面向数据重用的科学数据元数据描述规范研究

Research on Metadata Description Specification of Scientific Data for Data Reuse

付齐仙，西北农林科技大学图书馆

施栩婕，上海第二工业大学图书馆

李晨英，中国农业大学图书馆

2023年7月28日

目 录

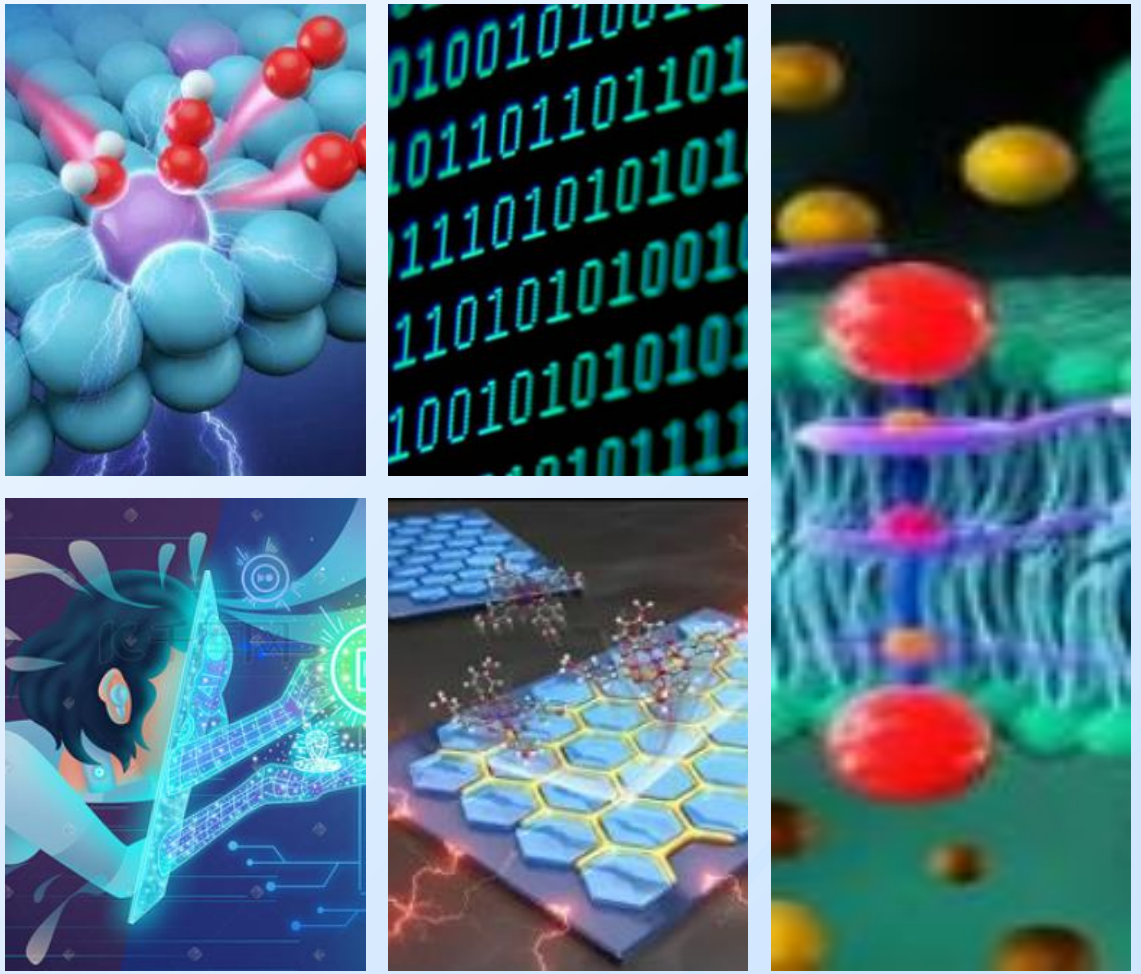
01 研究思路

- 研究背景
- 研究目标
- 研究框架与研究内容

1.1 研究背景

科学数据是加快建设科技强国，实现高水平科技自立自强的关键资源。

科学数据共享进程快速推进，数据重用需求进一步提升，促使科学数据规范化管理和提升科学数据的重用性越来越受到重视。



1.1 研究背景

FAIR原则：实现科学数据的Findable、Accessible、Interoperable、Reusable；FAIR原则共四项15条原则，其中关于元数据的就有13条，给科学数据赋予足够丰富的元数据描述是FAIR原则的核心内容。

结合数据重用的实际提供元数据描述规范，协助科研人员高效率制作高质量元数据是图情学科需要探讨的新问题。

1.2 研究目标

以推动数据重用为目的，开展科学数据元数据服务问题的方法性研究

2021年： 围绕**元数据应用方案设计**问题，提出了适用于高校和科研机构等特定应用场景需求的科学数据资源描述的元数据应用方案设计内容与方法。

2022年： 从**提高科学数据管理效能**的视角，提出了根据学科特点在科学数据管理的主要环节开展元数据服务和建设的实施路径。

元数据实体（metadata entity）是一组可以说明数据相同特性的元数据元素集合，NSTL统一文献元数据标准中也将其描述为“元素集”或“元数据实体对象”，同时元数据元素是元数据的基本单位。

2023年： 以**数据重用**为目标，以元数据实体为切入点展开科学数据元数据描述规范研究。

1.3 研究框架与研究内容

对**科学数据相关元数据标准规范**的内容分析

有哪些元数据实体需要描述?

明晰待描述的元数据实体

与**信息资源、科技资源和DC元数据标准**映射

核心元数据、通用元数据有哪些?

把握科学数据元数据描述重点

对**国家科学数据中心的元数据**进行分析

专用元数据有哪些?

结合科学数据的学科差异性和科研实际，满足描述需求

02 科学数据标准规范中相关元数据提取

- 科学数据标准规范筛选
- 科学数据标准规范中元数据实体提取
- 元数据描述原则分析

2.1 科学数据标准规范筛选

①国家标准全文公开系统，以“元数据”为关键词进行检索；

②在行业标准信息服务平台、国家科技资源共享服务平台检索；

③以2022年U.S.news世界大学排名前100高校的科学数据管理官网检索。

表1　科学数据元数据标准规范获取途径与结果

序号	标准规范类别	查询途径	检索结果	剔除重复项的结果
1	国家标准	国家标准全文公开系统	87	38
2	学科或行业标准	国家科技资源共享服务平台、行业标准信息服务平台	96	74
3	科研机构标准规范	科研机构科学数据管理页面	19	6

2.2 科学数据标准规范相关元数据提取

以上述118条科学数据标准规范（国际、国家、机构）为对象，通过文本内容分析，提取元数据实体并记录其定义和描述实体的子元素；最终提取出38个元数据实体及其所含的1127个子元素。

表2　科学数据标准规范中的元数据实体及频次统计表

序号	元数据实体名称	术语定义	所含元数据元素	频次
1	标识信息	唯一标识资源所需的基本信息	题名、标识符、摘要、关键词、……	9
2	质量信息	提供数据质量的评价信息	评价者、定性评价结果、评价日期……	8
3	空间表示信息	地理科学数据的空间格网、空间矢量表示方法	轴特征、维数、转换参数可用性……	5
……	……	……	……	……
38	数据卷信息	对描述数据集合的信息	名称、格式、说明、数据生成时间……	1

2.3 元数据描述原则分析

- **科学数据元数据框架应该包含核心元数据、通用元数据和专用元数据**

国家标准化管理委员会公布的《元数据标准化基本原则与方法 GB/T30522—2014 》中指出，适用于特定应用情境的元数据框架应该包含核心元数据、通用元数据和专用元数据，以实现不同的数据描述需求。

- **元数据之间应该相互独立**

元数据的设计应该遵循单一性原则，彼此之间相互独立，互不交叉包含，进而全面地描述数据文件或数据集的内容特征和外部特征。

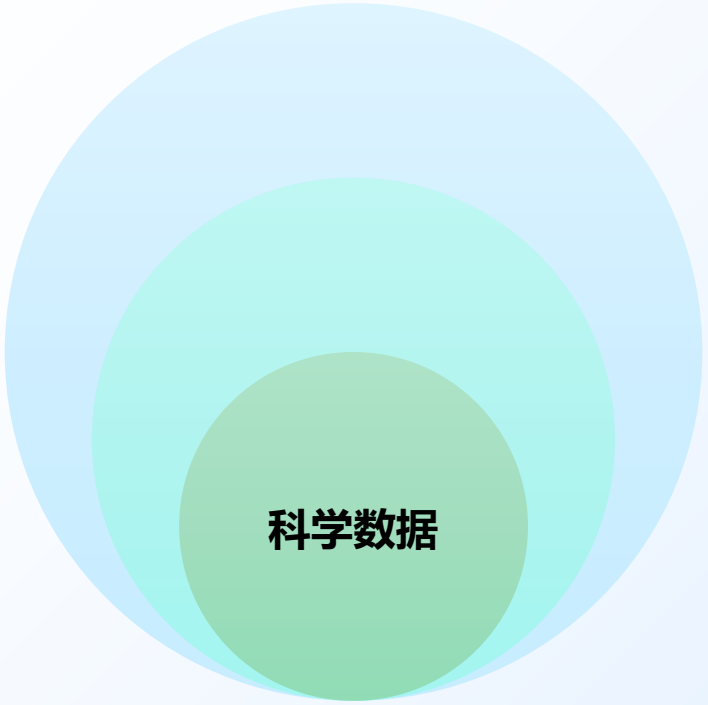
- **元数据描述应该充分考虑科学数据资源特征**

元数据设计与描述时要考虑到科学数据资源生产环境复杂、关联关系密切导致其多源、异质、异构的特点，同时还要尽可能地满足科学数据可发现、可获取、可重用、可互操作的需要。

03 科学数据相关元数据映射

- 科学数据与相关概念逻辑梳理
- 信息资源及科技资源元数据映射及核心元数据
- 都柏林核心元数据映射及通用元数据

3.1 科学数据与相关概念逻辑梳理



- **科学数据：**基础研究、应用研究和试验开发等过程中产生的各类原始数据及其衍生数据。

- **科技资源：**用于科技活动的人力、物力、财力以及组织、管理、信息等要素的总称，主要指研究实验基地和大型科学仪器设备、自然科技资源、科学数据、科技文献、科技成果等。

- **信息资源**：在政治、经济、科技和社会等各领域产生、使用、具有各种载体形式的信息内容。

3.2 信息资源及科技资源元数据映射及核心元数据

- 《信息资源核心元数据 GB/T26816—2011》
- 《科技平台资源核心元数据 GB/T 30523-2014》

表3　科学数据核心元数据实体汇总表

序号	核心元数据实体	元数据实体含义	所含元数据元素
1	日期	与数据发布共享、修改等相关的日期	最新修改日期、发布日期
2	数据类别	数据分类信息	类目名称、类目标准……
3	责任者	与数据相关的责任者信息	联系电话、邮政编码……

3.3 都柏林核心元数据映射及通用元数据

通用元数据实体是对核心元数据实体进行扩展得到的，用以描述科学数据的基本信息。

都柏林核心元素集（Dublin Core Element Set，DC）是对任意资源进行规范化管理的国际通用性元数据标准，不同类型的元数据标准规范都兼容DC标准；且DC核心元数据在DCMI术语表中可以找到修饰和限定的子元素，符合元数据实体的定义和特征；也能够与提取到的大多数元数据实体形成映射关系。

表4　都柏林核心元素集及其修饰元素

序号	DC核心元数据	修饰元素
1	Title	title，alternative title
2	Description	abstract、description、tableOfContents
3	Subject	/
4	Language	/
5	Source	/
6	Relation	references、references、replaces、requires……
7	Coverage	coverage、spatial、temporal
8	Data	dateAccepted、dateCopyrighted、dateSubmitted……
9	Identifier	bibliographicCitation、identifier
10	Type	accrualMethod、accrualPeriodicity、accrualPolicy……
11	Format	extent、format、medium
12	Contributor	/
13	Publisher	/
14	Creator	/
15	Rights	accessRights、license、rights、rightsHolder

3.3 都柏林核心元数据映射及通用元数据

表5　都柏林核心元素集与已有元数据实体映射表

序号	DC核心元素	子元素	核心元数据实体映射	已有元数据实体映射	通用元数据实体扩展
1	Title	√			题名信息
2	Description	√		主题信息	
3	Subject	×		主题信息	
4	Language	×		主题信息	
5	Source	×		项目信息	
6	Relation	√		相关资源信息	
7	Coverage	√		覆盖范围	
8	Data	√	日期		
9	Identifier	√			标识符
10	Type	√	数据类别		
11	Format	√			数据格式
12	Contributor	×	责任者		
13	Publisher	×	责任者		
14	Creator	×	责任者		
15	Rights	√		限制信息	

3.3 都柏林核心元数据映射及通用元数据

表6　科学数据通用元数据实体汇总表

序号	通用元数据实体	元数据实体含义	所含元数据元素
1	数据类别	科学数据的分类信息	类目名称、类目标准……
2	责任者	数据管理相关的责任者信息	名称、联系电话、邮政编码……
3	日期	与数据发布共享、修改等相关的日期	生产日期、发布日期……
4	题名信息	所描述数据对象的标题	中文题名、英文题名
5	主题信息	所描述数据对象的关键词等说明	关键词、摘要、语种
6	项目信息	作为数据生产来源的项目信息说明	项目名称、项目代码……
7	相关资源信息	与所描述数据相关的其他数据资源	数据代码、关联类型
8	覆盖范围	描述对象的时间空间等覆盖范围	时间覆盖范围、空间覆盖范围
9	标识符	所描述数据对象的全局唯一标识符	标识符代码、标识符引用来源
10	限制信息	与所描述数据知识产权等相关的说明	安全限制、法律限制……
11	数据格式	描述数据格式和存储介质的信息	数据量级、格式类型、载体类型

04 国家科学数据中心的元数据分析

- 科学数据中心相关元数据提取
- 科学数据专用元数据

4.1 科学数据中心相关元数据提取

大部分科学数据中心自2019年起为高校、企业和政府机构提供数据服务，提供了元数据模板和数据汇交说明，是目前较为成熟、访问量较高的科学数据共享平台。

微生物 科学数据中心	干细胞与再生医学 科学数据中心	植物 科学数据中心	基础学科 科学数据中心	高能物理 科学数据中心	基因组 科学数据中心
青藏高原 科学数据中心	生态 科学数据中心	冰川冻土沙漠 科学数据中心	地球系统 科学数据中心	海洋 科学数据中心	地球大数据 科学数据中心
化学化工 科学数据中心	凝聚态 科学数据中心	空间 科学数据中心	脑科学 科学数据中心	天文 科学数据中心	对地观测 科学数据中心

4.1 科学数据中心相关元数据提取

来源：18个科学数据中心的元数据模板和数据汇交说明。

表7　国家科学数据中心的数据描述内容汇总表

序号	科学数据描述内容	频次
1	标题、摘要、关键词、数据格式等基本信息	18
2	生产者或发布者的姓名、机构等责任者信息	18
3	项目、基金等来源信息	18
4	论文、专利等关联资源信息	11
5	访问政策、共享范围等访问限制	10
6	软件设备、温度湿度、生产时间等生产说明信息	8
7	数据评论、质量描述等质量说明	5
8	基础数据、实验数据等历史数据信息	5
9	资源下载链接等获取方式	5

4.2 科学数据专用元数据

专用元数据是除了通用元数据之外，结合描述数据资源特征扩展得到的。

科学数据专用元数据需要考虑科学数据资源本身具有的生产环境依赖性强、关联关系复杂等特征。

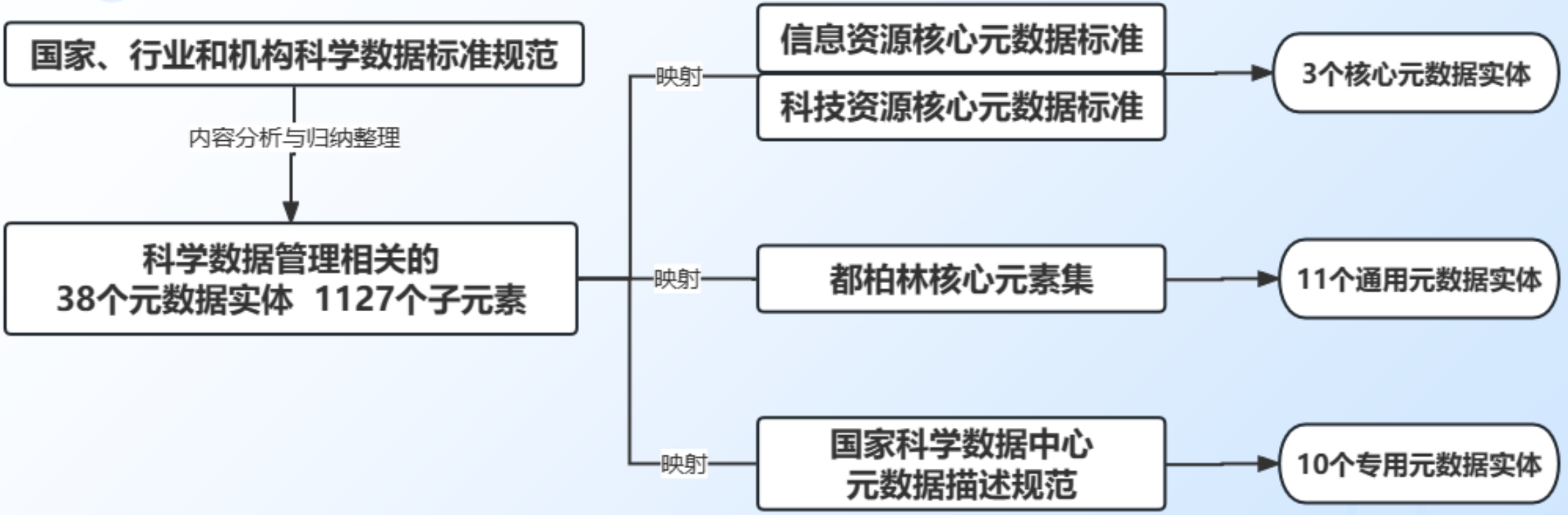
表8　科学数据专用元数据实体汇总表

序号	专用元数据实体	元数据实体含义	所含元数据元素
1	场地信息	生态学和土壤学采集数据时的场地名称、代码、面积等信息说明	场地类型、场地代码……
2	空间表示信息	地理学和生态学采集数据时的空间参照系统等相关信息说明	坐标系统名称、投影参数……
3	图示表达类目信息	地理学对本学科绘图符号及其应用表达的说明	图示表达类目引用……
……	……	……	……
10	地震数据附加属性信息	地理学中对地震数据特有的采样率、地震参数等附加信息的说明	地震参数探测时段……

05 研究总结与不足

- 研究成果小结
- 研究不足与展望

5.1 研究成果小结



研究得到了分层分级的科学数据元数据描述规范，是对繁杂的科学数据元数据的进一步梳理，为后续研制具有可操作性的科学数据元数据描述方法奠定基础。

5.2 研究不足与展望

研究不足：

缺乏系统性开展科学数据管理工作的经验，也未能借助科学数据管理平台进行实证，只是基于已有研究成果进行，有一定客观条件局限性。

研究展望：

1、基于Dataverse、Dryad等开源的科学数据管理平台，进一步验证其普适性。

2、结合数据生命周期、利益相关者等因素，明确不同层级元数据实体的著录时期和著录责任者。

3、2020年以来，国务院多次就“数据作为生产要素之一，正式纳入到国家所定义的要素市场化配置”做出指导意见，开放科学背景下对科学数据管理的要求更应该着眼于有效整合科学资源，为国家科技创新和经济社会发展提供支持，这也是下一步开展科学数据管理相关研究的思路。

敬请各位同仁批评指正！

付齐仙，fuqixian0218@163.com

施栩婕，18257554219@163.com

李晨英，licy@cau.edu.cn