CWM技术研发决策支持平台的元数据管理模型研究

赵 辉 张英杰 彭 洁
（中国科学技术信息研究所，北京 100038）

摘要：选择CWM规范作为多源数据整合的参考，构建基于CWM的统一元数据存储区，分析统一元数据存储区的数据冲突，提出解决方案，并以电动汽车产业为例，从该领域决策支持的业务需求分析出发，重点描述平台的数据类型及业务信息处理的主要逻辑，构建元数据管理模型及存储区。

关键词：决策支持；通用数据仓库；元数据模型；CWM；数据冲突
中图分类号：G356；G202；G203 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2014.04.003

Multi-Source Data Metadata Management Model Research of Decision Support Platform in Electric Vehicle Field

Zhao Hui, Zhang Yingjie, Peng Jie
(Institute of Scientific and Technical of Information of China, Beijing 100038)

Abstract: This paper studies how to build a unified metadata storage region based on the CWM specification, analyzed the data conflicts while building the unified metadata storage area, and put forward the solution. Finally selected the electric car field decision support platform as an example, analyzed business requirements, described the main data types and business information processing logic, built the unified metadata storage region.

Keywords: decision support, common warehouse meta model, CWM, data conflict

面向新兴产业科技研发方向服务的决策支持系统需要经济、社会、技术等多方面的数据支持，需要建立一种统一的且易于扩展的数据仓库环境。新兴产业科技研发方向的确定是复杂的，需要经济、社会、技术等多方面的信息，要支持科技管理人员、科技战略专家等对决策支持的不同要求，单一的工具很难满足这样的需求，因此只有建立面向电动汽车领域的科技决策支持系统就必须建立一种统一的且易于扩展的数据仓库环境，搭建统一的数据平台，提供一套通用、全面的接口体系，建立可持续的数据交换机制，才能为决策支持平台提供有效支撑[1]。

一般来说，数据仓库主要包括数据源、ETL工具、核心库、建模工具、数据仓库管理工具、前端访问和分析工具等。其中，元数据库及元数据管理工具是多源数据仓库管理的核心[2]。针对

作者简介：赵辉*（1971- ），女，中国科学技术信息研究所副研究馆员，研究方向：科技资源管理；**张英杰**（1979- ），男，中国科学技术信息研究所助理研究员，研究方向：信息资源管理；**彭洁**（1965- ），女，中国科学技术信息研究所研究员，博士，主要研究方向：科技信息资源管理。

基金项目：国家科技支撑计划课题“电动汽车多源信息决策支持系统研发”（2013BAG06B03）；国家科技支撑计划课题“电动汽车专题数据库”（2013BAG06B02）；国家自然科学基金项目“大数据挖掘在科技项目查重中的应用研究”（71303223）。
收稿日期：2014年5月19日。

基于多源数据技术研发决策支持系统的需求，本文借鉴公共数据仓库元模型标准CWM(Common Warehouse Metamodel)，构建多源元数据统一存储区体系架构，分析存储区构建过程中需要解决的数据冲突问题，提出应对策略，并应用于电动汽车决策支持系统。

1　元数据及通用元数据管理模型

关于元数据，目前还没有权威的统一定义。一般认为，元数据是关于数据的数据[3]。元数据是描述数据的标签或数据的上下文背景。在数据仓库中，经常被分为4种类型，即业务元数据、技术和操作元数据、流程元数据和数据管理制度元数据[4]。其中，业务元数据可以实现业务模型与数据模型之间的映射。业务元数据说明了业务数据的属性、范围、计算公示、业务规则等内容，帮助用户理解和使用业务数据，也便于进一步明确分析目标和挖掘对象。但是，在数据仓库中，由于进入数据仓库的数据来源多样，不同来源的数据，其元数据必然不同，甚至存在元数据的冲突。为了更顺利地整合、利用多源数据，就需要一个管理不同来源元数据的模型和工具。为此，OMG（Object Management Group）组织推出了一个公共数据仓库元模型的标准，提供描述数据源与数据目标之间的转换、分析、处理、操作等相关的元数据基础框架（构成规则集），作为在数据仓库和业务分析环境中进行元数据交换的指南[5-6]。CWM框架有两方面的作用，一方面是告诉数据仓库开发者，在整合集成多源数据信息时，元数据模型中要考虑哪些问题；另一方面，为开发者提供了非常具体的、可重用的、由UML语言给定的、由21个包组成的元模型框架。

在CWM中，分别从数据资源、数据分析、数据仓库管理层3个层面，将数据转化、分析、处理和操作相关的元数据划分成21个包（表1）[7]。其中，数据资源层可进一步细分为资源、基础和对象模型3个细分层。资源层定义了数据仓库从各类数据源（包括面向对象、关系数据库、记录、多维数据、XML数据等类型）中获取数据的元数据描述包。为了能更好地定义资源层，还需要基础层中的业务信息、数据类型、表达式、键和索引、软件部署、类型映射等相关基础数据的元数据描述。而对象模型层则是数据仓库的基础元数据描述规则，分别从核心包、行为包、关系包和实例包给出了参考描述。管理层和分析层是数据仓库自身管理所需的各种元数据模型参考描述。

2　CWM指导下的统一元数据存储区设计

从CWM元模型框架中可以看出来，数据资源层是整个数据仓库的基础，其目标是解决整个数据仓库系统如何从各个数据源获取数据的问题，为了对这些元数据进行有效管理，实现数据的复用，较为合理高效地做法是在系统中设立一个元数据统一存储区，实现数据仓库中的元数据与来源数据源的元数据的一一对应[8]。一旦来源数据源中的元数据发生变化，修改统一元数据存储区中的元数据映射关系，就可以完成更新。

统一元数据存储区主要分为两层：元数据交换层和元数据存储层。在元数据存储层，除了要保存标准元数据外，还要有两个数据处理模块：一个是为数据仓库存储和计算结果输出服务的元数据服务模块，一个是在整合不同来源数据的元数据后解决所产生冲突的数据冲突解决模块。在元数据交换层，则需要根据来源数据的不同情况，编制不同的元数据适配器，有针对现有数据

表1　CWM元模型框架

管理	数据仓库处理包			数据仓库操作包		
分析	转换包	联机分析处理包		数据挖掘包	信息可视化包	业务命名规则包
资源	对象包	关系型包		记录包	多维包	XML包
基础	业务信息包	数据类型包	表达式包	键和索引包	软件部署包	类型映射包
	对象模型（核心包、行为包、关系包、实例包）					

库的，有针对数据仓库建设的需求新建的，有针对已有数据分析工具的，有针对数据仓库分析进行特定抽取的（图1）。

在以上模型中，CWM元数据服务APIs主要从在线分析、可视化、数据挖掘、数据仓库管理和维护等方面，根据CWM模型框架中的管理层和分析层中的7个数据包规范，编制应用服务模块，而数据冲突解决模块较为复杂，需要重点考虑。

3　建设多源元数据存储区的数据冲突分析与解决

元数据存储区的作用是集中存储元数据，其物理特性也同时与元数据访问接口的实现方式相关，要建设基于CWM的元数据管理平台，必须首先确定基于CWM的元数据存储结构。在确定集成的元数据存储结构时，会出现多数据库的互操作冲突。这类冲突主要分为语义冲突、描述冲突、数据模型冲突和结构冲突等[9-11]。

（1）命名冲突。例如：①当期刊论文数据库中有作者的描述字段，字段名称被命名为author name，而在科研项目数据库中有项目承担人的描述字段，字段名称被命名为Investigator，这两个字段表征的是同一个概念，即研究者，因此发生同义词冲突，需要建立这样两个实体名称的映射关系。在电动汽车多源数据集成时，要将科技文献中的作者（author name）、科技项目承担人（Investigator）、机构中的关键研究者（researcher）与科技人才数据库中的人名（researcher name）建立映射，全部映射到科技人才数据表中。②当期刊论文题目被命名为title，而人的头衔也被命名为title时，就产生了两个元素有相同的名字却表示不同的实体或概念的情况，被称为元素命名同形异义冲突。此时，需要针对不同来源的数据表，建立“数据表+字段名”，与数据仓库中的元数据分别建立对应字段title和rank。

（2）属性集冲突。两个描述相同实体或概念的元素具有不同的属性集合时，则发生属性集冲突。例如，在期刊论文数据库中，作者的属性仅有姓名，而在电动汽车人才数据库中，人的属性有姓名、性别、年龄等，在元数据存储区就需要平衡两个数据集的属性冲突，决定使用哪个属性集。

（3）数据类型冲突。当不同的数据集在描述

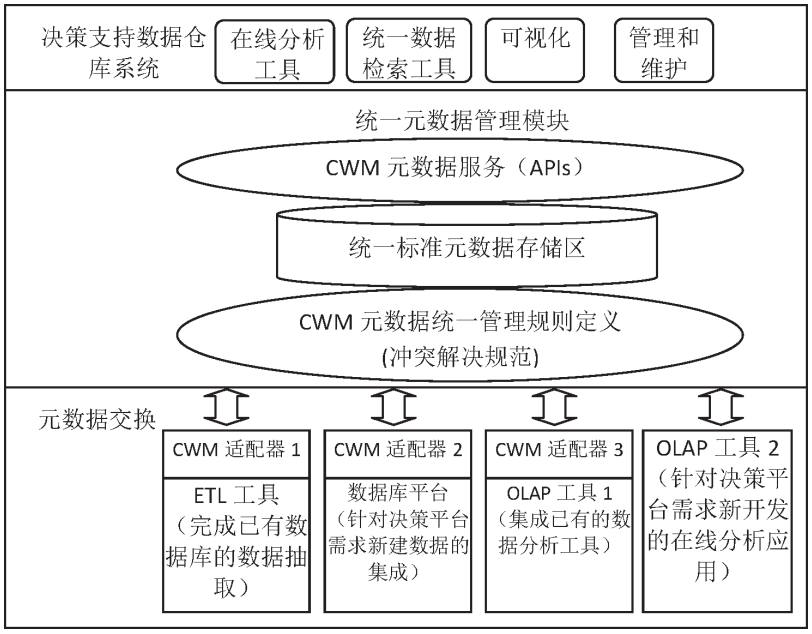

图1　电动汽车多源数据元数据体系架构

同一实体的同一特征时使用了不同的数据类型，则两个数据集间存在数据类型冲突。例如对于机构所在地区这一属性，在科研机构数据库中的数据类型是整数，而在厂商数据库中的数据类型是文本，则两个数据库在集成时产生了数据类型冲突。此时，根据主数据和参考数据的划分规则，统一使用整数类型，建立“地区”的参考数据表。

（4）数据量纲冲突。当两个数据集描述相同实体的同一特征时使用了不同的量纲，就造成了量纲冲突。例如，在车型数据库中，对于电动车整车性能指标之一是续航里程，其量纲是公里，而在零配件数据库中，电池性能的指标之一也是续航里程，其量纲却是英里，此时就需要把两个数据集中的续航里程量纲进行统一，最后确定为公里。

（5）数值范围和精度冲突。指两个数据集中的相关对象等价的数据元素有不同的范围和精度设置。例如在电动汽车项目数据库中的起止时间精确到年月，而在期刊论文数据库中的时间精度为年，则在分析应用时，只能选择年度分析。

（6）约束冲突。指相关对象中等价的数据元素有不同的实例约束。例如在期刊论文库中的作者是所有发表期刊论文的人，而在电动汽车人才数据库中，就将人限定为有一定影响力的、在领域做出重要贡献的研究者，其范围大大小于论文作者。在系统进行检索时，就需要选定二者的并集进行检索。

（7）键冲突。指对相关对象建立了不同的唯一标识。这在人、机构方面的冲突尤其突出，在期刊论文、专利、科技人才、科研机构、科技项目5个数据库中，在科技人才和科研机构数据库中对人和机构建立了唯一标识，而在期刊论文、专利和科技项目数据库中，没有建立人和机构的唯一标识，使得这5个数据库进行关联检索时会产生人的同名冲突、机构的异名冲突等，需要对期刊论文、专利和科技项目数据库中的人和机构名称进行标准化处理，以科技人才和科研机构数据库为基准建立唯一标识，对于无法判断是否是同名的人、是否是异名机构时，则先按同名人和异名机构处理，将数据归一化的工作留给使用者进行判定和整理。

4　案例研究

现将CWM决策支持平台的元数据管理用于电动汽车元数据管理模型及存储区建设。

4.1　电动汽车决策支持系统功能需求

电动汽车决策支持系统面向宏观技术决策提供服务，集成电动汽车技术进展信息、电动汽车市场及产销信息、电动汽车技术研发管理信息，为技术方向选择、技术方案制定提供基础信息、基于基础信息的多层次数据分析和多种数据分析产品（分析图、表、研究分析报告等）。

电动汽车决策支持平台可以提供的决策研究产品及功能包括以下几个方面。

（1）基础信息服务：提供对集成数据库的简单检索、高级检索、信息导航等服务。

（2）报表服务：提供电动汽车科技研发投入及产出统计表、电动汽车整车产销统计表、电动汽车关键零部件产销统计表、电动汽车配套基础设施建设状况等。

（3）多维度分析：从时间维、机构/人维、热点研究主题维、地区维5个维度对电动汽车科技研发投入及产出、整车产销量、关键零部件产销量、基础设施建设等进行多维度统计分析等。

（4）仪表盘：对关键指标制定状况监测面板，如技术成熟度监测、研发主题演化监测、科技资源（资金、机构、人）配置监测、研发能力监测、电动汽车市场需求与供给监测、电动汽车配套基础设施布局监测等。

（5）记分卡：对主要状况的评估评价，例如政策效果评估、电动汽车整车及关键零部件技术参数合理性评价、基础设施状况评价、电动汽车应用推广示范效果评价等。

4.2　电动汽车决策支持系统中的元数据模型

从上面的功能需求中可以看出，电动汽车决策支持系统中的数据主要围绕人、机构、主题、事项展开。其中，人主要指电动汽车领域的研究者，机构包括科研机构、大学、电动车及零配件

企业，主题指与电动车研发相关的各种词汇和短语，事项是研发项目、应用示范项目、电动车相关产品、研发活动产生的成果（论文、专利、研究报告等）。这四大数据类之间的关系如图2所示。

以上4类数据，用CWM模型来看，主要在分析层和资源层建立更加个性化的元模型规则，具体如表2所示。

在表2的关系包中，可以建立起多种关系，例如人与机构之间的聘用关系、人与人之间有合著关系、项目合作关系、上下级关系、师生关系等，机构之间有战略合作、业务合作、生产配套等关系，主题之间可以是共现、包含、并列等关系。

4.3　电动汽车决策支持系统元数据存储区

为了建立4.2节所述的电动汽车决策支持元数据存储区，需要建立如图1所示的4类适配器，即：（1）从已有数据库中抽取数据的适配器，针对学术文献、技术标准、专利、人力、机构和项目数据库。（2）按照电动汽车决策需求专门建立的元数据适配器，包括电动汽车厂商数据库、政策法规数据库、网络动态信息库、产品样品库（含整车和关键零部件）、基础设施库、示范运行项目库和资源环境库。（3）为数据分析、检索查询服务的数据适配器，如电动汽车词系统数据库。（4）为在线分析系统服务的数据适配器，如为社会网络分析专用的人－人关系、机构－机构关系、人－机构关系数据库等。

经过以上数据存储区的建设，同时处理好多源数据汇集过程中的数据冲突，使用方便的电动汽车多源数据决策支持系统的元数据管理模型全面构建完成。

5　结论与展望

本文主要结论：（1）由于CWM规范的通用性，其元数据模型定义相对抽象，在实际应用中难度较大；且规范定义的范围较广，实际使用时也只能选取其中的关键包参考使用。（2）统一元数据管理区构建过程中的4类适配器开发，在具体的系统中要根据实际需求进行筛选，相对而言，针对ETL的适配器开发遇到的情况比较复杂，工作量较大；针对OLAP的适配器则必须具有可扩展性，在实际使用中，会随时遇到增加新的OLAP应用的需求，管理难度相对较大。（3）多源存储区的数据冲突会分为两大类情况。一类面向数据的整合阶段，会遇到静态数据冲突；另一类面向应用，会遇到动态数据冲突。静态数据冲突可以使用本文提出的策略进行解决，动态数

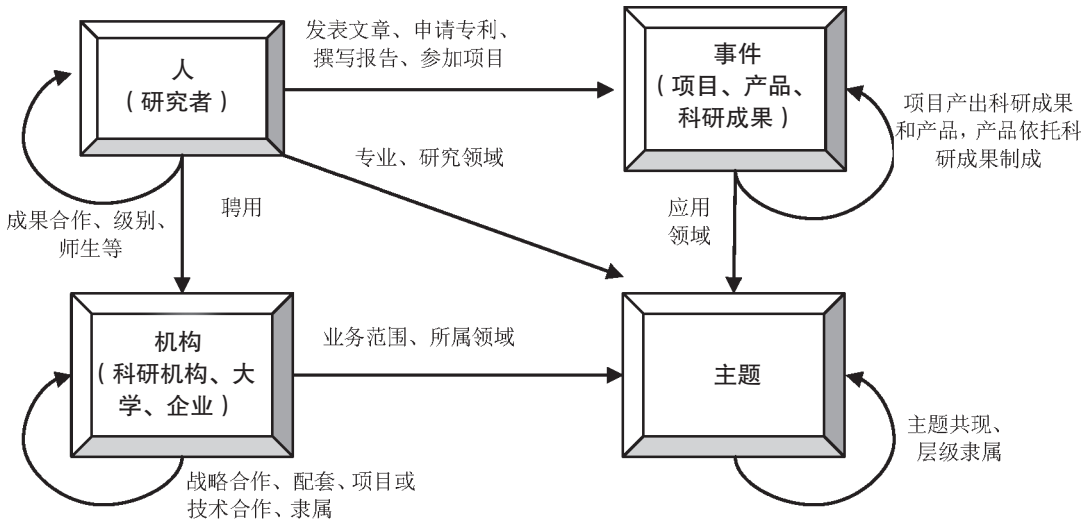

图2　电动汽车决策支持系统基础数据类型及其关系

表2　电动汽车决策支持系统的CWM元模型框架图

管理	数据仓库处理包			数据仓库操作包		
分析	已有数据库抽取的转换元数据包	基于论文、专利的联机分析元数据包		主题演化、成熟度挖掘元数据包	折线图、饼图、社会网络图、气泡图、地图等可视化元数据包	人、机构、主题、事件的命名规则
资源	人、机构、主题、事件的对象包	人−人关系、机构−机构关系、主题−主题关系、人−机构关系包		记录包	时间、地区等多维包	新闻动态整合的XML包
基础	业务信息包	数据类型包	表达式包	键和索引包	软件部署包	类型映射包
	对象模型（核心包、行为包、关系包、实例包）					

据冲突需要工作机制和流程的配合，将在后续研究中予以考虑。(4)对于电动汽车领域的四大类对象（人、机构、主题、事件）间关系的确定，是实际应用中的重点和难点，本文作者将在后续研究中进行深化。

参考文献

[1] 容会, 于勇涛, 陈震霆, 等. 元数据管理系统的研究与设计[J]. 价值工程, 2012, 31(13):171−172

[2] 郑洪源, 周良. 基于CWM的标准ETL的设计与实现[J]. 吉林大学学报:信息科学版, 2006, 24(1): 50−55.

[3] 维基百科.元数据[EB/OL]. [2014−07−24]. http://zh.wikipedia.org/zh/元数据.

[4] Inmon W H. 数据仓库[M]. 王志海,等,译.北京: 机械工业出版社, 2006.

[5] 李姗姗,宁洪,陈波,等. 通用数据仓库元数据模型的研究[J]. 计算机工程与科学,2004, 26(5): 52−55.

[6] 蒋楠,丁祥武. 基于模型驱动元数据管理策略的研究[J]. 计算机应用与软件, 2012, 29(1):188−190.

[7] OMG. Common Warehouse Metamodel(CWM) Specification (Version 1.1, Volume 1) [EB/OL]. [2014−04−21]. http://www.omg.org/spec/.

[8] 张明治. 基于CWM 规范设计的元数据管理系统[J]. 电脑知识与技术, 2014, 10(2): 254−258.

[9] 赵晓非,黄志球. 基于描述逻辑的CWM元数据冲突的检测和消解[J]. 计算机科学, 2010, 37(11):166−171.

[10] 杨俊. 非结构化数据信息提取的研究和实现[D]. 湖北:武汉邮电科学研究院,2008.

[11] 赵雨蒙. 基于模式映射的异构数据集成模型研究[D]. 山东:山东大学,2010.

(上接第13页)

[12] 新能源汽车技术创新将获资金支持[EB/OL].[2013−08−29].http://www.newenergy.org.cn/html/01210/10231249865.html.

[13] 安索夫战略[EB/OL].[2013−08−29].http://wiki.mbalib.com/wiki/%E5%AE%89%E7%B4%A2%E5%A4%AB%E6%88%98%E7%95%A5.

[14] Chandler, Alfred Dupont. Strategy and Structure[M]. Cambridge, MA: MIT Press, 1962.

[15] 马费成.数字信息资源的规划、管理与利用研究[M]. 北京:经济科学出版社,2012: 68.

[16] Jonkers, Henk, et al. Enterprise Architecture: Management Tool and Blueprint for the Organisation[J]. Information Systems Frontiers, 2006, 8(2): 63−66.

[17] Schekkerman, Jaap. How to Survive in the Jungle of Enterprise Architecture Frameworks: Creating or Choosing an Enterprise Architecture Framework[M]. Trafford Publishing, 2004.

[18] Zachman, John A. A Framework for Information Systems Architecture[J]. IBM Systems Journal, 1987, 26(3): 276−292.