搜索引擎结果的相关性排序往往依据某个关键词出现的频率,的确有人使用这一办法来提高网页在相关性排序中的名次。2)

元数据的总体质量值得怀疑。为了能够比较方便地叙述问题,笔者曾将元数据分为编目元数据和普通元数据。美国图书馆协会 1998 年完成的一项研究表明,只有充分地利用各种编目规则和其他一些质量保证手段(例如规范记录),普通元数据才能具有较高的质量,才能用于图书馆编目并成为编目元数据^[13]。

无论其初衷是什么,到目前为止,绝大多数元数据(包括普通元数据和编目元数据)仍然由专业人员提供,绝大多数有关元数据的讨论也只在专业人员范围内进行。因此,这样一些问题仍然存在:怎样让更多的非专业人员提供元数据?怎样管理和提高他们提供元数据的质量?

3.2 简单还是复杂

制订一种简单的元数据标准几乎成为更多非专业人员参加资源描述活动的一个必要条件。很难想象一个非专业人员会利用复杂的元数据标准进行资源描述。专业人员也欢迎简单的元数据标准,因为这意味着能够更快地掌握编目方法,更好地运用编目工具,更容易实现高质量的元数据,这属于前文所述第 3 种思路。

同样是基于 ISO 2709 的元数据格式,CCF 和 MARC 的情况很不相同。1992 年发布 CCF/B 和 CCF/F 之后,CCF 几乎没有任何进展,一个重要原因是研制小组一直将简单性作为 CCF 的一项根本原则。这和同一时期 MARC 发展(复杂化)的情况形成了鲜明对比。

TEI 头标是一种复杂的元数据格式。到目前为止,绝大多数 TEI 文本都是由专业人员创建,而 TEI 头标经常是图书馆员创建或维护。

DC 最初的定位是处于全文索引和精细的元数据格式(例如 MARC 和 TEI)之间的一种元数据标准,提供简明及易于理解的元数据格式。经过数次研讨会之后,Priscilla Caplan 发现 DC“不可避免地变得复杂起来”。一部分人坚持简单性是 DC 的最主要特征,另外一些人则不这么看。到 1997 年 3 月第 4 届元数据研讨会,这两种观点形成了两个学派,即最小主义(Minimalist)学派和结构语言(Structuralist)学派。前者几乎完全由图书馆界或者熟悉图书馆界复杂元数据标准(例如 MARC)的人员组成,后者则多来自技术界或非传统编目界。最小主义学派认为简单性不仅是非专业人员提供元数据的一个重要条件,对于那些不使用元数据复杂特性(例如修饰词)的索引引擎而言也很重要。并且不同的团体使用修饰词这些特性可能会引起元数据元素语义的偏移(drift),这将对语义互操作产生不利影响。结构语言学派则认为,只要能满足需求,更灵活地扩展和修饰元素,牺牲一些语义偏移也是值得的。

因此问题的关键应该是:为什么要使用扩展和修饰这些导致元数据格式复杂化的方法?显然,如果一个简单的元数据标准够用,是没有必要使用这些方法的。那么为什么简单的元数据标准不够用呢?有两种可能:应用的领域比较复杂,元数据标准没有提供足够的元素来描述资源的

各种重要特征;人们希望某种元数据标准应用于多个领域。

3.3 通用、专用和互操作

通用元数据标准的问题经常被人提起。“通用”意味着统一,编目人员不必学习不同的元数据使用方法,元数据研究人员不必考虑桥接各种元数据的方法,系统实现人员只需考虑单一的元数据格式。因为有这么多的好处,尽管目前通用元数据还没有成功的例子,并且有人将“通用元数据”比喻成一个陷阱,仍然不断有人在做这样的努力。

大多数人放弃通用元数据格式的一个原因是:它将是极其复杂的,因而研制和使用也将是极其困难的。DC 似乎是对这种观点的一种挑战:创建一种简单的、能够在网络中为各种用户团体接受的元数据元素集。问题是:人们发现,实际应用过程中,仅仅依靠 DC 的 15 个元素加上修饰词是不够的。那么为什么 DC 发布之后被广泛地试验和采用?其中的一个重要原因是它支持甚至鼓励人们进行局部扩展。如果有合适的 DC 元素就使用这个元素,如果没有就创造一个元素,创造出的这个元素的作用局限于某个具体应用或某个具体领域。相似领域或应用很可能会使用相同的元素约定,也可能使用同名但语义不同的元素。为了避免重复的甚至可能引起冲突的设计,名称空间(namespace)及注册机制(registry)应运而生。

Thomas Baker 将 DC 比作旅游者使用的皮钦语(pidgin)。人们去国外旅游问路或者点菜,在不熟悉当地语言的情况下,只能使用最基本的词汇,对方也会使用一些简单词汇来回答。这样的比喻非常恰当地描述了 DC 的简单性和通用性,也描述了不同领域和应用对 DC 进行局部扩展的必要性。DC 通用性不同于传统意义上的通用性,后者经常与完备性和封闭性同时出现,前者却依赖于开放性和可扩展性。

使用不同名称空间的 DC 应用怎样实现互操作是当前最热门的话题之一。早在第 2 届研讨会上,Carl Lagoze 就提出了 Warwick 框架作为一种解决办法。现在,人们的注意力主要放在了 XML(eXtended Markup Language:扩展置标语言)和 RDF(Resource Definition Framework:资源描述框架)上。与名称空间同时出现的还有注册机制问题,已经有多家单位遵循 ISO 11179 发布了自己的注册机制。ISO 11179《信息技术——数据元素规范和标准》分为 6 个部分:第 1 部分,数据元素产生和标准化框架;第 2 部分,域标识概念的分类;第 3 部分,数据元素的基本属性;第 4 部分,数据定义公式化规则和指南;第 5 部分,数据元素命名和标识原则;第 6 部分,数据元素注册。

名称空间和注册机制产生的好处及带来的问题是:1)可以将多种简单元数据标准结合起来,而不必研制一种复杂的元数据标准,但编目人员可能需要熟悉不同的元数据标准,因此编目工作的复杂程度没有降低反而增加;2)简单元数据标准能够方便地扩充,至少能做到应用间及领域

间互不干扰,并能提供一定的互操作,此时的确能降低编目工作的复杂程度,但即使同样是基于相同元数据标准(例如 DC)的应用,也可能存在兼容性和互操作问题。

3.4 标准和兼容

相对而言,MARC 格式是所有元数据格式中发展最完备、应用最广泛的一种。但 MARC 格式只是一种很笼统甚至会引起误解的说法,因为 MARC 格式不是指一种格式,而是一族格式的统称。例如 MARC 在美国国会图书馆和大英图书馆合作的时候就已经不是一种标准,而是两种标准,即后来的 USMARC 和 UKMARC。不少国家紧跟其后制定了各自的 MARC 标准,尽管他们也考虑了与其他国家标准的兼容性,但结果仍然是多种不尽相同的 MARC 标准。例如加拿大的 CAN/MARC 专门考虑了英法双语的名称款目和主题款目。这些 MARC 格式之间的转换并不像想象的那么简单。

所有 MARC 格式的公共部分是它们能够彼此交换的部分,这是 IFLA 设计 UNIMARC 的基本思路。理想的情况是,某个国家的图书馆只需要开发两个转换程序就能实现到其他任何一种 MARC 格式的转换:一个是从本国 MARC 格式转换到 UNIMARC 格式,另一个是从 UNIMARC 格式转换到本国 MARC 格式。这一思路和发明世界语的思路多少有些类似。但实际上 UNIMARC 格式不可能仅仅使用公共部分,因为许多书目数据元素并不存在国际认可的标准形式,此时 UNIMARC 就要扮演一种足够灵活的“中性”容器的角色,包容各国图书馆不同的处理惯例和方法。这样做的结果是 UNIMARC 不仅是一种交换用的格式(这是设计 UNIMARC 的初衷),而更是一种近乎独立的 MARC 格式。

根据 IFLA UBCIM 计划的调查,90 年代中期,全球 60 个国家的国家图书馆采用了基于 MARC 的格式。其中近半数采用基于 USMARC 的格式,另有 1/4 采用基于 UNIMARC 的格式。

USMARC、UKMARC 和 CAN/MARC 是英语书目记录的 3 种主要格式。1995 年 12 月美国国会图书馆、大英图书馆和加拿大国家图书馆正式启动了 3 种格式的协调计划,最终要形成一种统一的 MARC 格式,因此称为合并计划更为准确。1999 年 UKMARC 和 CAN/MARC 完成合并,新的 MARC 格式称为 MARC 21。美国国会图书馆 MARC 标准网站上几乎看不到 USMARC 这个词,取而代之的是 MARC 21。所以现在说 USMARC、UKMARC 和 CAN/MARC 是 3 种 MARC 格式并不准确。UKMARC 和 MARC 21 的差别较大,协调起来存在很多困难。因此原来 3 种格式最终合并成一种格式还需要一段时间,但人们已经看到了解决长期存在的 MARC 标准和兼容问题的一

缕曙光。

4 结束语

现在回过头来看看本文最初提出的两种思路。“人海”方案经常作为元数据标准提出的一种动机出现,但由于太理想化,在元数据发展的过程中逐步淡化。“简化”方案在某些方面已经看到希望,但很多问题仍然值得人们更深入研究。元数据世界复杂而多变,分析其中的主要矛盾,能帮助我们较为迅速和准确地辨明方向,不失为是一种可取的方法。

参考文献

- 1 CC:DA Task Force on Metadata: Final Report. <http://www.ala.org/alcts/organization/ccs/ccda/tf-meta6.html>
- 2 Rachel Heery. Review of Metadata Formats. <http://www.ukoln.ac.uk/metadata/review.html>
- 3 Metadata: an Overview. <http://www.nla.gov.au/nla/staffpaper/cathro3.html>
- 4 潘岩铭等. 从机读目录格式到通用记录格式:面向未来的信息服务. 中国图书馆学会 1999 年会论文
- 5 Lou Burnard. Text Encoding for Information Interchange: An Introduction to the Text Encoding Initiative. <http://www.uic.edu/orgs/tei/info/teij31/>
- 6 韩珏. 元数据标准概述. <http://www.istis.sh.cn/istis/dlib/report/metadata.htm>
- 7 Lorcan Dempsey. Specification for Resource Description Methods. Part 1. A Review of Metadata: a Survey of Current Resource Description Formats. <http://www.ukoln.ac.uk/metadata/desire/overview/>
- 8 都柏林核心修饰词. <http://www.istis.sh.cn/istis/dlib/report/DCQ.htm>
- 9,10 都柏林核心 (Dublin Core) 元数据发展简史. <http://www.istis.sh.cn/istis/dlib/report/dc.html>
- 11 Robert Saxton Taylor. Value-added Processes in Information Systems. Ablex Publishing, 1986
- 12 Cataloging at the Library of Congress. <http://lcweb.loc.gov/faq/catfaq.html>
- 13 真溱等. 未来图书馆编目技术. 情报理论与实践, 2000 (3)

真溱 高级工程师。通讯地址:北京 1229 信箱 7 分箱。邮编 100036。

(来稿时间:2001-01-10)