植物遗传资源学报 2012,13(5):704-708
Journal of Plant Genetic Resources

农作物种质资源调查数据标准制定与共享

司海平[1],刘俊辉[2],马新明[1],方　汐[3],曹永生[3]
([1]河南农业大学信息与管理科学学院,郑州 450002;[2]郑州牧业工程高等专科学校信息工程系,郑州 450011;
[3]中国农业科学院作物科学研究所,北京 100081)

摘要:系统提取并分析了农作物种质资源普查数据、调查数据、评价数据和保存数据等数据信息,采用基于数据元技术方法制定了农作物种质资源调查数据标准和数据元目录;定义了种质资源调查数据集以及对象和属性的映射关系;给出了基于XML数据标准存储及交换策略。标准的制定使农作物种质资源调查在"数据层"上达到统一,规范了数据库构建,促进了农作物种质资源调查数据的整合和共享。

关键词:农作物;数据元;XML;数据标准;数据共享

Establishment of Crop Germplasm Resources Investigation Data Standards and Sharing Strategy

SI Hai-ping[1], LIU Jun-hui[2], MA Xin-ming[1], FANG Wei[3], CAO Yong-sheng[3]
([1]*College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002*;
[2]*College of Animal Husbandry Engineering Department of Information Engineering, Zhengzhou 450011*;
[3]*Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081*)

Abstract: General survey data, investigation data, evaluation data, conservation data are abstracted and analyzed systematically in this paper. Data element technology is used to construct the data standards and data element catalogue of crop germplasm resources investigation; the complete data set of crop germplasm resources investigation and the mapping relationships between objects and attributes are defined later. And data storage and exchange strategies based on XML are given lastly. The data standards establishment is to achieve uniformity and standardization of crop germaplasm resources investigation on the "data layer", and regulate the database construction, promote the survey data of agricultural seed substance resources integration and sharing.

Key words: Crop; Data element; XML; Data normalization; Data sharing

农作物种质资源调查是获取种质资源分布和数量信息的主要手段[1]。20 世纪 50 年代以来,我国先后组织了多次农作物种质资源调查,在资源调查和信息共享方面取得了重大进展和显著成效,积累了大量的种质资源数据[2-4]。但在资源调查项目中,由于缺乏统一的调查数据规范标准,同一领域存在不同数据集,同一概念的数据元命名、定义、表达格式描述不一致等问题,增加了数据资源的整合和共享难度,造成资源的浪费。

随着数据标准化和规范化进程的不断开展,许多农业相关数据标准制定研究也相继呈现,取得了很多成果,如《农作物种质资源技术规范丛书》的编制,荔枝种质资源描述规范及数据标准研究[5]、农业统计信息数据元标准化[6]、药用植物资源的整合信息数据标准规范[7]、武夷山生物多样性数据共享标准体系[8]、数字农业信息标准[9-10]、热带农业科学数据标准化建设及对策研究等[11]。国际方面,作物数据标准制定方面研究比较早,农业标准化体系较

收稿日期:2011-12-06　　**修回日期**:2012-03-29
基金项目:农业部作物种质资源保护与利用专项(NB 2011-2130135-25-11)
作者简介:司海平,博士。研究方向:作物信息技术。E-mail:sihaiping@yahoo.com.cn
通讯作者:曹永生,研究员。研究方向:信息管理和信息系统研究。E-mail:caoys@mail.caas.net.cn

为完善[12],成立了一系列的标准化组织如:1947 年 FAO(联合国粮农组织)成立国际食品法典委员会(CAC)专门负责农业方面的标准化工作,标准化的理论研究方面,在《标准化的目的与原理》("The aims and principles for standardization" by Terrence Robert Beaumont Sanders,1972)和在《工业标准化原理》(松浦四郎,日本,1981)中详细地阐述了标准制定的目的、过程和方法等,为标准化的制定给以理论指导。

本研究依据数据元技术[13-14]原理,从农作物种质资源调查数据的值域、语义和句法 3 个层面出发,针对农作物种质资源领域不同数据集、不同领域相关数据集同一概念数据元的命名、定义、分类标识、表达格式等进行统一和规范,给出了农作物种质资源调查数据属性描述符,数据分组的 UML 模型和基于 XML/XSD 置换策略。农作物种质资源调查数据标准的制定对提高种质资源调查数据的质量和规范数据共享等方面具有十分重要的意义,同时也是农作物种质资源调查平台建设的重要内容,是国家自然科技资源共享平台建设的重要组成部分。

1　数据标准与数据元技术

1.1　数据标准

标准是对重复性事物和概念所做的统一规定,它以科学、技术和实践经验的综合成果为基础,经有关方面协商一致,由主管机构批准,以特定形式发布,作为共同遵守的准则和依据[15]。标准化是"在经济、技术、科学及管理等社会实践中,对重复性事物和概念通过制定、发布和实施标准,达到统一,以获得最佳秩序和社会效益"。标准化是人们在长期生产实践中探索创立起来的技术领域,是规模化工农业经济中的一种重要的应用技术。

1.2　数据元技术

数据元是通过一组属性描述其定义、标识、表示和允许值的数据单元[16],在特定的语义环境中是不可再分的最小数据单元,由对象类、特性和表示 3 部分组成:(1)对象类:是思想、概念和现实世界里事物的有限集合,具有清晰的边界和含义,也可以表现为需要对其进行研究、收集或存储相关数据的事物,例如种质类型、采集者、保存单位等。(2)特性:用来区分和描述对象,是对象类中的所有成员共同具有的,并且有别于其他对象类的显著特征,如生长特性、抗逆性等特性。(3)表示:用来描述数据被表达的方式,它与数据元的值域关系密切,数据元的值域是数据元的所有允许值的集合。表示有多种方式如货币、日期、图片等。

农作物种质资源调查数据元范围限定在农作物种质资源调查应用领域,是最小数据处理单元。数据元的表示规范是通过描述数据元的一系列属性来实现的,属性主要由包括标识类属性、定义类属性、关系类属性、表示类属性和管理类属性组成,一些属性可以根据需要作为可选项[17]。数据元属性依照一种标准方式来注册和控制,以便数据元字典中的数据元在信息交换中保持一致性,并且能够在不同的数据管理环境中进行数据元比较。

1.3　数据元的标准化

数据标准化是一种按照预定规程对共享数据实施规范化管理的过程,对象是数据元。数据信息系统中不同业务系统的数据整合不可避免会遇到异类数据库的协同工作,而统一的数据标准是协同工作的基础。数据元标准化即是数据自身的标准化,是对数据元的名称、定义、描述、分类、表示和注册等属性制定统一的标准,并加以贯彻、实施的过程。数据元的发展随着信息化的深入逐步开展,是按照制度原则编制数据元的标准化说明,并将标准数据文档提交到数据标准管理机构进行登记、审核、注册的过程,数据元标准化有助于构建高效的数据模型。

农作物种质资源调查数据元标准的建立是实现农作物数据标准化过程中的关键部分,研究对象是数据元和元数据,目标是建立标准化的信息表达方法和存储交换格式,以实现农作物种质资源信息正确表达及传输。

2　基于数据元技术的农作物种质资源调查数据标准制定

2.1　解决思路和方案

农作物种质资源调查数据标准针对农作物种质资源调查和数据整合共享需求而提出,通过数据元标准化的基本原则和方法,确定农作物种质资源调查基本数据元、类目分组、表示及存储交换等标准规范,实现农作物数据标准化。通过对数据元进行分类和编码并使之目录化,为农作物种质资源调查信息系统和其他数据库系统提供统一规范的接口。农作物种质资源调查数据标准制定

流程如图 1 所示:(1)借鉴和参考国内外相关数据标准和农作物种质资源技术规范丛书;(2)分析、归类、聚合现有种质资源数据库数据,进行属性语义分析、扩展和限定规则分析,制定农作物种质资源调查数据集;(3)分析和规范化数据属性,建立基于数据元的农作物种质资源调查数据标准,获得元数据的一致理解;(4)设计数据标准框架,形成数据标准框架初稿,然后试点应用,专家咨询穿插整个过程,同时标准的制定也是修改完善的循环过程。

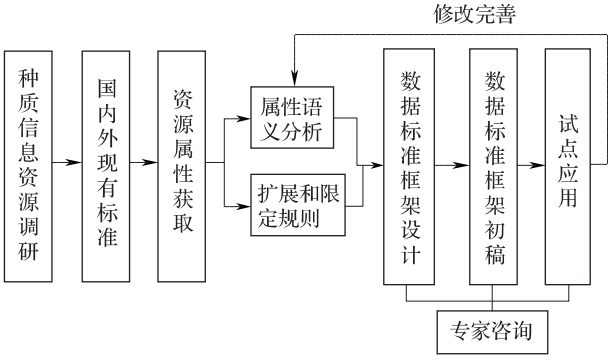

图 1 农作物种质资源数据标准制定流程

Fig. 1 The flow of crop germplasm resources data standards establishment

2.2 农作物种质资源数据元提取

主要采用自下而上提取法对数据元进行了提取并规范化描述。基于"数据元是自然界中存在的对象类的特性的表示"这种认识,识别出对象类,然后寻找对象类的特性并表示出来,如图 2 所示。数据的来源包括国家农作物种质资源数据库系统的 ERD[18] 图的实体对象、属性以及本领域的资源调查数据的数据项,抽取的数据项通过增加对象类词及表示类词形成了数据元。

农作物种质资源调查数据标准数据集可以表示为一个三元组(公式 1):明确需要收集数据的范围,定义数据元的完备数据集。

$$DE = \{U, A, V \mid U \cap A \cap V \neq \Phi\}\ f: u * a \rightarrow V \quad (1)$$

在该三元组中 U = {u1,u2,u3,…,un} 是有限非空集,U 为对象类,元素 u1,u2…un 称为农作物种质资源调查的对象。A = {al,a2,a3,…an} 是一个有限非空集,A 中的元素代表数据元的属性;V = {a(u) | u ∈ U} 是属性 a 的值域。f 表示一个信息函数,它为每个对象的属性赋予特定值,对于每一个 a ∈ A,有一个映射 f,即:

$$\forall a \in A, x \in U, f(a) \in Va \quad (2)$$

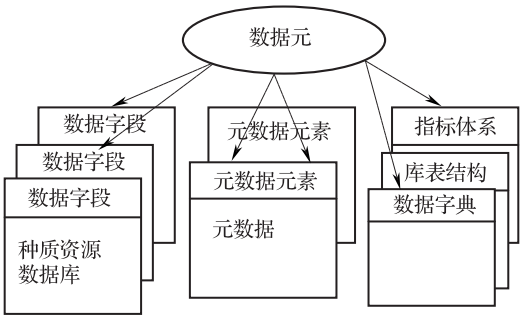

图 2 数据元提取领域范围

Fig. 2 Data element extract range

2.3 数据元表示的基本属性

数据元表示规范是通过数据元的一系列属性来实现的,这些属性实际上是数据元的元数据,表 1 给出了本标准中描述数据元属性的描述符集。

表 1 农作物种质资源调查数据元属性描述符

Table 1 Crop germplasm resources investigation data element attribution descriptor

序号 No.	名称 Descriptor	约束 Condition	定义 Definition
1	中文名称	M	数据元的中文名称
2	内部标识符	M	数据元分配与语言无关的唯一标识
3	英文名称	M	数据元的英文名称
4	定义	M	数据元基本含义说明
5	对象类	O	数据元所属类别
6	表示	M	数据元的表示格式
7	数据类型	M	数据元值域的表示形式
8	值域	O	数据元取值范围或相关内容的说明,如采取的代码标准,枚举类型给出的取值等
9	计量单位	O	如:cm,kg 等
10	同义名称	C	数据元的其它同义名称
11	备注	O	对数据元的补充说明

M:必选;C:条件选;O:可选 M:must,C:conditional,O:optional

2.4 数据元分类和表达格式

数据元分类可以帮助用户从众多的数据元中找出某个具体数据元,方便对数据元的管理。本标准按照数据元的属性成分分为以下 6 个类别,如图 3 所示:基本调查信息、采集信息、地理信息、环境信息、特征特性信息、其他信息。数据元标识符由 6 位组成,前 2 位为英文字母,标识数据元的分类,后 4 位数字为顺序号。其中 BZ:基本调查信息、CL:采集信息、GO:地理信息、EV:环境信息 、TZ:特征特性信息、OT:其他信息

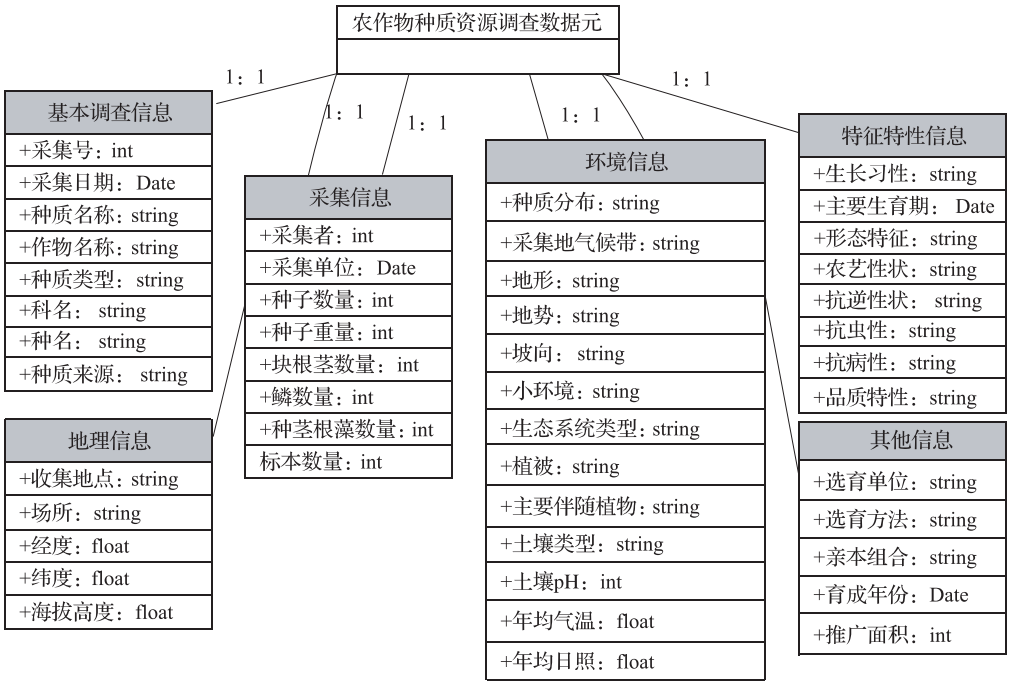

图 3　数据分组的 UML 模型

Fig. 3　UML model of data grouping

字符和数值数据格式长度采用固定长度和可变长度两种方式,固定长度是在数据类型表示符后直接给出字符长度的数目;可变长度分为不超过最大字符数和在定义的最小和最大字符数目之间。

数据类型是用于表示数据元的符号、字符等,取值包括字符、数值型、日期型、日期时间型、枚举型和布尔型等,字符和数值数据格式采用固定和可变长度两种方式。固定长度是在数据类型表示符后直接给出字符长度的数目;可变长度分为不超过最大字符数和在定义的最小和最大字符数目之间。

数据元描述实例如表 2 所示,数据元目录的建立是为了满足农作物种质资源调查信息系统和其他共享工程对数据元进行检索查看的需要。标准化后的数据元需上报有关部门进行审核,然后以数据元目录的形式发布。

表 2　数据元描述实例

Table 2　Instance of data element description

中文名称	收集地点
Descriptor in Chinese	Collection site
内部标识符	GO1001
英文名称	Collecting site
定义	采集地的省、县、乡、村的名称
对象类	地理信息
表示	名称
数据类型	字符型
数据格式	S100
同义词	采集地点
备注	

2.5　农作物种质资源数据标准元数据 XML/XSD 置换

XML(Extensible Markup Language)[19-20]是一种简单的数据存储语言,是 Internet 环境中跨平台的,依赖于内容的技术,是当前处理结构化文档信息的有力工具。使用 XML 方式来存储和备份农作物种质资源调查数据标准,可以灵活的将数据信息转移到其他平台和数据库系统中,同时占用的空间也非常小。通过直接存储为 XML 格式或将数据库文件用 XML 格式导出,数据可以很容易的导入到信息系统或其他数据库中,从而达到数据跨平台的不同数据库之间的数据交换目的。如使用 XML 技术将 oracle 数据库中的数据封装在一个 XML 文件中就可以被另外的数据库系统如 SQL Server 读入。

XML Schema 是使用一系列的元素来描述 XML 结构、约束等因素的语言,其中最重要的功能就是对数据类型的支持。通过对数据类型的支持,XML Schema 可更容易地描述允许的农作物种质资源调查数据标准文档内容,验证数据的正确性,与来自数据库的数据一并工作,定义数据约束(用来约束数据类型的容许值),可更容易地定义数据模型(或称数据格式)等,这使得农作物种质资源调查数据标准可以较容易的在不同数据类型间约束和转换数据。转换过程需将 UML 模型中每个类、类中的属性都转换为 XML Schema 代码段。

种质分布(类)转换 XML Schema:

```
1. <xs:element name = "种质分布" >
2. <xs:complexType >
3. <xs:sequence >
4. <xs:attributeGroup ref = "attrgroup"/ >
5. </xs:sequence >
6. <xs:simpleType name = "值域" >
7.       <xs:restriction base = "xs:string" >
8.       <xs:enumeration value = "广"/ >
9.       <xs:enumeration value = "窄"/ >
10.       <xs:enumeration value = "少"/ >
11.       </xs:restriction >
12. </xs:simpleType >
13. </xs:complexType >
14. </xs:element >
```

种质分布(属性)转换 XML Schema:

```
1. <xs:attributeGroup name = "attrgroup" >
2. <xs:attribute name = "中文名称"
3. type = "xs:string" use = "required"/ >
4. <xs:attribute name = "内部标识符" >
5. <xs:restriction base = "xs:string"
6. use = "required" >
7. <xs:length value = "6"/ >
8. </xs:restriction >
9. </xs:attribute >
10. <xs:attribute name = " 定义" type = "xs:
11. string" use = "required" >
12. <xs:attribute name = "表示"
13. type = "xs:string" use = "required"/ >
14. <xs:attribute name = "数据类型"
15. type = "xs:string" use = "required"/ >
16. </xs:attributeGroup >
```

通过 XSD(XML Schema Definition)对一类 XML 文档进行约束、确定其结构以及元素、属性及数据类型。在用 XSD 方法将数据元 UML 模型转换成相应的 XML 描述文档过程中,需要将 UML 模型中的类和属性都转换为 XML Schema 相应代码段,通过调用已定义的必选属性 attrgroup 和可选属性值域的扩展,来限定种质分布元数据的 XML 数据输入,值域定义为枚举类型,从"广"、"窄"和"少"之间进行选取。

3 结论

数据标准制定是信息化管理的重要组成部分,能够促进数据资源的整合、交换及共享。本研究从农作物种质资源调查实际出发,结合数据元技术特点,给出了农作物种质资源调查数据标准的具体内容和方法,对农作物种质资源调查数据进行了分类,建立了数据元目录,并阐述了基于 XML 数据标准存储和交换的策略。从农作物种质资源调查数据的分类与编码、数据格式、数据交换格式等方面进行标准化工作,确保农作物种质资源调查工作的一致性和规范性,为农作物种质资源数据的获取和利用提供在"数据层"上的统一。

参考文献

[1] 卢新雄,曹永生. 作物种质资源保存现状与展望[J]. 中国农业科技导报,2001,3(3):43-47

[2] 曹永生,方沩. 国家农作物种质资源平台的建立和应用[J]. 生物多样性,2010,18(5):454-460

[3] 王述民,张宗文. 世界粮食和农业植物遗传资源保护与利用现状[J]. 植物遗传资源学报,2011,12(3):325-338

[4] 司海平,方沩,唐鹏,曹永生. 基于 SOA 的农作物种质资源调查信息系统研究[J]. 植物遗传资源学报,2010,11(5):517-521

[5] 陈洁珍. 荔枝种质资源描述规范及数据标准研究[D]. 长沙:湖南农业大学,2005

[6] 侯振宇. 农业统计信息数据元标准化的研究[J]. 中国农业信息,2008(12):21-25

[7] 李丽玲,王雨华. 药用植物资源的整合信息数据标准规范探讨[J]. 云南植物研究,2008,30(5):597-602

[8] 蒋新华. 武夷山生物多样性数据共享标准体系的研究[J]. 福建电脑,2008(3):3-4

[9] 郭新宇,赵春江,王素英. 数字农业信息标准研究[J]. 中国农学通报,2005,21(5):404-407

[10] 李金才. 生态农业标准体系与典型模式技术标准研究[D]. 北京:中国农业科学院,2007

[11] 曾小红,王强. 热带农业科学数据标准化建设及对策研究[J]. 热带农业科学,2011,3(3):54-57

[12] 郭新宇,赵春江,王利文. 作物栽培信息标准化初探[J]. 作物研究,2003,17(3):133-134

[13] 林垚. 基于数据元技术的交通科学数据目录设计[J]. 科学技术与工程,2011,11(13):1671-1815

[14] 王丹,王文生. 元数据与数据元的内涵及其应用[J]. 农业网络信息,2005(11):27-30

[15] Ben L, Jing T, Duane C, et al. A metadata-driven approach to loading and querying heterogeneous scientific data [J]. Ecol Inform,2010,5(1):3-8

[16] 秦善华,史春波,邵庆. 基于数据元的数据模型语义描述[J]. 大庆石油学院学报,2009(3):100-103

[17] 高贵锦,龙翔. 基于数据元的交换数据标准维护[J]. 吉林大学学报:信息科学版,2005,23(1):37-41

[18] 刘庆河,郝文宁,韩宪勇,等. 基于数据元的数据交换规范研究[J]. 数字社区 & 智能家居,2010,6(10):2309-2310

[19] Sakr S. XML compression techniques: A survey and comparison [J]. J Comput Syst Sci,2009,75(5):303-322

[20] Ferraz C A, Braganholo V, Mattoso M. ARAXA: Storing and managing Active XML documents [J]. Web Semantics: Science, Services and Agents on the World Wide Web,2010,8:209-224