Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

国内外科技信息资源元数据框架比对研究*

□ 张英杰 彭洁 / 中国科学技术信息研究所 北京 100038

摘要：文章针对大数据时代的科技信息资源集成应用新需求，从Warwick框架和新加坡框架入手，概要介绍了国内外十余种科技信息资源元数据框架，并分别从框架设计范式、框架应用对象、框架应用阶段等维度，总结了各类元数据框架的发展特点和趋势。

关键词：科技信息资源，元数据框架，比对研究
DOI：10.3772/j.issn.1673—2286.2013.03.008

1 引言

伴随着互联网的发展，科技信息资源量呈指数增长，种类也日益丰富，这不仅表现为文本、语音、图像，更为熟知的还是体现为各类学术论文、图书、科学数据、科技计划项目、科技人才、科技成果、科技报告等各种信息。在这个以“大数据”为主要特征的时代，一方面科技信息资源粒度被切分得更小，另一方面科学活动所应用的科学数据集却呈快速膨胀的趋势。

元数据管理作为一种应对大数据的方式，它的产生和应用可以为科技信息资源的组织开发提供手段，更好地揭示科技资源特征内容，进而管理和利用资源。元数据框架不仅仅是将数据以表、字段的方式管理，更要为数据系统中描述一个基本的结构组织或纲要，提供一些事先定义好的子系统，给出把它们组织在一起的法则和指南[1]。

2 基本概念

元数据是“关于（数字）对象的（结构化的）信息”或者“与信息对象有关的结构化的信息”。框架是一种相对固定的设计模式，它是针对某种特定目标系统的具有体系性的、普遍性的问题而提供的通用的解决方案，架构往往是对复杂形态的一种共性的体系抽象。一个好的框架可以让数据管理员专注于业务逻辑的实现，并且把整个系统分成若干相互独立的层次，减少了构件的耦合性。

元数据框架（Metadata Framework）规范了设计特定资源的元数据标准时需遵循的规则和方法，它是抽象化的元数据，从更高层次上规定了元数据的功能、数据结构、格式设计、方法语义、语法规则等多方面的内容[1]。而在BI架构中，元数据框架定义涉及分析元数据的当前状态、处理过程，并为元数据管理系统提供一个开发蓝图，它从长远目标、具体目的和高层需求三个方面来描述[2]。一个健全完善的元数据框架应该包含以下五个特点：（1）可整合；（2）可扩展；（3）健壮性；（4）可定制；（5）开放性。

3 国内外科技信息资源元数据框架发展现状

最初，元数据的使用是与数据管理系统目录或知

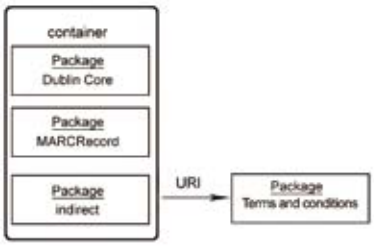
* 本文系中国科学技术信息研究所预研项目“科技信息资源集成应用元数据框架研究”（编号：YY—201120）成果之一。

识库的模式描述相关联的，主要提供有关存储数据和业务流程的相关信息。近来，人们已经普遍认识到，元数据已经成为促进异构数字信息整合和交换的主要因素。许多在研项目也说明，某些独特的元数据标准或格式不太可能普遍应用于所有资源，因此需要一个更高级的容器体系架构，能够适应各种已经在使用的不同元数据标准，建立兼容不同应用纲要的通用框架[3]。

3.1 国外科技信息资源元数据框架发展现状

（1）Warwick框架和新加坡框架

1996年，Warwick工作组提出了Warwick框架作为元数据应用的一般“容器（container）”和概念框架。该框架有两个基本组件，其中“容器”是一个整合多类型元数据集的基本单元，而这些元数据集也称作“包（package）”。“容器”可以是暂时的，也可以是永久的。每个“包”都是一种类型化的对象，包括三种形式：元数据集、间接参考和自身就是容器的包[4]。Warwick框架限制了DC本身的应用范围，局限于资源的发现。后期，为了更好地描述数字对象中包之间的关系，Warwick框架引入了一个新组件——目录（Catalog），用于具体说明“包”是数据还是元数据。



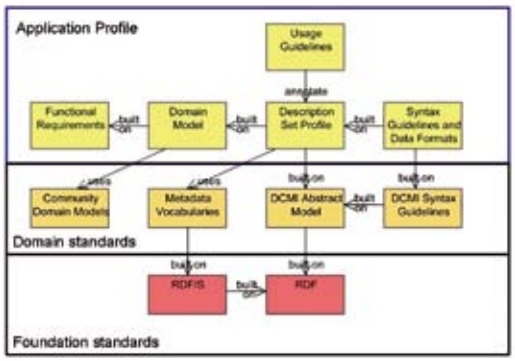


图1 Warwick框架和新加坡框架图

与此不同，新加坡框架（Singapore Framework）是设计元数据应用的框架，以帮助设计的元数据能获得最大程度的互操作性，同时通过这一框架使这样的设计文档尽可能被重用。新加坡框架定义了一系列描述组件，描述了这些描述组件所依据的标准规范与标准应用的领域模型及语义万维网的基础标准之间的关系，这些描述组件对于一个应用纲要来说，或是必须的或是有用的。新加坡框架形成了一个评估文档完整性和Web架构原则相符性的应用纲要[5,6]。

（2）Indecs元数据框架

Indecs[7]是原欧共体资助的欧共体信息2000计划的部分研究成果，代表了1998-2000年间的音乐界、知识产权界、文本出版界、作者、图书馆和其他领域的共同利益，现在已经被广泛应用到后续的各类元数据活动中。它是一个增强电子商业系统数据互操作性的元数据方案，它为实体勾画了三个既各自不同又相互重叠的基本视图：一般视图、商业视图与知识产权视图。一般视图适用于通用的环境，在此环境下实体分成三个基本大类：知觉对象（percepts）、概念（concepts）、关系（relations）。知觉对象是人的感官感知的客观事

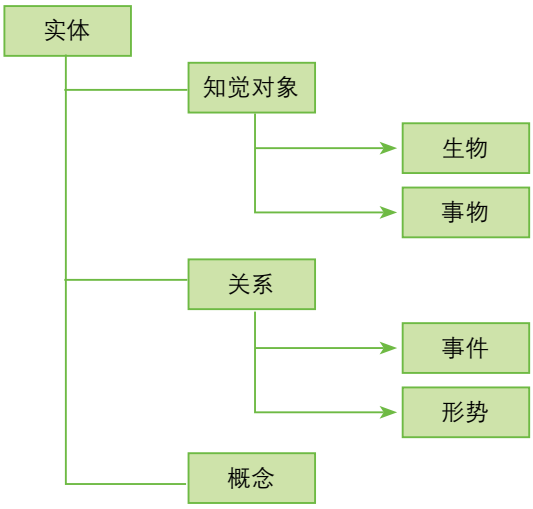


图2 Indecs元数据框架的一般视图

物；概念则是人对感知的事物进行概括、抽象的思维形式；关系反映知觉对象、概念之间的联系。知觉对象进一步划分为有生命的生物（beings）和无生命的事物（things）；关系可分为动态的事件（events）和静态的形势（situations）。

Indecs框架强调关系的重要性，它是Indecs框架分析的核心。Indecs框架没有预设任何具体的商业模型或法律框架，它可以用来描述知识产权交易、开放资源或免费存在的素材。该框架现状已经进一步发展成为一个通用本体方法，可以描述特定类型的实体和属性，以及在情景模型结构中联系它们的关联者。

（3）PREMIS元数据框架

2003年，OCLC和美国研究图书馆协会（RLG）联合开发了PREMIS数字资源长期保存元数据框架。该框架遵循了以下三个原则：（1）综合性，保存元数据应该涵盖数字保存过程的各个方面；（2）结构化，保存元数据框架应该是对数字保存系统主要功能组件/进程的高度补充；（3）普适性，保存元数据框架应该适用于各类数字对象、数字保存活动和机构。

PREMIS把涉及数字资源长期保存的实体分成五类：对象（objects）、知识实体（intellectual entities）、事件（events）、代理（agents）、权限（rights）。对象是以数字形态存在的信息独立单元；知识实体是一种可合理描述的内容聚集体单元；事件是保存系统中至少涉及一个对象或代理的行为；代理是在对象生命周期内，保存系统中的一个人、组织或软件程序；权限，也叫权限声明，是特定对象或代理的一项或多项权限或许可[8]。

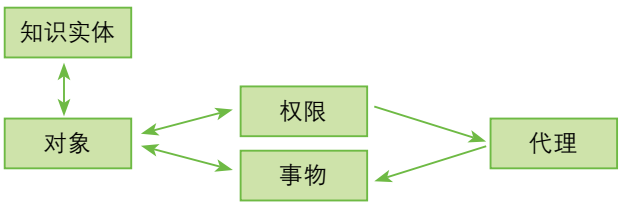


图3 PREMIS元数据框架视图

（4）统计数据通用元数据框架

统计数据通用元数据库框架起始于2004年2月由联合国欧洲经济委员会（UNECE）、欧盟统计局（Eurostat）、经济合作与发展组织（OECD）共同发起的统计元数据工作会，与会各方认为多年来，已经提出了多种多样的统计元数据模型、定义和概念，但依然缺少一个可以帮助各个统计机构建设元数据系统的通用框架。通用元数据框架是国际统计组织和各国统计机构共同努力的结果，由METIS指导小组和UNECE秘书处负责协调[9]。

2011年2月，METIS指导小组批准了通用元数据框架的维护战略，4月，批准了一份元数据框架的推广战略[10]。

通用元数据框架包括四部分，分别侧重于统计元数据系统的不同实践和理论维度。

部分A——公司情境下的统计元数据：描述了统计元数据系统项目中的管理和治理事宜。

部分B——元数据概念、标准、模型和注册：强化了定义良好的标准化术语的重要性。提供了相关概念、信息标准和模型的关联信息。

部分C——元数据和统计业务过程：提供了信息、最佳实践和其他材料，用以帮助统计组织中的元数据开发人员设计、开发一套满足业务需求的统计信息系统。

部分D——实施：关注相关国际和国家组织最近在实施或重构统计元数据系统的实践经验。

（5）电子文档元数据架构

Moura等人提出了一个描述网络电子文档的元数据架构[11]。在该架构的概念模型中，电子文档被表示为一种相互联系的层级。每个层级水平，应该有不同的元数据类型描述和展示该文档。

Collection层表示信息资源，如一份文档。知识产权层（Intellectual Content）表示某人或某组织对某一主体对象的想法和意见，文章的标题可以是知识产

图4 网络电子文档的元数据架构

权层的内容。内容表现层（Content Expression）规定了文章内容是如何组织的，文档内容表现层可根据内容进行分类，如技术报告、论文、手册和报纸等。物理层（Physical Embodiment）则用来展现内容传播的方式，如论文的格式就是一种物理体现。结构组件层（Structural Components）是最下面的一层，表示物理层分解后的形式，如一份Word文档，可以分解为页、段落和句子等。

上述元数据架构都强调数据描述和数据整合。Warwick框架主要集成多种元数据集。Moura的电子文档元数据框架可以从不同信息类目描述电子文件集和关联，并提供了将大量描述信息与DC元数据及其他元数据进行整合的一套方案。同时，我们也应该看到这些元数据框架在指导项目实践中的一些不足之处，如网络用户并不是信息专家，如何将用户所要表达的意思与元数据描述的数据进行匹配；管理用户在管理数据时，如何更好地跟踪特定数据的原始出处等等。

3.2 国内科技信息资源元数据框架现状

国内对元数据框架的研究起步较晚，主要受益于国内数字图书馆建设项目。如2000年6月的“中文元数据标准研究及其示范数据库”项目，2002年国家科技部重大基础课题“我国数字图书馆标准规范建设”和2006年的二期工程；2008年中国国家图书馆的“元数据总则”项目；2009年，中国国家图书馆的“古籍、拓片、舆图专门元数据规范”项目，迄今仍在不断研究和发展中[12]。

在各类项目实践中，2001年，肖珑等提出了中文元数据标准框架[13]，张晓林从Metadata生命周期的角度，提出了基于生命周期建立的Metadata开发应用框架[14]。2003年，清华大学图书馆牛金芳等人提出了清华大学图书馆保存元数据框架[15]。2008年，上海交通大学图书馆的王绍平根据本馆管理需求，结合已有的相关标准，提出了信息资源基础管理性元数据框架[16]。在国防科技信息领域，也提出了包括基本元数据、资源元数据、专用元数据的国防科技信息元数据标准框架[17]。

上述研究主要是在图书馆领域，围绕数字图书馆建设过程中的信息资源保存和管理开展的框架研究，资源对象主要是各类科技文献。与此同时，针对非文献类信息资源的元数据框架也开始蓬勃发展，如针对中国科学院科学数据库群数目庞大、内容丰富、学科领域广泛、系统异构和位置分散等特点和难点，中科院科学数据库中心主导制定了中国科学院科学数据库元数据框架，为科学数据库各学科数据集资源建立了统一的描述框架和通用的描述元素集合[18]。此外，在海洋数据[19]、测绘数据[20]等方面也都提出了各自的元数据应用框架。

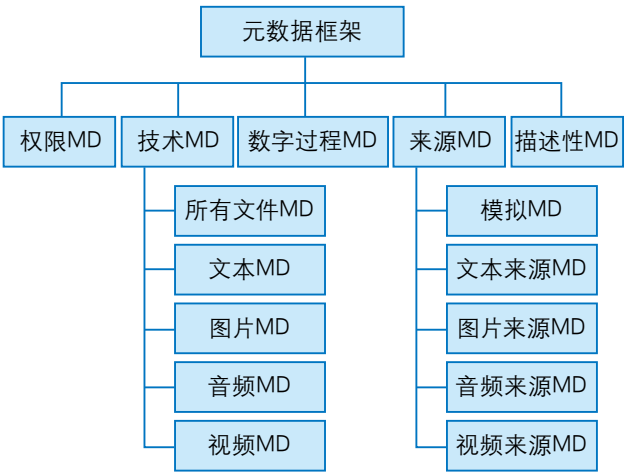


图5 清华大学图书馆元数据框架[15,21]

（1）清华大学图书馆元数据框架

清华大学图书馆元数据框架按资源的发展阶段组织元数据，反映了资源的发展过程。对于数字化的资源，它记录了数字化前模拟资源的特征、数字化过程以及当前数字资源的特征。对于born-digital的资源，它记录来源对象的特征和当前资源的特征。整个元数据框架包括5个部分：描述性元数据、技术元数据、权限管理元数据、来源元数据和数字化过程元数据。

（2）信息资源基础管理性元数据框架

信息资源基础管理性元数据框架基于信息资源的生命周期，是一个从数字图书馆中信息资源的全局，把握数字图书馆各个层面、各个运行阶段的管理机制，具有开放性、可扩展的元数据框架。它借鉴了Indecs和PREMIS，提炼了代理、信息资源、事件、权限等四个基本元素[16]，具体的框架如图6所示。

（3）科学数据库多媒体元数据框架

科学数据库多媒体元数据框架的顶层是科学数据库多媒体核心元数据，该核心元数据标准是发展后续具体类型多媒体资源、具体学科多媒体资源元数据的基础，是下面两层元数据标准的父级规范，后续所有的规范都是对该规范的扩展和应用；中间层是具体类型多媒体资源的核心元数据，是在多媒体核心元数据的基础上，参照国内外各类型多媒体资源元数据标准后，

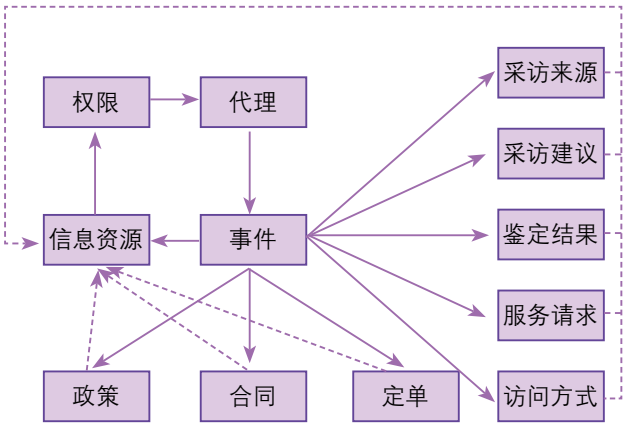


图6 信息资源基础管理性元数据框架

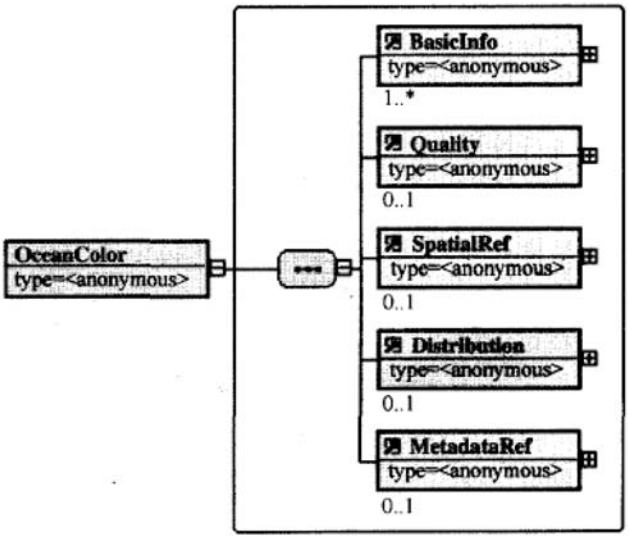


图8 海洋水色遥感元数据框架[19]

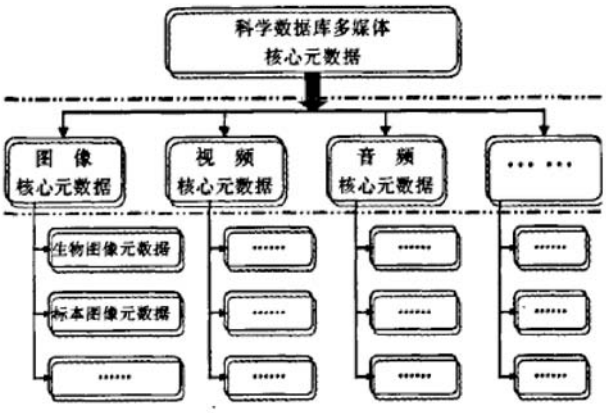


图7 多媒体元数据框架[18]

根据该类型资源的特点制定的；第三层是具体学科、类型多媒体资源元数据，是面向具体学科、专业、门类的具体需求和应用制定的，支持实现多媒体资源的细粒度描述，彼此之间可以根据需要确定互操作的基础。

该框架的设计充分考虑了现有资源的特点与现有标准之间的兼容，以及最大程度支持互操作。该框架是发展后续元数据标准的基础、原则和约束，后续标准的制定应充分体现该框架的优点，而基于该框架发展的元数据标准也一定具有强大的生命力。

（4）海洋水色遥感元数据框架

海洋水色遥感元数据框架的设计能够有效地解决海量海洋水色遥感数据的描述、组织、存储和管理问题，从数据基本信息、数据质量信息、数据空间参照信息、数据分发信息、元数据参考信息和数据结构信息6个方面进行描述，从海洋水色遥感元数据的产生、修

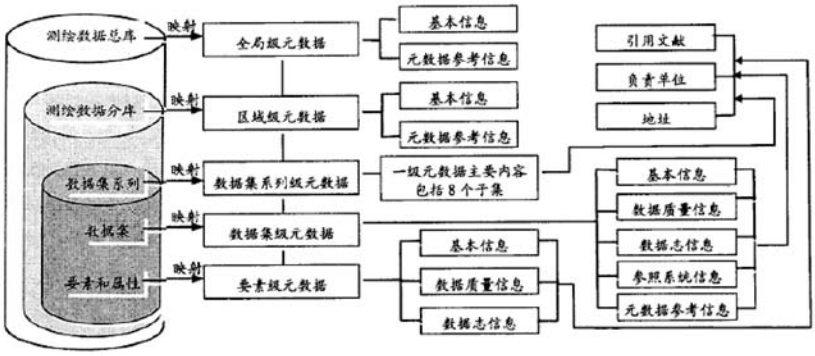


图9 测绘元数据框架[20]

改、发布等全过程进行描述。整个海洋水色遥感元数据框架由六大部分组成，其对应的XML Schema的根元素为OceanColor，对应的六大部分分别是BasicInfo、Quality、SpatialRef、Distribution和MetadataRef[19]。

（5）测绘元数据框架

测绘元数据框架把测绘数字产品按数据集系列、数据集、要素来划分，根据元数据描述对象的不同和测绘数据的组织结构特点，以地理单元为主线，设计了一种基于网络环境的多层次空间元数据框架，把测绘元数据分为5层。第1层为全局级元数据（Global Metadata），是对测绘数据总库中所有数据集的一个抽象映射；第2层为区域级元数据（District Metadata），是对测绘数据分库信息的总体描述；第3层为数据集系列级元数据（Dataset Series Metadata），是对一个具体空间数据库信息的总体描述；第4层为数据集级元数据（Dataset Metadata），这里的数据集是指一个具体空间数据库的逻辑组成部分；第5层为要素级元数据（Element Metadata），是描述数据集中各要素层数据特征的元数据[20]。

4 结语

通过综述国内外科技信息资源元数据框架概况，我们可以发现：

（1）从框架设计范式来看，各类元数据框架已经从传统的事物-属性-属性值的过程范式走向了面向对象的E-R设计，并正在向面向关系的R-R设计范式转变；

（2）从框架应用对象来看，从Warwrick框架所针对的类文本网络资源，到PREMIS的保存对象以及后续的统计数据、多媒体数据、科学数据和GIS数据等，应用对象变得日益广泛，元数据框架正在从抽象向泛化应用发展；

（3）从框架应用阶段来看，各类元数据框架正在从最初的描述性、保存性框架，向主题性、关系型、网络性框架发展，特别是伴随着语义网、BI、物联网等技术的发展，元数据框架已经逐渐覆盖了数字资源对象的全生命周期，不同框架也正在迈向交互兼容的新阶段。

总之，在元数据框架设计应用的过程中，大部分元数据框架的开发与应用过程，都采用跨部门、跨行业、跨国度的合作机制，最典型的就是新加坡元数据应用框架和Indecs元数据框架。同时，各类元数据框架在面向整合异构资源类型时坚持模块化、可扩展性和互操作性的基本原则，保持了不同元数据标准的复用和兼容。最后，面向即将到来的大数据e-Science时代，服务于科技信息资源集成的元数据框架也正在孕育着新的思路、新的突破。

参考文献

[1] 肖珑,陈凌,冯项云,等.中文元数据标准框架及其应用[J].大学图书馆学报,2001,19(5):29-35.
[2] DMBOK：元数据管理[EB/OL].[2012-08-20]. http://www.cnblogs.com/zhoujg/archive/2011/12/26/2301661.html.
[3] DENGY. The Metadata Architecture for Data Management in Web-based Choropleth Maps [EB/OL].[2012-06-20]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.7598.
[4] LAGOZE C. The Warwick Framework-A Container Architecture for Diverse Sets of Metadata[EB/OL].[2012-05-20]. http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.
[5] DCMI Specifications[EB/OL].[2012-05-20]. http://dublincore.org/specifications/.
[6] NILSSON M. DCMI Description Set Profile Model [R]. Working Draft, December 2007.
[7] Indecs Content Model[EB/OL].[2012-05-20].http://en.wikipedia.org/wiki/Indecs_Content_Model.
[8] Preservation Metadata for Digital Objects [EB/OL].[2012-05-20]. http://www.oclc.org/resources/research/activities/pmwg/presmeta_wp.pdf.
[9] A Strategy for Promoting the Common Metadata Framework-UNECE[EB/OL].[2012-05-20]. www1.unece.org/.
[10] Statistical Metadata in a Corporate Context: A guide for manager[EB/OL].[2012-03-20]. http://www.unece.org/fileadmin/DAM/stats/publications/CMF_PartA.pdf.
[11] DE CARVALHOMOURAA M, CAMPOSM L M,MARIABARRETO C. A Metadata Architecture to Represent Electronic Documents on the Web [C]// 3rdIEEE Metadata Conference NIH Campus, Bethesda, Maryland, U.S.A, April 1999.
[12] 中文元数据标准研究及其示范数据库[DB/OL]. [2012-05-20]. http://www.nlc.gov.cn/newgjgx/hzxm/201110/t20111031_54119.htm.
[13] 肖珑,陈凌,冯项云,等.中文元数据标准框架及其应用[J].大学图书馆学报,2001(05):29-35,91.
[14] 张晓林.元数据开发应用的标准化框架[J].现代图书情报技术,2001(2):9-11,15.
[15] 牛金芳,郑小惠,曾婷,等.清华大学图书馆保存元数据方案[J].大学图书馆学报,2003,21(2):22-25.
[16] 王绍平,郑巧英,孙华,等.信息资源基础管理性元数据框架的数据模型[J].情报杂志,2008,27(3):93-95,98.
[17] 梁珂.国防科技信息元数据设计研究[C]//第二十三届全国计算机信息管理学术研讨会论文集,2009:173-178.
[18] 陈峰莲,阎保平,黎建辉,等.科学数据库多媒体元数据框架[C]//中国科学院研究生院“21世纪计算机科学与技术”第八届研究生学术研讨会论文集,2004:98.
[19] 李学荣,李莎.海洋水色遥感元数据及其系统设计[J].热带海洋学报,2007(01):81-86.
[20] 付李红,周会群,徐士进,等.测绘元数据的框架设计和基于RDF/XML的表达[J].计算机辅助工程,2005(01):21-25.
[21] NIU J. A Metadata Framework Developed at the Tsinghua University Library to Aid in the Preservation of Digital Resources[EB/OL].[2012-03-20].http://www.dlib.org/dlib/november02/niu/11niu.html.

作者简介

张英杰（1979-），男，博士研究生，主要研究方向：未来技术分析，科技信息资源元数据。E-mail: zhangyj@istic.ac.cn

Comparison Study of S&T Information Resource Metadata Framework at Home and Abroad

Zhang Yingjie, Peng Jie / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Aiming at S&T information resource integration application demands in the big data era, it introduces more than ten kinds of S&T information resource metadata framework starting from Warwick framework and Singapore framework. Moreover, it summarizes the development characteristics and future trend of these metadata framework from the aspect of framework design pattern, framework application target and framework application stages.

Keywords: S&T information resource, Metadata framework, Comparison study

（收稿日期：2012-10-15）