第 36 卷 第 2 期 计 算 机 工 程 2010 年 1 月
Vol.36 No.2 Computer Engineering January 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)02—0272—03 文献标识码：A 中图分类号：TP311

矿业环境安全元数据标准研究

王 建，郝彦彬
(西北工业大学管理学院，西安 710072)

摘 要：针对矿业信息元数据标准不统一的问题，在参考相关元数据标准的前提下，考虑矿业环境安全数据的特点，采用 5W1H 方法确定矿业环境安全元数据核心元素集。参照美国联邦地理数据委员会元数据标准得到元数据框架，研究运用模型管理实现不同模式描述的元数据之间转换的算法。应用实例证明，该算法有利于矿业环境安全数据的合理组织和管理。

关键词：元数据；都柏林核心；5W1H 法

Research of Environmental Safety Metadata Standard in Mining Industry

WANG Jian, HAO Yan-bin
(College of Management, Northwestern Polytechnical University, Xi'an 710072)

【Abstract】 Aiming at the problem of not any uniform metadata standard to mining data, based on related metadata standards, takes the features of environmental safety data in mining industry into consideration. It uses Five Ws and One H(5W1H) method to determine metadata core element set for environmental safety data in mining industry. The framework of metadata is given based on Federal Geographic Data Committee(FGDC) metadata standard and research is done on the algorithm using model management to transform metadata described in different algorithms. Application example proves that this algorithm is in favor of the organization and management for environmental safety data in mining industry.

【Key words】 metadata; Dublin core; Five Ws and One H(5W1H) method

元数据是建立矿业信息数据仓库的重要基础，目前国内还没有针对矿业信息的统一元数据标准，且由于不同矿业部门元数据存储模式的异构性，如何将其集成、转换为统一的元数据模式有重要意义。模型管理解决的是元数据的互操作问题，它的目标是通过提供数据库构造中高层次的代数运算集(如 Match, Merge, Compose 等)，使其作为一个整体应用于模型和映射，而不是单独的构造模块，从而减少开发此类应用程序需要的编程数量。

1 矿业环境安全元数据核心元素集的确定及框架

1.1 元数据核心元素集的确定方法

文献[1]指出元数据核心元素集的确定是建立在对资源分析的基础上，它根据不同的应用需求有不同的确定方法。元数据核心元素集确定的主要方法如下：

(1)以现有某种国际元数据标准的某些元素为核心元素集；

(2)以现有某种国际元数据标准的全部元素为核心元素集；

(3)针对被描述资源的特征，取其有代表性的元素为核心元素集；

(4)以现有某种国际元数据标准中的某些元素再复用其他元数据标准中的元素组成核心集。

文献[2]从查询者的角度确定了元数据最小查询元素集。

综合上述观点，本文确定元数据核心元素集的步骤如下：

(1)以都柏林核心(Dublin core)标准的元素作为参照，对常用元数据标准进行比较分析，统计公共元素，并对 Dublin core 元素中没有而在其他元数据标准中出现较多的元素进行统计，这些元素组成元数据基本元素集。

(2)针对矿业环境安全数据的特点，对基本元素集进行扩充。

(3)从矿业环境安全数据使用者需求的角度，以上文得到的扩充后的元数据基本元素集为来源，采用 5W1H 方法进行分类并扩充，最终确定元数据核心元素集。

1.2 矿业环境安全元数据基本元素集的确定和扩充

本文首先以 Dublin Core 元素为参照，对 ANZLI 元数据元素、美国联邦地理数据委员会(Federal Geographic Data Committee, FGDC)元数据元素、DIF 元数据元素、GILS 元数据元素和 VRA 元数据元素进行比较分析，发现名称、主题、摘要、数据源、语言、时空覆盖范围、作者、出版者、其他生产者、日期、格式、标识等 12 个元素在这些元数据标准中多处出现，而关系、版权和类型这 3 个元素只出现了一次。

同时，本文对上述标准中含有的较通用的，而 Dublin core 没有的元素进行统计，得到获取和使用限制、联系信息、数据集状态、数据集用途和费用等 5 个元素为较常用的元素。

本文将前 12 个元素与较通用的 5 个元素合并为矿业环境安全元数据基本元素集。

但是，Dublin core 元素对著录对象的描述深度不够[3]，考虑到矿业环境安全数据与地理信息密切相关，具有很强的实时性和对数据质量要求较高的特点，本文把空间数据表示

基金项目：西北工业大学 08 年本科毕业论文重点扶持项目
作者简介：王 建(1984－)，男，学士，主研方向：信息管理与信息系统；郝彦彬，讲师、硕士
收稿日期：2009-06-16 **E-mail：**wangjianbisheng@163.com

信息、空间参照系信息、数据质量信息、实体和属性信息扩充到基本元素集，并称扩充后的基本元素集为首次扩充元素集。

1.3 矿业环境安全元数据核心元素集的确定

元数据方案设计中应遵循用户需求原则[4]，而用户需求可以从 What, Where, When, Who, Why, How 6 个方面进行比较全面的描述，因此，本文从矿业环境安全数据用户需求的角度，采用 5W1H 法对首次扩充元素集进行分类和再次扩充，最终确定矿业环境安全元数据核心元素集，5W1H 法的具体定义及其包含的元素如下：

(1)What：它的含义是数据集的内容和与其内容相关的特征的描述。What 中包含的元素见表 1。

表 1 What 中包含的元素

数据集专题分类	数据源
数据集名称	属性精度
数据集摘要	完整性
数据集主题	一致性
实体和属性信息	数据集的格式
数据集的大小	语言
数据集的粒度	数据集的标识符

其中，数据集专题分类是指数据集所属的专题划分，如测绘、地质等；实体和属性信息是数据集描述对象中的实体及其属性的信息；数据源是生产数据集的原始资料说明；完整性指数据集中关于遗漏、定义等数据集规则信息的说明；一致性指在相同背景下，同种信息保持一致的说明；数据集专题分类、数据集大小和数据集粒度是根据用户需求再次扩充的元素，其余的属于首次扩充元素集。

(2)Where：它指空间数据描述的相关信息，可以获得数据集的相关位置信息(如指向该数据集的有效链接)和数据集空间覆盖范围。Where 中包含数据集的位置信息、数据集的空间覆盖范围、空间数据表示信息、空间参考系信息和位置精度。其中，数据集的位置信息是根据用户需求再次扩充的元素，其余属于首次扩充元素集。

(3)When：它是指揭示数据集时间方面特征的信息。When 中包含数据集的日期和数据集的状态。数据集的状态是说明数据集生产进展情况及维护、更新的周期。它们都属于首次扩充元素集。

(4)Who：它是指数据集生产者、出版者和使用者的相关信息以及联系信息。Who 中包含数据集生产者、数据集出版者、数据集使用者和联系信息。其中，数据集使用者是根据用户需求再次扩充的元素，其余的属于首次扩充元素集。

(5)Why：说明数据集的用途。Why 中只包含数据集的用途，它是指生产者制定数据集的目的，如记录历史数据、为决策服务等，它属于首次扩充元素集。

(6)How：指数据集的使用限制、获取限制和费用信息。How 中包括数据集的使用限制、数据集的获取限制及费用。其中，数据集的使用限制是指数据集使用的规定，获取限制指获取数据集的规定或法律限制。它们都属于首次扩充元素集。

1.4 矿业环境安全元数据核心元素集的框架

FGDC 元数据标准除在美国国内广泛使用外，加拿大、印度等国也已采用，作为各自的国家标准[4]。ISO/TC 211 利用该标准文本作为基础，制定相应的国际标准。为了使本文确定的元数据标准具有通用性，在参照 FGDC 元数据标准框架的基础上确定元数据核心元素之间的层次关系，即核心元素集的元数据框架，见图 1。

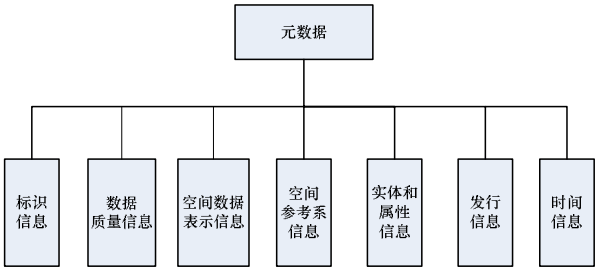

图 1 元数据框架

一层元素的下层子元素如下：

(1)标识信息：包括数据集专题分类、数据集名称、数据集摘要、数据集主题、数据集大小、数据集粒度、数据集标识符、语言、数据集位置信息、数据集空间覆盖范围、数据集用途和数据源。

(2)数据质量信息：包括属性精度、位置精度、完整性、一致性。

(3)空间数据表示信息：包括空间表示类型、矢量空间表示信息、栅格空间表示信息。

(4)空间参考系信息：包括平面坐标系定义、垂直坐标系定义。

(5)实体和属性信息：包括实体类型、它们的属性及属性的值域。

(6)发行信息：包括数据集生产者、数据集出版者、数据集使用者、联系信息、数据集的格式、数据集的获取限制、数据集的使用限制和费用。

(7)时间信息：包括数据集的日期和数据集的状态。

2 不同元数据存储模式的转换算法

由于不同矿业部门元数据存储模式的异构性，如何将其集成、转换为统一的元数据模式有重要的研究意义，本文运用模型管理实现不同模式描述的元数据之间转换的算法。

在模型管理中，模型表示所有可能实例 m 的集合。模型是应用的正式描述，例如关系模式、工作流定义、接口规范等。

本文用 S_1 表示元数据的原始关系模式描述模型，S_2 表示改变后的关系模式描述模型，d_1 表示元数据的原始 XML 模式描述模型，d_2 表示更新后的 XML 模式描述模型。d_1' 是 d_1 根据 S_2 剔除被删的元素后得到的子模型，C 由 S_2 转换而来，C' 是 S_2 中新添加的元素。

注意到 S_1 和 d_1 是用 2 种不同的模式语言表达的，因此，需要 C。与 S_1 相比，S_2 中新添加的元素在模式 d_1 中没有相对应的元素，因此，新添加的元素需要由源模式语言转化为目标模式语言。在解决方案中，假定存在一个转换工具，可将输入的关系模式转换为 XML 的输出模式以及源和转换模式元素之间的映射，因此，能得到模式 C 和映射 S_2_C。映射 $S_1\_d_1$ 由 $S_1\_d_1 = Match(S_1, d_1)$ 获得。C 与 d_2 是不同的，例如，C 包含 d_2 中不需要的表示 $m:n$ 关系的类型，它在结构上与 d_2 是不同的。

转换的具体算法为

Operator PropagateChanges(S_1,d_1,S_1_d_1,S_2,C,S_2_C)

S_1_S_2=Match(S_1_S_2);

$<d_1', d_1\_d_1'>$=Delete(d_1,Traverse(All(S_1)-Domain(S_1_S_2),S_1_d_1));

$<C', C\_C'>$=Extract(C,Traverse(All(S_1)-Domain(S_1_S_2),S_2_C));

C'_d_1' =Invert(C_C') × Invert(S_2_C) × Invert($S_1\_S_2$) × $S_1\_d_1$ × $d_1\_d_1'$;

<d_2,d_2_C',$d_2\_d_1'$>=Merge(C',d_1',C'_d_1');

$S_2\_d_2$=S_2_C × C_C' × Invert(d_2_C')+Invert($S_1\_S_2$) × $S_1\_d_1$ × $d_1\_d_1'$ ×Invert($d_2\_d_1'$);

return(d_2,$S_2\_d_2$)

该算法流程见图 2。

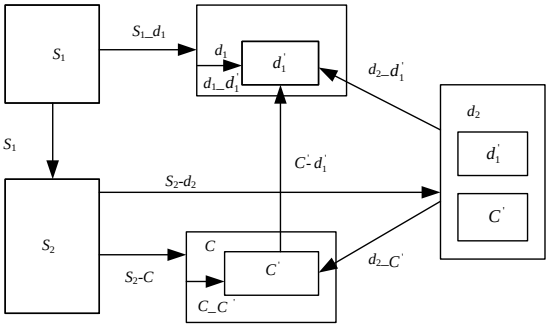

图 2　算法流程

3　元数据信息转换实例

下文以元数据中数据集的标识信息和联系信息为例，将它们的关系存储模式转化为 XML 模式，并假设它们对应的属性名也是不同的。其中，S_1是原始资源模型；S_2是改变后的资源模型；d_1是目标模型；d_2是更新后的目标模型。d_1'是 d_1根据 S_2剔除被删的元素后得到的子模型，C 由 S_2转换　而来，C'是 S_2中新添加的元素。元数据信息转换实例见图 3。

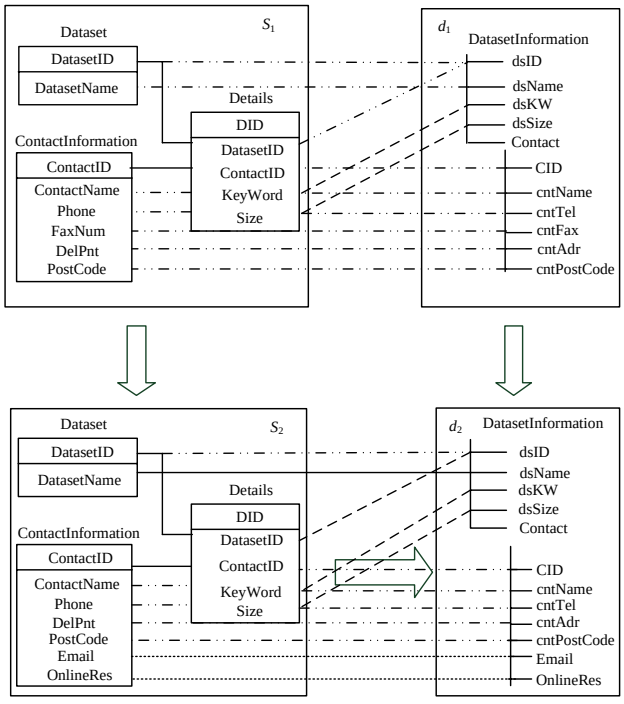

图 3　元数据信息转换实例

由上文的转换算法可知具体的转换步骤：

(1)模式 S_1和 S_2被“ Match ”来发现变化，它的结果是一个映射 $S_1\_S_2$。这个映射把 S_1和 S_2中相同的元素联系起来，S_2中新添加的元素(例如“ Email ”)和 S_1中被删除的元素(例如“ FaxNum ”)都没有相对应的元素，因此，它们并未被联系起来。

(2)为了获得新模式 d_1'(d_1的子模式)和映射 $d_1\_d_1'$。首先通过表达式 $All(S_1)-Domain(S_1\_S_2)$ 发现 S_1中删除的元素，也就是 S_1中所有未被匹配的元素。然后这些元素用来“ Traverse ”映射“ $S_1\_d_1$ ”。例如，被删除的关系属性“ FaxNum ”遍历映射 $S_1_d_1$并找到 XML 模式 d_1中的元素“ cntFax ”。最后利用运算符“ Delete ”从 d_1中剔除了 S_1中被删除元素在 d_1中的映像。

(3) S_2中的新元素，即那些在映射 $S_1_S_2$的值域中没有的元素，遍历 S_2_C，找到 C 中相对应的新添加元素。例如，关系模式属性“ Email ”遍历 S_2_C找到 XML 模式 C 中相对应的元素。利用运算“ Extract ”从 C 中提取只含新元素的模式“ C' ”，同时返回映射 C_C'。

(4)得到 C'_d_1'。

(5)模式 d_2通过“ Merge ”运算得到。

(6)计算映射 $S_2\_d_2$，映射 $S_1_d_1$也是输入的一部分。本文需要 $S_2_d_2$来保证若源模式变化后，此算法仍然可以复用。因为 d_2是 d_1'和 C'混合而来的，映射 $S_2_d_2$必然是 2 个映射的并，索引这 2 个映射分别是 S_2与 C'和 S_2与 d_1'。其中，第 1 个映射 S_2中元素“ Email ”和“ OnlineRes ”对应它们在 d_2中的元素，第 2 个映射 S_2中元素“ ContactID ”与 d_2中元素“ CID ”对应。

(7)返回模式 d_2与映射 $S_2\_d_2$。

4　结束语

本文确定了矿业环境安全元数据核心元素集及其元数据框架，同时运用模型管理实现了不同模式元数据之间的转换算法，该元数据标准中底层具体元素的确定有待于进一步研究。

参考文献

[1] 张春景. 信息系统元数据规范应用研究[D]. 上海: 华东师范大学, 2004.

[2] Chaitin K, Hirsch S, Mularz D. Identification of the Minimal Metadata Element Set for Search[EB/OL]. (2008-05-21). http://www.lic.wisc.edu/metadata/METAHOME.HTM

[3] 王卷乐, 游松财, 谢传节. 元数据标准结构分析与设计[J]. 地理与地理信息科学, 2005, 21(1): 16-18.

[4] 蒋景瞳, 刘若梅, 贾云鹏. 国际元数据标准的发展和研究现状[M]. 北京: 科学出版社, 1999.

编辑　陆燕菲