

# 科学数据仓储元数据标准研究与启示\*

崔佳伟 吴思竹 邬金鸣 修晓蕾 钱庆  
(中国医学科学院医学信息研究所, 北京 100020)

**摘要:** 科学数据的开放共享离不开数据仓储和元数据标准的支持, 为数据附加高质量的元数据描述, 是实现科学数据有效存储与管理的基础。目前国外的科学数据仓储元数据标准发展已经较为成熟, 但国内的相关标准还大多处于建设之中。本文拟对国外现有科学数据仓储相关的元数据标准的内容设计及实际应用情况进行系统梳理与深入探究, 以期为我国科学数据仓储元数据标准的构建及应用提供可资借鉴的参考。

**关键词:** 科学数据; 数据仓储; 元数据标准; 数据管理

**中图分类号:** G250

**DOI:** 10.3772/j.issn.1673-2286.2019.06.003

随着数据密集型科学研究成为常态, 各类科学数据呈爆炸式增长, 这为科研人员存储、管理以及共享数据带来了巨大的困难与挑战。科学数据仓储的出现虽然为科学数据的存储和共享提供了途径, 但如何实现仓储内数据进行有效管理和统一检索则成为另一个重要问题。科学数据元数据是对科学数据外部形式和内部特征的详细描述<sup>[1]</sup>, 能够为科学数据的组织、整合、交换、发现和获取等提供信息。为充分发挥科学数据仓储的功能与作用, 必须构建、应用能够满足仓储功能目标、资源特点和用户需求的科学数据元数据标准, 鉴于此类科学数据元数据标准主要应用于科学数据仓储, 故也可称之为科学数据仓储元数据标准。国外的科学数据仓储元数据标准发展已经较为成熟, 但国内的相关标准还大多处于建设之中。本文拟对国外现有科学数据仓储相关的元数据标准的内容设计及实际应用情况进行系统梳理与深入探究, 以期为我国科学数据仓储元数据标准的构建和应用提供可资借鉴的参考。

## 1 科学数据仓储元数据标准概述

在过去的几十年里, 为实现跨学科、跨领域的科学

数据发现与共享, 生物医学、物理学和社会科学等领域都建立了相应的元数据标准与规范。尤其在生物医学领域, 随着高通量测序技术的快速发展, 临床实验和人群队列研究的不断深化, 科学数据的膨胀驱动了该领域科学数据仓储的设计, 致使该领域科学数据仓储元数据标准的构建与应用均处于较为前沿位置, 对其他学科领域科学数据仓储元数据标准以及通用科学数据仓储元数据标准的建设具有较高的借鉴价值, 故本研究将生物医学领域科学数据仓储元数据标准与通用科学数据仓储元数据标准进行综合性分析。通过对国际科学数据仓储注册系统Re3Data.org中收录的仓储所应用的元数据标准进行统计与归纳<sup>[2]</sup>, 结合FAIRsharing<sup>[3]</sup>和英国数字监护中心(Digital Curation Center)<sup>[4]</sup>中收录的科学数据元数据标准, 并针对能够存储科学数据的相关仓储及仓储所使用的元数据标准进行调研, 在综合考虑每个标准的权威性、应用广泛性、领域影响力及内容覆盖力的基础上, 选取了3个通用的科学数据仓储元数据标准, 分别为Dublin Core、DataCite和Dataverse科学数据仓储元数据标准, 以及3个生物医学领域科学数据仓储元数据标准, 分别为DatA Tag Suite、W3C HCLS Dataset Description和Dryad科学数据仓储元数据标准。

\*本研究得到国家人口健康与科学数据服务平台后补助课题“国家人口与健康科学数据共享服务平台标准宣贯实施与推广”(编号: NCMI-MCE02-201801)和国家重点研发计划精准医学研究重点专项课题“重大疾病精准医学数据库群”(编号: 2016YFC0901602)资助。

## 1.1 Dublin Core

1995年3月,都柏林核心集(Dublin Core)诞生于美国俄亥俄州都柏林镇召开的第一届元数据研讨会上<sup>[5]</sup>。会议目的是希望建立一套简单通用的描述网络资源的方法,降低检索难度,从而提高网络资源利用率。而后经过多次修正和补充,逐渐形成现在包含15个元素的元数据标准。目前 Dublin Core已被多个机构作为正式标准发布(ISO15836、NISOZ3985、RFC5013),我国与其对应的标准为GB/T 25100—2010。

## 1.2 DataCite

DataCite Metadata Schema由DataCite国际联盟(the DataCite Consortium)制定<sup>[6]</sup>,该联盟创建的主要目标是支持科学数据存储并将科学数据的地位提升至合法的、可被引用的科学记录,使科学数据更易在网上获取。其创建的DataCite元数据标准包含一系列核心元数据元素,通过为数据集提供永久性唯一标识符(DOI)以及准确、一致性的描述,辅助科学数据的检索、共享、重用、应用和关联。

## 1.3 Dataverse科学数据仓储元数据标准

Dataverse是哈佛-麻省理工数据中心(Harvard MIT Data Center, HMDc)于2007年开发的一个科学数据管理系统,能够对科学数据进行发布、引用、存储、发现和在线分析<sup>[7]</sup>。Dataverse的元数据标准是以DDI(Data Document Initiative)元数据标准为基础扩展而成,根据不同的类型分为不同的区块,包括引用通用元数据区块和学科专有元数据区块。其中,引用通用元数据区块包含引用数据集所需的相关信息,是平台所有数据集的必备元数据区块,适用于描述所有类型 and 所有学科的数据集<sup>[8]</sup>;学科专有元数据区块则提供针对某一学科数据的元数据元素,覆盖生命科学、人文与社会科学、地理空间、天文与天体物理和政治学等多个领域。

## 1.4 Data Tag Suite

Data Tag Suite(DATS)是由NIH的bioCADDIE(biomedical and healthCare Data Discovery Index Ecosystem)开发的元数据模型<sup>[9-10]</sup>,设计初衷是满足

科学数据仓储DataMed的资源索引和检索需求<sup>[11-12]</sup>。DATS以实现跨数据仓储的数据集发现和获取为目的,其核心实体可用于描述任何类型的数据集,包括“数据集”“数据集分布”“获取”等实体,涵盖了数据集的基本信息;扩展实体则主要针对生物医学领域,包括“疾病”“研究”“生物”“分子”等实体,以期对生物医学领域的数据进行揭示。

## 1.5 W3C HCLS Dataset Description

W3C HCLS Dataset Description是由the W3C Semantic Web for Health Care and the Life Sciences Interest Group(HCLSIG)通过重用18个已有词表构建的RDF词表<sup>[13-14]</sup>,目的是为生物医学领域数据集提供一个高质量标准,以满足数据集的描述、关联、交互、更新、内容总结、索引和发现等功能需求。W3C HCLS Dataset Description对数据集的描述分为3个层级:①Summary层级,该层级对于数据集的描述独立于特定版本或格式;②Distribution层级,侧重于描述特定数据文件的格式及可下载位置;③Version层级,主要描述特定版本数据集的属性,并利用VersionNumber将Summary层级的描述与Distribution层级的描述相关联。除此之外,该标准中的元数据元素又分为5个专题模块,分别为:核心元素元数据;标识符;出处和变化;访问、获取;统计。

## 1.6 Dryad科学数据仓储元数据标准

Dryad是由美国国家进化分析中心等机构建立的科学数据仓储<sup>[15]</sup>,旨在实现对进化生物学、生态学及相关领域出版物的支撑数据的保存、发现、复用和管理。该仓储将元数据管理纳入科学数据管理的全过程,成为科学数据仓储元数据管理的典范,被称为元数据的“最佳实践”<sup>[16]</sup>。Dryad的元数据从描述内容来看可以分为3个模块(即出版物元数据、数据集元数据和文件元数据<sup>[17]</sup>),通过利用特定元数据元素值之间的继承,实现科学数据之间以及科学数据与期刊文章之间的相互关联。

## 2 科学数据仓储元数据标准分析

本文围绕科学数据仓储元数据标准的特点,提出分

析框架，从基础信息、内容设计和实际应用3个维度对这些元数据标准进行深入分析，以期为我国科学数据仓储元数据标准的构建和应用提供可资借鉴的参考。

2.1 基础信息

通过调研各元数据标准及其相关科学数据仓储，对各标准的基础信息进行整理，包括标准正文语种、发布机构/组织、发布国家、最初版发布时间、最新版发布时间、最新版版本号和可下载格式，具体内容见表1。通过比较可以发现，美国在元数据标准制定和发布方面发挥了重要作用，主要体现在其起步较早且发布的标准应用广泛。如Dryad科学数据仓储元数据标准和Dataverse科学数据仓储元数据标准的最初版均

于2007年发布，2015年发布的DATS更是被DataMed、OmicsDI、ICPSR等多个科学数据仓储采用。

随着科学数据仓储的持续发展，仓储收录的数据资源及服务类型不断丰富，这就需要仓储的元数据标准持续优化升级，实现多个版本的更迭。以DataCite为例，其最初版发布于2009年，而后每隔一年或两年便更新一次，2019年发布的最新版版本号为4.2。

同时，为适应不同类型科学数据的特点，支持数据的人机理解与处理，各元数据标准的格式也趋于多样化，从而满足用户多样性的需求。如JSON格式能够简洁清晰地揭示元数据标准内容元素的层次结构，XML格式的通用性有助于元数据标准在不同应用场景的共享与快速解析，RDF格式能够准确描述出标准中各元素间的关系等。

表1 元数据标准基础信息比较

| 标准名称                         | 缩 写  | 采用语种 | 标准级别 | 发布机构/组织                                | 发布国家 | 最初版发布时间/年 | 最新版发布时间/年 | 最新版版本号 | 可下载格式                         |
|------------------------------|------|------|------|----------------------------------------|------|-----------|-----------|--------|-------------------------------|
| Dublin Core                  | DC   | 英语   | 国际标准 | Dublin Core Metadata Initiative (DCMI) | 美国   | 1995      | 2012      | 1.1    | PDF                           |
| DataCite                     | -    | 英语   | 国际标准 | the DataCite Consortium                | 多国   | 2009      | 2019      | 4.2    | PDF、XSD、XML                   |
| Dataverse科学数据仓储元数据标准         | -    | 英语   | 平台标准 | Harvard University                     | 美国   | 2007      | *         | 4.8.4  | XMLX、TSV                      |
| DatA Tag Suite               | DATS | 英语   | 国际标准 | bioCADDIE                              | 美国   | 2015      | 2017      | 2.2    | JSON、XLSX、CMAP、PDF、EPS、SVG、PS |
| W3C HCLS Dataset Description | -    | 英语   | 国际标准 | HCLSIG                                 | 多国   | 2011      | 2015      | -      | RDF、PDF                       |
| Dryad科学数据仓储元数据标准             | -    | 英语   | 平台标准 | Dryad                                  | 美国   | 2007      | 2010      | 3.0    | XSD、PDF                       |

注：\*代表该项不确定

2.2 内容设计

在科学数据仓储元数据标准内容设计分析部分，分为设计概况分析和内容元素分析两个部分，逐层揭示现有标准值得借鉴之处与不足，辅助我国相关标准的制定。

2.2.1 设计概况

根据各元数据标准提供的创建指南或最佳实践，对标准的整体设计架构进行分析，主要包括元数据标准设计目标、元素总数、元素分类、核心/必备元素个

数、复用标准、是否提供受控词表和是否提供数据集引用标准或格式7个方面（见表2），进而从宏观层面对各元数据标准的内容设计进行较为直观的比较与分析。

在元素设置及分类方面，除Dublin Core外，其他5个标准都对自身的元数据元素进行分类，其中“Must”“Mandatory”“Requried”类元素均代表必备元素，（即在描述数据时必须出现的元素），占总元素的6%~80%，主要为10%左右。以DataCite为例，该标准共收录元数据元素75个，分为必备元素、推荐元素（在描述数据时推荐出现的元素）和可选元素（在描述数据时可不出现的元素）3类，其中必备元素9个，占总元素的12%。从理论上讲，元数据标准中设置的元素越多，其

表2 元数据标准内容设计比较

| 标准名称                         | 设计目标                                                                         | 元素总数                            | 元素分类                                                                              | 核心/必备元素个数                       | 是否提供受控词表        | 复用标准                                                                                                                        | 是否提供数据集引用标准或格式 |
|------------------------------|------------------------------------------------------------------------------|---------------------------------|-----------------------------------------------------------------------------------|---------------------------------|-----------------|-----------------------------------------------------------------------------------------------------------------------------|----------------|
| Dublin Core                  | 提高网络资源的利用率，降低其检索难度                                                           | 15个                             | -                                                                                 | -                               | 是<br>(提供已有词表链接) | -                                                                                                                           | 否              |
| DataCite                     | 满足用户科学数据检索需求，促进科学数据的出版与引用                                                    | 75个                             | Mandatory、Recommended和Optional                                                    | 9个                              | 是<br>(提供自建词表)   | *                                                                                                                           | 是              |
| Dataverse科学数据仓储元数据标准         | 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                                    | 78个(引用通用元数据区块)                  | Required和Optional                                                                 | 5个(引用通用元数据区块)                   | 是<br>(提供自建词表)   | DDI、DataCite、Dublin Core、ISA-Tab、Virtual Observatory (VO) Discovery and Provenance Metadata                                 | 是              |
| Data Tag Suite               | 为科学数据仓储DataMed资源描述服务，帮助用户跨数据仓储发现和获取数据集                                       | 180个(核心实体)                      | Must、Must Not、Required、Shall、Shall Not、Should、Should Not、Recommended、May和Optional | 14个(核心实体)                       | 是<br>(提供自建词表)   | DataCite、W3C HCLS Dataset Description、Common Metadata Elements for Cataloging Biomedical Datasets、NCBI BioSample等           | 否              |
| W3C HCLS Dataset Description | 提供针对生物医学领域数据集的高质量描述，解决数据描述不充分、没有任何一个元数据标准涵盖所有重要的元素以及不同的数据目录和数据仓储使用不同的元数据标准问题 | 61个                             | Must、Must Not、Required、Shall、Shall Not、Should、Should Not、Recommended、May和Optional | 10个                             | 是<br>(提供已有词表链接) | Citation Typing Ontology、Data Catalog、Dublin Core Metadata Types、Collection Description Frequency Vocabulary、PROV Ontology等 | 否              |
| Dryad科学数据仓储元数据标准             | 为科学数据仓储Dryad资源描述服务，实现数据与出版物之间的关联                                             | 17个(描述出版物)、16个(描述数据集)、19个(描述文件) | Required和Optional                                                                 | 10个(描述出版物)、10个(描述数据集)、14个(描述文件) | 是<br>(提供已有词表链接) | the Bibliographic Ontology、Dublin Core、Darwin Core                                                                          | 是              |

注：\*代表该项不确定

对数据集内容和特征的揭示也就越全面，更有利于促进科学数据的共享与发现。但在实际应用中，如果元数据标准中的元素过多，尤其是必备元素过多时，将导致数据著录过程过于烦琐和冗长，影响用户上传数据和仓储收集数据的效率，反而不利于科学数据的共享。而如果必备元素过少，虽然能够减少著录负担，但可能导致对数据资源的揭示程度不够，阻碍科学数据的发现与再利用。同时，通过对科学数据仓储进行调查发现，许多仓储虽然直接应用特定的元数据标准，但也会依据自身需求对标准元素进行增减或修改。以科学数据

仓储Zendo为例，其在应用元数据标准DataCite时，不仅使用了该标准的必备元素和推荐元素，还在此基础上进行一些额外的补充<sup>[18]</sup>，从而为用户提供丰富的资源描述和多角度资源服务，保证数据的可发现性和可重用性。

在提供受控词表方面，各标准均通过提供自建的受控词表或引用已有受控词表对部分元数据元素的值域进行限定。以Dataverse科学数据仓储元数据标准中的元数据元素“Subject”为例，该项可填入的内容只能从标准自建的受控词表中选择，即从“Agricultural

Sciences” “Medicine, Health and Life Sciences”

“Chemistry” 和 “Other” 等中选择一个或多个。通过此种方式, 有助于用户了解和选择恰当的词, 从而在创建数据集描述之始提高元数据质量, 确保元数据描述的规范性和一致性, 以便科学数据仓储未来提供更深层次的服务(如数据分析服务、知识发现服务等)。

元数据复用是指在一个元数据标准中复用一个或多个元数据标准中的部分元素, 利用不同元数据标准共同描述复杂资源, 以便扩展元数据标准的适用范围, 增强不同系统之间元数据的互操作性, 促进元数据的相互转换<sup>[19]</sup>。由“复用标准”列可见, 超过60%的元数据标准在构建时都会借鉴其他元数据标准, 复用已有元数据标准的部分元素, 从而在降低标准构建复杂性的同时提高自身实用性, 并为实现元数据标准间的互映射提供很好的基础保障。以Dryad科学数据仓储元数据标准为例, 其在构建时便复用了“the Bibliographic Ontology” “Dublin Core” 和 “Darwin Core” 这3个标准中的元素。

规范的数据引用标准或格式在数据认证、数据再利用以及追踪数据影响等方面均发挥了重要作用。由“是否提供数据集引用标准或格式”列可见, DataCite和Dataverse科学数据仓储元数据标准均提供数据集引用标准或格式, 而Dublin Core、DATS、W3C HCLS Dataset Description和Dryad科学数据仓储元数据标准并未提供。

### 2.2.2 内容元素

基于不同的设计目标和应用需求, 各科学数据仓储元数据标准的结构框架和内容元素设计均存在明显差异。Dublin Core、DATS和Dryad科学数据仓储元数据标准的设计目标较为类似, 主要为促进资源的存储、管理、发现与获取, 并不需要对资源内容进行详细的说明与揭示, 此类标准的元素相对较少。相比于Dublin Core, Dryad科学数据仓储元数据标准和DATS更加注重不同类型资源间的整合, 其中Dryad科学数据仓储元数据标准利用不同元数据模块中特定元数据元素值之间的继承关系将出版物与数据集相关联; 而DATS是基于“被引用”“使用”“存储”“符合”等实体间关系将科学数据资源与出版物、软件以及其他科学数据仓储和数据标准相联系, 同时还针对生物医学相关科学数据构建扩展实体。与上述3个标准不同的是DataCite、

Dataverse科学数据仓储元数据标准和W3C HCLS Dataset Description, 这3个标准的总元素较多, 必备元素却较少, 利用可选元素对数据资源进行灵活描述, 对于数据集的揭示更具有完整性, 尤其是Dataverse科学数据仓储元数据标准和W3C HCLS Dataset Description, 分别利用学科专有元数据区块和分层级的资源描述方法, 深入到科学数据描述的微观层面。

通过对各科学数据仓储元数据标准的内容元素展开分析, 探究各标准的元数据元素共有情况, 可以发现内容关联性和相似性较高的标准, 为实现标准间的映射及互操作奠定基础, 同时也为新标准的构建提供备选元素。为保证比较分析的元素在同一层面, 本文只选取各标准中描述数据集的元素进行分析。依据元素被各标准共有的情况, 可将元素分为3个大类, 分别为公共元素、基本元素和扩展元素, 每类元素具体涵盖的内容如图1所示。

虽然各元数据标准具体应用领域不同, 描述数据集的细粒度也不同, 但实现数据描述和数据检索都是其首要功能, 因此在各元数据标准中, 标题、标识符、描述、时间等记录数据集基本属性的公共元素都是必备的。在公共元素基础上, 各标准围绕自身构建目标及需求, 对数据集的内容和形式展开进一步说明, 归纳后形成基本元素和扩展元素。以DATS和Dryad为例, DATS标准的设计目标是帮助用户实现跨数据仓储的数据集发现和获取, 故其基本元素和扩展元素主要描述数据集的形式特征, 提供关于数据集的分发信息(如数据集的格式、版本、存储位置、获取方式等), 同时为满足生物医学科学数据的专业性, DATS还提供一些非核心实体, 用于记录数据集内生物体的分类情况、经历的研究过程和治疗过程等; Dryad科学数据仓储元数据标准为满足自身功能目标——实现生物医学科学数据之间以及科学数据与期刊文章之间的相互关联, 在基于公共元素对数据集的内容和形式进行简单描述的基础上, 利用扩展元素“Associated Dryad Publication Record Identifier”和“Associated Dryad Data File Record Identifier”记录数据集、出版物和数据文件间的联系。

## 2.3 实际应用

在上文对各元数据标准结构及内容进行分析的基础上, 结合各标准在科学数据仓储中的实际应用情

况,表3对它们的特点、不足、使用难度、适用范围以及应用实践进行总结,不仅为我国科学数据仓储提供选

择元数据标准的依据,也为新标准的构建提供参考和借鉴。

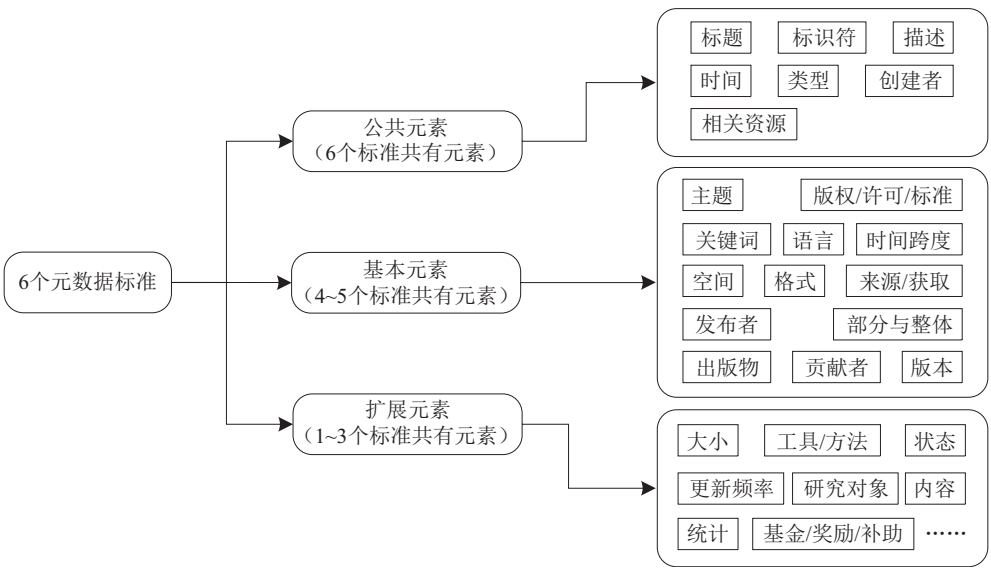


图1 元数据标准元素分类

表3 元数据实际应用比较

| 标准名称                         | 特 点                                                                        | 不 足                                       | 使用难度 | 适用范围                                              | 应用实践举例                                                                                 |
|------------------------------|----------------------------------------------------------------------------|-------------------------------------------|------|---------------------------------------------------|----------------------------------------------------------------------------------------|
| Dublin Core                  | 简易、通用、可扩展                                                                  | 在数据细粒度、数据关系、数据访问等方面的描述不成熟 <sup>[20]</sup> | ☆    | 对于数据集的一般性描述,主要适用于综合性科学数据仓储                        | 3TU.Datacentrum、Edinburgh Data Share、National Science Digital Library Data Repository等 |
| DataCite                     | 从数据的引用和复用出发,提供DOI (Digital Object Identifier) 作为数据集的唯一标识符                  | *                                         | ☆☆   | 对于数据集的准确、一致性描述,主要适用于综合性科学数据仓储                     | CancerData、FAIRDOMHub、Zenodo、Ag Data Commons等                                          |
| Dataverse科学数据仓储元数据标准         | 支持学科专有元数据                                                                  | 元素较多且顺序和层次划分欠清晰 <sup>[8]</sup>            | ☆☆   | 对一个研究课题或一项研究成果的数据和文件集合的一般性描述,主要适用于综合性科学数据仓储       | Dataverse、北京大学开放研究数据平台等                                                                |
| DatA Tag Suite               | 核心实体注重对访问和获取数据集所需相关信息的记录,并将仓储内数据集与出版物、软件以及其他科学数据仓储和数据标准进行关联,扩展实体主要围绕生物医学领域 | 数据集描述的准确性有所欠缺 <sup>[21]</sup>             | ☆☆   | 注重生物医学相关科学数据集的发现和获取,不对数据集进行详细描述,主要适用于生物医学相关科学数据仓储 | DataMed、OmicsDI、ICPSR、FITBR等                                                           |
| W3C HCLS Dataset Description | 描述范围覆盖数据集的各个角度,内容详细具体                                                      | *                                         | ☆☆☆  | 对于生物医学相关科学数据数据集详细、具体的描述,主要适用于生物医学相关科学数据仓储         | Identifier.org、Riken MetaDatabase、Bio2RDF、the EBI RDF Platform等                        |
| Dryad科学数据仓储元数据标准             | 将出版物与数据相关联                                                                 | *                                         | ☆    | 对于数据集的一般性描述,主要适用于综合性科学数据仓储                        | Dryad等                                                                                 |

注: \*代表该项不确定, ☆代表使用难度

基于上述元数据标准的层级深度、结构复杂度、元素丰富度和元素易理解度,可将它们的使用难度分为三级。一星级使用难度相对较低,包括2个标准,分别为Dublin Core和Dryad科学数据仓储元数据标准;二星级包括3个标准,分别为DataCite、Dataverse科学数据仓储元数据标准和DATS;三星级使用难度相对较高,仅有标准W3C HCLS Dataset Description。分析可发现,Dublin Core、Dryad科学数据仓储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,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。Dataverse科学数据仓储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。由此可见,对于科学数据仓储元数据标准的设计,应考虑著录者(包括专业和非专业著录人员及科学数据管理人员)和使用者(指共享科学数据的使用者)的需求、著录对象(被描述的各类科学数据)的特性,并在期间寻求最佳平衡和组配,兼顾元数据标准的描述准确性与应用便捷性,促进科学数据加工的规范化与标准化。

在标准的适用范围方面,Dublin Core、DataCite和Dataverse科学数据仓储元数据标准的元素设置均与学科无关,主要描述数据集的宏观层面,适用于综合性科学数据仓储。DATS和Dryad科学数据仓储元数据标准虽然都是为支持生物医学相关科学数据仓储而设计,但因DATS的核心实体和Dryad全部元素均与学科无关,故也可用于综合性科学数据仓储。W3C HCLS Dataset Description是基于对15个生物医学相关数据集实例的分析而构建的元数据标准,对生物医学科学数据集进行了深入内容层面的组织,主要适用于生物医学相关科学数据仓储。

结合现有元数据标准在科学数据仓储中的实际应用情况,可以分析出元数据标准在科学数据仓储中发挥的作用。本文以应用上述元数据标准的6个科学数据仓储为例,对目前元数据标准在科学数据仓储中发挥作用的功能模块进行分析与总结(见表4)。在数据上传和数据著录模块,科学数据仓储依据元数据标准设定用户上传数据时必须填写的数据的元数据描述,进而形成数据的著录信息,以使用户在不必浏览数据的情况下,能够对数据有基本的了解和认识。DataMed作为数据集检索系统,其并不提供数据上传功能,而是通过收录科学数据仓储,利用DATS元数据标准将各科学数据仓

储内数据集的元数据描述进行统一与规范化,进而实现跨数据仓储的数据集检索。在数据检索和数据浏览模块,各科学数据仓储依据元数据标准设定用户可选的检索项、检索结果筛选项以及分类浏览的分类依据。在数据获取和数据引用模块,许多科学数据仓储不仅提供数据集的获取方式及获取路径,还基于科学数据仓储的元数据描述提供规范的数据引用格式。但也有部分仓储并未提供数据引用模块,如DataMed和EBI RDF Platform。在普通数据转RDF模块,仅利用标准W3C HCLS Dataset Description的科学数据仓储EBI RDF Platform能够实现,即将ChEMBL、Ensembl、UniProt等数据库中的数据转换为RDF格式,从而使这些数据能够通过利用语义网技术进行访问。在API接口和OAI-PMH模块,部分科学数据仓储依据元数据标准为用户提供批量数据上传、数据下载、数据检索以及数据的元数据项下载等功能。图2从左至右分别展示了:①科学数据仓储Zenodo的数据上传界面,包括用户上传数据时必须填元数据描述项、推荐填写元数据描述项和选填元数据描述项,对应元数据标准DataMed的必备元素、推荐元素和可选元素;②科学数据仓储Dryad的数据著录信息详情页面,不仅提供了数据集的基本信息说明,还提供了数据的获取路径以及引用数据的标准格式;③科学数据仓储DataMed的高级检索界面,为用户提供了22个可选检索项,如“Title”“Author”“Description”等。

### 3 对我国科学数据仓储元数据标准构建及应用启示

2002年,在科技部的主导下,我国开始实施科学数据共享工程,制定了数据共享工程的核心元数据标准,并陆续启动医药卫生、气象、农业等9个学科领域科学数据共享中心的建设与共享服务试点。近年来,随着科学研究的不断深入,我国科学数据的数量和规模不断扩大,元数据标准也越来越多。但与国外相比,我国科学数据仓储元数据标准的规范化程度还不够高,体系有待完善,应用也未达到预期。在未来的发展中,我国科学数据仓储不仅要加强自身技术建设,更应增加与国外前沿机构的合作学习,深层次地扩展科学数据仓储元数据标准的构建与应用工作。

(1) 应用已有元数据标准,根据需求进行修改。不断涌现的科学数据仓储和管理平台亟需比较成熟的元

表4 各元数据标准在科学数据仓储中的应用情况

| 科学数据<br>仓储名称            | 元数据<br>标准                    | 数据<br>类别 | 元数据标准作用模块 |          |                |               |          |          |              |           |             |
|-------------------------|------------------------------|----------|-----------|----------|----------------|---------------|----------|----------|--------------|-----------|-------------|
|                         |                              |          | 数据<br>上传  | 数据<br>著录 | 数据检索<br>(高级检索) | 数据分类<br>浏览/筛选 | 数据<br>获取 | 数据<br>引用 | 普通数据<br>转RDF | API<br>接口 | OAI-<br>PMH |
| Edinburgh<br>Data Share | Dublin Core                  | 综合       | ✓         | ✓        | ✓              | ✓             | ✓        | ✓        | ×            | ×         | ×           |
| Zenodo                  | DataCite                     | 综合       | ✓         | ✓        | ×              | ✓             | ✓        | ✓        | ×            | ✓         | ✓           |
| Dataverse               | Dataverse科学数据<br>仓储元数据标准     | 综合       | ✓         | ✓        | ✓              | ✓             | ✓        | ✓        | ×            | ✓         | ×           |
| DataMed                 | DATS                         | 生物医学     | ×         | ✓        | ✓              | ✓             | ✓        | ×        | ×            | ✓         | ×           |
| EBI RDF<br>Platform     | W3C HCLS Dataset Description | 生物信息学    | ×         | ✓        | ✓              | ✓             | ✓        | ×        | ✓            | ✓         | ×           |
| Dryad                   | Dryad科学数据仓<br>储元数据标准         | 生命科学     | ✓         | ✓        | ✓              | ✓             | ✓        | ✓        | ×            | ×         | ×           |

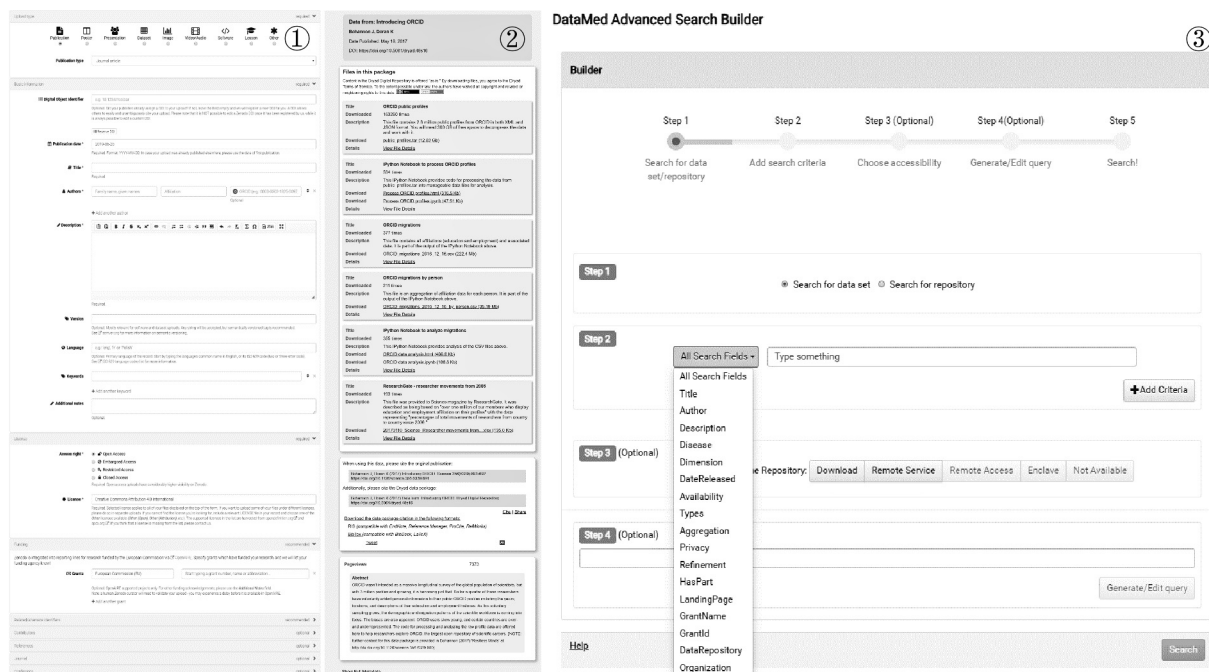


图2 科学数据仓储Zenodo、Dryad和DataMed的部分界面

数据标准进行管理,考虑到元数据格式规范设计和长期维护的复杂性以及国际化环境和互操作的需要,一般选择复用相关领域现有标准。在选择标准时,各平台和仓储应对自身功能目标进行探讨,明确著录科学数据的细粒度和层次,结合各元数据标准的特点与优劣,从而选择出适当的标准,并对选定的标准按需进行调整与修改。如仓储主要收录临床实验数据,并拟从微观层面对数据进行深入内容的描述与组织,则可选用W3C HCLS Dataset Description标准,并在其基础上增加相关专指性描述字段,从而对实验的操作流程、方法以及

时间进行详细说明。

(2) 参考已有元数据标准,构建新的标准体系。现有元数据标准虽种类多样,各具特点,但在实际应用中并不能满足所有科学数据仓储和管理平台的需求。这就需要部分机构、仓储或平台以解决仓储内科学数据管理、共享、应用等方面的问题为导向,考虑科学数据的特有属性、仓储的功能需求和服务对象范围,参考已有标准,构建具有实用性、准确性、可扩展性和前瞻性的科学数据仓储元数据标准,推进科学数据的全生命周期管理与规范引用。

(3) 规整元数据标准应用, 把控数据著录质量。数据著录是科学数据仓储元数据标准应用过程中的一个重要环节, 数据的著录质量严重影响数据在科学数据仓储内的管理效率与后期复用情况。因此, 在未来科学数据仓储元数据标准应用的过程中, 需要严格把控科学数据著录质量, 主要可从以下方面展开: 一是制定科学数据著录的质量控制方案和相关原则, 并以此为指导开展著录工作, 如为数据提交者制订元数据创建指南或者最佳实践等; 二是建立科学数据元数据描述质量评估指标, 将数据著录结果交由专业人员进行审核和评估, 并加强专业人员对元数据质量控制素养的培训, 保证他们对元数据标准以及每个元数据项有较为全面的理解。

## 4 结语

科学数据的开放共享离不开数据仓储和元数据标准的支持, 为数据附加高质量的元数据描述, 是实现科学数据有效存储与管理的基础。目前我国科学数据仓储元数据标准的构建与应用还处于探索阶段, 尚未形成良好的实践成果和合作机制。本文在概述国外典型通用科学数据仓储元数据标准和生物医学科学数据仓储元数据标准的基础上, 围绕它们的内容设计与实际应用开展多维分析, 以期为我国科学数据仓储元数据标准的构建和应用提供理论根基和实践参考。

## 参考文献

- [1] 赵华. 元数据标准与我国农业科学数据元数据[J]. 中国科技资源导刊, 2014 (5): 79-83.
- [2] Re3Data.org [EB/OL]. [2019-06-22]. <https://www.re3data.org/>.
- [3] FAIRsharing [EB/OL]. [2019-06-22]. <https://fairsharing.org/>.
- [4] Digital Curation Center [EB/OL]. [2019-06-22]. <http://www.dcc.ac.uk/>.
- [5] Dublin Core Metadata Element Set [EB/OL]. [2019-06-22]. <http://dublincore.org/documents/dces/>.
- [6] DataCite Metadata Schema [EB/OL]. [2019-06-22]. <https://schema.datacite.org/>.
- [7] Dataverse [EB/OL]. [2019-06-22]. <https://dataverse.harvard.edu/>.
- [8] 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.
- [9] SANSONE S A, GONZALEZBELTRAN A, ROCCASERRA P, et al. DATS, the data tag suite to enable discoverability of datasets [J]. Scientific Data, 2017, 4.
- [10] DataMed DATS specification v2.2 [EB/OL]. [2019-06-22]. <https://zenodo.org/record/438337#WqNqY2luZEY>.
- [11] DataMed [EB/OL]. [2019-06-22]. <https://datamed.org/>.
- [12] OHNO-MACHADO L, SANSONE S A, ALTER G, et al. Finding useful data across multiple biomedical data repositories using DataMed [J]. Nature Genetics, 2018, 49 (6): 816.
- [13] Dataset Descriptions: HCLS Community Profile [EB/OL]. [2019-06-22]. <https://www.w3.org/TR/hcls-dataset/>.
- [14] W3C HCLS Interest Group 2015. Community-Based Guidelines for Describing Biomedical Datasets [EB/OL]. [2019-06-22]. [http://pureapps2.hw.ac.uk/portal/files/9471219/hclsdd\\_revised.pdf](http://pureapps2.hw.ac.uk/portal/files/9471219/hclsdd_revised.pdf).
- [15] Dryad [EB/OL]. [2019-06-22]. <https://datadryad.org>.
- [16] JANE G, HOLLIE CW, SARAH C, et al. A Metadata best practice for a scientific data repository [J]. Journal of Library Metadata, 2009, 9 (3/4): 194-212.
- [17] Dryad Metadata Application Profile [EB/OL]. [2019-06-22]. <http://wiki.datadryad.org/images/8/8b/Dryad3.0.pdf>.
- [18] Zendo-FAIR Principles [EB/OL]. [2019-06-22]. <http://about.zenodo.org/principles/>.
- [19] 李丰梅, 盛梅. 基于XML语言的医学期刊元数据复用[J]. 中华医学图书情报杂志, 2003 (6): 36-38.
- [20] 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.
- [21] GONZALEZBELTRAN A N, CAMPBELL J, DUNN P, et al. Data discovery with DATS: Exemplar adoptions and lessons learned [J]. Journal of the American Medical Informatics Association Jamia, 2018, 25 (1): 13.

## 作者简介

崔佳伟, 女, 1995年生, 硕士研究生, 研究方向: 科学数据管理、医学知识图谱, E-mail: cui.jiawei@imicams.ac.cn。

吴思竹, 女, 1981年生, 博士, 副研究员, 研究方向: 医学科学数据管理与共享、医学信息系统与智能信息处理。

邬金鸣, 女, 1995年生, 硕士研究生, 研究方向: 科学数据共享。

修晓蕾, 女, 1993年生, 硕士研究生, 研究方向: 医学知识图谱、科学数据管理。

钱庆, 男, 1970年生, 硕士, 研究员, 硕士生导师, 通信作者, 研究方向: 医学知识组织、科学数据管理、数字图书馆, E-mail: qian.qing@imicams.ac.cn。

Research and Enlightenment of Metadata Standards for Scientific Data Repository

CUI JiaWei WU SiZhu WU JinMing XIU XiaoLei QIAN Qing

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China)

**Abstract:** The open sharing of scientific data is inseparable from the support of data repositories and metadata standards, and the addition of high-quality metadata descriptions for data is the basis for effective storage and management of scientific data. At present, the development of foreign metadata standards of scientific data repository is relatively mature, but most of the relevant domestic standards are still under construction. This paper intends to systematically summarize and in-depth study on the content design and practical application of metadata standards related to foreign scientific data repository, in order to provide reference for the construction and application of metadata standards for China's scientific data repository.

**Keywords:** Scientific Data; Data Repository; Metadata Standard; Data Management

(收稿日期: 2019-06-03)

■ 书 讯 ■

## 《汉语主题词表》

《汉语主题词表》自1980年问世以后,经1991年进行自然科学版修订,在我国图书情报界发挥了应有作用,曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要,由中国科学技术信息研究所主持,并联合全国图书情报界相关机构,自2009年开始进行重新编制工作,拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表(工程技术卷)》共收录优选词19.6万条,非优选词16.4万条,等同率0.84,在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表(自然科学卷)》共收录专业术语12.4万条,包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域,收词系统、完整,语义关系丰富、严谨,每条词汇都有相应的学科分类号表现其专业属性,并与同义英文术语对应。同时,建立《汉语主题词表》网络服务系统,提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘,是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版,分为13个分册,总定价3 880元。

《汉语主题词表(自然科学卷)》已于2018年5月由科学技术文献出版社出版,分为5个分册,总定价1 247元。两卷均可分册购买。