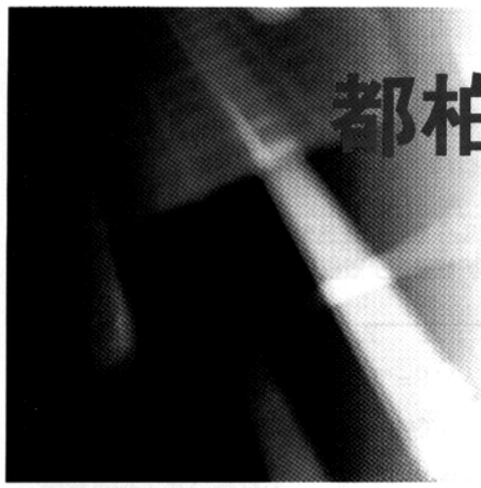

都柏林核心元数据及其应用

沈 艺 (南京师范大学图书馆计算机室 210097)

摘要：都柏林核心元数据(Dublin Core Metadata)作为互联网资源描述的有效手段，已受到广泛的关注。本文在对都柏林核心元数据进行较为全面分析的基础上，给出了都柏林核心元数据在HTML及XML中应用的实例。

关键词：Dublin Core 都柏林核心 元数据 情报检索

1 引言

互联网上信息资源丰富多采，不断增多的信息资源形成了信息的海洋。但随之而来的问题是，在海量信息环境中，通过搜索引擎进行信息检索，其准确率变得越来越低。为了有效地解决这一问题，引入了都柏林核心元数据(Dublin Core Metadata)。元数据在数据仓库中有广泛的应用，被称为是关于数据的数据，专门用来描述数据的特征和属性的。

都柏林核心元数据是由致力于规范互联网资源体系结构的国际性联合组织(Dublin Core Metadata Initiative)所定义，目的是解决网络资源的发现、控制和管理问题。其核心是一个精简的元数据集——都柏林核心元素集(Dublin Core Element Set)，简称为都柏林核心（DC）。1998年9月，互联网工程专题组（IETF）也正式接受了DC这一网络资源的描述方式，将其作为一个正式标准予以发布(RFC2413)。

2 都柏林核心元素集

都柏林核心元数据格式产生于1995年，由于它的简练、易于理解、可扩展、及能与其他元数据形式进行桥接等特性，使它成为了一个良好的网络资源描述元数据集。经过多次补充和修订，都柏林核心在结构和功能上逐渐的完善起来，能较好地描述网络信息资源。都柏林核心定义了十五项广义的元数据，

其内容详见下表：

元素标识	元素名称	定义	说明
Title	题名	分配给资源的名称	使资源为众所周知的有代表性的正规名称
Creator	创建者	制作资源内容的主要责任实体	创作、制作者包括个人、组织或机构，应该是用于标识创作、制作者实体的具有代表性的名称。
Subject	主题	资源内容的主题	用以描述资源主要内容的关键词语或分类号码表示的有代表性的主题词。
Description	说明	有关资源内容的说明	该说明可以包括但并不限于：摘要，内容目次，内容图示或内容的文字说明。
Publisher	出版者	制作资源有重要作用的责任实体	如包括个人、组织或机构的出版者。应是用于标识出版者实体的有代表性的名称。
Contributor	其他责任者	对资源内容负有责任的实体	其他责任者包括个人、组织或机构。应是用于标识贡献者实体的有代表性的名称。
Date	日期	与资源使用期限相关的日期、时间	资源产生或有效使用的日期、时间。推荐使用ISO 8601 [W3CDFT] 定义的编码形式，跟随的是YYYY-MM-DD形式。
Type	类型	资源内容方面的特征或体裁	类型包括种类、功能、体裁或作品集成级别等描述性术语。推荐从可控词表（如Dublin Core Types [DCT1] ）中选用有关术语。对于资源物理或数字化方面表示，采用“格式”项描述。
Format	格式	资源物理或数字化的特有表示	格式可包括媒体类型或资源容量。也可用于限定资源显示或操作所需的软件、硬件或其他设备，如容量包括数据所占空间和存在期间。
Identifier	标识	依据有关规定分配给资源的标识性信息	推荐使用依据格式化标识系统规定的字符或号码标识资源。如正规标识系统包括统一资源标识（URI），统一资源地址（URL）、数字对象标识（DOI）以及国际标准书号（ISBN）、国际标准刊号（ISSN）等。
Source	来源	可获取现存资源的有关信息	可从原资源整体或部分获得现有资源。建议使用正规标识系统确定的字符或号码标引资源来源信息。

Language	语言	资源知识内容使用的语种	推荐使用由RFC1766定义的语种代码，它由两位字符（源自ISO639）组成。随后可选用两字符的国家代码（源自ISO 3166）。如"en"表示英语，"fr"表示法语。
Relation	相关资源	对相关资源的参照	推荐用依据正规标识系统确定的字符或号码标引资源参照信息。
Coverage	范围	资源内容的领域或范围	范围包括空间定位（地名或地理座标），时代（年代、日期或日期范围）或权限范围。
Rights	版权	持有或拥有该资源权力的信息	版权项包括资源版权管理的说明。版权信息通常包含智力知识内容所有权（IPR）、著作权和各种拥有权。如果缺少版权项，就意味着不考虑有关资源的上述版权和其他权力。

以上15项元数据简洁、规范又较全面地概括了电子资源的主要特征，涵盖了资源的重要检索点、辅助检索点或关联检索点，以及有价值的说明性信息。这15项元数据依据其所描述内容的类别和范围可分为三组：1.对资源内容的描述；2.对知识产权的描述；3.对外部属性的描述。

资源内容描述类	知识产权描述类	外部属性描述类
Title	Creator	Date
Subject	Publisher	Type
Description	Contributor	Format
Source	Rights	Identifier
Language		
Coverage		

3 都柏林核心与HTML

都柏林核心在HTML中的应用根据HTML的版本不同略有不同，在HTML 2.0/3.2中是通过<META>标记来实现的。HTML4.0与HTML 2.0/3.2基本相同，只是对<META>标记作了扩展，主要是增加了SCHEME属性，以及增加了通用的Lang限定，并且加入了< LINK>标记来指向我们使用的元数据集及体系(SCHEME)限定的参考定义。下面给出一个例子。

```
<HTML>
  <META NAME="DC.Title" CONTENT=" 南京师范大学图书馆主页 ">
  <META NAME="DC.Date" CONTENT="(SCHEME=ISO8601) 2000-10-01">
  <META NAME="DC.Creator.CorporateName " CONTENT=" 南京师范大学图书馆计算机室 ">
  <META NAME="DC.Creator.CorporateName.Address"
     CONTENT="wxzx@pine.njnu.edu.cn">
  <META NAME="DC.Subject " CONTENT=" 读者指南，服务介绍，图书馆在线查询，图书馆在线资源，互联网资源，本馆主页集锦，相关链接 ">
  <META NAME="DC.Type" CONTENT="WWW 主页 ">
  <META NAME="DC.Identifier" CONTENT="(SCHEME=URL) http://lib.njnu.edu.cn/
    index.htm">
  <META NAME="DC.Language" CONTENT="(SCHEME=ISO 639) zh">
  </HEAD>
  <BODY>
  ...
  </BODY>
```

4 都柏林核心与RDF/XML

一个能对结构化的元数据进行编码，交换及再利用的体系框架。RDF本身并不对各种不同的元数据进行语义定义，而是提供一种框架体系，使不同的用户或团体能够在这一框架下定义他们自己的元数据的元素。RDF采用XML(eXtensible Markup Language，可扩展标记语言)作为交换和处理元数据的通用语法结构体系。

XML是用于网络环境下网页设计和数据交换、管理的新技术，具有很好的应用和发展前景。XML是国际标准SGML的一个子集、一种压缩形式，或者说是SGML一种实用形式。XML是用结构化的办法处理过去认为难以处理的非结构化的信息。XML是创建文档结构的工具，而不单单是将结构用于界面显示。它所创建的文档结构可以使管理系统精确地识别信息所在位置。它能提供数据库格式，通过交换格式以及其他应用走进所有数据处理程序。XML可以将数据的存储与数据的显现分开，即内容与形式分离。设计人员可以创建和管理自己定义的标记，语法是固定的，符号集是开放的。

XML的重点在于管理信息内容，包括超文本链接。XML的功能大大超过HTML（超文本标记语言）。XML语法：

<标记 属性＝值>信息内容</标记>

下面以中国首家网上中文期刊《神州学人》这个网络资源的都柏林核心元数据XML形式：

（下转第53页）

(上接第37页)

```
<?xml version="1.0" encoding="UTF-16"?>
<Bibliography><HEAD>,
<TITLE>都柏林核心形式</TITLE>
</HEAD><BODY>
<dc:Title>中华文化通志</dc:Title>
<dc:Creator.PersonalName="edt(主编)">张双鼓</dc:Creator>
<dc:Creator.CorporateName="bkp(制作)">清华万博网络技术有限公司</dc:Creator>
<dc:Subject>中国学者</dc:Subject>
<dc:Subject>中华学者</dc:Subject>
<dc:Description>该刊主要内容包括：中文报刊阅览室、CHISA(神州学人)周刊、《神州学人》月刊，以及若干资料库。有6个栏目，即：留学新闻、学人萍踪、学者论坛、留学生文学、服务资讯、教育国际交流。</dc:Description>
<dc:Publisher>上海人民出版社</dc:Publisher>
<dc:Contributor>上海人民出版社</dc: Contributor>
<dc:Date>1995-01-12</dc:Date>
<dc:Type>大型电子期刊</dc:Type>
<dc:Fomat>电子图书 (eBook)</dc:Fomat>
<dc:Identifier scheme="url">http://www.chisa.edu.cn/</dc:Identifier>
<dc:Language>chi </dc:Language>
<dc:Relation>清华万博网络技术有限公司</dc:Relation>
<dc:Coverage>中国 当代学者</dc:Coverage>
<dc:Righits> 神州学人编辑部 </dc:Righits>
</BODY>
</Bibliography>
```

5 结束语

都柏林核心研究是一项跨国家、跨学科的研究活动，根本目的在于促进网络资源发现，在国外已形成了一定的规模。都柏林核心研究由OCLC(美国联机图书馆中心)主要负责，并吸引了全世界计算机、网络、数字图书馆界等众多专家的参预。在我国也将有大范围的应用，期望有更多的成果出现。■

参考文献

1 *Dublin Core Metadata Element Set, Version 1.1: Reference Description http://dublincore.org/documents/1999/07/02/dces/。*

2 *Using Dublin Core http://dublincore.org/documents/2001/04/12/usageguide/。*

3 *Web编程指南，Marty Hall著，北京：清华大学出版社，1999.4。*