数据标签集及其适用性探析

宋宁远　刘晶
（武汉大学信息管理学院，武汉 430072）

摘要：科学数据的规范化和结构化描述是发展数据引用的基础，也是推动数据密集型科学研究的重要手段。数据标签集（DATS）是一套面向通用领域和生命科学、环境科学及医学等特定领域，用于描述科学数据的元数据标签集。本文通过分析DATS的元素、属性和关系，比较DATS与元数据框架DataCite和HCLS的异同，归纳DATS的特点，并从数据描述、数据归属、数据关联、数据访问四个方面探讨DATS在数据引用中的适用性。

关键词：数据引用；元数据；数据标签集；数据描述
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2018.06.005

随着科学技术的发展，科学研究范式正在逐步向数据密集型科学研究过渡。作为科研活动的基本素材及产出形式，科学数据既是推动科学研究的重要支撑，也是科学交流体系的关键载体。规范化的数据描述方案一方面可以为用户提供追溯数据源的有效路径，实现科学数据的有效访问，辅助研究者发现科学数据中潜在的科学知识[1]；另一方面，面向科学数据的元数据标准是实现科学数据有效复用与开放共享的基础，对促进以数据为主体的要素出版及语义出版的发展有十分重要的作用[2-3]。

目前，许多科研机构都发布了适用于特定领域的科学数据描述规范及元数据标准。其中，数据标签集（Data Tag Suite，DATS）是由美国国立卫生研究院开发的一套专门用于描述科学数据元数据及其结构的标签集[4-5]，旨在为医学研究数据提供一套通用的数据描述框架，便于科学数据的表示、存储、管理和复用，具有适用范围广、粒度细等特点，能够较规范地描述科学数据。对DATS基本结构进行分析，可以更加明确数据描述机制与方法，为提出更规范、全面的数据描述方案提供参考与借鉴，进而促进科学数据的访问、分享和复用，更好地支撑数据引用。

1 DATS标签集基本组成

DATS定义了用于科学数据集描述的元素（elements）、属性（property）及元素间关系（relations）。DATS元素可分为两大类，分别为核心元素和扩展元素。核心元素不受领域限制，具有普适性特征；扩展元素是根据特定领域（生命科学、环境科学及医学等）科学数据集的具体特点，制定的更为详细的描述方案，具备可扩展性，当出现新的数据描述需求时，可进一步根据具体领域的指定特征进行元素扩充。

1.1 DATS核心元素

DATS核心元素包含20个实体。按照DATS的定义，DATS核心元素被分为3类，分别是数字研究对象（Digital Research Object）、信息实体（Information Entity）和材料（Material）。数字研究对象类元素指科学研究过程中与科学数据集及其创建、存储、出版等活动相关的实体，信息实体类元素侧重对科学数据集的描述，材料类元素侧重描述与数据研究活动相关的机构、个人或物质性材料。DATS核心元素包含的子元素及其具体定义见表1。

DATS核心元素定义了179种属性，分为必选属性（MUST）、推荐属性（SHOULD）和可选属性

表1　DATS核心元素实体定义

元素类型	元素名称	定　义
数字研究对象	数据集（Dataset）	单个或多个代理出版或策划的数据集合
	数据集分布（DatasetDistribution）	数据集的特定可用形式。每个数据集会呈现不同的可用形式，这些形式可能代表数据集的不同格式或不同端点
	数据标准（DataStandard）	数据的格式、报告准则、术语
	数据仓储（DataRepository）	数据集的存储库
	软件（Software）	包含指令和操作集合的数字实体，用来支持计算机运行
	出版物（Publication）	由出版商提供的（数字）文档
信息实体	标识信息（IdentifiersInformatio）	主标识符信息
	替代标识信息（AlternateIdentifiersInformation）	替代标识符信息
	关联标识信息（RelatedIdentifiersInformation）	关联标识符信息
	注释（Annotation）	带有相应本体术语（IRI）的一对值（字符串或数值）
	访问（Access）	提供获取数据集或其他研究对象的方法信息
	基金（Grant）	政府或其他组织为特定目的分配的资金
	许可（License）	规定资源合法操作的法律文件。如重新分配、修改、复用等操作权限，以及资源引用的条件
	维度（Dimension）	实体特征，即被观察实体的个体可计量属性
	数据类型（DataType）	数据的性质
	时间（Date）	日历时间或时间戳
	地点（Place）	有边界的空间实体
材料	材料（Material）	物理实体（如病历）
	人物（Person）	人物实体
	机构（Organization）	机构实体

（MAY）。其中，必选属性23个，如标题（title）、类型（types）等；推荐属性62个，如标识符（identifier）、许可证（licenses）等；可选属性94个，如数据（dates）、空间信息（spatialCoverage）等。

1.2　DATS扩展元素

DATS扩展元素主要面向生命科学、环境科学及医学等领域对于科学数据集描述的特定需求，定义了13个实体。按照DATS标签集的定义，DATS扩展元素被分为4类，分别是流程（Process）、材料（Material）、疾病（Disease）和信息实体（Information Entity）。流程类元素旨在表征科学数据处理的一般过程，材料类元素旨在揭示产生科学数据的特定领域与具体来源，疾病类元素实现了生命科学、环境科学及医学等领域相关的疾病实体描述，信息实体类元素侧重对科学数据集的描述。DATS扩展元素包含的子元素及其具体定义见表2。

DATS扩展元素定义了137种属性。其包含17个必选属性，如名称（name）、输入（input）、输出（output）等；47个推荐属性，如初始数据（startDate）、结束数据（endDate）等；73个可选属性，如关键词（keywords）等。

1.3　DATS元素间关系

DATS定义了30种基本元素关系。根据DATS元素关系在描述实体对象深层语义信息时的功能差异，以及关系定义域（domain）和值域（range）的不同，本文将30种实体关系进行分类。第一类为引用关系。主要描述数据集与出版物间的参考引用。第二类为归属关系。明确了对数据集、出版物等有贡献及责任的人、机构和基金。第三类为流程关系。主要描述科学数据研究过

表2 DATS扩展元素实体定义

元素类型	元素名称	定 义
流程	活动（Activity）	在研究中预定的研究流程
	研究（Study）	从样本中获取数据并尝试得出结论，进而执行计划或设计的过程（或活动）
	数据处理（Treatment）	研究中将研究对象划分到不同情境，或基于特定标准将研究对象划分到不同类别，并比较不同结果的过程（或活动）
	数据获取（DataAcquisition）	通过特定技术测量获取数据的过程（或活动）
	数据分析（DataAnalysis）	转化数据或者生产数据的过程（或活动）
材料	生物实体（BiologicalEntity）	包括生物过程、分子功能或细胞成分
	研究组（StudyGroup）	基于一系列特定标准和规则的研究对象实体集合。同义词：种群、群落
	分子实体（MolecularEntity）	分子维度的物理实体，如蛋白质、核酸和化学物质。可以是非生物的、生物的或者合成的
	结构解剖实体（AnatomicalPart）	多细胞生物体的部分结构
	仪器（Instrument）	帮助代理执行活动的实体
疾病	疾病（Disease）	在人类、动物或植物中结构或功能紊乱，产生特定症状或影响特定部位
信息实体	类别值对（CategoryValuesPair）	数据的扩展机制，允许为实体添加额外属性
	分类信息（TaxonomicInformation）	有机体或生命体的分类和命名

程中不同实体间的关系，揭示科学数据研究活动的执行过程；科学数据研究活动应用的仪器、软件和材料；科学数据研究活动输出的数据集。第四类为存储关系。主要描述数据集在数据仓库中的存储信息。第五类为许可与标准化关系。主要描述数据集、出版物和软件的许可信息。具体DATS元素关系见表3。

DATS在详细定义数据描述的元素同时，还对元素间关系进行较全面的描述。借助这些规范化的关系定义，更有利于实现科学数据间的关联共享。

2 DATS特征分析

为更好地对DATS进行分析，总结归纳其特征，本文通过与DataCite、HCLS等科学数据元数据描述方案及框架进行对比，在对比分析的基础上，总结DATS的特点。

2.1 DATS、DataCite及HCLS的对比

DataCite元数据框架[6]和HCLS元数据模型[7]是目前相对主流的数据描述元数据框架。DataCite元数据框架由推进数据引用的规范化机构DataCite提出，是出于促进数据引用和数据检索的目的，为信息资源描述提供精确统一的标识而创建的科学数据元数据元素集合，适用对象为广义的科学数据集合，即涵盖各种类型的科学数据，而不局限于传统数字型数据。用户可以通过数字资源标识符（Digital Object Identifiers，DOI）永久访问数据资源。HCLS元数据模型由W3C开发，旨在描述健康与生命科学领域的数据集，强调与数据版本、来源、交换、查询和检索等相关的元素及其属性，对数据的描述分为概要层面（Summary Level）、版本层面（Version Level）和分布层面（Distribution Level），借助资源描述框架（Resource Description Framework，RDF）对数据集进行描述。表4从元素适用范围、元素属性、元素关系对DATS、DataCite和HCLS进行比较，总结DATS描述科学数据的特征和优势。

从表4可以发现，DATS、DataCite和HCLS都关注科学数据集的描述与引用问题，但是适用范围和组件结构各有不同，实际应用场景存在一定差异。通过对比分析发现，DATS的适用场景更灵活，组件定义更丰富。

2.2 DATS的特点

通过分析DATS的元素、属性和关系，比较DATS、DataCite及HCLS的适用范围和组件结构，本文总结得出DATS描述科学数据集时具有以下4个特点。

（1）适用范围灵活，能够适应不同领域科学数据描述需求。DATS一方面通过核心元素集定义了通用领域

表3　DATS元素关系

关系分类	关系标签	定义域	值　域
引用关系	Cites	Publication	Publication
	isCitedBy	Publication; DataSet	Publication; DataSet
归属关系	relatedTo	Dataset	Dataset
	hasAuthor	Publication	Person
	acknowledges	Publication	Grant
	awardedBy	Grant	Organisation
	hasAwardee	Grant	Person
	funds	Grant	Study
	isFundedBy	Study	Grant
	isGranteeOf	Person	Grant
	affiliatedTo	Person	Organisation
流程关系	performedBy	Activity	Person
	performedOn	Activity	Date
	executesScheduledEvent	Activity	Design
	uses	Process	Instrument; Software
	isUsedBy	Instrument; Software	Process
	outputOf	Dataset	DataAcquisition; DataAnalysis
	measures	DataAcquisition; DataAnalysis	Dimension
	recruits	Study	Material
	usesReagent	Study	Material
	derivesFrom	Material	Material
	hasCharacteristic	Material	Dimension
	appliedTo	Treatment	Material
	assignedTo	Material	StudyGroup
	bearerOf	Disease	Material
	isAbout	Study; Dataset	BiologicalProcess
存储关系	stores	DataRepository	DataSet
	storedIn	DataSet	DataRepository
许可与标准化关系	hasLicense	Software; Publication; Dataset	License
	conformsTo	Dataset	DataStandard

的数据描述方案，另一方面通过扩展元素集实现对特定领域科学数据的表征。因此，DATS不仅可用于通用领域科学数据集的描述，也能够满足生命科学、环境科学及医学等领域科学数据描述的特定需求。

（2）元素粒度丰富，能够更加精准地描述科学数据。相较于其他数据集描述方案，DATS定义了更丰富的元素和元素间关系，DATS通过对数据集和数据仓储等元素的定义，实现对独立数据、数据集、数据仓储、数据仓储集合等不同存在状态的科学数据规范化描述，并通过hasPart、aggregation、aggregatorOf等多种关系的定义，描述不同粒度科学数据间的语义关系。总体来说，DATS更加适应多粒度科学数据资源描述需求，描述能力更强，能够更精确地实现科学数据的描述与定义。

（3）关联外部资源，全面揭示科学数据的情境信息。DATS以数据集实体为核心，同时定义了用以描述外部资源的实体（如科学研究过程、所用材料等），并通过元素属性和关系实现数据集与外部资源的关联，共同揭示科学数据的语义信息和情境信息，便于研究人员深入理解科学数据的研究过程。

（4）描述更规范，与主流元数据标准进行映射。DATS标签集通过复用已有科学数据描述方案中的部

表4 元数据描述方案对比

		DATS	DataCite	HCLS
提出时间		2015年	2009年	2011年
适用范围	适用领域	通用领域/特定领域（生命科学、环境科学及医学等领域）	通用领域	健康与生命科学领域
	适用对象	数据仓储中任意粒度数据单元，通常是科学数据集	适用于科学研究中任何类型的对象；通常为数据集	健康与生命科学领域科学数据集
元素	元素数量	33个	68个	62个
	必选元素属性	Title；Name；type（s）；creators；access；part of；schedulesDataAcquisition；input；measures；output	Identifier；Creator；Title；Pubilisher；PublicationYear；ResourceType	Type declaration；Title；Description；Publisher；Creators；Version identifier；Version linking；License；File format
	元素类型	Digital Research Object（数字研究对象）、Information Entit（信息实体）、Material（材料）、Process（流程）、Disease（疾病）	–	Core Metadata（核心元数据）、Identifiers（标识符）、Provenance and Change（起源与发展）、Availability/Distributions（可用性与分布）、Statistics（统计）
	元素关系数量	30种	25种	–

分元素及属性，实现DATS标签集与现有元数据模型的映射。复用的元数据既包括通用领域的元数据框架，如通用标记词汇元素标签集（schema.org）[8]和DataCite元数据框架；又包含特定领域科学数据的元数据框架，如健康与生命科学领域描述数据集（HCLS）和ga4gh metadata model[9]。通过与不同元数据框架的映射，DATS实现了对现有资源的复用，对元素的定义更规范，可扩展性更强。

3 DATS适用性分析

通过以上对比分析不难发现，DATS在数据描述、资源关联等方面具有较强的表达能力，为进一步明确DATS的适用性，本文分别从数据描述、数据归属、数据关联、数据访问四个方面对DATS使用情况进行讨论。

3.1 数据描述

实现科学数据准确复用的前提是对科学数据进行准确描述，包括对科学数据的类型、标题、关键词等详细定义与表征。DATS围绕科学数据，定义数据标准、数据类型、维度、注释、时间和地点等元素，同时对包括标题、类型等在内的属性信息进行表征，多维度地定义了科学数据及数据集。此外，DATS还分别定义了数据、数据集、数据仓储、数据仓储集合等元素，多粒度地揭示了科学数据的深层语义特征，奠定数据引用的基础。另外，通过扩展元素，DATS还可以更精确地表征科学数据的领域特征。因此，DATS多维度、多粒度、可扩展地定义领域科学数据的特点，能够更好地支撑数据引用。

3.2 数据归属

Parsons[10]和Borgman[11]将归属与数据问责联系起来，从而明确“谁创造了数据价值”及“谁应该为数据负责”。有研究进一步指出，单一的数据归属机制并不适用于所有科学数据，科学数据应该归属于所有对数据有贡献的个人或机构[12]。对科学数据的归属信息进行准确定义，有助于对数据贡献者研究工作的认可与嘉奖，推动科学数据创建、出版和管理的科研信用及奖赏机制的完善。

DATS定义了机构类元素，详细定义了人物、机构、基金、出版商等实体，并对数据集与机构类元素间的归属关系进行了清晰明确的定义，能够准确表示科学数据的来源信息。DATS不仅能够满足科学数据贡献者的描述需求，而且是对数据贡献者研究工作的认可，是对贡献者所承担责任的监督，能够促进研究机构、个人更加规范地引用和分享科学数据。

3.3 数据关联

数据关联侧重描述科学数据与外部资源，如出版物、数据集等多类型资源的关联。实现广泛的数据关联是支撑数据引用的关键，借助数据间的关联关系，可以准确地定位数据归属，并为数据访问提供路径。

DATS定义了数据集的citation和primaryPublicationg关系，实现了科学数据与出版物实体的关联，增强了科学数据和科技文献的互操作性；DATS核心元素定义了数据集与软件、材料等实体的关联，描述管理或获取科学数据的软件及与科学数据相关的物质实体；DATS扩展元素定义了医学领域的特定资源，揭示医学领域科学数据研究过程中可能涉及的资源实体。DATS通过科学数据与外部资源的关联描述，很大程度上增强了科学数据的情境信息，能够帮助用户获取与科学数据所属领域和研究活动相关的知识，从而更好地促进用户对科学数据的理解和复用。

3.4 数据访问

数据访问侧重科学数据的检索、发现、访问和复用。通常情况下，科学数据就像科学文献中的一个隐藏知识实体，很难被研究人员发现和检索。因此，为更好地实现数据引用，科学数据描述方案尤其需要重视对数据访问的支持。

数据访问主要包括访问限制、载体媒介、存储位置和唯一资源标识符。在访问限制方面，DATS定义了许可元素，描述科学数据的合法操作类型，如重新分配、修改、复用等；DATS还定义了访问元素，如描述科学数据的访问类型（下载、远程访问、局部访问、不可访问等）、科学数据的授权类型（无授权、点击同意授权、注册授权等）、科学数据的物理访问路径（登录页面和标准URL）。在载体媒介方面，DATS定义了数据集分布的formats属性，描述了科学数据的载体媒介，如PDF、XML、Application等。存储位置和唯一资源标识符方面，DATS定义了storedIn属性，描述了科学数据的物理存储位置。然而，对于科学数据而言，物理存储位置（如URL）存在易变和缺乏持久性的问题，而唯一资源标识符能够为资源提供唯一的、独立于物理存储位置且持续不变的标识符，解决物理存储位置的问题。目前比较具有代表性的唯一标识符包括DOI[13]、统一资源标识符（Uniform Resource Name）[14]、Handles[15]等。DATS采用DOI作为数据集的唯一资源标识符，用户可以通过DOI地址持续稳定地访问科学数据，推动用户对科学数据的引用和复用。因此，借助DATS可以更准确地定义科学数据的存储位置及访问方式，更好地适用于数据引用。

4 总结

本文分析了DATS的元素、属性及其关系，并将其与DataCite、HCLS元数据框架进行比较，分析总结DATS在描述科学数据时的特征和优势，最后详细阐述DATS的适用性情况。研究表明，DATS标签集具有强大的数据引用能力，能够满足通用领域和特定领域对于数据管理和数据复用的需求。本研究的不足之处在于对DATS在数据引用中的适用性研究侧重理论层面，未来将继续对DATS在数据引用中的适用性进行实证研究。

参考文献

[1] IQSS. Data citation principals [R]. Boston: Harvard University, 2012: 1-21.

[2] 黄如花，李楠. 国外科学数据引用规范调查分析与启示 [J]. 图书馆学研究，2016（10）：2-9.

[3] 李丹丹，吴振新. 研究数据引用研究 [J]. 图书馆杂志，2013，32（5）：65-71.

[4] HUNGER C，VILANOVA L，PAPAMANTHOU C，et al. DATS-Data Containers for Web Applications [C] // International Conference，2018：722-736.

[5] Working Group 3: descriptive metadata for datasets-dats model [EB/OL]. [2018-05-08]. https://biocaddie.org/group/working-group/working-group-3-descriptive-metadata-datasets.

[6] DataCite Metadata Schema 4.0 [EB/OL].（2016-09-19）[2017-11-13]. http://doi.org/10.5438/0012.

[7] Dataset Descriptions: HCLS Community Profile [EB/OL]. [2017-11-13]. https://www.w3.org/TR/hcls-dataset/.

[8] Schema.org [EB/OL]. [2017-11-13]. https://fairsharing.org/bsg-s000593.

[9] ga4gh-schemas [EB/OL]. [2017-11-13]. https://github.com/ga4gh/ga4gh-schemas.

[10] PARSONS M A. How to cite an earth science data set the National Snow and Ice Data Center [C] // AGU Fall Meeting，2012.

[11] BORGMAN C L. Big data, little data, no data [J]. Big Data Little Data, 2015.

[12] 宋宇，真溱，汤珊红. 数据引用的共同原则 [J]. 情报理论与实践，2015，38（8）：145.

[13] 吴立宗，王亮绪，南卓铜，等. DOI 在数据引用中的应用：问题与建议 [J]. 遥感技术与应用，2013，28（3）：377-382.

[14] Uniform Resource Names（URN）Namespaces [EB/OL]. [2017-06-14]. http://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml.

[15] PANGAEA [EB/OL]. [2018-03-01]. http://www.pangaea.de/.

作者简介

宋宁远，男，1991年生，博士研究生，研究方向：语义出版、知识组织、数字人文，E-mail：songny_whu@126.com。
刘晶，女，1993年生，硕士研究生，研究方向：语义出版、数字阅读。

Data Tag Suits and the Applicability Analysis

SONG NingYuan LIU Jing
(School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: The structuration of scientific data is the foundation of the development of data citation, and it is an important means to promote the data-intensive scientific discovery. Data Tag Suite is a set of tag sets to describe the metadata and structure of the scientific data for general domain and life, environment and biomedical fields. By analyzing the elements, attributes and relationships of DATS and comparing the similarities and differences between DATS, DataCite and HCLS, this paper aims to discusses the applicability of DATS in data citation from the aspects of data description, data attribution, data connection, data access.

Keywords: Data Citation; Metadata; DATS; Data Description

（收稿日期：2018-06-07）

书讯

《中国高被引分析报告2017》

《中国高被引分析报告2017》按理、工、农、医、人文、社科等领域划分为50个学科，综合分析了各个学科的高影响力论文、研究热点与前沿、高影响力期刊、高影响力作者和高影响力科研机构，并以关联图谱的方式展现了多种学术关系，有助于科研人员及时发现并跟踪研究热点，有利于期刊编辑部监测本刊学术影响力，有利于科研管理机构评估科研能力，是高等院校、科研院所及期刊编辑部等相关单位和人员的参考工具书。

该书以“中国知识链接数据库”为依托，数据覆盖我国6 000余种期刊的论文及引文。书中分学科揭示了高影响力的学者、研究机构（大学、研究所、医院等）、地区（省/自治区/直辖市）、学术期刊、图书、外文期刊和会议录，并采用共词分析、共被引分析和合著分析等方法绘制出各学科的前沿主题分布以及作者、机构和期刊间关联的知识图谱。

《中国高被引分析报告2017》由中国科学技术信息研究所编制，曾建勋主编，科学技术文献出版社出版。欢迎业界同仁鉴阅订购。