知识组织:图书馆职业的核心能力

刘 炜

摘 要 图书馆职业在历史上一直承担人类知识有序化的工作,只是由于知识生产的批量化和工业化,缺乏必要的工具,在速度和深度上无法根据知识的内容进行整序,而只能根据载体形态进行组织加工。数字图书馆技术的发展将有可能借助语义技术和许多新的工具,使图书馆行业数千年来积累起来的知识组织经验,在万维网时代发扬光大。参考文献4。

关键词 知识组织 MARC 编目 元数据 语义技术 互联网

2009年是MARC正式应用40周年,也是互联网发明40周年。MARC的应用可以看成是图书馆行业迈向自动化、规范化和全球化的起点。MARC和互联网同岁,但它们的普及程度却完全不同,40年的发展历程中它们可以算作两条平行线,而如今随着OCLC的Open WorldCat提供全球化的服务,以及各类型图书馆普遍提供基于Web的OPAC检索,这两项技术已经开始交汇并轨了。笔者在这里通过对MARC和互联网发展历程的深入探讨,试图说明,借助于最新的数字图书馆技术,数字化的知识组织在互联网时代也是必要的和可行的。语义万维网的提出和研发正走在这个方向上。图书馆行业数千年来积累起来的知识组织经验,在网络时代,完全可以脱胎换骨、焕发青春。

1 从MARC到RDA:图书馆风光不再?

1969年3月的一天,美国国会图书馆首次启用了磁带格式的MARCII书目系统,每周发行大约包含1000条英文图书记录的磁带。记录格式是遵照前一年出版的《Subscriber's Guide to the MARC Distribution Service》手册而编制,后来该手册修订为《Books:a MARC Format》,是MARC格式的奠基之作。MARC格式研制起始于20世纪60年代初,该手册是近十年研究成果的结晶。

用今天的眼光来看,MARC是什么?它是一种行业性的元数据标准,正是它,使得图书馆行业在信息技术应用方面至少比其它行业领先20年。它规定了图书馆进行馆藏描述的属性字段、结构和编码方式等,早期的馆藏主要是图书,后来扩大为期刊、报纸、手稿、缩微品、音像资料等所有文献类型,用MARC来描述馆藏的目的是"揭示"馆藏,使馆藏更加"有序"和"结构化",从而便于检索查找和获取文献。这个标准一下子使图书馆有了中心工作,使图书馆的业务流程规范化,提高了效率,并使联机、联合、合作编目成为可能。MARC的制订者希望用它来描述馆藏对象所有方面,包括各种物理性状、责任实体、内容属性(分类、主题、人地时事等等)等等。当时由于条件所限,大量的图书馆还用不起计算机,有限的计算和存储资源需要物尽其用,尽可能发挥计算机信息处理快速和准确的优势,因此MARC最初的基本作用只有两个:打印卡片和馆藏检索。

从总体上看,一个图书馆的MARC数据库可以看成是该馆所有馆藏的指代物,全球图书馆的MARC库可以认为是整个社会知识产出的指代物,是一个有序化知识体系的缩影,也是索引。我们能够通过书目控制①,获得人类所有社会性知识的大致情况,从而能够进行所谓的"知识组织"。

美国国会图书馆作为美国国家图书馆之一,面临数字时代的到来,已公开宣布放弃承担一国书目控制的责任[1],因为它感到力所不逮。OCLC正在接过这个接力棒,成为一头MARC数据的巨兽,竭力实现全球图书馆书目控制的历史重任。根据OCLC 2008/2009年度报告[2],它已有书目数据1.391亿条,馆藏数据超过10亿条,参与馆72,035家,涉及国家超过112个。

MARC产生的时代是信息匮乏的时代,当时许多国家还在为消除文盲而努力,信息不对称、存在巨大的知识鸿沟是普遍现象。几乎所有的外化知识(或称社会性知识),例如书刊报手稿等一切文字载体,都还存在于纸质媒体中,而且几乎所有的纸媒体介质,都收藏于图书馆中。并且,MARC还包括了大量视频音频非书资料。这些东西所形成知识宇宙,都是MARC的编码对象。它使得图书馆能够傲立于这个时代,成为大规模社会知识交流必须经过的中枢。OCLC如果早20年达到世界级规模,它就相当于现在的Google,整合所有人类知识,成为我们全球图书馆行业的OPAC,以及全人类知识的前台、组织者和控制者。

而现在,Google正在步OCLC后尘,扮演数字时代知识巨兽的角色。

人类历史上出版过的图书保存至今的也就1亿多种,就书目而言,OCLC已经几乎搜罗完备,接下去Google正在择其精华,进行宏大的数字化计划。目前已经扫描了逾1千万种,凭借谷歌的实力,完成扫描只是时间问题。古埃及亚历山大图书馆的理想不就是"搜罗人间所有图书"吗?这个梦想Google不久即将实现。但Google所要做的,恐怕远不止于此,因为传统的纸媒正在逐渐淡出知识宇宙的中心,地心说已让位于日心说,亚历山大图书馆的理想已经被赋予了新的含义:应该包括所有"数字化知识"。Google要做数字时代的亚历山大图书馆。

面临Google的汹汹来势,图书馆依然雄心不泯,从国际编目界最倾力推出的RDA(Resource Description and Access,资源描述与检索)可以依稀看出,图书馆界希望重温MARC时代的辉煌。虽知其不能,但仍在努力。

RDA是数字时代的编目规则,是英美编目条例第三版(AACR3)的正式名称。它有如下特点:

(1)它是数字图书馆的MARC著录规则。扩展了适用对象,不只是针对纸质资源,而是定位于描述和检索所有资源②,从而能够作为复合型图书馆的"大一统"的元数据规范,并适用于互联网环境。

(2)它是全世界各族人民的MARC。适应了全球化浪潮,加强了国际性,消除了英美国家特有的内容,是第一部全球大同的编目规则。

(3)它引入了一套本体,即首次全面实践了FRBR,突破了传统MARC数据的扁平结构,为MARC数据进入互联网,以及Web时代的信息资源描述和规范控制提供了概

① 书目控制是1949年谢拉和他的同事提出的概念,意为"从书目的目的出发,控制人类已出版的全部文献"。

② 这里对于"资源"的定义与万维网协会的定义是一致的,即"具有标识的一切信息单位"。

念模型。

(4)它不再仅仅是一套文本,同时也是一套Web工具;不仅具有方便的、"事件敏感"型查索功能(包含词表登记注册管理等功能),同时支持开发商集成到管理系统中,提供各类商业性的Web服务。

(5)它独立于编码和数据格式,定位于"内容"规范,从而能支持MARC、DC、MARCXML、MODS、RDF/XML等等众多输出格式。

(6)它成为连接过去与未来的桥梁。吸收了大量DC元数据的研究成果,使它能够"兼容"互联网,并能把各类遗留系统中的书目数据(即各种MARC数据),带入到Web中继续发挥价值,并为互联网提供一套经典的"书目控制"手段。

最关键的,是它承载了我们图书馆人的理想:整序知识,书目控制。要做到这一点,在当今数字世界,对于图书馆行业来说,是非常困难的。目前国际编目界围绕RDA的无休止的争论,以及RDA正式版本迟迟不能按期推出,也的确让人怀疑图书馆行业是否还能实践理想。但这项社会职能在历史上一直是赋予图书馆行业的,属于一种"制度安排",因此无论如何我们总要努力。即使图书馆一个行业做不到,也可以联合其它行业或企业,或借助于一些新生力量,贡献自己的智识,推动"知识整序"目标的最终实现。

2 知识的组织者:让历史告诉未来

图书馆员历来是知识的组织者,并且经常是同时代最为睿智、博学的学者。老子做过东周国家图书馆馆长[3];孔子删定六经,至少是其私家馆藏的"研究馆员";亚里士多德是最早试图对古希腊人们关于自然界和人类社会所有知识进行分类的人,他画了一棵著名的知识树。试想如果没有接触到当时大量的知识素材和知识载体,他如何能进行分类?亚述巴尼拔国王在尼尼微建立了藏有大约2.5万块泥板文献的皇宫图书馆[4],如果没有合理有序的仓储系统,亚述王在重要时刻找不到他所需要的资料,馆长的性命恐怕都保不住。公元前250年古希腊诗人卡利马科斯为著名的亚历山大图书馆编制的分类目录Pinakes,应该能印证分类与图书馆是如影随形、须臾不离的。据说亚历山大图书馆正是亚里士多德的弟子法拉雷乌斯(底米特瑞斯)向埃及国王托勒密一世建议而建,他自己也当仁不让成了馆长。这所图书馆不仅在形制上极尽所能,模仿雅典的博学园(又称智慧宫),在功能和管理上也照搬前贤,吸引了当时为数众多的一流大师,如阿基米德、阿历斯塔克斯、欧几里得、希罗菲卢斯等等,他们在这里的研究、讲学、翻译、著书、立说,或许还有分类、编目、征订、外采、咨询、整架、上架、装订等,成为当时的科学、文化、哲学、艺术的知识宝库与传播站,奠定了该城当时世界文化和科研中心的地位,使得亚历山大图书馆很早就成为业界楷模,蜚声遐迩,流芳百世。

我国古代负责管理图书馆的也无一例外都是大学问家。当然,我国古代没有"图书馆"的称谓,这个称谓1894年才从日本传入。我们当初称为府、阁、观、台、殿、院、堂、斋、楼等,体制也有所不同,但以内容为中心,形制为内容服务的思想还是一致的。中国古代的封建王朝一直有盛世修典的传统。历朝历代,只要皇帝想起做学问了,就召集最有学问的大学士们操刀,皓首穷经,为往圣继绝学,将所能收集到的所有书籍打散,按照当时所认识到的知识的内容体系(如七略、四部等)重新建构,校雠辑录,查考版本,择其精华,编成"类书"。因此刘向、刘歆、班

固、王俭、郑樵之类,首先并不是目录学家,而是历史学家、文学家。他们一般多注重内容的考据和思想的源流,对于如何方便地查检,往往不是十分关心,造成方法论的缺失。类书的编撰整理组织方法,已经不能看作是在编纂一本著作,而是在创建知识的集大成,这应该是一所图书馆才能承担的功能。这使得我国古代的图书馆学家(应该称为目录学家更合适)首先都是内容专家,而不是分类编目专家。但是其背后"上穷碧落下黄泉"、企图穷尽知识的思想,也算是朴素的书目控制吧。

历史上两个事件使图书馆的知识组织工作产生革命,影响了图书馆的业务模式和管理形态。

一个是区别于财产账册的查检式目录的产生。16 世纪以前的目录基本上都是财产账目,当时的图书馆应该都是开架阅览,当然只对特定读者开放,如有需要可以抄录。西方最早诞生的查检式目录通常都有字顺和分类两种方法,如 1545 年盖世纳出版的《世界书目》有 21 个大类和 250 个细目,同时有字顺标题索引,他还建议图书馆在款目上加上排架号(这本书目甚至具有了"总书目"甚至联合目录的作用)。卡片目录于 1861 年在哈佛大学图书馆的产生更具有划时代的意义,造成了实体馆藏与虚拟馆藏的分离,发明了各类的排列、查检形式,大大扩展了目录的检索功能,不仅能进行特性检索,还有族性聚类和各类参照入口;目录是实体馆藏的缩影、映射,但是比实体馆藏灵活,可以任意编排,例如字顺、主题、分类等组织方式,只需要有"链接"(索书号)指向实体馆藏的具体位置即可。

一个是内容著录和形式著录的分离,导致了描述性编目与主题编目的分离,由于印刷术的产生使出版产业化,也扮演了知识贫民化大众化的推手,也使得图书馆以书为载体的组织不可能像以往那样追根溯源、考根究底。各类图书馆越来越倾向于形式编目,而越来越不可能进行主题揭示了。这也是为什么早期的图书馆学家都是哲学家或大学问家,而后来者不再能进行"辨章学术、考镜源流"的原因。

现在看来,不论是 1841 年大英博物馆馆长潘尼兹提出著名的《编目九十一条》,还是杰维特规则(美国第一部编目规则)、卡特的字典式目录、后来的英美规则和普鲁士规则,以及柳别斯基的编目思想,都受限于手工目录的技术条件,纠结于"鸡毛蒜皮的小事",例如讨论是将哪些内容作为标识款目,是按照字顺、分类还是按照著者排列,哪些字段应该来自于书皮而哪些不能相信书皮,等等,提出繁复的原则和莫名其妙甚至经常是互相矛盾的规则,并为了一点点细小的不同而发生分歧,成为不同"体系",互相批判彼此是实用主义、理想主义、教条主义、书目主义等等。有人认为正是这些特性造成了图书馆职业的性别特征,并且使我们的职业在工业化的同时,变得琐碎、无趣和工匠化,站到了学术研究的对立面。

3　我们能做什么:掌握利器,决胜未来

同样的 40 年前,不过是在秋天,伦纳德·克莱罗克教授站在一个电冰箱大小的灰色金属盒子旁,开心得像刚做了父亲,他把这个盒子叫作"界面信息处理器"(IMP)。克莱罗克教授用一根长 4.6 米的灰线将两台计算机连接起来并传输了数据,后来又与 643 公里外斯坦福研究所的一台计算机成功连接。从此,各地的计算机被迅速连接起来,直到遍布全球的互联网时代到来,影响了数亿人。

MARC 与因特网今年同庆 40 周年,但它

们两者的影响力和普及程度是不能同日而语的。计算机技术是从处理信号、字符、数据、信息、语义、知识、智能一层层地发展上来,MARC直接就作为一种行业性的内容描述标准,创建时显得太过于领先当时的时代。20世纪60年代末70年代初,计算能力还极其有限,字符集问题尚未解决,字符集标准尚不统一(英文字符就同时有ASCII和EBCDIC),关系型数据库技术才刚刚萌芽,信息检索还停留在倒排索引提高磁带文件系统的查询效率方面,全文检索也是闻所未闻的事情,这个时候,跨越多个发展阶段,直接关注对于知识载体的内容进行编码,实在是匪夷所思。因特网经过40年的发展,经历了Email、FTP、Gopher、Web等百花齐放,到目前才开始关注语义和知识的表达,才进入MARC40年前所关注的领域。

万维网(World Wide Web)是目前因特网上占据主导地位的一种应用,它的基本思想,是应用超文本传输协议(HTTP),通过在全局网络中具有唯一性的标识符:统一资源定位符(URL),对遍布互联网各处的,以一种特殊排版语言——超文本编码语言(HTML)——编码的文件进行存取、呈现和其它服务。当然,现在HTTP、URL、HTML都出现了很多扩展和变化,能够支持多媒体信息,实现多种动态内容的发布。我们可以发现,万维网的这种机制非常类似于图书馆:统一资源定位符相当于索书号,排版语言相当于著者、分类、主题等不同标目的卡片格式,超文本传输协议相当于有序的目录体系。通过一整套目录查询机制,定位或汇集馆藏。

从宏观上看万维网也完全可以与图书馆类比,万维网上的文献空间可以看成是一个巨大的知识空间,但是它的编码目前还只是停留在字符和字符串的编码,缺少结构化信息,更没有任何知识或语义的标注。搜索引擎尝试了各种排序算法,试图在以字词匹配为基础的全文检索之外,能够自动地建立起一定的信息结构,猜测出哪些信息更加符合我们的提问。这种猜测是以各种概率或向量空间算法为基础的,并不具有人工进行概念标注的确定性,因此万维网的知识空间还没有任何的知识标注,严格说来它并不是知识空间,而只是一个信息空间。目前语义万维网要做的,是想通过类似于图书馆员进行编目一样,给这个信息空间进行标引,使其成为名副其实的知识空间。

对整个万维网进行语义标注并使其结构化,被认为是一个乌托邦,DCMI早期希望发明一种简单的元数据语义描述格式,让人们能够方便地用来标注网页,但后来的实践证明没有人愿意费事这样做,相反这样做的人大都出于希望提高网页在搜索引擎中的关键词排名,于是造成了元数据的滥用,反而让搜索引擎避之不及。

那么如何做才是可行的?目前一般认为有两个途径:一是,资源在产生之时就附带了语义描述信息。如数码相机存储的图像格式都包含EXIF信息,甚至可以包含GIS信息。许多电子书格式也内嵌了元数据信息。将来越来越多的元数据信息是在资源生产或加工过程中,通过软件或管理流程自动产生的,这种方式的前提是,在应用领域中必须要有获得一定公认的元数据格式标准。二是,采用大规模自动处理(例如自然语言处理、文档分析等方法),批量获取结构性语义信息。这种方式也是需要预先定义一定的模型,或者找到一定的规律。

大量"前Web时代"的科学数据库、书目数据库(MARC)或商用数据库,本来已经是结构化的信息,包含丰富的语义,是最容易上网的、现成的、最宝贵、最有价值、经过规范控制的权威数据。

上述的工作,实际上都是建设数字图书馆所应做的工作,对这些数据进行加工整理,或制定标准规范,提供加工整理方法、流程,让计算机来帮助进行加工、整理、转换并提供 Web 服务,都是未来的图书馆员应该承担的、最合适的工作。

我们有哪些知识组织和整序工作可以在数字时代继续发扬光大,作为我们的核心竞争力,可以帮助我们在 Web 时代继续引领潮流呢?

首先,当然是将传统编目中属于"主题编目"的一类技能和成果,也就是目前统称为"知识组织系统"(包括分类法、叙词表一类)的规范体系,转化为 Web 可以读懂和处理的格式。这就要弄懂 Web 知识组织编码语言(SKOS)和本体语言(OWL)。由于传统书目系统或情报检索系统采用这些概念体系都是给人用的,不具有严格的机器编码,机器对这些词表和分类法并不可读。

其次,进行描述性编目仍然是图书馆员无法逃避的责任,这个岗位可能会改换名称,例如叫做"元数据编目员"之类。这项工作可能更多的是进行元数据标准规范(领域应用的元数据规范一般称为元数据应用纲要)及著录规则的制定、流程设计、培训推广、质量控制等,需要支持自动或批量的元数据编目。

再次,为进行更多的主题编目或者开发主题编目算法而研制各类本体,以及研发各种支持 SKOS 和 OWL 的 KOS 编目工具。并从事资源的主题标注、批处理或互操作开发工作。

最后,所有的标准规范或者方法工具最终都将体现于元数据编码中。因此制订、审核并应用各类元数据编码方案(主要是基于 RDF/XML 的编码模式:Encoding Scheme)也应该是这个环节中的重要内容。

目前,实现所有上述工作的技术架构已经建立,这就是 RDF、SKOS、OWL 等一系列描述规范,以及以 Linked Data 为代表的语义存取机制。我们只需要掌握它们,就能在未来的数字世界中,继续做好本该属于我们的知识组织工作。

参考文献

1　美国国会图书馆《书目控制的未来》报告[OL].[2009-12-20]. http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

2　http://www.oclc.org/news/publications/annualreports/2009/2009.pdf[OL].[2009-12-20].

3　朱建亮.我国东周著名国家图书馆馆长老子其人其书——第一部管理学著作述评[J].图书馆论坛,2009(1)

4　互动百科"图书馆"词条[OL].[2009-12-20]. http://www.hudong.com/wiki/%E5%9B%BE%E4%B9%A6%E9%A6%86

(刘　炜　研究馆员　上海图书馆)

收稿日期:2009-12-27