

· 学位论文摘登 ·

网络信息组织：模式与评价

黄如花

(武汉大学信息管理学院, 武汉 430072)

【摘要】对现有的网络信息组织模式的优缺点进行了系统的分析与评价,指出图书情报界应该在积极参与网络信息组织中体现学科的开放性。

【关键词】网络信息组织; 数字图书馆; 学科信息门户; 搜索引擎; 网络资源指南

【中图分类号】G 254 G 302

Models of Information Organization on Networks

HUANG Ruhua

【Abstract】The article systematically analyzes characteristics and problems of current information-organization models on networks. The author concludes that librarians should take an active part in theory and practice in information organization on networks so as to make the best of our superiority.

【Keywords】information organization on networks; digital library; subject-based information gateway; search engine; web directory

关于网络信息组织的方式与方法,国内外学者依据不同的标准划分出了不同的类型。有的学者将网络信息的组织方式划分为网络一次信息组织、网络二次信息组织和网络三次信息组织^[1]。有的学者从信息的存贮形式出发,认为目前网络信息的组织方式主要有文件方式、数据库方式、主题树方式、超媒体方式等^[2]。实际上,网上的信息组织并不是采用单一的方式,而往往将多种方式结合使用,如搜索引擎和学科信息门户采用了上面划分出的超媒体方式和数据库方式,而在超媒体方式中,文件和数据库均可作为链接的节点。

从网络信息组织对象的范围看,网络信息组织的模式可以划分为4个层次:第一个层次为微观的组织模式,包括文件、超媒体、数据库与网站方式,它们在下面几个层次中都得到了广泛的应用;第二个层次为中观的组织模式,主要有按照一定的著录格式对网络信息进行重组的编目和针对特定用户、汇集某一学科或专题领域资源的学科信息门户,它们收集的资料经过人工筛选和处理,会有所取舍,而且,对资源的整序与控制有一定的深度;第三个层次为宏观的组织模

式,主要指网络资源指南与搜索引擎,它们力图对整个网络的资源进行控制,倾向于做整个网络范围资源的索引,广泛地汇集网络资源;第四个层次为对网络信息进行分布式组织的数字图书馆,其组织的资源已经远远超出网上信息的范围。同时,尽管编目和学科信息门户组织资源的范围小于网络资源指南和搜索引擎,它们的某些做法是针对网络资源指南和搜索引擎的某些缺点而做的改进,其中,编目属于一种重组,而学科信息门户是更深层的组织。尽管这样的划分不一定准确,但是,这些组织方式都是我们在Internet上所能看到的信息的呈现方式,也便于理解和评价。需要说明的是,这里的所指并不局限于只对网上信息的组织,因为网上信息与图书馆馆藏实体信息整合的趋势已日趋明显。本文便是依据这些信息组织模式之间的关系来组织的。

1 网络信息的微观组织模式

1.1 文件(File)

因特网的信息组织,首先是要将网外丰富的信息资源电子化,投入网上,形成网上的各种文件。文

件是一种历史较长的信息组织方式,其优点是简单方便,除文本信息外,还适合于存储程序、图形、图像、图表、音频、视频等非结构化信息或多媒体信息。因特网的文件提供了非常丰富的信息资源,方便了用户从网上查看、下载或打印。

以文件为单位对信息资源共享和传输的方式使得网络负载越来越大。文件系统只涉及信息的简单逻辑结构,而当结构较为复杂时,就难以实现有效的控制和管理,文件只能是海量信息资源管理的辅助形式,或者作为信息单位成为其他信息组织方式的管理对象。其次,网络文件的语言、编码与格式多样,且解读方式迥异,给网络信息的组织与交流带来了不少的困难,对一些非英语国家来说,这个问题显得尤为突出。由于技术上的原因,要对这些不同格式的内容进行组织,涉及到多方面的技术,不同的文件格式需要用不同的软件来显示。不同的文件格式之间并非都可以相互兼容,不同的格式之间转换后会发生变化。

文件的制作者有责任采用通用的标准,努力克服不同操作平台的浏览方式与显示的差异,其适用的平台与软件应为通用的并易于获取(为共享软件或提供下载地址)。

1.2 超文本/超媒体(Hypertext/Hypermedia)

超文本既是一种新型的文本信息组织方式,也是一种有别于传统检索技术的新型信息获取方式,它将网络上相关文本的信息存贮在许多节点上,节点间以链路相连,节点表示信息单元、片段或组合,而链表示结点间的同义、反义等关系,使用户可以从任一节点开始,根据信息间的联系,从不同角度浏览和查询信息。随着计算机技术的不断发展,图像、声音、视频、动画等多媒体信息逐步进入超文本系统中,使得超文本进一步发展为超媒体。

超媒体组织信息的优势表现在:信息的非线性编排、信息表达形式的多样性、伸缩性强(互相链接的文件可多可少,可随时增删)、能体现文献间的引用与被引用关系。

但是,用超文本(超媒体)组织信息会引起信息迷航、偏离主题、侵犯知识产权和增加信息资源的有序化整理与组织的难度等问题,而且,单纯基于超文本的导引浏览式检索只能靠浏览发现相关主题、扩大检索范围或调整检索主题,不能直接对所需信息进行查找。

1.3 数据库(Database)

数据库是对大量的规范化数据进行管理的有效

方式,它将数据经合理分类和规范化处理之后,以记录的形式存储于计算机中,用户通过关键词及其组配查询,就可以找到所需信息线索(即相关站点链接),并通过信息线索直接连接到相关的网络信息资源。

数据库对网络信息组织有以下优势:对大量的结构化数据的处理效率高;数据的最小存取单位是信息项(字段),可根据用户需求灵活地改变查询结果集的大小,从而大大降低了网络数据库传输的负载,在数据量大的情况下,这一优点更为突出;以数据库技术为基础已建立的信息系统,为网络信息系统的建立提供了现成的经验;数据库技术与网络技术的融合使得上网后的各个单独的数据库之间通过链接被联系起来,为跨库检索与资源整合奠定了基础。但是,数据库对网络上大量存在的非结构化信息的处理难度较大,无法处理日益复杂的信息单元,缺乏直观性和人机交互性。数据模型的扩充更新、知识发现技术的应用以及数据库技术与超媒体技术的结合,将是数据库用于网络信息资源组织与开发的发展趋势。

1.4 网站(Web site)

网站是网络信息资源的重要组成部分,是一种用标记语言(描述性语言)将信息组织好,再经过相应的解释器或浏览器翻译出的包括文字、图像、声音、动画等多种信息的组织方式。从网络的组织结构可以看出,信息资源主要是分布在网站上,互联网提供的信息服务在很大程度上也是依靠网站来实现的,因而网站集信息提供、信息组织和信息服务于一身。

网站组织信息最突出的问题是信息的保存与信息质量问题。尽管网上信息的更新是必须的,但是旧的版本应被保存并且可以被参照以便维护资源的永久性。大量散布虚假信息、淫秽信息、色情信息、暴力信息和进行煽动性政治宣传的网站充斥着因特网。

2 网络信息的宏观组织模式

文件、超媒体、数据库和网站的丰富信息入网后,快捷、高效地从中找到所需要的信息,一直是用户的愿望。随着网上信息的急剧膨胀,对它们进行组织与控制的工具——网络资源指南和搜索引擎便应运而生,它们是目前 Internet 上两种很流行的信息组织方式和重要的检索工具,组织的是整个网络范围的信息,提高了网上资源的序化程度,在一定程度上满足了人们网上信息查询的需求。当然,这样的划分只是为了介绍的方便,并不是绝对的区别,而是针

对检索工具的主要功能而言的,因为二者有相互融合渗透的趋势。

2.1 网络资源指南(Web directory)

网络资源指南是基于人工建立的网站分类目录,是网站的分类链接列表。它们通过人工浏览 Internet 页面,根据一定的标准(不同指南的选择标准有别)来挑选所录用的链接资源,然后将各种资源按一定的分类体系(自己设计的分类体系或已有的分类法)组织,并辅之以年代、地区、主题等分类,形成分类树状结构目录。因而,网络资源指南方式有时又被称为主题树(subject tree)方式。Yahoo! 是最典型的网络资源指南。

2.1.1 网络资源指南的优点

网络资源指南组织网络信息的优点在于:组织的信息专题性较强,且能较好地满足族性检索的要求,用户按规定的分类体系,逐级查看,按图索骥,目的性强,查准率高;采用树状结构组织网络资源具有严密的系统性;屏蔽了网络资源系统相对于用户的复杂性,提供了一个基于树状结构的简单易用的网络信息检索与利用界面。最为突出的是,网络资源指南的类目是完全根据网上信息的类型和特征及一般用户查询的重点设置的,并且具有动态性,因此,有很高的适应性和实用性。同时,充分利用计算机操作环境与技术而建构的分类体系,可揭示多维知识空间的联系。参见(如 Yahoo! 中以@ 标明的类目)和分析方法也有不同程度的使用;利用超文本技术把类与类、类与记录链接起来,有友好的用户界面。这些都是传统分类法所无法匹敌的。

2.1.2 网络资源指南的问题与对策

网络资源指南组织信息的问题有:收录范围与新颖性不够、对资源的描述不够、分类体系与方法不尽合理(如:分类大纲的设置不够科学、类名不规范、类名用语模糊、类目排列随意性强且不够合理、分类缺少提示、重复列类和各大类下细分的程度差异太大等),它们也未能对网络信息资源进行有效地整序和控制,需要利用分类法与主题法的成果。20 世纪 90 年代以来,国际著名的分类法(DDC, LCC 和 UDC 等)和主题词表都纷纷被改造成为适合于网络信息资源组织的工具,已建立起多个比较典型的综合性网络资源系统。

2.2 搜索引擎(Search engine)

2.2.1 搜索引擎信息组织的特点

搜索引擎通过在互联网上提取各个网站的信息来建立自己的数据库,并向用户提供查询服务,它在

组织信息中具有以下特点:定期自动搜寻有关 Web 站点,以采集各类信息资源;自动对这些资源进行标引、著录,并将标引结果组织到数据库;提供基于 Web 的检索和各种检索限制,并可按相关度、时序等标准输出检索结果。

2.2.2 搜索引擎在信息组织中的问题与对策

作为一种信息组织方式,搜索引擎最突出的问题是检索效率不高。其他的问题有:命中的文献间相互孤立或关系模糊不清(找出的各文献是相互孤立的,不能体现信息之间主题的相关性、不同版本间之互动性与关联性等);各搜索引擎标引方式也没有统一的规范;检索界面各不相同,增加了用户的认知负担;检索结果的排序机制有待改善。不同的搜索引擎对搜索结果排序的标准不同,即便是同样按相关度排序的各个搜索引擎,它们在相关度的计算上也是差别很大。近年来,由于商业利益驱动,百度等搜索引擎推出的“竞价排名服务”,使得搜索结果的排序依据就是竞价的多少,而一些优秀的网站往往排在后面。

面对 Internet 上纷繁复杂且杂乱无章的庞大的信息资源,信息的整序显得日益重要,主要由 IT 界设计研制的网络资源指南和搜索引擎存在许多问题,而有效的优化办法尚处于探索之中。网络信息的组织和管理仍然需要以信息组织的理论来指导,才能达到高度的有序化,因为信息不论以什么形式传播,其内容属性都是不变的。一些专家呼吁用“图书馆员的思维”管理网上信息资源。在这种背景下,网络资源编目方式得以产生。

3 网络信息的重组模式——编目

20 世纪 90 年代以后,图书情报界开始将传统的编目格式进行改造,以适应变化频繁的网络信息资源的描述和组织。机读目录格式的扩充和元数据的产生就是基于这样一种思路。

3.1 MARC 格式的扩充

随着信息存储技术的发展,越来越多的资源可以以电子的形态获取。因特网的发展,使图书馆的资源已不再局限于存在于该组织的实体的物理资源,还包括因特网上数以万计的虚拟的网络信息资源,因此,图书馆除了使用 MARC 来描述馆藏印刷型文献的书目信息之外,还必须对网上的信息资源进行揭示。

为了能在机读目录中揭示电子资源(包括光盘、网络资源等),MARC 的制定者 LC 早在 1991 年就

提出以 U S M A R C 为主要框架,制定囊括网络信息的相应字段,并和 O C L C 等单位不断对 U S M A R C 进行了多次局部修订以适应网络资源编目发展的需要,在其中加入、更新或扩充定义了一系列新的字段。例如,用 856 字段(电子资源定位与检索)用于著录网络信息资源的存取方式及其他必要信息,如 U R L、存取方式、主机名称、路径、文档名称等;用 256 字段表示计算机文件特征;用 516 字段反映计算机文件类型或数据附注;用 753 字段对检索计算机文件所使用的系统细节进行描述。目前,我国的西文机读目录格式基本依照 U S M A R C,而中文机读目录格式一直以 U N I M A R C 为基础,依据 U N I M A R C 增加了相关字段。

用 M A R C 进行的网络信息组织实际上属于受控编目,要求编目人员严格遵守编目规则(如 A A C R 2),利用受控词表和标准分类体系等,能对信息进行较完整层次的分析描述,编目数据质量高,提高了用户发现资源的能力。但是,属于精致编目的 M A R C 格式仅仅用于图书馆系统间数据交换,不能取代系统的内部格式,而且,在反映结构化数据方面,缺乏可扩充性。格式本来就十分复杂的 M A R C 进行了扩充后,更使非专业人员难以参与进来,而图书馆员又难以完成数量极为庞大的网络信息资源的编目,也无力应付网页的动态变化。现实的需求促进了元数据的发展。

3.2 元数据(Metadata)

元数据最通用的定义是“关于资料的资料”,是因特网中组织信息的重要工具,由一组有关资源的各个方面的属性(attribute)组成,每个属性包括一个属性类型的一个或多个属性值(value)。目前出现了很多种元数据规范,如:都柏林核心元数据(Dublin Core,DC)、更结构化的文本输入创始计划(Text Encoding Initiative,TEI)等。还有许多应用于各个专业领域的元数据标准。其中,DC已成为国际上最通用的元数据,也是万维网联盟(W3C)推荐的标准,DC由15个基本元素构成,其应用领域涉及政府、教育、管理、地理、图书馆等。由于DC本身还处于不断发展的过程中,至今并不完备,体系也逐渐庞大。

M A R C 格式与 D C 等元数据格式在网络资源组织中各有优势与不足,在网络资源组织上两者应共存互补。前者用于组织专业性较强的学术性信息,后者则用于组织普通的网络资源。当然,也可以同时使用两种格式,O C L C 的合作联机资源目录就同时

用 U S M A R C 和 D C 进行编目。

3.3 网络资源编目问题

编目方式的优点是每条记录都经过了严格的选择,具有较强的针对性和较高的可靠性,但其中的问题也不容忽视。由于网络资源数量庞大且变化多端、编目的范围与繁简程度难以确定、各种编目规则与格式的兼容、其他元数据与 M A R C 之间的映射等问题制约着网络资源编目的发展。

4 网络信息的深层组织模式——学科信息门户

针对网络资源指南和搜索引擎的不足,图书情报界除了对网络资源编目外,还积极将图书馆传统的信息采集、标引和组织优势扩展到 W e b 空间,试图提高网上资源的序化程度,以弥补搜索引擎的不足和改善目前检索低效的状况,开发出了“基于学科的信息门户”(Subject-Based Information Gateways,即 S B I G s),以下简称为学科信息门户。S B I G s 曾经是“提供可检索和可浏览的因特网资源目录的联机服务系统,集中于某一相关的学术领域,提供对经图书馆工作人员遴选和按学科组织的因特网资源的利用^[3]。”在开放式数字信息服务环境下,学科信息门户将成为特定学科领域的信息资源、工具与服务的整合平台。

4.1 学科信息门户信息组织的特点

学科信息门户按照某学科(专题)用户的要求对网络中相关的信息资源进行了更有针对性、更深入的揭示,在给用户“指路”的同时提供更专门、更专深的信息检索服务,有助于专业用户在本领域的“信息超市”中选择高质量的资源和对信息“一站式获取”,从而保证用户获得“所得即所要”的信息。对图书馆来说,学科信息门户拓宽了本馆的“馆藏”,而对整个网络信息系统来说,其信息的序化程度得以提高。正是由于学科信息门户具有这些优于搜索引擎的特点,新的学科信息门户近几年在国外不断涌现。2001年底正式启动的中国国家科学数字图书馆已建立起图书情报学等多个学科信息门户。

4.2 学科信息门户的问题

学科信息门户有效地弥补了搜索引擎检索结果冗余量过大、检准率不高的不足。但是,因为学科信息网关主要甚至全部靠人工、使用受控语言来组织网络资源,却不具备搜索引擎索引资源涉及面广、检索覆盖率高的特点,相比之下收集的资源相当有限,学科信息门户建立所需的经费和建立之后的更新与

维护等问题都不容忽视。

5 网络信息的分布式组织模式——数字图书馆

如前文所述,文件、超媒体、数据库、网站、网络资源指南、搜索引擎、编目和学科信息门户等信息组织方式已运用于 Internet 信息的组织。从某个局部来看,如某个文件、利用超文本链接的相关资源、某个网站、某个数据库、某个编目记录集合或某个学科信息门户,是有控制的、相对集中的、有序和规范的。但从总体上看,由于互联网上的信息没有统一的控制,信息的质量参差不齐,网上的信息是分散、无序、不规范的;由网络互联在一起的分布信息仓储是异构的,这些各自独立的信息仓储具有各自不同的组织、描述和检索方式,难以实现跨仓储的统一利用;对知识的运用还远远不够,尤其是面向需求的用户知识和领域知识。人们需要一种跨仓储的、统一的、高效的访问和利用工具,以及高质量信息的生成、组织和提取途径,数字图书馆正是迎合了这种需要。

5.1 数字图书馆信息组织的特点

5.1.1 以数字对象为组织单元

与以网站(网页)或文件为对象的组织相比,这种组织更为深入,可以实现对知识内容的标引而不仅限于文件标题或关键词。

5.1.2 信息海量性

数字图书馆面临的数据是多类型和海量的,它存储的信息是以 TB (Terabyte, 1 TB= 1000GB) 为最低限度的,如“美国的记忆”已有 104TB 的存储量。

5.1.3 资源分布化

数字资源存放在不同结构的不同空间,构成数字图书馆数据层的各个“存储小间”(storage cells)有着不同的目标和存储对象,每个仓储在本地对各自的信息进行组织,并施以相应的筛选、索引、联合等控制,由不同的单位(机构)建设或管理,在此基础上,借助于数字图书馆的开放框架,在各信息仓储间进行互联,在总体上构成一个分布式的系统。该系统要涵盖多个分布式的、超大规模的、具有可互操作的异构多媒体资源库群,通过因特网对全球用户提供高效、跨库、无缝连接的信息服务。

5.2 数字图书馆信息组织的难点

由于数字图书馆出现的时间还不长,在信息组织中存在许多有待研究与解决的问题。目前面临的主要挑战是适用工具与平台的开发,包括总体结构

标准、软硬件技术、信息录入工具、知识挖掘工具等,标准与规范的建立、合作建设、各国在对本国文化进行“数字勘探”中对信息技术发达国家的依赖。虽然数字图书馆尚处于探索阶段,在现有的各种网络信息组织模式中,它最有发展前景,将成为下一代因特网上信息资源的组织与管理模式。

6 结束语

图书情报机构在传统的文献信息资源组织中积累了丰富的经验,但在以网络为主要形式的信息资源迅猛发展的时代,如果忽视对网络信息、网络环境、网络用户的研究,如果不以积极的态度在新的组织对象和环境中推广自己的成果并加以发展,图书情报界难免会在本属于自己“领地”的知识信息组织大舞台上显得有些黯然失色。

各种分类法与主题词表在网络信息组织中的运用、MARC 格式的扩充、OPAC 对网络信息资源的揭示与提供检索、元数据的提出、学科信息门户和数字图书馆的出现,无不包含着图书情报工作者的智慧。Yahoo! 和 Google 能成为各自领域的佼佼者,与前者对分类法思想的借鉴而后者对引文索引理念的运用不无关系。这些事实无可争辩地说明,图书情报界在网络信息组织中尚有巨大的潜力。

鉴于以上情况,如何针对网络信息资源的特点,特别是针对网络信息超链接、媒体多样、动态性强、可以全方位地多维揭示与描述的特征,改造传统的分类法与主题词表用于组织网络信息资源,并吸收计算机科学等其他学科的营养,这是摆在新世纪信息工作者面前的重要课题。时代的发展要求我们以开放的思维积极参与网络信息组织的实践与理论研究,图书情报学与现代信息技术的结合是学科发展的新的生长点,也是发挥我们知识信息组织的特长和提升学科在网络空间地位的必然要求。

注释:

- 1 藏国全. 论网络信息组织. 图书情报知识, 2002(3): 2-5
- 2 张俊. 略论网络信息资源的组织. 图书情报知识, 1998(2): 32-35
- 3 Subject Gateways. URL: <http://www.desire.org/html/subjectgateways/subjectgateways.html> (访问日期: 2003-3-26)

作者简介

黄如花, 武汉大学信息管理学院副教授、博士。已出版专著 1 部, 参编 2 部, 发表论文 30 余篇。