基于本体的军用元数据模型研究

王亚宁[1],齐玉东[1],程继红[1],杨 斌[1,2]
(1.海军航空工程学院,山东 烟台 264001;
2.西北工业大学,陕西 西安 710072)

摘 要:元数据是"关于数据的数据",是用来描述资源属性的信息。使用元数据的目的是为了使资源更容易被发现、获取和利用。然而由于不同的部门采用的元数据模型不同,导致了数据不能在部门间共享。为了解决这一问题,文中将本体的思想引入到元数据模型中,并参照军用标准建立了基于本体的军用元数据模型。与传统的元数据模型相比,这种模型能从语义层面更好的表达领域中的各个概念以及概念之间的各种关系,为解决信息系统之间的异构问题提供了相应的理论基础。

关键词:本体;元数据;元数据模型;GJB/Z 139-2004
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2011)04-0227-04

Study on Ontology Based Military Metadata Model

WANG Ya-ning[1], QI Yu-dong[1], CHENG Ji-hong[1], YANG Bin[1,2]
(1. Naval Aeronautical and Astronautical University, Yantai 264001, China;
2. Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Metadata is "the data about data", it's the information that is used to describe resources' properties. The purpose of using metadata is to find, get and use resources easily. However, different department has different metadata, which made data can not share among departments. For the sake of solving the problem, introduce the thought of ontology into metadata model, and then present an ontology based military model refer to the military standard. Comparing with the traditional one, this model can express domain concepts and the relationships between them from semantic level, which provides theoretical basis for solving the isomerism between information systems.

Key words: ontology; metadata; metadata model; GJB/Z 139-2004

0 引 言

元数据是"关于数据的数据(data about data)"[1],它是对其他数据的特征进行描述的数据,是一种对信息资源进行有效组织和管理的工具,能帮助数据使用者查询所需要的信息。但是由于不同的信息系统所采用的元数据标准不同,各个信息系统之间存在异构性,信息不能共享和交换,这种现象是无法避免的,是元数据本身无法克服的,这就需要在元数据层次之上再建立一种方案,在不同的元数据标准之间进行语义映射,从而实现异构系统之间的互操作,这种机制就是本体。"本体是概念化的、明确的规范说明"[2,3],是一种提取、理解和处理领域知识的工具,它能够对不同实体对象之间错综复杂的关联关系进行很好的描述,从而为信息的组织、管理以及检索、查询提供模型和方法。

收稿日期:2010-08-02;**修回日期**:2010-11-21
基金项目:海军装备部军械科研项目(415147W8)
作者简介:王亚宁(1982-),男,山东烟台人,硕士生,工程师,研究方向为海军航空、导弹装备综合保障技术;程继红,教授,研究方向为海军航空、导弹装备综合保障技术。

1 元数据的作用及类型

1.1 元数据的作用

元数据是数字信息组织和处理的基本工具,它们为各种形态的数字化信息单元和资源集合提供规范、普遍的描述基准和方法,在数字化网络化信息服务中正发挥着日益重要的作用,其作用大体分为以下几种[4]:

1)发现和确定。用于帮助用户发现并确定所需信息资源,数据元素一般仅限于标题、作者、主题等一些简单信息,如DC元数据。

2)著录描述。用于全面、详尽的描述数据并进行著录。元数据包含内容、位置、获取方式、制作与利用方法等。数据元素的数量一般比较多,如MARC元数据。

3)资源组织描述,包括对资源最基本的描述和对

知识组织系统的描述,如RSLP元数据。

4)资源管理。包括利用资源和对管理过程进行控制,例如权限管理、数字签名、资源评鉴、访问管理、支付统计等。

5)资源保护与长久保存。能够长久保存数据资源,元数据除了能够对信息进行有效的描述和确定外,通常还包括详尽的格式信息、迁移方式、保存责任等内容,如CEDARS元数据。

6)系统功能与进程。对系统的具体功能要求进行描述,能够自动识别和匹配功能及其执行的过程。

7)系统建模,对系统的总体流程进行描述,能自动定义和识别系统流程,对模块进行搜索、嵌套和匹配,例如UML。

1.2 元数据的类型

根据元数据作用的不同可以将其划分为以下4种类型[5]:

1)管理型元数据是用来对数据资源进行组织和管理的元数据,如信息采集、数字化标准选择、排架信息、版本控制等;

2)描述型元数据是用来对数据资源进行描述的元数据,支持对数据资源的识别和发现;

3)保存型元数据是与数据资源长期保存相关的元数据,通常包含结构、格式、更新、转换等信息;

4)技术型元数据是用来描述与系统运行技术相关的元数据。

2 本体的概念及建模元语

2.1 本体的概念

本体是一个哲学概念,最早来源于哲学领域,是研究事物存在的本质和组成的哲学问题。随着社会的进步,信息系统需要对世界结构进行建模,本体逐渐发展成为了信息系统和哲学相互沟通的桥梁[6]。

近几十年,本体这个词被引入到了计算机界,并被赋予新的定义。Gruber认为“本体是概念模型的明确规范说明”[8],1997年,Borst在此基础上又稍作调整,提出:“本体是共享概念模型的明确的形式化规范说明”[9]。1998年,Studer在前人的基础上给出了一个本体较为明确的解释,即“本体是共享概念模型的明确的形式化规范说明”[10],其中包含了本体的四层含义:

1)概念化:对现实世界的一些现象涉及到的相关概念进行抽象,得出概念化模型。

2)明确性:对相关概念以及概念的描述是明确的,无歧义的。

3)形式化:计算机可以理解。

4)共享:本体中的概念及概念约束都是在相关领域内达成共识的,可共享的。

在知识系统、信息系统领域,越来越多的人研究本体,对本体的定义还有很多,在此不一一赘述。虽然每个定义的表达方式不同,但是它们对于本体的理解却大致相同。

2.2 本体的建模元语

对本体进行组织分类,可以总结出本体的基本建模元语,主要包括5个部分[11]:

1)类(classes)或概念(concepts):指客观存在的任何事物,如行为、任务、功能等。本体中的类往往具有层次关系。它一般采用框架结构来表示对象的集合,其中包括概念的名称,概念间存在的各种关系以及概念的自然语言描述等。

2)关系(relations):领域中概念间的联系和交流,如部分关系(PartOf)、非交关系(DisjointWith)等。

3)函数(functions):是一种关系,表示多个元素只能唯一确定一个元素的特殊关系。如FatherOf就是一个函数,FatherOf(x,y)表示y是x的父亲。在这个关系中,孩子x可以有多个,但父亲y只能有一个。

4)公理(Axioms):表示在领域内得到共识的永真式断言,用来解释关系之间或者函数之间存在的关联关系或规范约束。

5)实例(instances):用来表示属于某个概念的个体。

3 《数据标准化管理规程(GJB/Z 139-2004)》

《数据标准化管理规程(GJB/Z 139-2004)》[12]是军用数据标准制定和管理的顶层指导性技术文件,描述了数据标准要求以及数据标准化管理规程。数据标准要求涉及实体和标准数据元素的命名、定义及元数据等内容;数据标准化管理规程涉及军用数据标准的需求确定、制定、批准和实施。该文件适用于军用数据标准的制定和管理。数据标准为信息系统如何实现数据格式化提供框架。

该文件定义标准数据元素的名称由三部分组成:实体名(必需的)、特性修饰符(可选的)和类属元素,其中类属元素由类词修饰符(可选的)和类词(必需的)组成。数据元素名至少包括实体名和类属元素,可选的修饰符可用于进一步表明数据元素的内容。标准数据元素命名格式如表1所示。

表1 标准数据元素命名格式

实体名	特性修饰符	类属元素	
		类词修饰符	类词

军内各单位因其特定的工作和作战环境的差别,

可能产生大量的元数据。其中有些是特有的,有些是各个单位共同需要的。这些具有广泛共性的元数据是支持建立数据标准的重要基础。文件中描述了为支持全军范围数据标准化所要求的实体元数据、数据元素元数据和类属元素元数据。每个元数据的描述主要包括三部分:元数据名称、元数据约束程度、元数据的定义。

军用元数据模型中各元组元素的确定过程正是在此军用数据标准的框架内展开,充分考虑文件中制定的军用数据标准以及所描述的各类元数据,对基于本体的元数据模型的各个元组进行分析研究,确定各元组中的元素及元素间的相互关系。

4　基于本体的军用元数据模型

定义 1:基于本体的元数据模型是一个 7 元组,记作 $M = \langle T, X, Td, Xd, Ra, Tc, Tr \rangle$ [13]:

其中,T 为术语集,表示概念术语和属性术语的集合;X 为实例集,表示个体的集合;Td 为术语定义集,表示用构造符来实现多个术语定义一个术语的集合;Xd 为实例声明集,是类的实例声明的集合,其中类术语来自于术语集,个体来自于实例集;Ra 为属性声明集,是声明类术语的属性的集合;Tc 为术语注释集,是用自然语言对术语进行描述的集合;Tr 为术语约束集,描述了类术语之间的约束关系,主要是两个术语之间的关系,其中约束关系有三种:等价关系($\equiv$)、包含关系($\sqsubseteq$)和非交关系($\sqcap \varnothing$)。

定义 2:给定元数据模型为 $M = \langle T, X, Td, Xd, Ra, Tc, Tr \rangle$,模型的解释为一个二元组,记为 $I = < \Delta^I, \bullet^I >$,简称解释。

其中 $\Delta^I \neq \varnothing$ 为 M 的论域,$\bullet^I$ 是解释函数,它将 T 中的每个原子类 C 都映像为 Δ^I 的一个子集 $C^I \subseteq \Delta^I$,将 T 中的每个原子属性 P 都映像为一个二元关系 $P^I \subseteq \Delta^I \times \Delta^I$,将 X 中的每一个体 a 映像为 Δ^I 中的元素 $a^I \in \Delta^I$。

在对《数据标准化管理规程(GJB/Z 139-2004)》中描述的元数据进行分析研究的基础上,现给出一个参照《数据标准化管理规程(GJB/Z 139-2004)》的基于本体的军用元数据模型:

M = < {military concepts and attributes}, {military instances},
{Data element ≡ Entity ∩ ∃Generic element}, {C(a), C ∈ T, a ∈ X},
{Entity ⊆ ∃name. String ∩ ∃shortName. String ∩ ∃definition. String ∩ ∃note. String
∩ ∃editionId. String ∩ ∃countId. String ∩ ∃stateCode. String ∩ ∃departmentId. String
∩ ∃departmentName. String ∩ ∃modelName. String,
Data element ⊆ ∃name. String ∩ ∃countId. String ∩ ∃stateCode. String ∩ ∃creatorCode. String
∩ ∃shortName. String ∩ ∃dataType. String ∩ ∃sqlDataType. String ∩ ∃departmentId. String
∩ ∃security. String ∩ ∃maxLength. String ∩ ∃standardOrgId. String ∩ ∃departmentName. String
∩ ∃domainDataTypeId. String ∩ ∃definition. String ∩ ∃note. String ∩ ∃modelName. String
∩ ∃unitName. String ∩ ∃accuracyId. Integar ∩ ∃accuracyPercent. Decimal,
Generic element ⊆ ∃name. String ∩ ∃countId. String ∩ ∃stateCode. String ∩ ∃creatorCode. String
∩ ∃shortName. String ∩ ∃dataType. String ∩ ∃security. String ∩ ∃maxLength. String
∩ ∃standardOrgId. String ∩ ∃domainDataTypeId. String ∩ ∃definition. String ∩ ∃note. String},
{the definitions of military concepts}, {relationships between two concepts} >

术语命名规则:

1)概念术语为一个单词时,开头字母大写,其余字母小写,如:Missile;

2)概念术语为多个单词时,单词之间无空格,每个单词开头字母大写,其余字母小写,如:NavyMissile;

3)属性术语为一个单词时,全部字母小写,如:name;

4)属性术语为多个单词时,单词之间无空格,第一个单词全部小写,其余单词开头字母大写,其它字母小写,如:maxLength。

基于本体的军用元数据模型是一个 7 元组,各元组的组成如下:

术语集为军用概念术语(包括实体、数据元素和类属元素)和属性术语(包括实体元数据、数据元素元数据和类属元素元数据)的集合,如 Missile, name, countId 等;

实例集为军用实例术语的集合,如 yj-83, hy-1 等;

术语定义集为用实体和类属元素来定义数据元素的集合,用等价关系来表达,如 NavalMissile ≡ Missile ∩ ∃armyServices. {navy},其中 Missile 为实体,表示导弹;armyServices 为类属元素,表示军种,navy 为军种的一个实例;NavyMissile 是一个数据元素,表示海军导弹;

实例声明集用来声明 X 中的实例,基本形式为:C(a),表示实例 a 是类 C 的一个实例,其中 C ∈ T, a ∈ X,如 Shore-to-shipMissile(hy-1),表示 hy-1 是岸舰导弹的一个实例;

属性声明集描述军用概念术语属性,如 Missile ⊑ ∃name. String ∩ ∃shortName. String ∩ ∃definition. String ∩ ∃countId. String ∩ ∃stateCode. String 表示导弹的属性有名称、访问名称、定义、计数标识符、状态代码等,这些属

性的值类型均为字符型。属性声明集中的属性在《数据标准化管理规程(GJB/Z 139-2004)》中都有详细描述,根据描述对象的不同分为三个部分:实体元数据、数据元素元数据和类属元素元数据。表 2 列出各属性及其在元数据模型中的表示形式和值类型。

表 2 《数据标准化管理规程(GJB/Z 139-2004)》中定义的元数据(部分)

元数据名称	模型中的表示	属性值类型
1)实体元数据		
a. 实体名称	name	String
b. 访问名称	shortName	String
c. 定义	definition	String
d. 注释	note	String
e. 版本标识符	editionId	String
f. 计数标识符	countId	String
g. 状态代码	stateCode	String
2)数据元素元数据		
(1)数据元素通用元数据		
a. 数据元素名称	name	String
b. 计数标识符	countId	String
c. 状态代码	stateCode	String
d. 访问名称	shortName	String
e. 数据类型	dataType	String
f. 安全性	security	String
g. 最大字符数	maxLength	Integer
h. 定义	definition	String
i. 注释	note	String
(2)数据元素定量元数据		
a. 计量单位名称	unitName	String
b. 定量准确度标识符	accuracyId	Integer
(3)数据元素定性元数据		
a. 定性准确度百分比	accuracyPercent	Decimal
3)类属元素元数据		
a. 类属元素名称	name	String
b. 计数标识符	countId	String
c. 缩写名	shortName	String
d. 数据类型	dataType	String
e. 安全性	security	String
f. 最大字符数	maxLength	String
g. 定义	definition	Stirng
i. 注释	note	String

术语注释集将术语集中的术语用自然语言进行描述,其目的是为了实现术语的用户可读性,以便更好地支持元数据模型的后续维护。如对导弹的描述 tc(Missile) = the weapon which can control its own track and destroy the target depond on its own power.

术语约束集用来限定术语之间的关系,主要是两个术语之间的关系。假设 D 和 E 为两个类术语,术语约束有以下三种形式:

(1)D f E:表示 D 是 E 的子类;

(2)D ºE:表示 D 与 E 是等价类;

(3)D ∅ ∩E:表示 D 与 E 是非交的。

如 NavalMissile f Missile 表示海军导弹是导弹的子类,Missile ∅ ∩ Mine 表示导弹与水雷之间不存在共同的实例。

5 结束语

文中在对《数据标准化管理规程(GJB/Z 139-2004)》中的数据标准要求以及数据标准化管理规程进行分析研究的基础上,利用本体理论建立了基于本体的军用元数据模型,这种元数据模型是面向语义的,可以将元数据中的术语以及术语之间的关系更加清晰明确地表达出来。

参考文献:

[1] Metadata infrastructures seminar preparation[EB/OL]. 2008. http://colab. mpdl. mpg. de/mediawiki/Metadata_Infrastructures_Seminar_Preparation.

[2] Gruber T R. A translation approach to portable ontologies[J]. Knowledge Acquisition, 1993, 5(2): 19-20.

[3] Gruber T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal of Human-computer Studies, 1995, 43: 907-928.

[4] 张晓林. 元数据研究与应用[M]. 北京:北京图书馆出版社, 2002.

[5] 吕琼芳. 元数据与网络信息资源的组织开发[J]. 图书馆研究与工作, 2005(3): 6-8.

[6] 邓 志, 唐世渭, 张 铭, 等. Ontology 研究综述[J]. 北京大学学报(自然科学版), 2002, 38(5): 730-738.

[7] Neches R, Fikes R E, Gruber T R, et al. Enabling Technology for Knowledge Sharing[J]. AI Magazine, 1991, 12(3): 36-56.

[8] Gruber T R. A Translation Approach to Portable Ontology Specification[R]. Technical Report of Knowledge System Laboratory(KSL), 1993.

[9] Borst N. Construction of EngineeringOntologies for Knowledge Sharing and Reuse [D]. Enschede: University of Twente, 1997.

[10] Studer R, Benjam ins V R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering, 1998, 25(12): 161-197.

[11] Perez A G, Benjamins V R. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem - Solving Methods[C]//Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods, 1999: 1-14.

[12] 徐冬梅, 汪嘉颐, 张展新, 等. 数据标准化管理规程[M]. 北京:总装备部军标出版发行部, 2004.

[13] 王洪伟, 吴家春, 蒋 馥. 基于本体的元数据模型及 DAML 表示[J]. 情报学报, 2004, 23(2): 131-136.