●萧德洪

网络资源的著录单元探讨

摘 要 提出发现、确认、选择适合于学术和知识门户收入的网络资源需要考虑的问题。讨论了针对适用的网络文献和资源系统,通过分析其 URL 链、文件格式、系统架构、逻辑组织以及其他特征等,确定其为"独立资源单位"而在学科导航系统中予以著录。参考文献 8。

关键词 网络资源 资源著录 目录实体 学科导航
分类号 G254

ABSTRACT In this paper, the author summarizes some problems in the discovery, identification and selection of network resources for academic and knowledge portals, discusses the description of network resources as "independent resource units" in subject navigation systems. 8 refs.

KEY WORDS Network resource. Description of resource. Bibliographical entity. Subject navigation.
CLASS NUMBER G254

目前大多数图书馆对传统载体的文献资源与网络信息资源的处理有着显著的区分,甚至可以说无法统一。虽然我们提供了网络资源服务,也只是简单地将其作为一种新的服务而很少用传统图书馆学的知识去解决这些资源的利用问题。例如,我们在采购或租用一个数据库时,通常是按整个库或者包(package)采购或租用的,虽然这是提供商的销售手段,但它确实在被大量采用,但在传统的文献采购中,我们是不能容忍书商根据他的意愿把大量的书刊捆绑销售给我们的。在文献加工领域,我们也以元数据的处理工具取代传统的 MARC 来标引网络资源,这使网络资源与图书馆的传统馆藏在信息系统上处于分离状态。接下来就是对电子资源进行采购验收,当我们发现电子资源的文献单元很多情形下竟然按章节提供,如一本《红楼梦》的电子书被拆分成 120 个单元卖到图书馆,我们似乎很难接受这种现实。

实际上是,这种情况是正常的。

美国南加州大学科学与工程图书馆在对网络资源评价时考虑 4 个方面:(1)来源:谁或哪里提供的资源?(2)效用:资源指向的用户或其他的目标是什么?(3)内容:如覆盖的主题;单主题或跨学科?覆盖的语种;覆盖的时间;包涵的出版物类型;综合性的资源还是选择性的资源?普通的还是学术性的?(4)结构:主要指建立在资源类型(如事实型、书目型、文摘型或全文型)之上的关于资源的组织(含数据结构、索引体系、检索策略以及帮助信息)方面的特性。

1999 年,美国 Health Summit Working Group 提出了一个网络资源评价政策书(*Criteria for Assessing the Quality of Health Information on the Internet-Policy Paper*),列举 7 条标准:(1)可靠性:包括资源来源、现时性、相关性和同行评价。(2)内容的准确与完整。(3)公开:有资源介绍和如何使用的信息。(4)链接:包括链接的选择、构建和第三方对这一资源的链接。(5)设计:指网站布局,含可获得性、逻辑组织与站内搜索能力。(6)互操作性:包括反馈机制和提供用户之间进行信息交换的手段。(7)申告:有产品与服务营销,也有信息的最初提供商的申明。从中看出,网络资源评价和选择是讲究内容与形式并重的。因此,我们在进行页面标引和收藏时,要分析出逻辑的独立资源单位,以达到准确表达单个网络文档的目的。

1 资源与实体

从传统的文献编目概念出发,"文献"这个术语在著录和标引中是指一部文献、一组文献或文献的一部分。那么,在网络资源方面,因为没有物理形式,"文献"一词难以引申过来,该由什么术语来表达这一网络资源的存在形式和资源著录的基础呢?

国际图联用了"实体"(Entity)的概念,这一概念在 ISBD (International Standard Bibliographic Description,国际标准书目著录)中也出现过,但未得到充分阐述。国际图联在 1996 年的 FRBR (Functional Requirements for Bibliographic Records)草案中提到以用户任务划分的检索过程:发现实体、确认实体、选择

实体、获得实体的存取。发现,即根据用户检索标准来发现实体,用实体的属性和关系在一个文件或数据库中找到检索的结果;确认,在于辨别相同特征的两个或两个以上的被著录的实体;选择,就是选取在内容和载体形式上满足用户本身要求的实体;获得,指通过购买、借阅或者通过网络链接得到这些被著录了的实体。编目必须尊重这些用户任务原则,比如,在发现阶段,编目原则要提供不同检索途径,并提供权威档控制机制,以将用户的检索指向所标引的实体。这样就达成了每一个实体通常只有一个名字(题名、责任者等),而名字的其他变形也可检出,保证检全率。在辨认和选择阶段,编目后所提供实体的各种属性帮助用户辨认这一实体是否是最符合自己需求的。比如一个相同的作品,不同的版本、印次就可以提供给用户作为判定的依据。同时,其他的特征也可以得到,如实体的内容部分,像插图、索引和参考书目等的特征。对电子实体而言,由于电子实体的拷贝成本极低,足够的特征信息尤为重要,这些将影响到用户对资料的选择。在获得阶段,用户依靠目录中提供的服务信息,在传统的馆藏资料中,索书号即是标识符,起到指向馆藏地点的作用。在互联网资源中,统一资源标识就成了读者获得资料的标记,但极不稳定。因而互联网资源编目中的回访和重新修订记录显得极其重要。

2 简单分类

互联网资源的特征是其广泛的信息和提供着相关资源的超链接,这些对图书馆的用户来说很有价值。网站由网页构成,网页又是由文本和图片(有时甚至有声音和影像)组成。目前网页是建立在HTML语言上的,因此"网页"和"超文本标记语言页面"两个术语在某种程度上可以互换。在以往关于网页资源编目的探讨中已经明确:网页指的是在对网站或URL进行存取时,呈现给用户可视的、用眼可读的、所谓的"表面"数据。

ALA的下属委员会MARBI早在1991年就提出第49号讨论稿,讨论网上信息资源的利用问题,揭示了网上资源的某些不确定性,比如,应该著录些什么?在线的特性,在线的涵义,在线的网络可获得性及其电子格式的属性等等。次年,讨论获得了一些进展,比如,远程存取是这类资料的规定属性和统一属性,因为这些资源不像实体馆藏一样,馆藏在架,伸手可触,并可为读者借出。当时也认定了这些远程存取的实体有两大类:一是数据资源如软件,文本和数据文件,目录数据库;二是系统与服务类,如校园信息系统,图书馆目录系统,公告板等。这些实体可以提供FTP或Telnet登录。

沙(Sha, Vianne)在其建议的Guidelines for Cataloging Internet Resources中,也将互联网资源分为两类:一是系统与服务,包括所有的目录和子目录,这些目录和子目录中包含不只一个目录实体。二是独立的目录、子目录和文件,这些只包含单一的目录实体。第一类以一个实体进行著录,类似传统编目中的专著;第二类类似传统编目中的丛刊,如果它们确实是电子丛刊,或者各自的内容具备各自更新的倾向,就应该各自著录。假如电子资源的地址总在不停地变换,则只著录这个"系统与服务",标引其最为专指的目录或文件集即可。

3 DC已有成果的运用

DC核心元数据标准给我们提供了资源区分的一些线索。比如,"语言"元素的使用,使我们在标引时注意到不同语言的材料构成的是否是一个资源实体,从而区分确属不同的资源。又如,"时空范围"、"权限管理"等给一个资源的界限画出了一定的轮廓。下面就一些重要的元素加以说明。

题名与题名屏不是所有的资源都有,有的资源并没有这一页面。如果没有,在前页的有关栏目中可以找到题名线索,比如主菜单、纲要说明、标头部分或者出现在其他页面区域的内部链接。

但如果页面上没有可见的文档题名,而源文件的元数据头标区里给出了资源题名,也是不可靠的。这时需要借助其他元素的特征综合考察。

创作者,这是个很重要的概念,有的一个网站的不同资源是同一个机构的产品或服务,有些则是与其他机构或个人合作的产物。不同的创作者往往可以区分资源的独立性。另一方面,"创作者"在检索时是个重要的存取点。与创作者类似的是,贡献者也有这样的特性。

资源类型是区分资源单位的有效办法,假设我们有一个好的资源类型控制表,就可以给资源标引一个好的指导。

格式也很重要,不同的资源在数据格式上表现不同。比如DC中规定用的text/html, ASCII, Postscript file, executable application,或JPEG image等等。与"资源类型"一样,把格式设定了,可以在一定程度上把

不同的资源分别开来。

资源标识可以看做是网络资源的 ISBN，全球惟一。不同的是它容易被迁移。

来源，即衍生出特定资源的资源。换句话说，所标引的独立资源是“来源”的衍生物，来源是独立资源的根节点或枝节点，而独立资源是叶子节点。

4 关于网上资源存在形式的进一步分析

内容自然是很重要的因素，除内容的准确、可信、及时等必须具备的特性外，网站资源在内容上的主题集中度和明晰的主题界限显然是判断是否可以作为独立资源的基础。在形式方面的判定依据有：

(1)链接。重要的是向外链接，联向站内其他资源或其他资源站点。这些链接尤其是外部链接的内容是经过专业化过滤选择的，也与提供该链接的本站的内容特征相符。链接还包括回退链接。我们做“学科导航”，是根据内容相关性来组织网页资源的。当我们要对“关联”资源做出存取点时，很自然地要单独取标引这一相关资源，而不管这个相关资源的主机上的所有资源又是如何相互掺杂在一起的。

(2)体系结构。指服务于特定内容资源由不同的应用子系统构成的有机体系。比如在一个服务系统中，可能有站内搜索引擎、咨询台、图形用户界面(GUI)或者交互式用户界面，也可能有用户认证系统、购物车、支付系统等。重要的是网站的体系结构对用户来说是否透明，用户是否可以很容易地进入和退出，这些资源单位是否有明显的逻辑特征，而且易于和其他资源区分开来。

(3)设计。主要指页面的设计和布局，也就是对文字、图形和超链接的安排。这些安排可以给标引者和用户提供网站组织的逻辑线索，即使普通的用户也能便捷地在同一站内得到并列提供的不同独立资源。

(4)逻辑组织。指适航性而言，好的网站总是瞄着它的潜在点击者，有明确目的。每个资源的内容一致，相互的参照也在于帮助浏览者更好地理解整个资源的信息结构。但站内搜索不能算是一种逻辑组织，它犹如链接一样，功能通常用于跨越多项独立资源的大范围和深度检索，所以站内搜索对独立资源的判别而言，不一定有用。

(5)申告信息。网站一般有有关的声明，这类声明通常是针对整个网站的，也有的是针对网站内的某一特定资源的。还有一类申告是通知，通告内容提供者提供的内容或服务的计划或启事，这类信息对确定资源独立性与否也有帮助。

5 独立资源单位

独立资源单位，指在因特网上可以通过统一资源标识直接进行存取的、字串匹配的网络资源单位，不包括通过人机交互(查询或者搜索)获得的文档。独立资源单位遵循两项原则：

(1)因为是因特网中的资源，所以它强调“所见”和视觉呈现，强调资源内容与文档格式不可分离的原则，它必须是借助于特定的浏览器或工具来进行阅读的。

(2)字符串匹配原则。首先，必须有协议标志字串，以“http”、“ftp”等为引导，随之以文件的路径和属性，最好是具有文件扩展名(如.htm,.html,或.NET/ASP/JSP/PHP/Perl/CFML等)的字符串。它排除计算机文档分级目录中的“文件夹”或者“属类”作为资源标引的对象。

比照 Sha 的概念，我们也把网络资源分为两类：一是系统与服务，包括所有的目录和子目录，无论这些目录和子目录中包含多少个目录实体，她的处理是按照集中的原则一次著录。但我们要求将这个系统与服务的目录和子目录尽可能地分析出来，形成一个个独立的目录实体，以尽可能回避上位的、不够专指的标引结果。二是独立的目录、子目录和文件，这些有的只包含单一的目录实体。对于后者，处理原则相同。

具体文档的 URL(Uniform Resource Locator)一般都嵌入了级次不同的根目录，这种根目录，如果被单独标引出来，则成了族 URL(home URL)，通常以斜杠“/”为后缀。作为所有的 URL 下级 RRL 的路径时，这一“族 URL”就成了所有 URL 的前缀，包含在下级 URL 字串中。

事实上我们不能根据 URL 字串的外形来断定以“/”为字串结尾的是根目录或根节点，是不是我们所要的“独立资源单位”。因为我们给定一个 http://www.greatdocs.com/foo/bar/这样的 URL 字串后，我们就获得了所要的资源，但我们无从知道在这/foo/bar/之后给我们的那个文件到底是 foo/bar/index.html、foo/bar/bat.htm 或是 foo/bar/home.html，还是 foo/bar/default.html 或 foo/bar/default.htm，一切都是由该地服务器事先设定的，作为用户是很难找到踪迹的。另外，因为一种同样内容的资源可能在网络上存在不

同的载体或地点，也可能是不同的格式，它必须确定资源"题名"与URL相对应的重要性。在URL上可以是"多对一"的关系。但从逆向看，也就是从资源本身出发指向URL时，不应出现"多对一"的关系。简而言之，一个资源可能有多种物理形式或者数据格式的存在，但一种物理形式或数据格式的资源只能指向一个URL。

6 资源类型控制的意义

图书馆在浩如烟海的图书中进行采访选择，必须有一个采访大纲做指南，其目的是规定本图书馆藏书建设的主题范围和非主题因素。在互联网资源中也存在同样的要求，所以导航库建库方针应事先确定好适合目标用户的主题类目、资源类型和资源水平级别。

资源类型控制的意义在于，它是资源选择过滤器的重要组成部分，因此成为最为通用的准则。主题类目、资源级别只要不符合目标用户需要的就不在收录之列。但有时我们无法判定某一资源是大众性的还是学术性的，如政府组织的、财经商业机构的、工业贸易机构的、非营利组织的乃至于个人提供的资源既面向学术也面向大众，就必须有专门的资源类型控制方案进行过滤。

属学术性的导航系统，大约有以下几方面的资源类型：(1)参考性资源：资源导航、辞典与百科全书、文献与索引、统计资料、标准与规范、专利文摘；(2)全文性资源：数据库、电子期刊、研究报告、政府出版物、独立性文本资源；(3)行业性资源：协会学会、大学院系、研究机构、专家学者、教学资料；(4)交互性资源：邮件列表、论坛/讨论组、新闻组、搜索引擎；(5)多媒体资源：图像资料、音频资料、视频资料；(6)事件与其他：重要会议、研究项目、学术动态、专业动态、专业网站、工具软件、产业、产品、市场、案例、仪器、图书馆、博物馆等。

网络上还存在着大量的服务性资源，这些资源在系统的某个主题方面，或者针对某些特定用户，可能上升为学术性资源，但在一般情形下，这些归为大众性资源。因此在著录时一般予以排斥，如确实需要，应当结合其他因素慎重选择。服务性资源包括：新闻、市场行情、专家咨询，天气预报、地区代码、邮政编码、电话号码、电子邮件地址，地理地图、航班时刻表、火车时刻表、汽车时刻表、机场、银行、租车、旅馆、饭店、公寓，政府、国家、领事馆、大使馆，创业指导、投资资本、小企业指南、工商行政等资讯。

资源类型区分的另外一个意义在于，我们在做好类型划分的基础上，可以有选择地规定哪些类型为必选，哪些类型为可选。比如，导航系统如以内容全文提供为重，就势必把索引型的资源划为可选，而不是将所有相关的索引都收入。如果导航系统以提供链接列表为主，则索引型的资源，如搜索引擎、资源导航、文摘与索引等就成了必选项。

参考文献

1 Taylor, Arlene G. The Information Universe: Will We Have Chaos or Control? American Libraries 25(1994): 629 - 632
2 Caplan, Priscilla: http://info.lib.uh.edu/pr/v4/n2/caplan.4n2
3 Sha, Vianne: http://www.itcompany.com/inforetriever/catinpol.htm
4 http://www.usc.edu/isd/locations/science/sci/pubs/criteval.html
5 http://www.ala.org/Content/NavigationMenu/ALCTS/Division_groups/MARBI/Next_Section_3.htm
6 Cataloging Internet Resources: A Manual and practical Guide. http://www.oclc.org/support/documentation/worldcat/cataloging/internetguide/
7 http://hitiweb.mitretek.org/docs/policy.html
8 萧德洪等.学术图书馆学科导航门户资源类型表的设定.大学图书馆学报,2004(4)

萧德洪 厦门大学图书馆副馆长，副研究馆员。通信地址：福建省厦门市。邮编361005。

(来稿时间：2004 - 07 - 13)