

面向多源异构科技信息治理的元数据标准规范体系构建*

罗婷婷¹ 赵瑞雪^{1,2} 李娇¹ 傅智杰³ 武丽丽³ 黄永文¹ 鲜国建^{1,2}

(1. 中国农业科学院农业信息研究所, 北京 100081; 2. 农业农村部农业大数据重点实验室, 北京 100081;
3. 中国工程院战略咨询中心, 北京 100088)

摘要: 为开展大数据环境下多类型、多来源、异构化科技信息的汇聚治理, 实现大数据资源的规范化描述与互联互通, 提高数据资源的可发现、可利用和开放共享能力, 本文基于元数据理论和知识对象建模思想, 构建了一套广泛适用的、可扩展的元数据标准规范体系, 覆盖13类通用容器、24类资源元素集描述规范及28个规范编码表, 并编制了配套的XML Schema形式化描述规范, 实现对多类多源异构元数据向统一的XML格式转化、验证和解析等自动化处理。该规范体系已在中国工程科技知识中心开展了应用验证, 指导30余个分中心完成24类数据资源超过亿级数据的转化汇交, 有力支撑了工程科技“元数据海”的建设, 快捷、高效地实现了近百类专业领域特色资源元数据标准规范的定制与应用。

关键词: 多源异构; 科技信息; 元数据; 标准体系

中图分类号: G254.36 **DOI:** 10.3772/j.issn.1673-2286.2021.04.009

引文格式: 罗婷婷, 赵瑞雪, 李娇, 等. 面向多源异构科技信息治理的元数据标准规范体系构建[J]. 数字图书馆论坛, 2021 (4): 58-67.

随着大数据智能时代的到来, 大数据已成为新型战略资源、生产要素和驱动创新发展的新引擎。大数据环境下, 科技信息的类型多样, 数量以前所未有的速度增长, 且数据结构、存储方式存在很大差异, 造成数据集成困难、交互复杂, 很大程度上影响了数据服务、知识发现服务的效果, 因此数据治理应运而生。数据治理是提升数据价值的基础^[1], 旨在通过相应的标准、规范、流程和方法等, 确保数据统一管理、高效运行, 并在数据使用过程中充分发挥数据价值的过程^[2]。想要实现多源异构数据的汇聚融合、互联互通, 数据治理是必不可少的课题, 而资源的标准化, 即构建多源异构科技信息元数据标准规范体系是提升数据治理效果的首要环节, 同时也是数据治理面临的难点之一。

本文面向多源异构科技信息的数据治理需求, 在广泛参考借鉴国内外现有标准规范基础上, 兼顾人和

计算机多种场景应用需求, 研究制定了一套适用于多种类型、多源异构科技信息的元数据描述标准规范体系, 从数据资源中抽象出相应的特征属性, 组成特征元素集合(元数据元素), 来描述特定类型的资源^[3], 初步实现多源异构数据资源间关联、深层次、细粒度的规范描述, 为各领域、多类型资源的规范描述、交换复用、开放共享和关联互通提供指导和支撑。

1 国内外元数据标准规范发展现状

当前, 描述数据资源的元数据标准规范日渐增多。国外具有代表性的有: 都柏林核心元数据元素集(DC)^[4]具有较强通用性; 美国国家生物技术信息中心(NCBI)的JATS作为美国国家标准得到了广泛应用和认可^[5]; Dryad元数据标准规范^[6]为Dryad科学数

*本研究得到国家社会科学基金项目“科技论文全景式摘要知识图谱构建与应用研究”(编号: 19BTQ061)资助。

据仓储资源描述服务,实现数据与出版物的关联,被称为科学数据仓储元数据的最佳实践^[7];Dataverse^[8]元数据标准是为科学数据仓储Dataverse资源描述服务,以降低查找、获取和复用科学数据的难度;Web of Science^[9]、Scopus^[10]作为具有较大影响力的数据库,其元数据规范已成功应用于数字化生产过程。国内具有代表性的有:国家科技图书文献中心(NSTL)编制的统一文献元数据标准^[11]支持NSTL多类型文献数据资源的统一描述;国家标准《科技平台 资源核心元数据》^[12]共设计了7个元数据元素(标识符、资源名称、最新提交日期、描述、关键词、访问限制、资源信息链接地址)和2个元数据实体(提交单位、资源类别);期刊论文^[13]、电子图书^[14]、学位论文^[15]、网络资源^[16]、电子连续性资源^[17]等元数据行业标准规范也分别设计了元素、属性等用来描述、利用和规范相应的数字资源。

但上述标准规范均存在一些不足。如DC、Dryad元数据标准描述的内容相对较少,虽然增加了标准的易用性^[18],但同时也导致它们存在数据细粒度、数据关系等方面揭示不足等问题^[19]。Dataverse元数据标准虽然描述的内容较多,使数据描述的准确性与全面性得到保证,但由于元素的顺序和层次划分欠清晰,对其应用范围的扩大造成限制^[20]。Web of Science、Scopus及NSTL采用的元数据标准局限于对文献资源的规范、描述与应用。《科技平台 资源核心元数据》虽然描述对象为多源异构数据资源,但其重点是从资源名称、标识符以及资源管理角度对其进行的规范与描述,其数据元素比较简单,没有对各类资源深层次、细粒度的描述规范。期刊论文、电子图书、学位论文等元数据行业标准规范局限于某一类型资源的规范与描述。因此,在国内外元数据标准规范基础上构建一套适用于多源异构科技信息的通用元数据标准规范体系,实现多源异构科技信息数据资源的规范化描述与互联互通,提升数据治理效果,提高数据资源的可发现、可利用和开放共享能力。

2 元数据标准规范体系的构建

2.1 构建目标、原则与思路

元数据标准规范体系的构建基于元数据理论和科技信息数据资源对象化关联建模思想,该元数据标准规范体系面向人和计算机两类应用场景,适用于多类、

多源、异构数据资源,可以支持海量科技信息的统一描述、交换、复用、转换、整合,从根本上实现科技信息数据资源的汇聚融合。

随着科技信息的大量涌现和相关技术工具的成熟,科技信息的组织和利用呈现精细化、颗粒化和语义关联化等特征,元数据描述也呈现出细粒化、模块化、关联化等发展趋势。因此,构建多源异构科技信息元数据标准规范体系要遵循以下原则。

(1) 模块关联化。模块化是元数据最重要的特征,是数据管理规范的基础,其实质是根据实际需要,将资源对象区分为若干个实体,从而将对信息资源的描述转化为对多个不同实体的组合和关联的描述。通过对所包含的各类资源进行调研分析和综合对比,本标准规范将其内容设计为通用容器、资源元数据元素集和规范编码体系3个模块。

(2) 知识单元最小粒度化。描述对象粒度越小越精确,元数据标准规范体系的构建将知识单元最小粒度原则贯穿各个层面,按最小粒度设计元素或属性,尽可能细致地描述资源对象各个层面的信息,为下一步的数据分析和评价打下良好的基础。如机构字段,可细分为上级单位、主办单位、所在国家、省份、城市等,方便精确定位机构和统计分析机构的成果产出。

(3) 可复用、可扩展。可复用、可扩展是指既兼容国内外现有标准规范,又支持根据实际情况做出相应扩展。本标准支持对通用容器的组装复用,也可在本规范制定的元数据基础上,进行相关通用容器和元素集的复用、继承和扩展,支持根据专业特色资源和个性化应用需求进行其他类型资源元数据规范的扩展。

(4) 计算机可识别、可理解。元数据不仅面向人,在数据交换、互操作过程中,更多地是由计算机来实现的,因此需要面向计算机和网络化应用,将元数据规范文本转化为XML Schema形式化描述文件,并基于XML Schema将各类资源元数据转化、封装为XML文件,可以支持计算机对XML文件的自动识别、理解和验证。

多源异构科技信息元数据标准规范体系构建的总体思路是以实现多源异构科技信息的统一描述、规范、汇交、互联为目标,构建集通用容器、资源关联模型、元素集元素、规范编码体系、XML形式化描述规范“五位一体”的元数据标准规范体系,保证数据描述的准确性与全面性,揭示数据的细粒度与关联性,尽可能扩大该规范体系的应用范围。

2.2 构建流程

构建多源异构科技信息元数据标准规范体系的核心

心流程包括5个步骤,即资源类型分析与确定、资源关联模型构建、元数据描述框架设计、元数据元素集元素设计、元数据形式化描述及数据转换(见图1)。

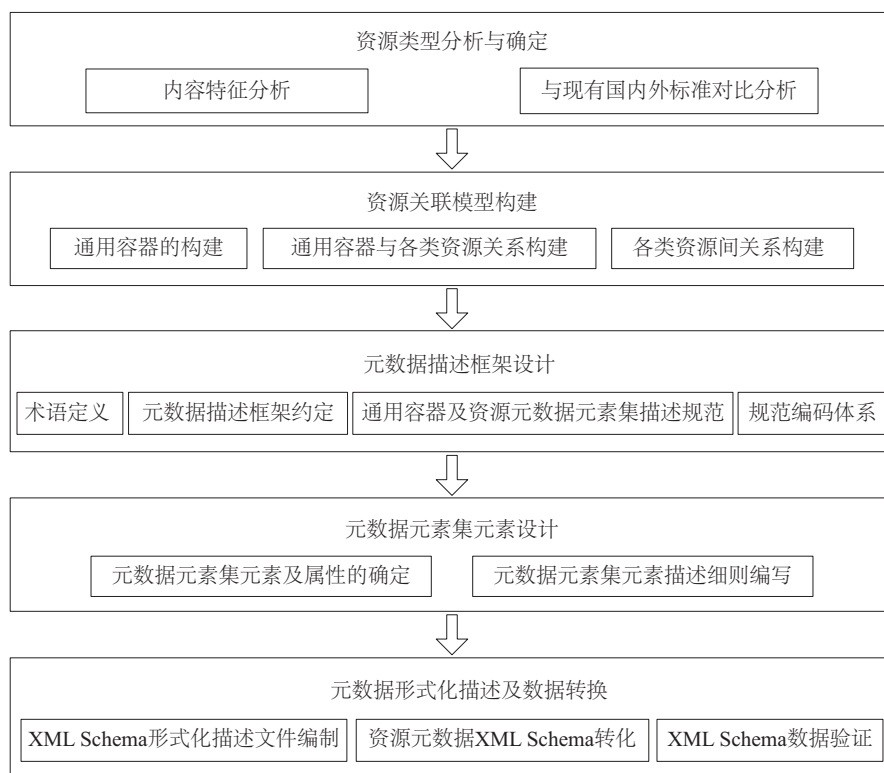


图1 构建多源异构科技信息元数据标准规范体系的核心流程

2.2.1 资源类型分析与确定

(1) 内容特征分析。开展各类多源异构科技信息内容特征分析,是元数据标准规范制定的基础与前提,即明确需要制定元数据标准规范的资源类型。分析的要素主要包括资源名称、类型、简介、体量、更新频率、是否有共性需求等。对于更新频率较高、数据质量较好、用户群体较广的科技信息数据资源,建议制定元数据标准规范;反之,建议不予制定。通过调研分析,最终选取24类数据资源制定其元数据标准规范,分别为期刊、期刊论文、图书、学位论文、会议论文、科技(咨询、行业)报告、科技机构、专家学者、科研项目、知识应用、新闻资讯、产业政策、专利、标准、图片、音频、视频、科技成果、技术工具、课件、统计数据、地理信息数据、百科、学术活动。上述资源来自多个领域,且包含文献、报告、项目、成果、政策、音视频、图片、数值等多种类型。

(2) 与国内外现有标准对比分析。资源类型确定

后,需要调研分析国内外是否已有同类标准,若有同类标准,需要将其与计划制定的资源数据进行对比分析,最大范围内吸收、复用现有标准,在此范围外的,则根据资源具体情况进行扩展。

2.2.2 资源关联模型构建

资源关联模型由通用容器的构建、通用容器与各类资源关系构建以及各类资源间关系构建三部分组成,具体如图2所示。

(1) 通用容器的构建。通用容器是指各类资源元数据元素集中涉及的共性元素,将这些共性元素提取并形成通用集合,以供资源元素集组合使用。通用容器可被各类资源元素所引用。如各类资源均有系统唯一标识符以及增、删、改等管理层面的元素信息,可将这些共性元素提取形成通用集合,命名为“管理通用容器”,各类资源均可调用该容器,避免共性元素重复制定,提高制定效率以及避免冗余,且可实现模块化管

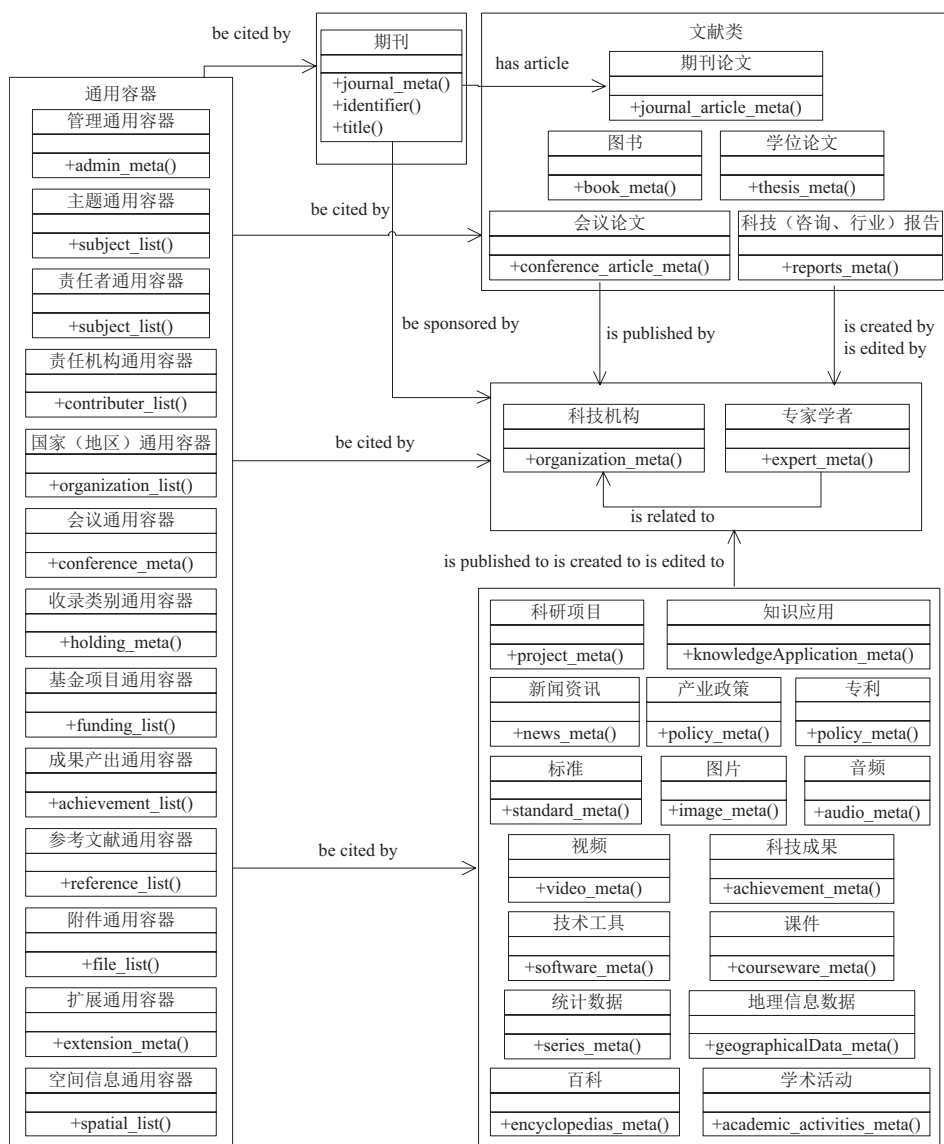


图2 资源关联模型

理。本规范提取24类资源中涉及的共性元素, 构建了13个通用容器, 分别为管理通用容器、主题通用容器、责任者通用容器、责任机构通用容器、国家(地区)通用容器、会议通用容器、收录类别通用容器、基金项目通用容器、成果产出通用容器、参考文献通用容器、附件通用容器、扩展通用容器、空间信息通用容器。

(2) 通用容器与各类资源关系构建。通用容器与各类资源元素之间的关系主要是引用关系, 如主题通用容器、责任者通用容器、国家(地区)通用容器、管理通用容器和参考文献通用容器均可被期刊论文元素、会议论文元素等资源引用。

(3) 各类资源间关系构建。各类资源元素之间的关系有来源关系、从属关系、沿革关系等。例如, 一篇

期刊论文来源于一本期刊; 一篇期刊论文、一篇学位论文、一篇会议论文、一本图书可以有一个或多个责任者(作者、作者机构); 一位专家学者可来自一个科技机构, 可获得(参与)一项或多项科技成果、科研项目、专利; 一个科技机构可主办一种或多种期刊, 可获得一项或多项科技成果、科研项目、专利, 可发布一项或多项新闻资讯、产业政策、行业标准、视频、音频, 可研制一项或多项知识应用。

2.2.3 元数据描述框架设计

(1) 术语定义。术语定义用于描述本规范的元素属性涉及的术语及其定义, 是制定资源类型元数据规

范的前提和基础。本规范对44个术语进行了定义,如“元素”是元数据的基本单元,代表资源的一部分数据内容;“属性”是对元素进行的描述、限定、说明;“容器类元素”是包含其他元素的元素,与其他元素的关系为父子或祖先/后代的关系;“通用容器”是指数据集和其他类型资源元素集中涉及的共性元素,将这些共性元素提取并形成通用集合;“编码体系”是元数据修饰的一种方式,用来规范元素或属性取值范围的受控词表或规范名称列表。

(2) 元数据描述框架约定。元数据描述框架约定

是整个元数据描述框架体系的核心。通用容器及资源元数据元素集中所有元素术语的定义借鉴DCMI术语的定义方法以及ISO/IEC 11179-3: 2013标准“Registry meta model and basic attributes”^[21],需遵循表1所示的元数据描述框架约定。该框架从标识符、名称、出处、定义、数据类型、最大长度、频次范围、描述规则、数据样例9个方面来描述,并详细说明了每个描述项的定义和约束。其中,定义是指对每个描述项含义的解释说明,约束是指该描述项的取值是否必备。必备是指该描述项必须有值,可选是指该描述项取值可为空。

表1 元数据描述框架约定

序 号	元数据描述项的名称	定 义	约 束
1	标识符 (Item Identifier)	元素的唯一标识符,以URI的形式给出。如: http://spec.ckcest.cn/core/elements/1.0/	必备
2	名称 (Name)	赋予元素的唯一标记	必备
3	出处 (Defined By)	一般给出定义元素的来源名称及来源的URI。如无来源名称与URI,也可以是定义元素或维护元素的机构名称	必备
4	定义 (Definition)	对元素概念与内涵的说明	必备
5	数据类型 (Datatype)	元素允许取值的数据类型	必备
6	最大长度 (Maxlength)	元素取值的最大限度	可选
7	频次范围 (Occurrence)	元素使用的频次范围采用区间表示方法:(min, max),同时包括对必备性和最大使用频率的定义min=0表示可选;min=1表示必备;max=∞表示最大使用频次没有限制	必备
8	描述规则 (Described Value Domain)	元素的取值规则、取值范围等	必备
9	数据样例 (Data Element Example)	该元素的样例	可选

(3) 通用容器及资源元数据元素集描述规范。这是基于元数据描述框架,对通用容器及资源元数据元素集的描述规范进行分层和细化,主要包含元素集元素简表的编制及元素描述细则的编制两方面。

第一,元素集元素简表的编制。它是从资源内容、字段约束等多方面对资源进行的全方位描述,具体描述项如表2所示,包括中文名称、名称、数据类型、频次范围、最大长度、复用标准。

第二,元素描述细则的编制。为了让数据操作人员或使用人员更清晰、准确、直观地理解每个元素的规范与约束,针对元数据集内每个元素从出处、定义、描述规则、样例数据等方面进行的详细阐述,具体描述项

表2 通用容器及资源元数据元素集元素简表的字段结构说明

序 号	字段名称	字段说明	字段内容约束
1	中文名称	元素的中文名称	必备
2	名称	赋予元素的唯一标记	必备
3	数据类型	元素允许取值的数据类型	必备
4	频次范围	元素使用的频次范围	必备
5	最大长度	元素取值的最大限度	可选
6	复用标准	标注是否引用现有标准或通用容器	可选

如表3所示,包括标识符、名称、出处、定义、频次范围、最大长度、描述规则、数据样例。

表3 元素描述细则的字段结构说明

序 号	字段名称	字段说明	字段内容约束
1	标识符	元素的唯一标识符, 以URI的形式给出	必备
2	名称	赋予元素的唯一标记	必备
3	出处	来源名称与URI, 也可以是定义元素或维护元素的机构名称	必备
4	定义	对元素概念与内涵的说明	必备
5	频次范围	元素使用的频次范围	必备
6	最大长度	元素取值的最大限度	可选
7	描述规则	元素的取值规则、取值范围、是否配合属性值使用等	必备
8	数据样例	该元素的样例	可选

(4) 规范编码体系。“编码体系”是元数据修饰的一种方式, 用来规范元素、属性取值范围的受控词表或规范名称列表, 通过代码规范元素或属性取值范围, 如表4所示, 包括代码、中文名称、英文名称。

表4 元数据规范编码表的字段结构说明

序 号	字段名称	字段说明	字段内容约束
1	代码	赋予该取值的唯一标记, 一般为英文或数字, 用以规范元素或属性的取值	必备
2	中文名称	元素或属性取值的中文名称	必备
3	英文名称	元素或属性取值的英文名称	可选

根据实际需求, 本规范制定了资源类型、来源机构、分类主题词等28个规范编码表(见表5), 用以规范元素或属性取值范围。以“唯一标识符编码表”为例(见表6), 枚举了24类资源的唯一标识类型, 其取值可根据实际情况进行增删改等操作。

2.2.4 元数据元素集元素设计

根据多源异构科技信息的特征、约束条件等多方面内容, 开展元数据元素集元素设计, 其内容主要包括通用容器及各类资源元数据元素集的元素、属性的确定及元素描述细则的编写。在元数据表达过程中, 元素与属性是配合使用的, 如xml:lang属性是指语种, 该属性与题名、摘要等元素配合使用, 表达其不同语种的元素内容。

资源元素集元素的确定即表2资源元数据元素集

表5 规范编码表清单

序 号	名 称	序 号	名 称
1	资源类型编码表	15	使用许可类型编码表
2	来源机构编码表	16	收录类别编码表
3	分类主题词来源编码表	17	人员(机构)角色编码表
4	专利分类号类型编码表	18	主题容器类型编码表
5	语言代码编码表	19	参考文献类型编码表
6	唯一标识符编码表	20	开放用户类型编码表
7	其他题名类型编码表	21	附件存放路径编码表
8	资源更新频率编码表	22	数据集类型编码表
9	学历学位编码表	23	分中心编码表
10	职称编码表	24	地理信息成果类型编码表
11	学术成果类型编码表	25	科技成果类别编码表
12	国家/地区/组织代码表	26	标准类型编码表
13	机构类型代码表	27	专利类型编码表
14	作者标识符类型代码表	28	学术活动类型编码表

表6 唯一标识符编码

序 号	代 码	中文名称
1	DOI	数字对象唯一标识符
2	ISSN	国际标准连续出版物编号
3	EISSN	国际标准电子连续出版物编号
4	ISBN	国际标准图书编号
5	CN	国内统一刊号
6	Project Code	项目编号
7	Project Application Code	项目申请代码
8	Achievement Code	科技成果编号
9	SICI	连续出版物及其文章、卷期等独立描述单元编码
10	BICI	图书及其组成单元唯一标识符
11	PII	出版物唯一标识符
12	ISNI	国际标准名称标识
13	NACAO	全国组织机构代码
14	RP-ID	报告号
15	URI	统一资源标识符
16	KID	浙江大学专家注册系统唯一标识
17	Others	其他类型标识符, 可根据需要扩展编码

元素简表内容的设计, 分为3个部分: ①设计一个容器类元素用来对该资源数据所有元素进行封装, 如achievement_meta用来封装科技成果元素集中所有元素, 一般为元素简表的第一个元素; ②资源元素集主体元素的设计, 即能够全面描述或体现该类资源特征的元素, 主要包括基本信息和关联信息; ③管理操作类元

数据的设计,即描述与该类资源元素集相关的管理信息,主要包括系统唯一标识符、数据创建时间、修改时间、删除标识、使用权限等,在本规范中通过引用管理信息通用容器来实现,一般位于元素简表末尾。

通用容器元数据元素集的设计只包含前两部分,即设计一个容器类元素用来对该通用容器所有元素进行封装、元素集主体元素的设计。

基于确定的资源元数据元素集元素,按照表3的字段结构说明,编写每个元素的描述细则。

表7 管理通用容器元数据元素集元素简表

序 号	中文名称	名 称	数据类型	频次范围	最大长度	复用标准
1	管理容器	admin_meta	容器类元素	[1, 1]		
2	数据集唯一标识符	dataset_id	String	[1, 1]	100	
3	系统唯一标识符	guid	String	[1, 1]	100	
4	资源类型	resource_type	String	[1, 1]	50	
5	知识中心分类号	cae_classification_no	String	[0, 1]	1 000	
6	是否删除标识	is_delete	String	[0, 1]	2	
7	初始创建时间	created_time	DateTime	[1, 1]		
8	最新修改时间	updated_time	DateTime	[1, 1]		

表8 元素“是否删除标识”的描述细则

描述项	描述内容
标识符	http://spec.ckcest.cn/core/is_delete
名称	is_delete
出处	中国工程科技知识中心元数据规范
定义	记录数据集中的每条数据是否被逻辑删除
频次范围	[0, 1]
最大长度	2
描述规则	描述记录是否做了逻辑删除标记,若取值为Y,则为已逻辑删除,取值为N,为未做删除标记,如果该值为空或其他字符,则该记录为正常状态(未删除)
数据样例	<is_delete>Y</is_delete>

(2) 资源元数据元素集元素设计。资源元数据元素集元素设计需要在详细深入了解资源特征的基础

(1) 通用容器元数据元素集元素设计。通用容器元数据元素集元素设计需要保证其元素具备通用性和可复用性,以供资源元素集组合使用,以“管理通用容器”为例,按上述通用容器元素的设计步骤,依次确定了管理通用容器的封装容器类元素及其主体元素(见表7)。

在此基础上,按照表3编写每个元素的描述细则,以“管理通用容器”中的元素“是否删除标识”为例(见表8)。

上,对资源内容进行分类、提取、组合,从而使得设计的元素能够准确、全面地描述该类资源。如表9所示,以科研项目为例,依次确定科研项目封装容器类元素、科研项目元素集主体元素以及管理操作信息,在主体元素部分,主题、项目人员、项目机构、科研产出、附件等元素均复用通用容器,项目人员及项目机构为关联信息,分别通过责任者唯一标识符、责任机构唯一标识符关联专家学者及科研机构两类资源。

在此基础上,按照表3编写每个元素的描述细则。以“科研项目”中的元素“唯一标识”为例,如表10所示,该元素需要搭配属性type使用,即唯一标识符所属的类型,其详情在“描述规则”中进行了阐述。

表9 科研项目元数据元素集元素简表

序 号	中文名称	名 称	数据类型	频次范围	最大长度	复用标准
1	科研项目元素	project_meta	容器类元素	[1, 1]		
2	唯一标识	identifier	String	[0, ∞)	100	
3	项目名称	title	String	[1, 1]	1 000	
4	其他项目名称	alternative	String	[0, ∞)	1 000	
5	摘要	abstract	Text	[0, 1]		
6	其他语种摘要	abstract_alternative	Text	[0, ∞)		

续表

序 号	中文名称	名 称	数据类型	频次范围	最大长度	复用标准
7	主题	subject_list	容器类元素	[0, 1]		参见主题通用容器
8	项目人员	contributor_list	容器类元素	[0, 1]		参见责任者通用容器
9	项目机构	organization_list	容器类元素	[0, 1]		参见责任机构通用容器
10	项目金额	funding	String	[0, 1]	50	
11	起始日期	project_start	String	[0, 1]	20	
12	结束日期	project_end	String	[0, 1]	20	
13	科研产出列表	achievement_list	容器类元素	[0, 1]		参见成果产出通用容器
14	详情地址	detail_URI	String	[1, ∞)	1 000	
15	附件	file_list	容器类元素	[0, 1]		参见附件通用容器
16	扩展元素	extension_meta	容器类元素	[0, 1]		参见扩展通用容器
17	管理信息	admin_meta	容器类元素	[1, 1]		参见管理通用容器

表10 元素“唯一标识”的描述细则

描述项	描述内容
标识符	http://spec.ckcest.cn/core/identifier
名称	identifier
出处	中国工程科技知识中心元数据规范
定义	在特定范围内给予科研项目的一个明确标识
频次范围	[0, ∞)
最大长度	100
描述规则	采用符合正式标识符体系的字符串进行标识, 其类型编码参见“附录: 表46唯一标识符编码表”。 ①type=ProjectCode, 表示项目号, 为重要非必备字段; ②type=ProjectApplicationCode, 表示项目申请代码, 为重要非必备字段
数据样例	<identifier type="ProjectCode">CK20160523-1</identifier> <identifier type="ProjectApplicationCode"> 20160567-1</identifier>

2.2.5 元数据形式化描述及数据转换

元数据形式化描述是将数据以计算机可读方式进行描述与规范, 强调数据的输出或存储方式。本规范采用XML语言实现元数据的形式化描述。XML语言包含一组定义语义标记的规则, 将元数据规范文本转化为XML Schema形式化描述文件, 并基于XML Schema将各类资源元数据转化、封装为XML文件, 支持计算机对XML文件的自动识别、理解和验证。

(1) 基于XML Schema的元数据形式化描述。采用XMLSpy软件实现13类通用容器及24类数据资源描述元数据的XML schema的编制并生成其XSD文件。图3描述了科研项目信息资源描述元数据的XML Schema,

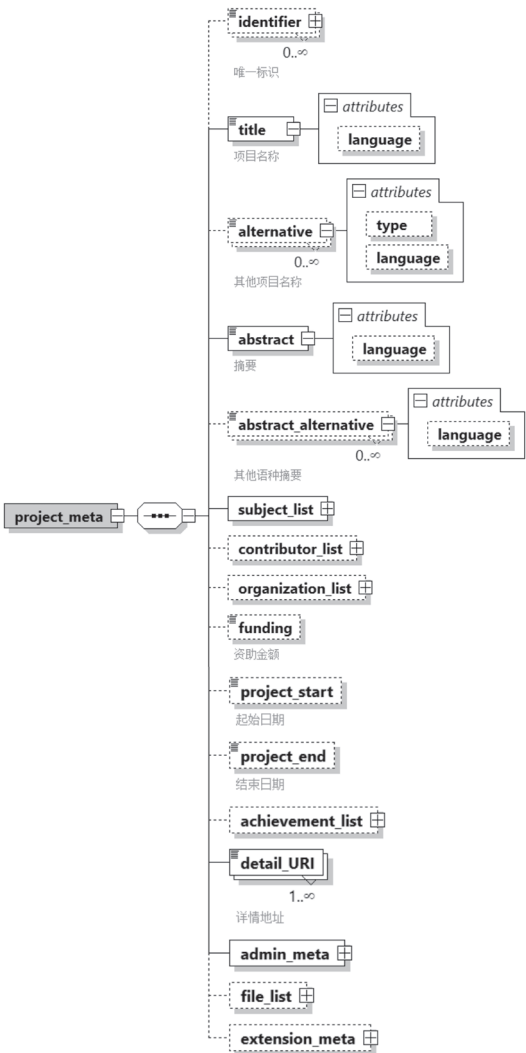


图3 科研项目信息资源描述元数据的XML Schema

显示了科研项目信息资源描述元数据的树形结构,呈现了科研项目信息资源元素集所有元素,以及搭配使用的所有属性,同时也显示了每个元素的必备性。

(2) 元数据描述实例。在编制的XML Schema基础上,生成包含具体信息资源元数据内容的XML数据文件。

3 典型应用

目前,制定的科技报告、科技机构、专家学者、科研项目等24类通用资源元数据标准规范,已在中国工程院组织建设的中国工程科技知识中心开展了广泛应用。中国工程科技知识中心自2012年建设以来,已建立化工、农业、医药、地质、能源和卫生等近30余个专业知识服务分中心系统,通过自建、联盟、采购、网络开放获取等方式,汇聚了20多个主题领域的资源,包括文献、数值数据、工具(事实)、行业报告和政策法规等,数据资源建设总量为68亿条,体量达到100TB,本规范指导了来自30余个专业领域的24类亿级别数据资源元数据的规范、描述与汇交,为实现元搜索服务、系统互操作,从根本上实现工程科技领域资源的汇聚融合和互联互通,实现海量资源一站式、全方位搜索和发现服务奠定了坚实的基础。

随着各类科技信息数据资源,尤其是专业性较强的特色资源不断增加,已制定的24类通用资源元数据规范不能适用于各类特色资源。为进一步提高这些特色资源的可发现能力和利用率,有必要制定各专业领域的优势特色资源元数据标准规范,从而实现各类特色资源的汇聚融合、互联互通和网络共享服务。因此,基于24类通用资源元数据规范的制定经验和总体设计思路,结合专业资源特色和数据描述规范,又指导中国工程科技知识中心21个分中心制定了国家地质公园、地下水资源图件、卫星发射记录、电子元器件产品等100类特色资源元数据标准规范,并实现XML Schema形式化描述、转换、验证与汇交。

4 结语

本文构建的元数据标准规范体系初步实现了多类型、多来源、异构资源的统一描述、规范及互联互通,且已在工程科技领域开展了广泛应用,并以此有效指导了各领域特色资源元数据规范的制定与应用,但随着资源

数据量激增,类型多样,面临的问题也会层出不穷,仍需要在应用实践中进一步检验其可行性、适用性和扩展性,不断优化完善多源异构科技信息数据资源元数据标准规范体系。为实现更深层次的关联互通并支撑智能化服务应用,还需要将现有元数据规范体系向富含更多语义关系的RDF Schema、OWL本体模型转换,从而支持大规模基于科技知识图谱的表示、转换与关联。

参考文献

- [1] 周振国. 治理框架视域下的数据治理研究[J]. 农业图书情报学报, 2020, 32(7): 57-62.
- [2] 数据融合之数据治理[EB/OL]. [2021-04-01]. <https://www.talkwithtrend.com/Article/243133>.
- [3] 王国复, 涂勇, 王卷乐, 等. 科学数据共享中的元数据技术研究[J]. 中国科技资源导刊, 2008, 40(1): 30-36.
- [4] DC metadata[EB/OL]. [2021-04-01]. <http://dublincore.org/>.
- [5] NISO JATS Version 1.1 (ANSI/NISO Z39.96-2015)[EB/OL]. [2021-04-01]. <http://jats.nlm.nih.gov/archiving/tag-library/1.1/index.html>.
- [6] Dryad[EB/OL]. [2021-04-01]. <https://datadryad.org>.
- [7] GREENBERG J, WHITE H C, CARRIER S, et al. A metadata best practice for a scientific data repository[J]. Journal of Library Metadata, 2009, 9(3/4): 194-212.
- [8] Dataverse[EB/OL]. [2021-04-01]. <https://dataverse.harvard.edu/>.
- [9] ISI Web of Science-Science Citation Index Expanded[EB/OL]. [2021-04-01]. <http://www.webofknowledge.com/WOS>.
- [10] Scopus[EB/OL]. [2021-04-01]. <http://www.scopus.com/>.
- [11] 张建勇, 于倩倩, 黄永文, 等. NSTL统一文献元数据标准的设计与思考[J]. 数字图书馆论坛, 2016(2): 33-38.
- [12] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. 科技平台 资源核心元数据: GB/T 30523-2014[S]. 北京: 中国标准出版社, 2014.
- [13] 中华人民共和国文化部. 期刊论文元数据规范: WH/T 67-2014[S]. 北京: 中国标准出版社, 2014.
- [14] 中华人民共和国文化部. 电子图书元数据规范: WH/T 65-2014[S]. 北京: 中国标准出版社, 2014.
- [15] 中华人民共和国文化部. 学位论文元数据规范: WH/T 68-2014[S]. 北京: 中国标准出版社, 2014.
- [16] 中华人民共和国文化部. 网络资源元数据规范: WH/T 50-2014[S]. 北京: 中国标准出版社, 2014.

- [17] 中华人民共和国文化部. 电子连续性资源元数据规范: WH/T 64-2014 [S]. 北京: 中国标准出版社, 2014.
- [18] 崔佳伟, 吴思竹, 邬金鸣, 等. 科学数据仓储元数据标准研究与启示 [J]. 数字图书馆论坛, 2019 (6): 19-27.
- [19] 赵龙文, 莫荔媛, 陈明艳. 面向政府数据开放的资源描述方法 [J]. 图书情报工作, 2017, 61 (6): 115-121.
- [20] 朱玲, 聂华, 崔海媛, 等. 北京大学开放研究数据平台建设: 探索与实践 [J]. 图书情报工作, 2016, 60 (4): 44-51.
- [21] ISO/IEC 11179-3:2013 [EB/OL]. [2021-04-01]. <https://www.iso.org/standard/50340.html>.

作者简介

罗婷婷, 女, 1985年生, 硕士, 助理研究员, 研究方向: 知识组织、大数据融汇治理、信息管理与信息系统。
赵瑞雪, 女, 1968年生, 博士, 研究员, 研究方向: 信息管理与信息系统、数字图书馆、知识组织与知识服务。
李娇, 女, 1989年生, 博士研究生, 助理研究员, 研究方向: 知识组织、知识图谱。
傅智杰, 女, 1978年生, 硕士, 副高级工程师, 研究方向: 知识组织、大数据管理、信息系统。
武丽丽, 女, 1983年生, 硕士, 助理研究员, 研究方向: 数据分析、数据管理、知识服务。
黄永文, 女, 1975年生, 博士, 副研究馆员, 研究方向: 知识组织与知识服务。
鲜国建, 男, 1982年生, 博士, 研究员, 通信作者, 研究方向: 大数据融汇治理、知识组织、知识图谱, E-mail: xianguojian@caas.cn。

Construction of Metadata Standard System for Multi-source and Isomerized Scientific and Technological Information Governance

LUO TingTing¹ ZHAO RuiXue^{1,2} LI Jiao¹ FU ZhiJie³ WU LiLi³ HUANG YongWen¹ XIAN GuoJian^{1,2}

(1. Agricultural Information Institute of CAAS, Beijing 100081, China; 2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China; 3. China Academy of Engineering Strategic Consulting Center, Beijing 100088, China)

Abstract: Based on the theory of metadata and the idea of knowledge object-oriented modeling, this paper constructs a flexible and widely applicable metadata standard system, covering 13 kinds of generic containers, 24 kinds of resource element set description specifications and 28 specification coding tables, to carry out the aggregation governance of multi-type, multi-source and isomerized technology resources in the big data environment, realize the standardized description and interconnection of big data resources, and improve the discoverability, utilizability and sharability of data resources. Moreover, this paper compiles a supporting XML Schema formal description specification to support the computer's automatic processing of conversion, verification and parsing from multi-type and multi-source heterogeneous metadata to unified XML format. The standard system has been applied and verified in China Knowledge Centre for Engineering Sciences and Technology, and has guided more than 30 sub-centers to complete the transformation and collection of 24 kinds of data resources with more than 100 million-level data, strongly supporting the construction of "metadata sea" of engineering sciences and technology. In addition, the customization and application of metadata standards and specifications for characteristic resources in nearly 100 kinds of professional fields have been further realized quickly and efficiently based on the system.

Keywords: Multi-source and Isomerized; Scientific and Technological Information; Metadata; Standard System

(收稿日期: 2021-04-01)