数据仓库中的元数据模型研究

曾志勇,余建坤

云南财经大学　计算机科学系　云南　昆明　650221

摘　要：OMG（对象管理组织）提出的通用数据仓库元模型 CWM（Common Warehouse Metamodel）已成为数据仓库元数据方面的唯一标准，它使数据仓库工具和元数据库之间交换元数据变得容易和方便。本文介绍了 CWM 的重要性、由来和结构，并给出了一个利用 CWM 进行元数据交换的应用实例。

关键词：数据仓库；元数据；通用数据仓库元模型

Research on Metadata Model of Data Warehouse

Zhi-yong ZENG, Jian-kun YU

Dept. of Computer Science, Yunnan University of Finance and Economics, kunming Yunnan 650221, China

Abstract: *OMG's Common Warehouse Metamodel is now a single data warehouse metadata standard. Interchange of metadata between data warehousing tools and metadata repositories made easy and convenient by using CWM. In this paper, we present the origin, importance and architecture of CWM, and offer an application case.*

Key words: *Data Warehouse; Metadata; CWM*

1 引言

随着信息技术在企业的迅速应用，表达信息的数据也随着时间和业务的发展不断膨胀。这些数据往往分布在异构且不兼容的分布式环境中，冗余度和不一致性高，领导和决策者难以从这样复杂的数据环境中得到有用的决策信息。数据仓库作为一个将各种复杂异构的数据按统一规则存储起来，并将数据整理转换为有价值信息的技术，提供了一个很好的解决方案。

但是数据仓库的建设过程却很少一帆风顺，相反它是早期数据仓库建设者们夜间的噩梦。一个重要的原因是操作数据源和数据仓库工具的多样性使得数据仓库的构建和维护变得极其具有挑战性。这些数据和工具的多样性不仅是在语义上（例如核心数据模型），而且也是在结构上的（例如怎么抽取和导入数据的操作细节）。

由于缺乏一个对数据和数据仓库工具之间接口的公共、可共享的描述模型，在数据仓库数据集成时，要进行集成的情况就如图 1 所示：

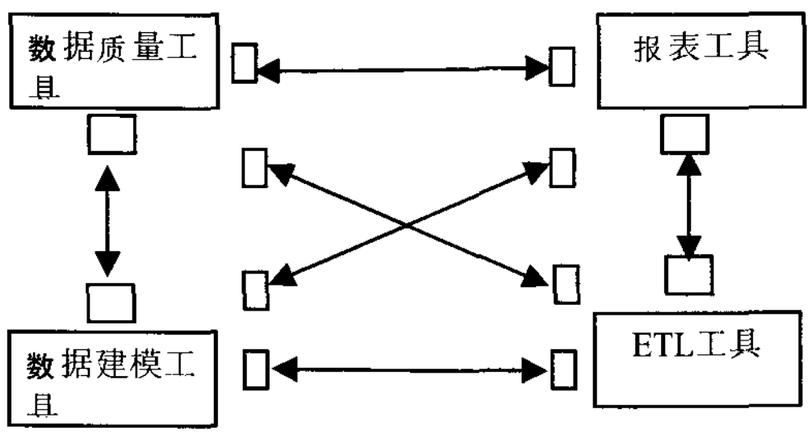

图 1　没有标准的数据集成情况

为了解决对数据的统一描述问题和数据在工具间的交换、迁移的困难，于是人们使用了元数据，并经过努力，最终统一了元数据标准——通用数据仓库元模型 CWM（Common Warehouse Metamodel）。

2 元数据

按照传统的定义，元数据（Metadata）是关于数据的数据。它是描述数据仓库内数据的结构和建立方法的数据，可将其按用途的不同分为两类：技术元数据（Technical Metadata）和业务元数据（Business Metadata）。

技术元数据是描述关于数据仓库技术细节的数据，这些元数据应用于开发、管理和维护数据仓库；业务元数据从商业和业务的角度描述数据仓库的数据，提供了良好的语义层定义，业务元数据使业务人员能够更好的理解数据仓库分析出来的数据。

3 元数据标准——CWM

由于元数据是如此的重要，不同的数据仓库生产商为了定义出元模型的行业标准版本，结成了两大联盟并开发出了各自的标准。第一个标准是元数据联盟（Meta Data Coalition， MDC）的开放信息模型（Open Information Model, OIM）。该计划最初由微软提出，后来提交到 MDC。第二个是由 Oracle、IBM、Hyperion 以及 NCR 所倡导，并随后提交给对象管理组织（Object Management Group, OMG）的通用数据仓库元模型 CWM（Common Warehouse Metamodel）。

这两大标准在竞争中逐步完善，2000 年 MDC 和 OMG 两大组织合并，新的 OMG 在 2001 年 2 月发布了 CWM 1.0，为数据仓库厂商提供了统一的元数据标准，从而为元数据管理的发展铺平了道路。

CWM 是基于 UML, MOF, XMI3 个标准来设计、操作、交互数据仓库元数据。其中，UML(标准建模语言)提供了一个强大的建模语言用来描述数据仓库中的各种元数据，MOF 提供操纵元数据的接口，XMI 提供利用 XML 交换元数据的机制。CWM 元模型的组成如图 2 所示。

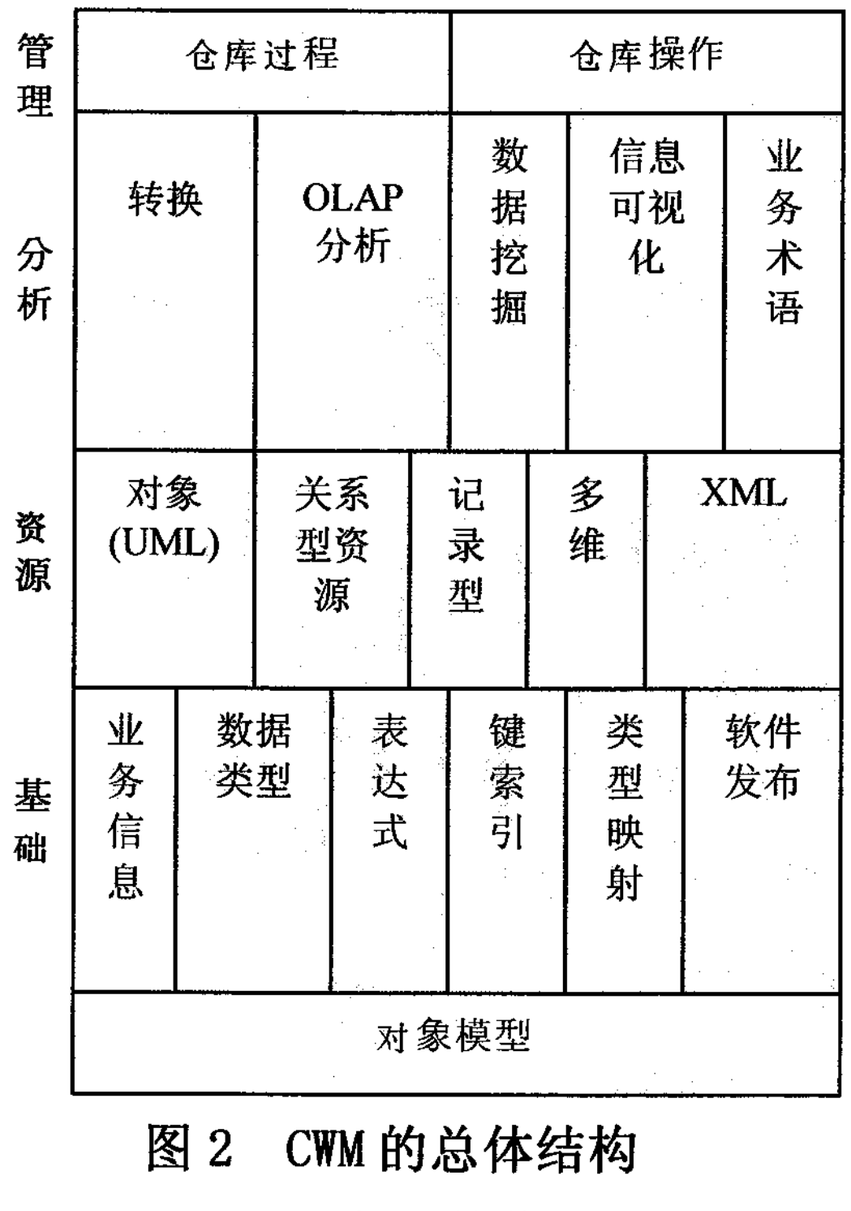

图 2　CWM 的总体结构

CWM 各层的作用如下：

对象层：　这个 UML 的子层用作 CWM 的基本元模型。它实际上是 UML 标准的一部分。CWM 从 UML 标准中选取了一些关键的对象，从而构造高层的对象。

基础层：　基础层为更高层次提供 CWM 特定的服务。它包含了很多基本的元模

型包，它们所表示的概念和结构能够被上层的 CWM 包所使用。

资源层：　资源层定义了各种不同类型的数据资源。如对象型数据源、关系型数据源、记录型数据源、多维数据源、XML 数据源。

分析层：　该层主要是说明如何对数据源中的数据进行分析处理,包括数据转换(Transformation) 、在线分析处理(OLAP) 、数据挖掘、信息可视化等多个方面。

管理层：　该层的主要目标是支持数据仓库的日常操作和管理，如 ETL 过程的管理。

由上，CWM 提供了基于模型的元数据集成体系结构所需的、用于描述问题域的语义完整的公共元模型。如果构建数据仓库用到的各种软件产品、工具和数据库产品就 CWM 元模型达成一致，它们就都能理解 CWM 元模型的实例（模型或者元数据），元数据很容易在各部分之间进行交换和共享。

目前，主要的数据仓库厂商 IBM、Oracle、Hyperion、SAS、Dimension EDI、Genesis IONA、HP、NCR 和 Unisys 等都宣布支持 CWM。以 CWM 为元数据标准的数据集成情况就如图 3 所示。

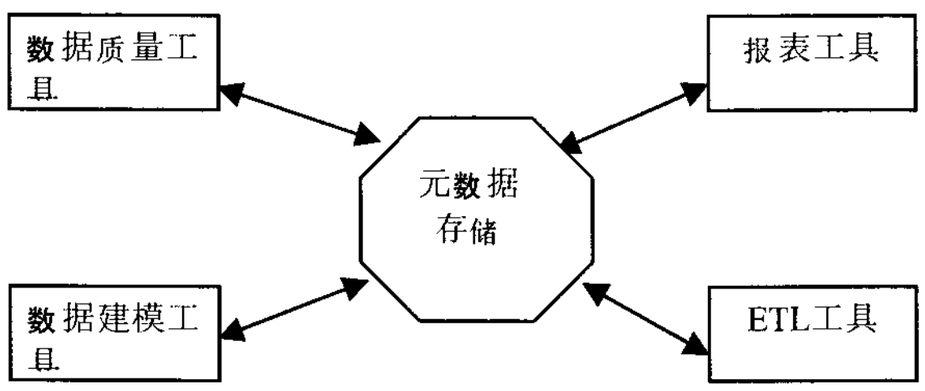

图 3　以 CWM 为元数据标准的数据集成情况

4 应用实例

在实际的数据仓库建设过程中，人们常常希望将设计阶段产生的数据模型直接无缝地应用到后面阶段的工作中，如 ETL 和元数据管理。但在 CWM 出现之前，这样的数据交换非常困难甚至不可能，因为我们使用的数据建模工具和 ETL 工具、元数据管理工具往往是由不同数据仓库厂商所提供。

为了验证 CWM 在信息共享方面的能力，我们选取了常用的数据建模工具 ERWIN 和领先的数据仓库软件平台 SAS 9 进行互通实验。也就是把在 ERWIN 中设计好的数据模型通过 CWM 导入到 SAS9 中去。

SAS9 通用的元数据管理构架是 SAS Open Metadata Architecture，它为所有 SAS 应用提供公共的元数据服务，从而改进了应用之间的通信方式。它支持业界标准 OMG 的 Common Warehouse Metamodel（CWM）。通过使用新的 SAS 工具软件，如

Management Console 和 SAS Cube Studio，实现了 SAS 元数据的统一管理，规范了 OLAP 的开发和维护。

SAS Metadata Server 是 SAS 存储元数据的集中的、共享的场所，而元数据管理器提供了对元数据的管理。对于元数据的导入，我们关心的是能否直接将数据建模工具的结果正确地导入 SAS 的元数据存储库。CWM 模型使得这种导入成为现实，当然，这要求所使用的数据建模工具也支持 CWM 模型。如果数据建模工具不直接支持 CWM 模型，则可以通过第三方的工具，如 meta integration 公司的 model bridge 将其先转换为 CWM 模型。

在我们的例子中，由于使用的 ERWIN 版本（4.0）暂时还不支持 CWM，所以我们先将 ER 图存为 xml 格式，接着用 model bridge 将其转化为 CWM 模型，再在 SAS9 的 management console 的元数据管理器里导入。测试使用的 ER 图如图 4。

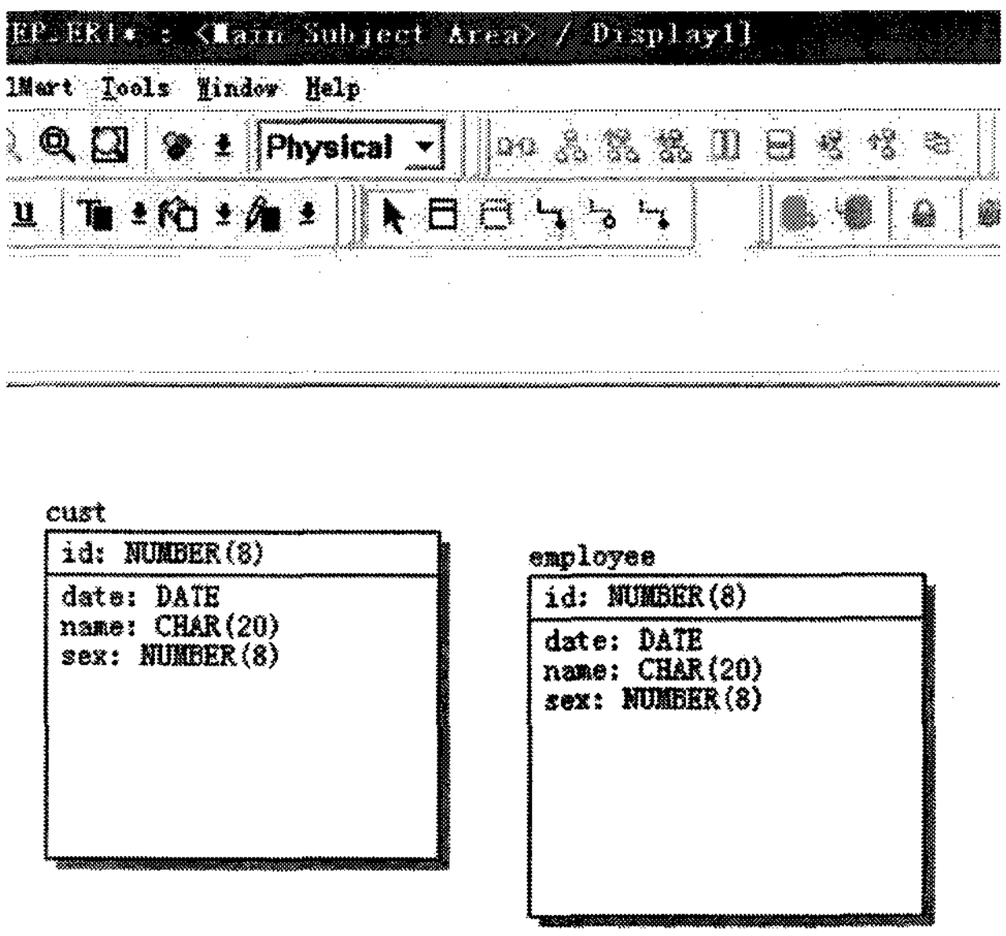

图 4　要导入的 ER 图

导入过程如下：

1．将 ER 图存为 XML 文件。

2．在 model bridge 里导入 xml 文件。Import Bridge 选 CA Erwin 4.0 SP1 to 4.1，选择导入的 xml 文件，按 Import Model 按钮（见图 5）。

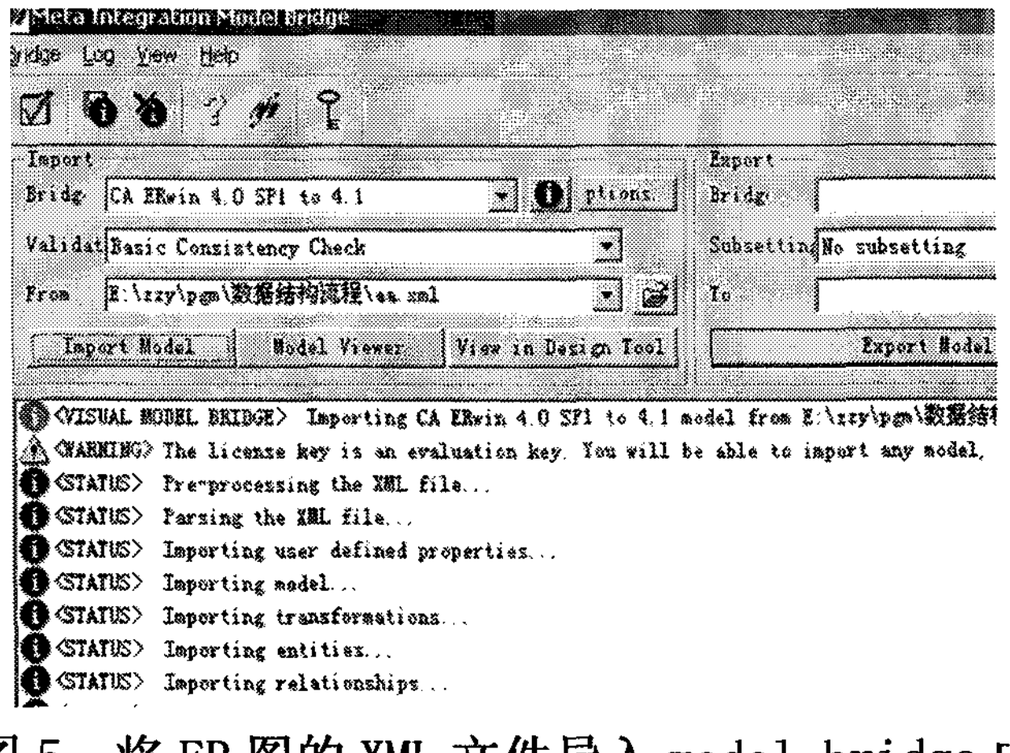

图 5　将 ER 图的 XML 文件导入 model bridge 中

3．在 model bridge 里导出 xml 文件。Export Bridge 选 SAS ETL Studio，选择导出的 xml 文件名，按 Export Model 按钮（图略）。

4．在 SAS9 的 management console 的元数据管理器里选中元数据存储库，点右键，选导入元数据（图略）。

5．选择导入类型----CWM 导入，按下一步按钮（图略）。

6．指定要导入元数据的文件，也就是 model bridge 导出的 xml 文件，按下一步按钮（图略）。

7．选择导入的元数据要连接的目标服务器，按下一步按钮（图略）。

8．选择将指向要导入的表的 SAS 逻辑库，按下一步按钮（图略）。

9．选择完成。

10．在 Management Console 里可以看到刚导入的数据模型（见图 6）。

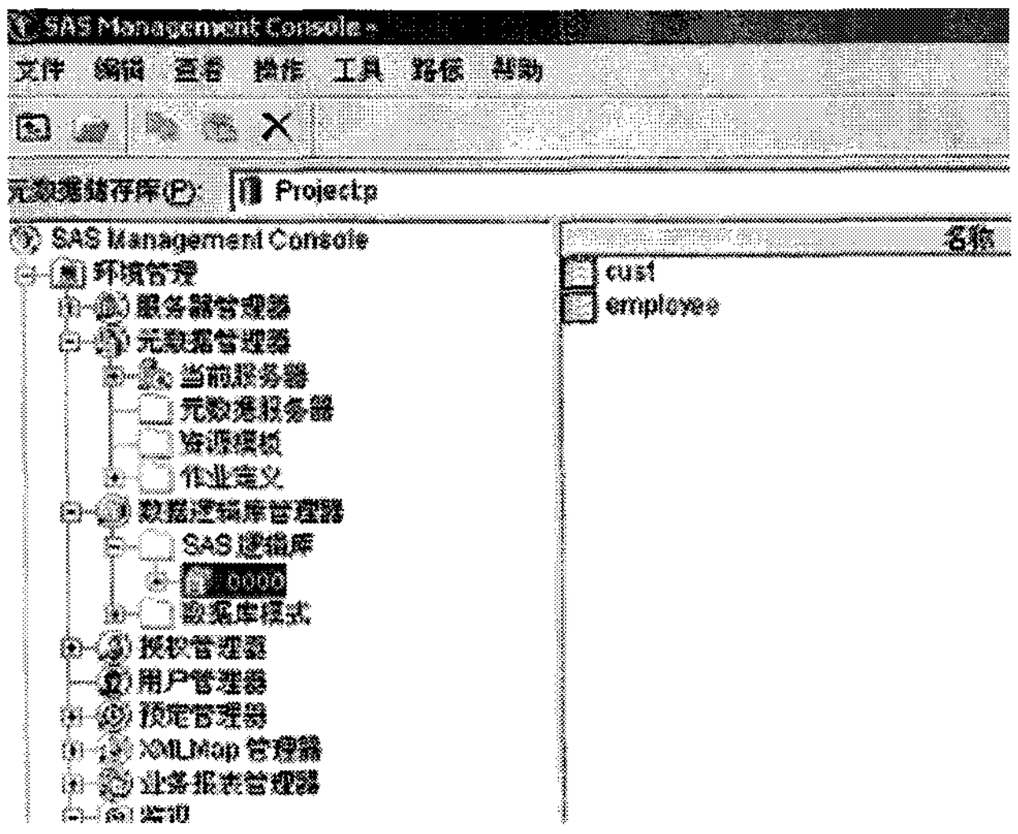

图 6　导入成功

11．可继续看导入的结果，双击 employee 表，点列标签（见图 7）。

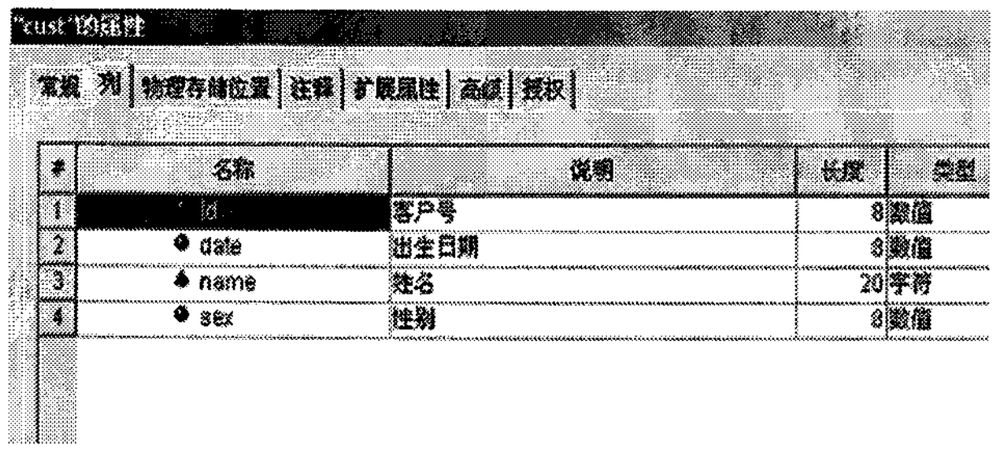

图 7　已导入到 SAS 逻辑库中的数据模型

5 总结

CWM 规范已成为业界公认的关于元数据的主流统一标准，越来越多的数据库厂家支持 CWM 规范。CWM 的优势在于它完全独立于任何具体实现的元模型，任何支持 CWM 的工具都能够相互理解其元数据实例。它不仅可以使软件厂商最终拥有建立真正可互操作的数据库、工具及应用程序所需要的公共的元模型和交换机制，也可以使用户受益，使得他们能够在选择最佳产品的同时，保证了不会因为工具之间的不可互操作性而导致投资的浪费。从实际应用的角度上来讲，由于 CWM 为数据仓库和业务分析领域中的各类软件组件的元数据定义了通用的语言，使得它们可以在不需要知道彼此的专有信息结构和接口的前提下，实现了在元数据级别上的有效集成。

参考文献:

[1] Common Warehouse Metamodel(CWM) Specification [S]. version 1.1, March 2003

[2] 王强,刘东波,王建新.数据仓库元数据标准研究 [J].计算机工程,2002,28(12):123-125.

[3] 王裕明,吴忠.商务智能中元数据管理模型研究 [J].计算机应用与软件,2005,22(8):34-35.

[4] Dr D,Chang T.CWM Enablement Showcase: Warehouse Metadata Interchange Made Easy Using CWM. [EB/OL].http://www.cwmforum.org/poperpresent.htm, 引用日期（2006-09-09）.

[5] David Marco.元数据仓储的构建与管理 [M].张铭,李钦等译.北京: 机械工业出版社,2004.

[6] 倪晟,基于 CWM 数据源的 XML 搜索引擎[D].硕士学位论文,国防科技大学,2004.

作者简介：

曾志勇(1974－)，男，贵州望谟人，高级工程师，博士，主要研究领域为数据仓库、数据挖掘、分布式计算；余建坤(1962－)，男，硕士，教授，硕士生导师，主要研究领域为数据仓库、数据挖掘。