图书馆实体信息资源组织的两大发展路径

王松林

摘　要　随着搜索引擎的日益学术化,图书馆非但没有对互联网信息资源进行有效的书目控制,而且其实体信息资源的组织也有被搜索引擎资源组织所替代的趋势。针对搜索引擎资源组织的优缺点,提出图书馆实体信息资源的两大发展路径——章节化组织和 FRBR 化组织。前者可以解决网络阅读"短、平、快"的问题,后者则可解决检全率尤其是检准率的问题。图 1。参考文献 15。

关键词　图书馆　实体信息资源　信息组织　章节组织　FRBR

分类号　G254

ABSTRACT　As the search engine gradually becomes academic, library has not effectively taken bibliographic control of the networked information resource. Instead, there is a trend that its physical information resource organization will be replaced by the search engine's resource organization. Considering the merits and drawbacks of the search engine's resource organization, the author proposes two ways and means of the library's physical information resource organization, which are piece-analytical organization and FRBR(Functional Requirements for Bibliographic Records) organization. The first deals with the problems of being short, easy-understanding and quick in the networked reading, and the second addresses the problems of recall ratio, and especially of precision ratio. 1 fig. 15 refs.

KEY WORDS　Library.　Physical information resource.　Information organization.　Piece-analytical organization.　FRBR.

CLASS NUMBER　G254

信息资源可分实体信息资源和网络信息资源两大类,如果说对网络信息资源的组织是搜索引擎的强项,那么对实体信息资源的组织则是图书馆的强项。如同现在搜索引擎力图将实体信息资源组织纳入自己的范围一样,图书馆最初也曾想对互联网上的信息资源进行书目控制,并在 1997 年创办了网络编目的专业期刊《因特网编目杂志》(Journal of Internet Cataloging)[1]。但随着搜索引擎的日益学术化,图书馆非但没有对互联网信息资源进行有效的书目控制,其实体信息资源的组织也逐渐被搜索引擎资源组织所替代。因此,如何利用现有技术来组织实体信息资源,以扬搜索引擎之长而避其之短,就成为图书馆人不得不思考的一个问题。本文在此抛砖引玉,希望引起图书馆人的进一步思考。

1　扬搜索引擎之长,使实体信息资源组织章节化

从 1999 年到 2008 年,中国出版科学研究所发布了五次调查结果。其中,第四次调查统计结果表明:纸书阅读率 6 年来持续走低,而网络阅读率 6 年间却增长了 6.5 倍[2];而最近一次调查统计结果则表明:网络阅读率已以 36.5% 的比率首次超过了 34.7% 的图书阅读率[3]。2007 年,OCLC 的成员调查报告《网络世界的共享、隐私和信任》数据显示,与 2005 年相比,现今用户使用搜索引擎的比例由 71% 上升至 90%,使用网上书店的比例由 50% 上升至 55%,只有图书馆网站的使用比例由 30% 下降至 20%[4]。另据中国互联网络信息中心(CNNIC)2009 年 1 月 13 日发布的《第 23 次中国互联网络发展状况统计报告》数据显示:截至 2008 年底,我国互联网普

及率已达22.6%(首次超过全球平均水平的21.9%),网民人数已达2.98亿(其中2.7亿人使用宽带上网),未来还有增长的趋势。因此,图书馆需要研究阅读及网络阅读的特点并找到适合网络阅读的实体信息资源组织方法。

1.1 阅读及网络阅读的特点

从读者主体看,读者分生活读者、学习读者和工作读者三类;从阅读取向看,阅读分资讯阅读、修养阅读和愉悦阅读三种。其中,最广泛的读者群不是文学读者,而是文章读者。因为生活、学习和工作中的读物绝大多数是实用文章(新闻文、应用文、教科书、学术文等),即使是消闲性、学习性、专业性文学阅读,也伴随着大约一半的文章阅读(文学评论、文学史等)。概言之,文章读者的实用阅读是国民阅读的主流,更是人类阅读的主体[5]。

世界上的书,主要是具有千百年历史的纸本书(p-book)和只有十多年历史的电子书(e-book)。当印刷文本被大量转换成视觉文本、电子文本时,国民阅读便进入了读图时代和读网时代,因此,国民阅读率的统计应该包括读电子书的网民。归纳起来,网络阅读的主要特点是"短、平、快"。所谓"短",是指网络阅读的文章短小精悍;所谓"平",是指网络阅读的文章浅显易懂;所谓"快",是指网络阅读的文章获取快捷。而这其中,又以"短"最为突出,因为短小精悍的网络文章比较适应人们现今生活、工作的快节奏,也符合在现今的信息海洋中获取知识的需求[6]。

1.2 实体信息资源组织的章节化

面对如今网民喜爱短小精悍的文章阅读,图书馆又该如何进一步进行载体表现层面上的实体信息资源组织呢?笔者认为,图书馆在对实体信息资源进行组织时,可参照国家图书馆的做法,在建了中文图书书目数据库后再对它们建立一个中文图书目次信息数据库,但是更加提倡像亚马逊的书内搜索(Search Inside)以及Google、读秀知识库等检索系统那样,在将书作了整体描述后再对其章节进行组织。

对书的章节进行组织,其实就是将书"打碎"后进行组织的一种方法。而将书"打碎"组织将利用CNMARC中的327内容附注字段及其相应的检索字段。原来CNMARC的327字段第二个指示符未定义,第一个指示符定义为"完整程度指示符"(即赋"0"表示内容附注不完整,赋"1"表示内容附注完整),其唯一一个可重复的$a子字段记录附注内容(即章节信息)。由此看来,原来CNMARC中的327字段简单等同于其464单册分析字段[7]。

在《新版中国机读目录格式使用手册》中,CNMARC的327内容附注字段除了保留其第一个指示符的定义外,另外还将其第二个指示符定义为"结构指示符"(即赋"#"表示非结构式附注,赋"1"表示结构式附注)。其次,如果内容附注是非结构式的,全部附注文字均记入可重复的$a子字段;如果内容附注为结构式的,全部附注文字则分别记入可重复的$a(最高一层章节)、$b(一级子章节)、$c(二级子章节)、$d(三级子章节)、$e(四级子章节)、$f(五级子章节)、$g(六级子章节)、$h(七级子章节)和$i(八级子章节)子字段,以及$p(页码)和$z(其他信息)子字段。可见,现在CNMARC中的327内容附注字段要比其464单册分析字段更具层次感,也更能满足读者对图书章节的组织要求[8]。

CNMARC的327内容附注字段,无论在单馆计算机编目还是在联机联合编目中,各馆过去普遍都不重视,更不要说那些已将编目业务外包出去的图书馆了。但在不断变化的信息和技术环境下,图书馆若要有效地支撑其实体信息资源的管理与访问,必须要向亚马逊的书内搜索以及Google、读秀知识库等检索系统学习,否则会因全文搜索的检索效果越来越差而落伍。除此之外,Google另一想法是要为数字图书馆资源建立高水平的逐词索引(word-by-word index,即以文献全文的所有有效词作为索引标目并注明其出处)。与人工索引相比,Google的逐词索引更加深入,可最大限度地挖掘出值得寻找的信息,并使社会上的信息原子极快地发生裂变。这些动向都需引起图书馆人的足够重视。

2 避搜索引擎之短,使实体信息资源组织 FRBR 化

使用搜索引擎的人都有这样的一种体验,即每次搜索的结果少则数十上百条、多则数千上万条,其有用信息常被无用信息所淹没。而随着实体信息资源的增长,图书馆的 OPAC 书目检索也有这种日益"搜索引擎化"的趋势,即搜索的结果有时也会成百上千,常常使检索者感到无所适从。如用国家图书馆的 OPAC:①输入检索词"曹雪芹",设定作者检索,命中记录竟有 584 条之多;设定所有字段检索,命中记录竟有 734 条之多;②输入检索词"红楼梦 + 石头记 + 金玉缘",设定题名检索,命中记录竟有 1955 条之多;③输入检索词"红楼梦 + 曹雪芹",设定所有字段检索,命中记录竟有 562 条之多;④输入检索词"红楼梦",设定主题检索,命中记录竟有 1201 条之多;设定题名检索,命中记录竟有 2124 条之多;设定所有字段检索,命中记录更是高达 2279 条之多[9]。众所周知,《红楼梦》是曹雪芹身后唯一留世的作品,如果曹雪芹像英国文学家哈代那样是位多产作家并被译为多个语种,那各馆 OPAC 书目显示又该呈现一种什么情景?上个世纪末,国际图联颁布的 FRBR(Functional Requirements for Bibliographic Records: final report,书目记录功能需求)[10],可以化解上述馆藏书目数据库日益"搜索引擎化"的趋势。

2.1 FRBR 的主要内容

FRBR 既不是一个新的 ISBD,也不是一部编目规则,而是一个实体 - 关系模型(以下简称"FRBR 模型")[11]。FRBR 模型的最核心部分是定义了一系列与图书馆目录相关的事物类别(实体)、从属于每个类别的特征(属性),以及可能存在于各种类别之间的关系。首先将与图书馆目录相关的实体定义为三组,其中,第一组实体包含从属于文献的四个受编实体,从内容到载体分别为作品(Work)、内容表达(Expression)、载体表现(Manifestation)和单件(Item);FRBR 模型定义的第二组实体是能创造一个作品,实现一个内容表达,产生或订购一个载体表现,修改或处理一个单件的个人(Person)和团体(Corporate body);FRBR 模型定义的第三组实体是用于反映一个作品的主题实体。除了以上一、二组实体外,另外主要使用的实体是概念(Concept)、物体(Object)、事件(Event)和地点(Place)。

以上 FRBR 模型中的每一个实体都由一系列的"属性"表征,如作品的属性有作品题名(Title of the work)、作品形式(Form of work)、作品日期(Date of the work)、其他识别特征(Other distinguishing characteristic)、预期的结果(Intended termination)、预期的受众(Intended audience)和作品的背景(Context for the work)等;个人的属性有个人名称(Name of person)、日期信息(Dates of person)、个人头衔(Title of person)和其他相关标识(Other designation associated with the person)等。其次,第一组实体(作品、内容表达、载体表现和单件)之间的关系是"结构关系";第二组实体(个人和团体)和第一组的任何实体之间的关系是"责任关系";FRBR 模型中的任何实体与独立实体"作品"之间的关系是"主题关系"。此外,FRBR 模型的实体之间还存在一些更加"微妙"的关系,比如两个不同作品间或同一个作品的两个不同内容表达间的"整/部关系"(整/部关系也存在于两个不同载体表现之间或同一载体表现的两个不同单件之间),两个不同作品间或相同作品的两个不同内容表达间的"智力关系",以及两个不同载体表现、同一载体表现的两个不同单件、或一个载体表现与另一个不同的载体表现的单件之间的"再生关系"。

2.2 FRBR 的实现方式

如上所述,FRBR 仅是一个实体 - 关系模型,而不是一个数据模型。因为 FRBR 为每个实体所定义的属性在很多情况下都太一般化,以致于如果不加提炼就无法将它像一般的模型那样实现。例如题名可以有不同的性质,尽管 FRBR为作品、内容表达和载体表现这三个实体的每一个实体都定义了一个题名属性,但是这

种“题名概念”的分类还不足以覆盖实际的需要和目前在用的题名类型。

FRBR 既然不是一个数据模型,那它又是如何被“实现”的呢? 最好的情况是基于它设计一个中间数据模型,最差的情况是就将它错当成一个数据模型。但不管是哪种情况,不是将一个现存的格式映射到 FRBR 上,就是将 FRBR 映射到一个新的格式上;后者将直接影响现存的机读目录格式,而前者则直接影响现存的 OPAC 检索界面。由于在短时期内改变现存机读目录格式的可能性不大,所以国外的研究之前大都放在将一个现存的格式映射到 FRBR 上[11],其原理就像剥笋那样将一个作品的内容表达、载体表现和单件层层进行剥离,如图 1 所示:

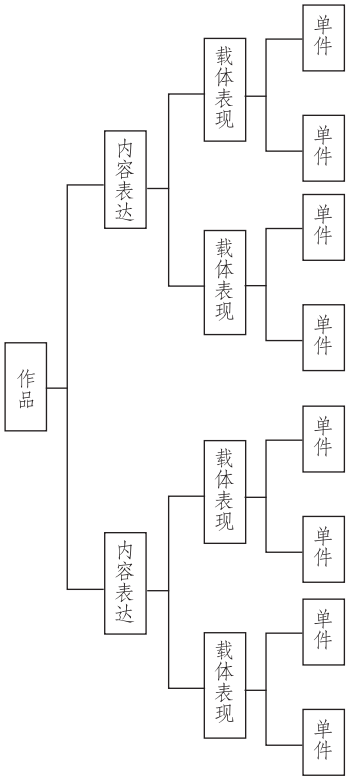

图 1 基于 FRBR 模型的检索

基于 FRBR 模型的分面检索结果,其结构层次鲜明,可以显示各书目记录间的关联性,且便于读者/用户辨别和理解检索结果中的各实体间的关系,从而大大节省了其检索时间与精力。前述《红楼梦》若基于 FRBR 模型来检索,其结果肯定不会一次出现上述成百上千条记录。当然,这种前台整合的显示需要后台信息资源组织的配合。在这方面,美国的做法是在内容表达和/或载体表现的数据记录上增加 004 和 990 字段。其中,004 字段用来链接内容表达和/或作品记录,990 字段则用来反映本记录的实体类型(即作品、内容表达或载体表现)。

笔者认为,将一个现存的格式映射到 FRBR 上来的做法是种被动做法、权宜之计,而将 FRBR 映射到一个新的格式上来的做法才是一种主动进取、长效之计[12]。随着时间的推移,目前采用将 FRBR 映射到一个新的格式上来的呼声越来越高,并进一步影响到对编目条例和机读目录的修改[13]。因为现存的编目条例和机读目录格式还不能很好地适应 FRBR 结构化或层次化的信息组织要求,尤其还都缺乏对内容表达的记录基础[14]。

3 结语

为扬搜索引擎之长、解决网络阅读“短、平、快”的问题,图书馆可利用机读目录对实体信息资源进行章节化组织;为避搜索引擎之短、解决检全率尤其是检准率的问题,图书馆可利用机读目录对实体信息资源进行 FRBR 化组织。虽然机读目录目前在 OPAC 上检索不成问题,但这种存贮在特定数据库中的机读目录数据目前还无法被搜索引擎搜索和索引。为此,图书馆还需不断优化软硬件设备,加强网站建设,或将书目记录交付给搜索引擎,以方便读者/用户从搜索引擎上获取,从而增加图书馆网站的访问量。如 OCLC 的 Open WorldCat 就计划将其书目数据送给 Google 和 Yahoo,使习惯利用搜索引擎检索书目信息的用户成为自己潜在的用户。2006 年 5 月 23 日,北京大学图书馆与百度签定了独家战略合作框架协议,即北京大学图书馆将其图书书目数据授权给百度,使读者/用户可以通过百度来检索北京大学图书馆的馆藏书目。技术力量较弱的图书馆,甚至可以考虑将自己的书

目记录以博客等形式让网络蜘蛛自动抓取。如此,读者/用户就可以通过搜索引擎来检索和利用图书馆的书目数据。加上以 Google 为代表的搜索引擎均以相关度排序,图书馆的书目数据在搜索结果中肯定排序靠前。此外,不靠搜索引擎而主动将机读目录数据 XML 化放在网上发布,也不失为一种好方法。但是,无论是前面一种被动方式,还是后面一种主动方式,如果图书馆的实体信息资源组织与其数字图书馆资源的建设结合起来,那将起到事半功倍的效果。

图书馆的地位与作用随着信息资源普遍可获得性的程度提高在不断下降[15]。现在,Google 等搜索引擎又开始独领风骚,图书馆更需考虑自己的生存空间和发展前途。以往搜索引擎的信息组织对象一般是普通的网页,这对图书馆不构成致命的威胁;而今有些搜索引擎已经转向学术搜索领域,并用自己的先进技术与其他信息机构进行合作,这使图书馆真正到了"狼来了"的时代。一旦 Google 等搜索引擎可以担负起图书馆的使命——组织世界文献信息,那图书馆还有多大的生存空间和发展前途?图书馆的正确做法是在感到危机的同时,积极考虑扬长避短,真正做到"与狼共舞"。

参考文献:

[1] 胡小菁. 编目的未来[J]. 大学图书馆学报, 2008(3):18-22, 37.

[2] 王余光, 李雅. 图书馆与社会阅读研究述略[J]. 山东图书馆季刊, 2008(2):4-12.

[3] 朱光, 陈斯斯. 网络浏览取代"青灯黄卷"[N]. 新民晚报, 2008-08-23(A6).

[4] Sharing, privacy and trust in our networked world [EN/OL]. [2009-01-13]. http://www.oclc.org/reports/sharing/default.htm.

[5] 曾祥芹. 用科学阅读观引领大众阅读新潮[J]. 山东图书馆季刊, 2008(2):1-3, 25.

[6] 董一凡. "浅阅读"不应遭遇"深谴责"[J]. 图书馆杂志, 2009(1):26-29.

[7] 潘太明等. 中国机读目录格式使用手册(修订版)[M]. 北京:科学技术文献出版社, 2001:280-281.

[8] 国家图书馆. 新版中国机读目录格式使用手册[M]. 北京:北京图书馆出版社, 2001:231-235.

[9] 富平. 按照 FRBR 模型构造书目检索体系的思路[J]. 数字图书馆论坛, 2008(2):28-39.

[10] IFLA Study Group on Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records: final report. [R/OL]. [2009-01-13]. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

[11] Patrick Le Boeuf. 美好的 FRBR 新世界[J]. 王松林, 译. 国家图书馆学刊, 2006(4):82-86, 96.

[12] 王松林. 从 FRBR 看编目条例和机读目录格式之变革路向[J]. 中国图书馆学报, 2004(6):21-25.

[13] 顾犇. 国际文献编目领域标准规范发展进展[G]// 中国科学技术信息研究所, 全国信息与文献标准化技术委员会编. 信息资源组织及其标准规范学术研讨会论文集. 北京:编者, 2008:84-92.

[14] 胡晓鹰. FRBR 概念模型与 CNMARC 之比较研究[J]. 图书馆论坛, 2007(5):110-114.

[15] 程焕文. 关于改变图书馆学研究立场的思考:从"用户永远都是正确的"说起[J]. 中国图书馆学报, 2008(3):89-93, 102.

王松林 南京政治学院上海分院教授、博士生导师。通讯地址:上海。邮编:200433。

(收稿日期:2009-02-19)