

多实体的国家科技重大专项元数据框架研究

刘春燕¹ 刘 敏¹ 安小米²

(1. 中国科学技术信息研究所, 北京 100038; 2. 中国人民大学信息资源管理学院, 北京 100872)

摘要: 简述国家科技重大专项项目元数据特征, 分析国家科技重大专项元数据框架的构建原则、框架内容及要素、框架构建模型, 探讨国家科技重大专项元数据框架的多实体领域模型、多实体国家科技重大专项元数据元素集设计、国家科技重大专项应用三大核心要素, 总结国家科技重大专项项目多实体元数据框架的科学性和适用性。

关键词: 国家科技重大专项; 元数据; 元数据实体; 多实体; 元数据框架

中图分类号: G250

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2021.01.001

Metadata Framework Research of National Science and Technology Major Project Based on Multi Metadata Entity

LIU Chunyan¹, LIU Min¹, AN Xiaomi²

(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. School of Information Resource Management of Renmin University, Beijing 100872)

Abstract: This study briefly describes the National Science and Technology Major Project (NMP) metadata characteristic, on the basis of the demand analysis of the construction principle of NMP metadata framework, content and elements of NMP metadata framework, building model of NMP metadata framework, it lays special stress on the three core elements of NMP metadata framework: multi-entity metadata domain model, metadata element set design for NMP. Finally it summarizes the scientificity and applicability of multi metadata entity framework design on NMP metadata.

Keywords: National Science and Technology Major Project (NMP), metadata, metadata entity, multi-entity, metadata framework

0 引言

2006 年颁布的《国家中长期科学和技术发展规划纲要(2006—2020 年)》^[1]确定了国家科技重大专项的实施。国家科技重大专项^[2]聚焦国家重大战略产品和产业化目标, 为实现国家目标, 通过核心技术突破和资源集成, 在一定时限内完

成重大战略产品、关键共性技术和重大工程, 具有研究规模大、管理内容复杂、组织形式多样、人员分工专业化、利益相关者众多、成果类型多等特点, 是典型的“跨学科、跨领域、跨机构”大科学项目。国家科技重大专项过程性数字化数据和结论性成果性数据体现了国家重大科技历史, 价值极大, 需要有效地梳理、共享和保存。

作者简介: 刘春燕(1974—), 女, 中国科学技术信息研究所副研究馆员, 主要研究方向为信息与文献标准化、信息资源管理(通信作者); 刘敏(1980—), 女, 中国科学技术信息研究所助理研究员, 主要研究方向为信息素质、科技资源; 安小米(1965—), 女, 中国人民大学教授, 主要研究方向为知识管理。

基金项目: 2013 年度国家社会科学基金项目“面向共享的科技计划项目元数据框架研究”(13BTQ018)。

收稿时间: 2020 年 9 月 16 日。

作为描述科技重大专项项目背景、业务流程及成果等多层次对象的结构化描述语言和工具,在E-science环境下,国家科技重大专项元数据既是促进国家科技重大专项全生命周期产生、累积和共享信息和知识资源的重要技术方法和手段,也是以全新理念、方法和手段管理国家科技重大专项时组织战略层面的优先选择。但是,结合国家科技重大专项项目特点,采用平面的元数据元素集却难以满足项目过程描述需求和共享功能。

实体(entity)是指任何存在、业已存在或即将存在的有形或无形的事物,包括这些事物间的关联^[3]。在元数据领域,元数据实体是指具有某种共同特征的一组元数据集合,ISO 23081^[4]采用文件元数据领域中的文件、人员、业务和法规要求四类实体及其相互关系的元数据概念模型揭示文件管理生态环境中的元数据概念体系框架。在科研领域,欧盟向其成员国推荐的常用欧洲研究信息格式CERIF^[5]通过项目、人、组织等元数据实体显示研究对象、研究活动、研究结果及相互关系。英国科研委员会中心实验室(Council for the Central Laboratory of the Research Councils, CCLRC)的科学元数据模型^[6]构建了包括政策(policy)、项目(programme)、研究(study)和调查(investigation)(包括实验、测量、模拟不同类别)等实体的多层级科学活动数据模型。

为了满足国家科技重大专项过程描述需求和实现共享功能,本文拟采用元数据框架设计思路,利用多实体元数据实现国家科技重大专项元数据领域模型描述,从宏观层次规范元数据功能、数据结构、格式、语义、语法等内容,并在此基础上借鉴DC元数据^[7]等通用描述元数据从微观层次对国家科技重大专项元数据元素进行定义和描述。

1 元数据框架构建原则及其内容要素

1.1 构建原则

国家科技重大专项元数据作为描述国家科技重大专项背景、业务流程及成果的数据,具有如下特征:一是国家科技重大专项元数据本身也

是一种重要科技信息资源,需要采取技术手段和组织保障对其进行描述、组织和共享;二是国家科技重大专项元数据以科技重大专项资源共享作为需求出发点和应用目的,其元数据构建以资源共享为中心展开;三是国家科技重大专项元数据设计满足全面描述重大专项全生命周期的重点要素,具有丰富的语义互操作等功能要求。根据上述国家科技重大专项元数据特征,其元数据框架构建应遵守如下原则。

(1)以需求为导向。元数据需求分析是确保满足用户需求,实现设计功能达到预期目标的重要前提。张茜^[8]认为,元数据需求分析方法有演绎法、归纳法、分析法及功能结构法。朱淑丽^[9]认为,数字化档案管理元数据需求分析的信息来源范围包括权威性文献、数字化档案及其用户等。与之类似,国家科技重大专项元数据需求分析也应结合元数据功能目标,确定需求分析的信息来源范围,从用户、国家科技重大专项相关政策规范等权威性文件、国家科技重大专项项目活动过程3个方面综合分析元数据需求。

(2)借用国内外相关元数据研究成果。为确保共享和互操作,国家科技重大专项元数据元素应尽量复用已有元数据元素,如DC核心元数据等。正如Kieron Niven等^[10]认为的那样,项目、资源、文件等不同对象的一般元数据或特定元数据,从根本上来说,都将提供关于项目或文件的“who, what, when, where and how”等信息。因此,国家科技重大专项元数据元素也应重点揭示这些信息。

(3)以多实体元数据框架模型构建为中心。可采用需求分析方法、流程分析方法、调查法、文献保证法、标准化法、复杂系统方法、设计科学研究方法、书目记录功能分析法等方法或其组合的方式构建元数据框架。国家科技重大专项元数据框架可以采用业务流程分析法中的元数据实体—关系法,即借鉴信息系统设计时常用的需求分析结构化方法,确定在不同视角下、不同生命周期业务过程中的国家科技重大专项元数据实体、关系及其属性,建立元数据实体—关系

(E-R) 模型。

(4) 科技重大专项元数据术语来源于科技重大专项相关政策法规、标准规范、文献资源等。元数据术语对理解元数据元素的含义具有重要意义。一般包括标签、定义和注释 3 个部分。DC 维护机构 DCMI 维护的元数据术语集中了每一个术语包括以下部分或全部的属性：名称（赋予数据元素的唯一标记）、URI（用于唯一标识该术语的统一资源标识符）、标签（分配给术语的标签）、定义（对术语概念和性质的明确说明）、术语类型（如元素、编码体系等）、状态（由 DCMI 应用委员会分配给术语的状态）、发布日期（术语第一次公布的日期）。科技重大专项元数据术语集应遵循从科技重大专项相关政策法规、标准规范、文献资源等收集整理的原则，以确保科技重大专项元数据框架的应用一致性和应用长期有效性。

1.2 内容及要素

刘春燕^[1]认为，科技计划项目元数据框架是为跨系统、跨平台和跨组织的元数据资源共享而确定的元数据构建原则和方法，其目的是从宏观层次规范元数据功能、数据结构、格式、语义、语法等内容，从微观层次为元数据的定义、描述、发布和维护等组成部分提供连贯一致的指南。2007 年，DCMI 制定了实现最大化互操作和最大可用性的 DCMI 新加坡框架^[12]。借鉴 DC 新加坡框架，国家科技重大专项元数据框架也应包

含功能需求、领域模型（多实体元数据框架）、元数据描述集、元数据术语集、使用指南、信息系统、评价指标体系等要素。其框架模型见图 1。

2 元数据框架构建模型研究

国家科技重大专项元数据框架构建思路可从科技重大专项项目全生命周期视角分析梳理不同科技计划项目的研究成果及项目决策的运动流程，明晰重大专项业务活动中的研究成果管理需求和流程，按照用户需求映射到元数据，元数据管理整合到信息流，信息流进一步整合到业务流的可逆需求，构建捕获基于决策共享和成果共享的、支持各种类型的、覆盖科技重大专项全生命周期的国家科技重大专项理想元数据框架。此外，分析梳理当前各科技计划政策法规、标准规范、项目申报书等计划项目信息，建立当前科技计划项目元数据实体概念模型。在此基础上，提出我国科技重大专项元数据改进措施。

国家科技重大专项元数据框架构建过程主要是对图 1 中的功能需求、领域模型、元数据描述集、元数据语法及数据格式等模块具体化的过程。其构建模型如图 2 所示。

图 2 中的领域模型及元数据描述集是国家科技重大专项元数据框架的核心模块，其他模块如术语集主要利于元数据概念理解和互操作，元数据管理系统主要提供元数据功能实现的数字应用环境，使用指南主要提供元数据实际应用的可操

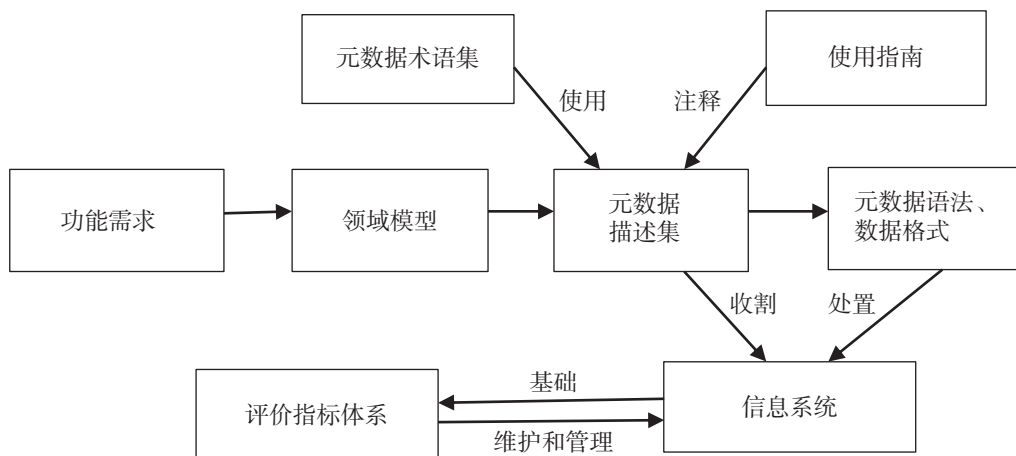


图 1 多实体国家科技重大专项元数据框架

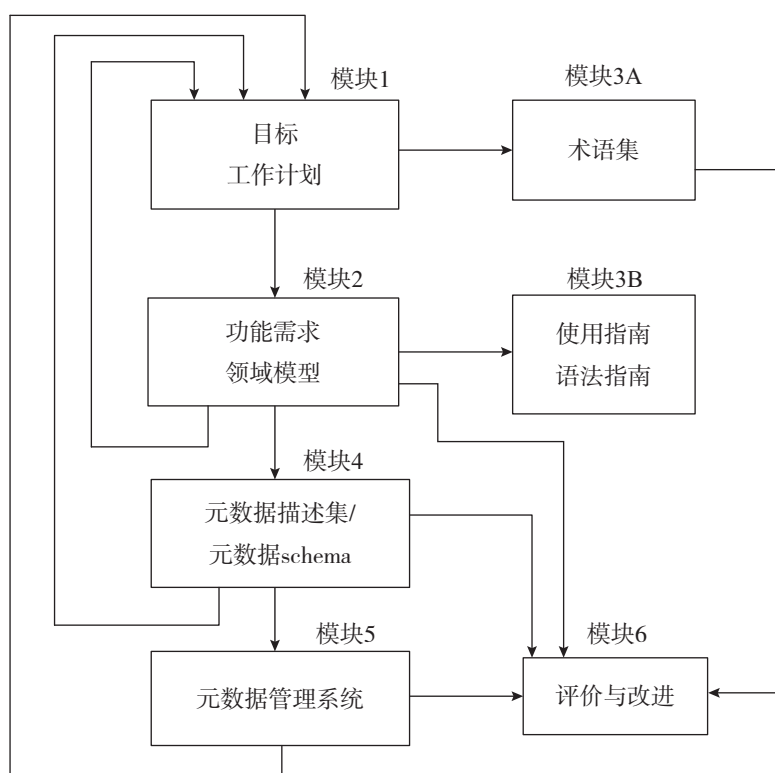


图2 国家科技重大专项元数据框架构建模型

作性指南。图2中的各阶段模块之间可以迭代，并相互呼应，互相借鉴，形成一个统一的整体。

3 多实体的元数据领域模型

国家科技重大专项管理过程中相继提出了成果转化和产业化、科研信用、绩效评估、资金管理等多方面的要求。为满足国家科技重大专项的上述多种管理功能和信息共享功能的需求，本研究中的国家科技重大专项元数据领域模型采用多实体元数据框架模型。

为避免Ann Asserson等^[13]指出的通用欧洲研究信息格式元数据模型1991版（Common European Research Information Format 1991, CERIF 91）以单一实体为中心出现的各种应用问题，国家科技重大专项元数据实体应采用通用欧洲研究信息格式元数据模型2000版（Common European Research Information Format 2000, CERIF 2000）的设计思想，设置实体并通过标识符、时间等元素进行实体间的关联。由于国家科技重大专项项目周期长、管理层级复杂、成果类型多样化，只

采用CERIF 2000项目、责任者、组织机构3个基本实体（project, person, organizational unit）不能完全满足元数据设计的功能需求。因此，应借鉴通用欧洲研究信息格式元数据模型1.6版（Common European Research Information Format 1.6 Version, CERIF 1.6）^[14]的多实体及实体结构化（基本实体、成果实体、2级实体、链接实体等）设计思想，开展国家科技重大专项元数据实体类型设计。

参照ISO 23081-2^[15]的多元数据实体模式，结合国家科技重大专项项目全生命周期活动及信息共享要求，国家科技重大专项元数据实体可划分为基本实体、成果实体、保障实体、评估实体及链接实体五大类，其类型及含义见表1。国家科技重大元数据实体框架模型见图3。

从图3可以看出，国家科技重大专项元数据包括基本、成果、保障、评估、链接共五大实体（用编号①至编号⑤分别在图3中标注），基本涵盖了表1中的国家科技重大专项元数据功能需求。图3中的链接实体包括实体内部链接（小圆

表 1 国家科技重大专项元数据实体类型

中文名称	英文名称	实体内涵
基本实体	Basics entity	以国家科技重大专项的项目、责任者、机构、信息（文件、档案）基本要素为描述对象的元数据集合
成果实体	Results entity	描述国家科技重大专项的专利、产品、出版物及相关奖励等战略性成果的元数据集合
保障实体	Supporting entity	描述规范和保障国家科技重大专项顺利开展的政策法规、经费、活动过程的服务及事件的元数据集合
评估实体	Evaluation entity	描述国家科技重大专项风险、综合绩效及信用等评估类元数据集合
链接实体	Link entity	描述各实体之间的关联关系的逻辑类实体

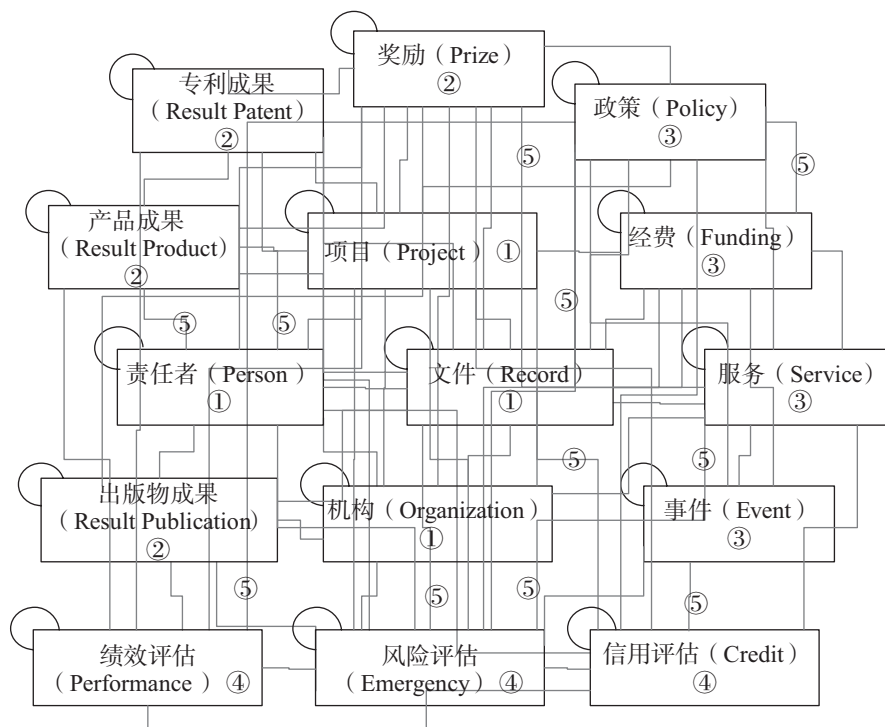


图 3 国家科技重大专项元数据实体框架

弧部分)和实体间链接(折线部分)两种链接类型。其中,实体内部链接可以表示为nmpProject_Project(项目与项目间链接)、nmpEvent_Event(事件与事件间链接)等,实体间链接可表示为nmpResultProduct_Person(专利成员与责任者间链接)、nmpFunding_Record(经费与文件间链接)等,这种柔性的链接方式能最大程度地体现国家科技重大专项层级复杂、项目元数据功能需求众多的特点。

与平面的元数据方案相比,国家科技重大专项多实体元数据框架采用图2所示的多实体领域模型,具有模块化、关联性、可扩展等优点。关于模块化:图3所示的国家科技重大专项元数

据实体都是按照同样的原则分析功能需求而构建的,每个元数据实体都可按照DC元数据元素的方式开展具体元数据设置,并可以按照功能需求进行增减,如修改具体元数据元素层级则不会对整个元数据框架产生影响。关于关联性:国家科技重大专项元数据实体间通过链接实体发生关联,如责任者实体和项目实体,通过责任者在项目中的角色、责任者参加项目的开始时间和结束时间发生关联(图4)。关于可扩展:国家科技重大专项元数据框架采用多实体柔性化设计方案,由于通过具有语义含义的链接实体发生实体间的关联,因而可以像搭建积木一样,根据分类、角色、任期、类型等不同含义进行组合,具体实体

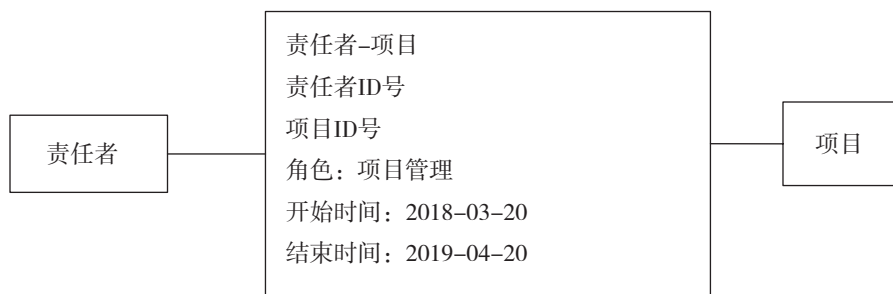


图4 关联实体链接两个实体示例

也可以根据元数据需要实现的功能进行增减。

4 多实体的元数据元数集设计研究

国家科技重大专项元数据可以有效实现对国家科技重大专项数字资源信息的描述，确保数字资源能够被准确及时识别、定位及获取。多实体国家科技重大专项元数据元数集设计时必须满足对国家科技重大专项多实体的全方位信息描述。

考虑国家科技重大专项元数据框架多实体特征，其元数集设计首先应将各个实体需要实现的元数据功能进行梳理，并对各个实体间的交叉关联进行分析。如项目实体包括经费、合作、项目团队、项目成果等元数据项，责任者包括项目成果、科学数据、项目、实践等元数据项，出版物包括作者、出版者、项目等元数据项。表2是基础实体元数据功能梳理后形成的元数据表。

表2中的元数据表各实体元数据复用了identifier（标识符）、name（名称）、time（时间）、description（简介）等都柏林核心元数据及CERIF中的通用元数据元素，便于国家科技重大专项元数据的共享和互操作，表2中元数据的描述属性包括英文小写便于计算机编码识别的元素名、便于阅读的标签、来自术语集中的定义集注释等。

表2中国家科技重大专项子元素名复合词元数据元素名采用小骆驼拼写法（lowerCamelCase）标识，即第一词的首字母大写，后面每个词的首字母大写，部分太长的英文可用缩写，如国家科技重大专项事件元数据元素名为nmpEvent、产品成果元数据元素名为nmpResultProduct、责任者出生日期元数据元素名为nmpPersBirthdy。

另外，多实体国家科技重大专项元数据集设计时，表1中的链接实体中的元数据通过起始时间和终止时间进行标记，这样即使发生某些改变，如成员变更等，也不需要变更关系，只需将新关系标记开始时间插入进去即可。这种处理方式可确保国家科技重大专项历史元数据的保存和可追溯。表3以表2中的部分链接元数据为例，对部分链接实体开展了元数据描述。

从表3中可以看出，国家科技重大专项元数据间的链接主要基于链接类型标识实现语义链接的，为有效规范链接实体类型，需要国家重大专项链接类型规范文档，对所有链接类型标识进行层级化规范，如项目间链接类型设置“项目层级”（project structure）大类，项目与责任者链接类型设置“责任者角色”（person role）大类，并在大类下列出具体的链接类型。链接实体元数据通过元数据标识符在语义层实现其功能。

5 元数据框架应用研究

国家科技重大专项元数据框架应用应重点研究如下内容：一是国家科技重大专项元数据应用推荐设计研究。元数据应用推荐是为满足元数据互操作性而建立的，通常由一个或若干个元素集组成。应用推荐通过“强烈推荐”“任选”等来描述实施过程中必备、可选等元素。应用推荐揭示了组织或机构优选哪种元素集，提供对每一个元素的指导和最优方法，对特定领域内变化的可能性作出判断。二是国家科技重大专项元数据著录指南编制。著录指南主要是为了消除元数据著录应用时的不一致、模糊性而设计的。如“国家科技重大专项管理信息系统”是所有责任者都

表 2 基础实体元数据元素表

元素	子元素名	子元素标签	定义
Project (项目)	nmpProjName	项目名称	项目的正式名称
	nmpProjID	项目标识符	用以标识项目的明确唯一的字符串
	nmpProjDescri	项目简介	项目基本信息的描述
	nmpProjURI	项目网站	项目开展活动的 URL
	nmpProjStartdate	项目开始日期	项目立项时间
	nmpProjEnddate	项目结束日期	项目结题时间
	nmpProject_Project	项目与项目间链接	项目之间的关系描述, 如子项目、子课题等
Person (责任者)	nmpPersName	责任者名称	责任者姓名信息描述
	nmpPersID	责任者标识符	用以标识责任者的明确唯一的字符串
	nmpPersBirthdy	出生日期	责任者出生日期
	nmpPersGender	性别	责任者性别
	nmpPersPhone	电话	责任者联络的电话
	nmpPersEamil	电子邮箱	责任者联络的 E-mail 地址
	nmpPerson_Person	责任者与责任者间链接	责任者在项目活动中的关系描述
	nmpPerson_Project	责任者与项目间链接	责任者在项目中的任务分工描述
	nmpPerson_Org	责任者与机构间链接	责任者在机构中的职务或岗位描述
Organization (机构)	nmpOrgName	机构名称	机构的正式名称
	nmpOrgID	机构标识符	用以标识机构的明确唯一的字符串
	nmpOrgDescr	机构简介	机构基本信息描述, 包括机构类型、发展历史等
	nmpOrgURI	机构网址	机构主页的 URL
	nmpOrg_Org	机构与机构间链接	机构与机构在项目中的角色关系描述
	nmpOrg_Project	机构与项目间链接	机构在项目中的任务分工描述
Record (文件)	nmpRecoID	文件标识符	用以标识文件的明确唯一的字符串
	nmpRecoDescr	文件简介	文件基本信息描述
	nmpRecoRight	文件权限	文件获取范围等信息描述
	nmpRecord_Record	文件与文件间链接	文件与文件间替代、关系描述
	nmpRecord_Project	文件与项目间链接	文件与相关项目的所属或关联等关系描述
	nmpRecord_Org	文件与机构间链接	文件与相关机构的发布或实施对象等关系描述
	nmpRecord_Event	文件与事件间链接	文件与相关事项产生于或规范等关系描述

表 3 国家科技重大专项链接实体元数据描述示例

链接元数据元素	实体 1 标识	实体 2 标识	链接类型标识	开始时间	结束时间
nmpProject1_Project2	Project1-ID	Project1-ID	Subitem (子项目)	2007-01-01	2010-12-31
nmpPerson1_Person2	Person1-ID	Person2-ID	Supervisor (主管)	2004-01-13	2011-11-27
nmpPerson1_Project2	Person2-ID	Project1-ID	Participant (参与人)	2007-01-01	2010-12-31
nmpOrg1_Org2	Org1-ID	Org2-ID	Part of (从属)	1990-12-01	-
nmpRecord1_Event2	Record1-ID	Event2-ID	Generation (产生)	2009-01-13	2010-11-12

认为很重要的元素, 管理信息系统能够提供关于科技计划的很多相关信息, 并具有很多功能。例如, 国家科技重大专项申报功能、国家科技重大专项结题验收功能等。但对实际科技计划进行计划级元数据著录时, 可能会遇到很难找到或需要权限获取网址等问题。

Getaneh Alemu 等^[6]认为, 元数据简单化是用户的要求, 但元数据的丰富性则是为满足大量的语义关联和越来越与内容相关的实际需要。在实际应用中, 多实体的国家科技重大专项元数据元素需要在实践中反复验证和修改, 并从多实体层面考虑为国家科技重大专项信息系统的用户界

面简洁性方面提供可操作的技术保障。

6 结语

本文研究的多实体、多层级的国家科技重大专项元数据框架适应了国家科技重大专项层级多、项目复杂等特征,为分散、异构的国家科技重大专项项目数据的数字化管理提供了有效的工具。本文研究的国家科技重大专项元数据框架融合了都柏林核心元数据框架、通用欧洲研究信息格式元数据模型CERIF等国内外元数据研究和实践成果,研究采用的语义层链接实体方案拓展了元数据框架模型构建思路,采用的具有子元素层级的国家科技重大专项元数据元素设计在实践中具有可操作性。

参考文献

- [1] 国务院. 国务院关于印发实施《国家中长期科学和技术发展规划纲要(2006—2020年)》若干配套政策的通知[EB/OL]. [2019-09-09]. http://www.gov.cn/gongbao/content/2006/content_240246.htm.
- [2] 国家科学技术部重大专项司. 国家科技重大专项[EB/OL]. [2019-09-09]. <http://www.nmp.gov.cn/>.
- [3] ISO/IEC JTC1. Information technology— Metadata registries (MDR) —Part 1: Framework: ISO/IEC 11179-1: 2004 [S]. Geneva: International Organization for Standardization, 2004.
- [4] Technical Committee ISO/TC46. Information and documentation—Records management processes—Metadata for records Part 1: Principles: ISO 23081-1: 2006 [S]. Geneva: International Organization for Standardization, 2006.
- [5] The European Organisation for International Research Information. CERIF data model[EB/OL]. [2019-09-11]. <http://www.eurocris.org/Index.php?page=CERIFreleases&t=1>.
- [6] Shoaib Sufi, Brian Mathews. CCLRC scientific metadata model: Version 2[EB/OL]. [2019-09-11]. http://www.researchgate.net/publication/30405864_CCLRC_Scientific_Metadata_Model_Version_2.
- [7] Dublin Core Metadata Initiative. Dublin core metadata element set [EB/OL]. [2019-09-15]. <https://www.dublincore.org/specifications/dublin-core/dces/>.
- [8] 张茜. 电子文件管理元数据需求分析理论研究[J]. 浙江档案, 2018(3): 36-39.
- [9] 朱淑丽. 试论数字化档案管理元数据需求分析[J]. 速读(下旬), 2016(3): 359.
- [10] NIVEN K, WATTS J. Archaeology data service/digital antiquity, guides to good practice [EB/OL]. [2014-07-25]. http://guides.archaeologydataservice.ac.uk/g2gp/CreateData_1-2.
- [11] 刘春燕. 资源共享视角下的科技计划项目元数据框架构建研究[D]. 北京: 中国人民大学, 2011.
- [12] Dublin Core Metadata Initiative. The singapore framework for dublin core application profiles[EB/OL]. [2020-09-01]. <http://dublincore.org/documents/singapore-framework/>.
- [13] ASSERSON A, JEFFERY K, LOPATENKO A. CERIF: Past, present and future: An overview[EB/OL]. [2019-09-10]. https://xueshu.baidu.com/usercenter/paper/show?paperid=de2127b6f9b236aba4dd8e240fb9c069&site=xueshu_se.
- [14] euroCRIS. CERIF-1.6[EB/OL]. [2020-09-01]. <https://www.eurocris.org/cerif/feature-tour/cerif-16>.
- [15] Technical Committee ISO/TC46. Information and documentation—Managing metadata for records—Part 2: Conceptual and implementation issues: ISO 23081-2: 2009 [S]. Geneva: International Organization for Standardization, 2009.
- [16] ALEMU G, STEVENS B, ROSS P. Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries a social constructivist approach[EB/OL]. [2019-09-10]. https://www.researchgate.net/profile/Brett_Stevens2/publication/241890605_Towards_a_conceptual_framework_for_user-driven_semantic_metadata_interoperability_in_digital_libraries_A_social_constructivist_approach/links/53e0cba90cf2d79877a500e3/Towards-a-conceptual-framework-for-user-driven-semantic-metadata-interoperability-in-digital-libraries-A-social-constructivist-approach.pdf.