

元数据标准与我国农业科学数据元数据

赵 华 王 健

(中国农业科学院农业信息研究所, 北京 100081)

摘要: 以阐述国外科学数据元数据标准及其内容为出发点, 重点围绕我国农业科学数据元数据的发展现状与存在问题展开分析, 针对存在的问题, 在借鉴国外先进元数据标准的基础上, 提出完善与改进我国农业科学数据元数据的建议。

关键词: 科学数据; 元数据; 元数据标准; 农业科学元数据

中图分类号: G250

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2014.05.013

Analysis of Scientific Metadata Standard and Agricultural Scientific Measures in Our Country

Zhao Hua, Wang Jian

(Agricultural Information Institute, CAAS, Beijing 100081)

Abstract: The paper summarizes several scientific metadata standards and their contents, focuses on analyzing the development and current issues about scientific metadata in agricultural field in china. According to existing problems, puts forward the corresponding measures to improve and perfect current agricultural scientific metadata based on learning from foreign advanced metadata standards.

Keywords: scientific data, metadata, metadata standard, agriculture scientific data

科学数据是指人类社会科技活动(实验、观测、探测、调查等)或通过其他方式所获取的反映客观世界的本质、特征、变化规律等的原始基本数据以及按照不同需求进行系统加工的各类数据集和数据产品^[1]。科学数据不同于一般数据,其来源相对规范,产生于人类认识自然、发展科技的活动过程中,其内容是人类长期科学活动的知识积累,除了具备一般科技资源所具有的专业性、成套性以及系统性等特点外,还具有较高的科学研究价值。因此科学数据的管理与共享成为

了科学界的头等大事,人们经过长时间的研究探索,发现运用元数据来实现科学数据的管理与共享是目前最有效的一种方式。在过去的几十年里,为实现不同学科、不同领域间的数据发现和数据共享,物理学、生命科学和社会科学等领域都建立了相应的元数据标准与规范。本文将对国外科学数据元数据标准及其内容进行分析,重点对我国农业科学数据元数据标准现状与问题进行探讨,并提出相应的改进措施。

作者简介: 赵华*(1980-),女,管理学硕士,农业信息研究所助理研究员,研究方向:科学数据管理;王健(1971-),男,理学博士,农业信息研究所副研究员,硕士生导师,研究方向:信息处理。

基金项目: 国家社会科学基金项目“科学数据用户相关性标准与使用模式”(14BTQ056)。

收稿日期: 2014年3月24日。

1 科学数据元数据概述

1.1 科学数据元数据含义

元数据英文名称Metadata,定义为“关于数据的数据”,或是描述和限定其他数据的数据。元数据作为描述信息资源的特征和属性的结构化数据,具有定位、发现、证明、评估、选择信息资源等功能^[2]。作为一个专用术语,元数据现已广泛应用于各个领域。

科学数据作为一种特殊的信息资源,一方面包括通过科技活动或其他方式所获取到的原始基本数据,另一方面是根据不同科技活动需要加工整理的各类数据集。用于描述此类信息资源的元数据被称为科学数据元数据(简称科学元数据)。科学数据元数据对科学数据形式和内部特征进行详细的描述,为科学数据共享提供信息,其主要目标是提供科学数据资源的全面指南,以使用户对数据资源进行准确、高效与充分的开发与利用。

科学数据元数据通过回答用户的一系列问题:有什么?怎么样?如何获取?怎么使用?使用户可根据需要正确选择、使用、交换数据。同时,元数据也方便了数据管理机构管理海量数据,实现数据库的集成,对数据集进行管理维护和数据目录服务。另外,通过元数据,数据生产者对数据进行生产、加工、更新、归档等工作变得更容易,体现为数据集建立后,随着数据生产人员的变化及时间的流逝,后期接替人员虽对先前数据了解较少,但仍可依据元数据组织数据的生产、更新、加工与增值等工作^[3-4]。

1.2 国外科学数据元数据标准

科学数据与科学研究密切相关,不同领域的科学数据在数据结构、格式、存储和数据处理上存在很大区别,为实现跨领域跨学科的数据共享,需要对元数据进行规范,建立相应的标准。目前,国外对科学数据元数据的认识存在两种观点:一种是图书情报领域的科学数据元数据,另一种是具体研究领域的科学数据元数据,如地理学、生物学、气象学等,不同学科领域的元数据

标准内容存在着很大的差异。

(1)图书情报领域的科学数据元数据标准与规范

图书情报领域把科学数据当作一类特殊的信息资源进行组织与管理,关注的是科学数据的标识和引用信息,认为元数据主要向用户提供科学数据的一些基本属性的描述,目的是方便用户检索数据,并最终发现所需的研究数据。因此数据发现、数据检索是该领域元数据标准实现的首要功能。DataCite核心元数据标准^[5]是典型的元数据标准。建立该标准是为了促进研究数据的出版与引用,其明确规定了数据集标识、责任者、标题、出版单位、出版年份等5项信息在任何情况下都属于元数据的必选内容,而主题信息、贡献者、日期、资源类型、摘要、地理位置等信息在特定条件下属于必选元数据内容。除此之外,元数据内容还可根据需要扩展为对科学数据集的更为详细的描述。

(2)具体研究领域科学数据元数据标准

具体研究领域的元数据功能更为复杂,Qin Jian等结合元数据操作、用户任务和数据研究需求,总结出科学数据元数据数据管理、数据质量控制、数据再利用和数据发现4个功能,其中数据管理功能是所有功能的基础^[6]。比较常见的元数据标准包括地理学科领域FGDC元数据标准^[7]和ISO/TC 211^[8]、生物多样性领域的Darwin Core元数据标准^[9]、生态学领域的Ecological Metadata Language(EML)^[10]、气象学领域元数据标准CF(Climate Forecast)^[11]和社会科学领域的元数据标准Data Documentation Initiative(DDI)^[12]。由于不同研究领域对科学数据的关注点不同,元数据描述的侧重点不同,其内容具有很大的差异(表1)。

其中,FGDC标准和ISO/TC 211标准对科学数据集的描述最为详尽,除了应用于地理数据的描述外,其他许多领域也在应用该标准。EML标准主要围绕与生态数据密切相关的一系列信息展开描述,由EML制定的生态元数据种类包括一般数据集元数据、地理元数据、空间元数据、分类元数据、方法元数据和数据表元数据。CF元

表1 国外几个科学数据元数据标准内容

ISO/TC 211	FGDC标准	DarwinCore标准	EML标准	CF标准	DDI标准
标识信息	标识信息	标识信息	基本信息（关键词、摘要、责任者、使用限制等）	维度描述（时间、空间、数量等）	文档描述信息
数据质量信息	数据质量信息	事件信息	地理信息	变量描述	研究描述信息
数据志信息	空间数据组织信息	生物发现信息	时间信息	坐标描述	数据文件描述
空间数据表示信息	空间参照系统信息	生物分类信息	分类信息	数据描述	变量描述信息
参照系统信息	实体和属性信息	地质环境信息	方法信息（抽样设计、测量方法、质量控制等）	整体属性描述（数据集标题、责任机构、数据来源、参考信息、加工历史、评论信息等）	其他相关材料描述信息
要素分类信息	分发信息	位置信息	结构信息		
分发信息	元数据参考信息	资源关系信息	其他相关信息（数据引文等）		
元数据参考信息		参考信息			

数据标准侧重于描述每一个数据变量的值、数据的时间空间属性和数据处理的方法等信息，该标准更适合于描述观测数据。DDI标准对与数据相关的文档、研究、数据文件和变量等信息展开描述，由于其非模块化的结构，导致其可扩展性差，应用范围较窄，通常只适用于对调查研究数据的描述。秦健对科学数据元数据做过调查研究，发现各标准制定的初衷都没有局限于领域范围内，都强调了标准在其他领域的可扩展性^[13]。但现状是不同的研究领域和团体还是更倾向于制定自己的标准，而没有彻底照搬已有的标准。

2 我国农业科学数据的元数据

2.1 现状

与国外相比，我国科学数据元数据起步较晚，但发展迅速。自2002年在科技部主导下启动实施科学数据共享工程以来，广泛开展我国科学数据元数据研究，各个学科领域的元数据标准相继建立。其中在农业领域，主要是中国农业科学院农业信息研究所、中国科学院区划研究所等研究单位先后提出了包括农业科技信息核心元数据标准框架与农业资源空间信息元数据的行业规范与标准等。随着农业科学数据共享平台的搭建，为了整合我国农业领域科学数据资源，提高数据库建库质量，提升农业科学数据加工的规范化、

标准化，制定了农业科学数据元数据标准和核心元数据标准，主要应用于国家农业科学数据共享中心，适用于农业科学数据共享、编目、元数据交换和网络查询服务。

农业科学数据元数据标准中包含有元数据实体信息和数据集引用信息两类元数据格式，元数据实体信息中规定了必选模块为数据集标识信息、内容信息、分发信息、限制信息和维护信息，对于数据质量信息定义为可选信息。农业科学数据共享核心元数据是唯一标识一个数据集所需的最少元数据内容。核心元数据为用户提供数据的最基本信息，包括数据内容、数据分类、数据存储与访问信息、数据提供单位信息以及数据更新等信息，便于用户查询检索。核心元数据内容由全集元数据内容中的必选项构成^[14]。

2.2 存在问题

(1)元数据标准体系不健全。目前，农业领域只是在通用层面上建立了科学数据元数据标准，尽管包括了全集元数据和核心元数据，但对于该领域专用元数据标准与规范的建设还相对欠缺。农业科学是一个庞杂的学科群，涵盖了生物、环境、经济等学科领域，导致农业领域科学数据数量庞大、种类繁多、内容复杂且具有交叉性，如有关农业生物多样性、农业生态环境、农业土壤肥料等研究领域的科学数据与其他学科的

研究存在着明显的交叉重叠。因此,农业领域要充分实现与其他领域的数据交换与数据共享,需要不断完善元数据标准体系,特别是农业领域专用元数据标准与规范的建立十分重要。

(2)元数据内容不够全面。农业领域的科学数据具有连续性、时间性、空间性、地域性以及种类和要素多样性等特点。这就要求科学数据元数据的内容必须能够向用户提供数据的这些属性方面的信息,如科学数据的采集方法、数据的使用情况等,然而现状是这些属性并没有体现在元数据内容中,不能满足需求。现有的元数据标准中反映数据内容的要素有数据集标题、数据集关键词、摘要等,而且元数据实体中包含了数据内容信息模块,作为数据共享层面的元数据,其对于科学数据内容的体现,只有资源域一项元数据元素。对于了解农业领域的用户来说,很容易理解数据内容,但对领域外的用户可能会需要额外的信息来理解数据内容,如数据集标识使用的要素类目信息或描述数据集数据层内容的信息等。

(3)元数据功能不完善。科学数据元数据的功能已经不仅仅局限于对资源的简单描述或索引,其实现的功能已经发生变化。除了承担描述、定位、搜索、评价和选择资源的作用外,还承担着管理科学数据,维护数据安全和控制数据质量的功能,其中除了数据发现功能外,用户最为关注的功能是评价和验证科学数据。现有的农业科学数据元数据从信息资源发现与使用的角度,对资源的标识、分发、使用和维护限制等一般信息进行了描述,其功能偏重于数据发现和数据使用,并不能满足数据用户评价数据,甚至验证科学数据的需求。

(4)元数据应用不规范。在元数据实际应用中存在随意性,元数据著录者无视元数据记录的完整性,只是站在自身角度上完成元数据元素项的内容,甚至因怕麻烦而省略一些项目。这势必导致元数据的质量和内容的达不到用户的需求。最为典型的是,元数据内容中规定了如何描述数据质量,通过数据志来反映,其中包括了数据源和数据处理步骤。但在实际应用中,元数据著录者

在很多情况下省略了该项内容,事实上数据质量信息是用户评价和使用数据的重要参考依据,尤其对于加工处理过程十分复杂的数据,用户对该项信息尤为关注。

3 改进对策与建议

3.1 加强元数据标准体系建设,满足用户需求

为促进农业科学数据的管理与共享,需要不断地完善农业领域元数据标准体系,目前已经制定了农业科学数据共享公共元数据标准,还需要加强农业领域专用元数据标准建设,在对已有的元数据标准进行整理与分析的基础上,结合实际需要,扩展建设专用元数据标准。比如农业遥感影像专用元数据、土地利用专用元数据、农业气象专用元数据、农业生态专用元数据等标准的构建,使元数据能充分体现出农业科学数据的诸多特性,在适应农业专用领域数据资源的管理与共享的同时,更方便了与农业外其他领域,如地学领域、生态领域、气象领域等进行数据交换。以农业生态数据元数据标准的构建为例,可以借鉴国外生态领域元数据标准EML,对生态数据的一系列必要信息,包括变量定义,方法单元,数据收集时间、地点,数据收集者标识信息,抽样设计等进行描述,以此建立的元数据标准更能满足用户的需求。

3.2 完善元数据内容,提升元数据功能

元数据的内容取决于元数据所发挥的功能,相比其他领域元数据,科学数据元数据发挥的功能更为复杂,包括数据管理、数据发现、数据评价和数据再利用等。目前,农业科学数据元数据在数据管理和数据发现方面发挥了较好的作用,但在数据评价上还有待加强。因此需要对元数据内容加以完善,来提升元数据功能。

完善农业科学数据元数据内容的具体措施包括:(1)对农业科学数据的特征与属性进行分析是完善元数据内容的必要步骤。元数据通常向用户提供科学数据的外在特征(大小、格式、语言等)、知识产权、内容信息等方面的描述外,也包括学科情境方面的信息。由于农业科学数据种

类多, 涉及面广, 与生态、生物和经济等多个学科领域相交叉, 数据类型既包括实验数据又包括观测数据和统计数据, 数据既具复杂性又具交叉性。在元数据对科学数据各方面属性描述时, 考虑增加描述项。例如在对实验数据进行描述时, 其数据内容方面可以增加实验背景、方法和仪器等方面的描述, 对于观测数据而言, 其数据内容可以增加数据采集方法、仪器、观测时间和空间等描述信息, 通过这些内容, 用户可以对数据内容有更深的了解。(2) 增加第三方评价元数据。目前, 在图书馆领域元数据中, 专门出现了评价类元数据, 其内容涉及了同行专家对信息资源的评价、用户使用信息后的反馈与评价信息等, 尤其是在教育信息资源方面, 用户在评价和选择资源时往往很依赖这些评价类元数据内容, 也叫第三方元数据或注释类元数据^[15]。实际上, 对于科学数据而言, 同行专家评价信息和用户对资源的使用评价信息, 在用户做出选择使用数据前通常也会发挥着一定的作用。因此, 元数据内容可以考虑增加用户评价类元数据。(3) 从农业科学数据用户角度着手, 分析用户的信息需求, 深入研究用户查找数据的行为, 分析用户在查找和评价科学数据时的关注点, 研究用户判断科学数据的相关性和可用性的标准, 明确元数据的各个内容项在用户发现和评价数据时如何发挥作用。只有在回答了这些问题的基础上, 才能更好地改进现有的农业科学数据元数据标准, 完善元数据内容。

3.3 规范元数据的应用, 提高元数据质量

元数据著录是元数据应用过程中的一个重要环节, 为保证元数据质量, 需要对元数据著录进行规范。具体的措施包括3个方面: 一是要制定详细的著录规则, 指导元数据的著录; 二是要加强元数据标准的宣传, 对元数据著录人员进行培训, 使元数据著录人员深刻理解每个元素项的意义和作用, 以避免著录过程中的随意性; 三是要开发元数据著录管理系统, 提高元数据著录效率。元数据著录管理系统可考虑提供各级检索功能, 实现从元数据值到数据实体的定位, 一改以

往枯燥乏味的元数据著录过程, 在不同的区域显示不同的著录元素, 并支持专家对元数据信息进行审核。此外, 由于科学数据元数据的自动生成功能还未能实现, 随着这方面研究的不断深入, 相信势必会改进元数据著录工作, 提高元数据的质量。

参考文献

- [1] 何建邦, 曹彦荣, 马立广, 等. SDS/T 2122-2004, 科学数据共享工程技术标准[S]. 2007.
- [2] 刘嘉. 元数据导论[M]. 北京: 华艺出版社, 2002.
- [3] 戴爱德, 李德兴. W D C 中国地质科学数据网的核心元数据[J]. 国土资源信息化, 2004(4): 8-10.
- [4] 马伯元. 元数据目录技术在科学数据整合中的应用[J]. 石家庄铁路职业技术学院学报, 2007, 6(4): 90-93.
- [5] Starr J, Gastl A. isCitedBy: A metadata scheme for DataCite[J]. California Digital Library, 2011, 17: 1-6.
- [6] Qin J, Ball A, Greenberg J. Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data[C]//DC-2012, Kuching, Sarawak, Malaysia. 2012
- [7] Federal Geographic Data Committee. FGDC Metadata Quick Guide[EB/OL]. [2014-03-20]. <https://www.fgdc.gov/metadata/documents/MetadataQuickGuide.pdf>.
- [8] Di L. The Development of Remote-Sensing Related Standards at FGDC, OGC, and ISO TC 211[C]//Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International. IEEE, 2003, 1: 643-647.
- [9] Wiecezorek J, Bloom D, Guralnick R, et al. Darwin Core: An Evolving Community-developed Biodiversity Data Standard[J]. PLoS ONE, 2012, 7(1): 1-7.
- [10] Eric H Fegraus, Andelman S, Jones M B, et al. Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation Read[J]. Bulletin of the Ecological Society of America, 2005, 86(3): 158-168.
- [11] Gregory J. The CF Metadata Standard[J]. CLIVAR Exchange, 2003, 8(4): 1-5.
- [12] Ryssevik J. The Data Documentation Initiative (DDI) Metadata Specification[J/OL]. Ann Arbor, MI: Data Documentation Alliance, 2001[2014-01-01]. http://www.ddialliance.org/sites/default/files/ryssevik_0.pdf.

(下转第103页)

- [21] Wartena C. HsH: Estimating Semantic Similarity of Words and Short Phrases with Frequency Normalized Distance Measures[C]// Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Atlanta, Georgia, USA, 2013: 48.
- [22] Lau J H, Cook P, Baldwin T. Unimelb: Topic Modelling-based Word Sense Induction[C]. Second Joint Conference on Lexical and Computational Semantics (*SEM), 2013, 2: 307-311.
- [23] 孙辉. 事件时序关系识别的研究与实现[D]. 哈尔滨: 哈尔滨工业大学, 2010.
- [24] 王凤玲. 基于条件随机域模型的英语时间表达式识别研究[J]. 电子技术, 2012, 39(5): 8-10.
- [25] Frank P D D A. Projecting Temporal Annotations Across Languages[D]. Saarbrücken, Germany: Universität des Saarlandes, 2007.
- [26] Caselli T, Prodanof I. TETI: A TimeML Compliant TimeEx Tagger for Italian[C]// Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on. IEEE, 2009: 185-192.
- [27] Sauri R. Annotating Temporal Relations in Catalan and Spanish. TimeML Annotation Guidelines (Version TempEval-2010)[J/OL]. [2014-03-22]. http://comunicacio.barcelonamedia.org/technical_reports/BM2010_04.pdf.
- [28] Bittar A, Amsili P, Denis P, et al. French TimeBank: An ISO-TimeML Annotated Reference Corpus[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Short Papers), 2011: 130-134.
- [29] Spreyer K, Frank A. Projection-based Acquisition of a Temporal Labeller[C]. IJCNLP, 2008: 489-496.
- [30] Sauri R, Badia T. Spanish TimeBank 1.0 Corpus documentation[J/OL]. [2014-03-22]. https://catalog.ldc.upenn.edu/docs/LDC2012T12/README_spanish-TimeBank.pdf.
- [31] Jarzębowski P, Przepiórkowski A. Temporal Information Extraction with Cross-Language Projected Data[M]// Advances in Natural Language Processing. Berlin, Germany: Springer Berlin Heidelberg, 2012: 198-209.
- [32] Li W, Wong K F. A Word-based Approach for Modeling and Discovering Temporal Relations Embedded in Chinese Sentences[J]. ACM Transactions on Asian Language Information Processing (TALIP), 2002(3): 173-206.
- [33] Li W, Wong K F, Cao G, et al. Applying Machine Learning to Chinese Temporal Relation Resolution[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 582.
- [34] Cheng Y, Asahara M, Matsumoto Y. Constructing a Temporal Relation Tagged Corpus of Chinese Based on Dependency Structure Analysis[C]// Temporal Representation and Reasoning, 14th International Symposium on. IEEE, 2007: 59-69.
- [35] Cheng Y, Asahara M, Matsumoto Y. Use of Event Types for Temporal Relation Identification in Chinese Text[C]. IJCNLP, 2008: 31-38.
- [36] Llorens H, Saquete E, Navarro B, et al. Data-driven Approach Based on Semantic Roles for Recognizing Temporal Expressions and Events in Chinese[M]// Natural Language Processing and Information Systems. Berlin, Germany: Springer Berlin Heidelberg, 2011: 88-99.
- [37] Entrikin R. Applying Machine Learning to Usage of Aspect Markers in Chinese Text[D]. Massachusetts: Brandeis University, 2012.
- [38] Zhang X, Liu Z, Liu W, et al. Chinese Event Classification for Event Ontology Construction[J]. Journal of Computational Information Systems, 2013, 9(9): 3511-3519.

(上接第83页)

- [13] Qin J, Li K. How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure[EB/OL]. [2013-10-28]. <http://dcevents.dublincore.org/IntConf/index/pages/view/2013-peer-Abstracts#Qin>.
- [14] 孟宪学, 赵瑞雪, 周国民, 等. 农业科学数据核心元数据[M]. 北京: 中国农业科学技术出版社, 2005.
- [15] You S. Evaluative Metadata in Educational Digital Libraries: How Users Use Evaluative Metadata in the Process of Document Selection[J]. TCDL Bulletin, 2010, 4(2): 1-11.