

·实践平台·

分类法在元数据体系构建中应用的理论分析

吴飞翔 (浙江经济职业技术学院 浙江杭州 310018)

摘 要: 文章对元数据内部体系和外部体系的建立分别进行了讨论,阐述了分类法在构建元数据框架体系中的有效应用模式。

关键词: 分类法 元数据 元数据体系

中图分类号: G254.1

文献标识码: A

文章编号: 1003-6938(2009)06-0131-03

Theories Analysis of the Classification Application in Constructing Metadata Frame Systems

Wu Feixiang (Zhejiang Technology Institute of Economy, Hangzhou, Zhejiang, 310018)

Abstract: This paper discusses the construction of interior system and exterior system of metadata and the two aspects in which traditional taxonomy is applied to metadata framework construction effectively, and the influence of the metadata and the concrete request of the metadata project establishment.

Key words: classification; metadata; metadata systems

CLC number: G254.1

Document code: A

Article ID: 1003-6938(2009)06-0131-03

1 元数据体系的建立

元数据体系是一种用来描述数字化信息资源的编码体系,特别是网络信息资源的基本特征及其相互关系,从而确保这些数字化信息资源能够被计算机及其网络系统自动辨析、分解、提取和分析归纳(即所谓机器可理解性)的一整套编码体系。^[1]一般而言,元数据体系由两部分组成,即外部系统及内部系统。所谓外部系统即各种独立于具体系统的、被广泛承认的、通用的元数据标准的总和;内部体系是指系统本身的元数据处理方法和体系结构,即元数据管理系统。

为了实现数字化图书馆和外界信息环境的沟通,元数据内部系统和外部系统必须是同构的。这种同构关系实际是将外部元数据系统映射到数字化图书馆的内部体系中的方法,为了建立同构关系,首先要定义元数据方案的体系结构,一般包括元数据的语义、语法和结构的规定。语义是元数据互操作的本质,语法是表现形式。结构是描述框架。这三方面被视为解决元数据互操作的技术途径。^[2]语义问题即是要提出一套应用于本项目资源对象描述的核心数据要素集;语法和结构问题即是要提供元数据的置标方案以及可供元数据进行语义交互的“包”和“容器”。^[3]

就以上对元数据语义、语法和结构的要求可以分为更为具体的6个部分:

(1)基准元数据系统。基准元数据系统是指某个数字化图书馆标准的元数据系统。它的作用是:作为基准元数据,组织标识数字化图书馆中的数字化信息资源;以标准形式描述用户的查询提问;为各种网络信息发掘工具提供数字化信息。

(2)元数据字典。元数据字典是一种用于各种元数据体系到系统基准元数据系统相互转换的对照表,它描述了各种元数据的基本特征,构建了各种元数据与基准元数据系统的对应关系,其基本作用是为系统的转换模块提供转换依据。

(3)数据属性集。数据属性集是指数字化图书馆存储数据的属性总和。元数据管理系统可通过数据属性集将数字化图书馆的数据结构和基准元数据相对照,保障它们之间的可互换性。

(4)数字化信息资源集。数字化信息资源集描述的对象是信息源。数字化图书馆系统可以通过信息源特征集来确定某个信息源所采用的元数据体系,将用基准元数据表达的查询式转换成各信息源所采用的元数据表达式,从而决定各个信息源的检索方法并解释检索结构。

(5)转换模块。转换模块主要是提供实现各种元数据之间相互转换、翻译的方法。

(6)维护模块。维护模块可以对各种对照表进行添加、删除、修改等动态管理,保证元数据管系统的可扩展

展性和可维护性。

2 分类法在元数据体系构建中应用的模式

由上述分析得知,在元数据体系的六个模块中,基准元数据系统、数据属性集是针对内部体系而言的;元数据字典和转换模块是相对于外部体系而言;剩下的是维护和查询体系的有效组成部分。总体而言,元数据体系的建立与否取决于这六个模块的建立,抽象来讲,就是内部系统和外部系统的建立及维护内外系统三者的统一。为了实现不同元数据体系之间的互操作,需要对其进行规范控制以促进数字信息资源共享,数字信息资源管理的重要工具——分类法的应用则是规范控制中重要的环节。^[4]在内部和外部系统建立时,分类法充当了不同的角色和作用。

2.1 内部元数据系统构建中的分类法

内部系统的建立主要取决于基准元数据系统的建立,进一步的基准元数据的建立其实质就是采取统一的元数据格式,例如DC或者MARC格式。再更进一步的分析,不论是DC还是MARC元数据的建立,首先要明确元数据中元素的建立,对于DC而言就是15个元素的基本信息建立,对于MARC就是数据区中元素信息的建立。^[5]

虽然MARC和DC在著录内容上有区别,但是在本质上都是对原信息资源的一种外在描述。为了简便起见,本文只是针对DC元数据的元素建立进行详细分析(DC元数据的元素建立表见表1)。^[6]

从表不难发现,在建立基准元数据体系的过程中,分类法作为元素编码体系主要是针对在Subject主题这一元素,也就是说在内部体系的建立中,分类法主要是应用于基准元数据的主题词建立。目前DC主要应用于数字图书馆,笔者调查发现,都是采取这一典型的方式,即利用分类法作为对主题词的编码体系。

目前各国的数字图书馆或信息中心,都纷纷利用了分类法对其内部数据库中的元数据进行分类。^[7]例如利用UDC的Arup-Information Management Group, Hilary Doughty Research Resources Unit的ISER-Institute for Social and Economic Research,都是利用UDC进行元数据分类。利用DDC对其内部数据库中的资源进行分类的也不少,如CIBS-Canadian Information by Subject。

2.2 外部元数据系统构建中的分类法

外部系统的主要作用在于建立和连接两个或两个以上的内部系统,同时提供一种相交换的可能性,使得元数据在网络交流中可以畅通无阻。资源需求者可以有效快捷地搜寻到所要的信息资源;还可以保证这一信息资源的可用性,这里的可用性是指数据在不同的计算机能否同样的读取,对于后者而言,现在主要致力

于RDF、XML等数据格式标准化。本文主要是对于前者的研究,即在检索过程中如何有效地使信息需求者找到所需资源。这一方面,目前主要的研究方面在于概念网络的建立。^[8]

表1 DC元数据的元素建立表

元素名称	元素精确限定	元素编码体系
Title 题名	Alternative 交替提名	-
Creator 创建者	-	-
Subject 主题	-	LCSH 国会图书馆主体词表 MESH 医学主体词表 DDC 杜威分类法 LCC 国会图书馆分类法 UDC 国际十进制分类法
Description 描述	Table of contents 目次 Abstract 文摘	-
Publisher 出版者	-	-
Contributor 其他责任者	-	-
Date 日期	Created 创建 Valid 生效 Available 可获得 Issued 发行 Modified 修改	DCMI Period W3C-DTF
Type 类型	-	DCMI Type Vocabulary
Format 格式	Extent 范围	
	Medium 媒体	IMT
Identifier 标识符	-	URI 统一资源定位符
Source 来源	-	URI 统一资源定位符
Language 语种	-	ISO 639-2RFC 1766
Coverage 覆盖范围	Spatial 空间	DCMI Point ISO 3166 DCMI BOX TGN

以分类法和主题词表为基础,集成语义元数据构建概念网络的方法如图1所示。

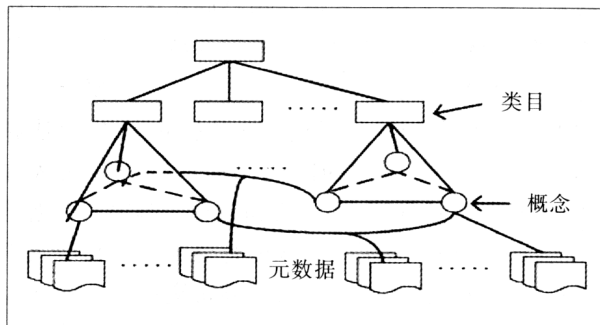


图1 集成语义元数据概念网络^[9]

整个系统由两层构成。分类法和主题法结合起来构成系统的上层——Ontology, 下层是语义元数据。首先将主题词改造成一个概念网络, 一个主题词和它所有的同义的非正式主题词(Used For)所组成的集合构成一个概念节点。概念网络中的概念有两类基本关系

(从主题词表继承而来):层次关系和相关关系。层次关系对应于主题词表中的上下位关系(BT,NT),是概念网络中的纬线。所有通过层次关系联系起来的概念群,构成一个相对独立的概念族,对应于传统词表中的词族。相关关系对应于主题词表中的相关关系,是概念网络中的经线,表达了概念间的横向联系,可以存在于分布在不同概念层次和不同概念族的任何两个概念之间。这样,层次关系和相关关系把所有的概念编织成一个纵横交错的概念网络。在此基础上,将分类表嵌入到概念网络中。对于分面主题词表,类目已经和主题词融为一体,只需要依据分类层次建立这些类目性概念的分类等级关系。

系统的关键目标是将特定应用的语义元数据集成到概念网络中。以OPAC的书目数据为例,根据一条书目记录的主题词字段的值把它分配到概念网络中相应的概念节点下,作为它所对应的概念的一个数据实例。如果一条书目记录的主题是用多个概念组配标引的,就把它作为所对应的这些概念的相关关系的数据实例。这样就把所有的元数据根据他们的主题分别组织到概念网络中的节点(概念)和边(概念间的联系)中去了。如果把分类法和主题法看成是树干的话,将元数据集成进去,就相当于让这棵树长满了叶子。

概念网络不再是一个抽象概念的集合,而是包含了具体数据实例的知识网络。系统中的这两个层面——上层的Ontology和下层的元数据紧密地结合在一起。

这样的概念网络是一个资源的组织框架。它充分利用了语义元数据中的内容标引信息,将离散的元数据单元连接起来组织到一个一致的知识体系中,为利用数字图书馆中大量的语义元数据提供了一个新途径。^[10]类似于图书馆中排列有序的书架,为用户提供了一个知识浏览和检索的信息空间。通过浏览概念的层次和联系,用户得以明确自己的信息需求。用户提出查询后,不是一头钻到数据库中直接搜寻文献,而是在概念网络中找出对应的概念,保证了检准率,而所有包含此概念的文獻都已经组织在这个概念之下,这样保证了检全率,同时通过概念间的层次关系和相关关系,很容易实现扩检和缩检。另外,由于组织有序的概念网络是一个有效的知识导航系统,也更利于支持用户的学习。

从分类法在构建外部元数据体系的应用分析来看,分类法在此提供了一个相当于表层搜索功能的引擎。在实际应用中,OC LC的Connexion就是提供这样一种浏览功能,Connexion网络浏览器中的交互提供一种基于DC元数据的WebDewey的搜索功能。例如国内的上海数字图书馆同样也是利用DC元进行数据检索交换的。

3 综合分析

从上文的分析可见,分类法在元数据体系构建的过程中分为两个部分,第一,即内部元数据体系的基准元数据体系的建立,在这一过程中主要针对了元数据的主题词的应用进行元数据的分类,这是一个内部的、基础的。第二,构建概念网络对不同元数据体系间进行查询、交换、使用。相对于第一过程而言,这一过程分类法的应用是在元数据外部形成的,只是在构建概念类目层面提供了相应的需求。主要原因在于一个信息资源的利用即要对使用者适用也要适用于组织者。正如在一个图书馆的内部,我们针对图书的主题词或者别的什么表征向量按一定的标准——在这里就是我们所说的各种分类法,将图书归门别类,放到书架上,这一过程也是针对组织者而言的;但是我们的目的并不在于将书归类好,真的目的在于为了方便查找而进行归类,所以我们在把书放到书架前会为其定义一个索书号,当读者查找时不再是一个书架一个书架地查找书名而是直接找到索书号并根据这个索书号进行查找,当然这一过程中是针对使用者而作的。在这整个的过程中前一个过程就是我们所说的内部系统的建立,即分类上架的过程,分类法就是为图书分类提供了一个标准,过程二就是外部系统的建立,即索书号的建立及录入,分类法的作用表现在为索书号的分类提供标准。

参考文献:

- [1]刘炜等.DC元数据的历史、现状及未来[M/OL].《图书馆杂志理论学术年刊2005:元数据与图书馆》.上海:上海科技文献出版社,2005.[2007-09-16]http://eprints.rclis.org/archive/00003408/.
- [2]孔庆杰,宋丹辉.元数据互操作问题技术解决方案研究[J].情报科学,2007,(5):754-758.
- [3][6][9]刘炜,张亮.数字图书馆的体系结构和元数据方案[J].情报学报,2003,(4):19-22.
- [4][7]黄如花.数字信息资源管理的重要工具——分类法在构建元数据框架体系中的应用调查及建议[J].情报科学,2007,(11):1601-1608.
- [5]陈军.我国元数据研究状况分析[J].江西图书馆学刊,2008,(1):121-125.
- [8]王军.VISION:集成分类法、主题词表和语义元数据的概念网络[J].情报学报,2003,(8):412-418.
- [10]张静,江汇泉.知识本体在分类法叙词表元数据组织中的应用研究[J].现代情报,2006,(10):164-166.

作者简介:吴飞翔(1978-),男,浙江经济职业技术学院图书信息中心馆员。