开放信息模型的 DataCleaning Schema 设计

The designs of XML Schema of an Opening Information Model DataCleaning

隆益民　陈立仪　（广东女子职业技术学院 广东广州 511450）

摘要：基于 XML 技术，经过对信息交换平台的信息描述深入的研究，提出了开放信息模型（OIM），对信息进行统一的描述，使信息可以跨平台发布。本文介绍数据清洗模型的设计。

关键词：信息交换平台　信息模型　信息描述　数据采集　XML　数据清洗

1 前言

信息模型是描述某个工具、应用程序、数据结构或信息系统的元数据类型集合。在信息交换平台里，信息模型定义存储在计算机系统的信息（文件、文档、数据表、短消息等）里，并由应用程序使用元数据类型。与信息交换平台一起使用的信息模型采用可扩展的标记语言（XML）进行描述，所以称它为开放的信息模型（OIM：Opening Information Model）。开放的信息模型使信息的交换和发布变得更加便利和规范。

数据清洗的 XML 模型（DataCleaning Schema），是信息交换平台新增加的功能，由于信息交换平台是基于 P2P 的架构，数据源的差异性很大，为了使各个不同来源的数据有一个统一的格式，就必须增加数据清洗的功能，把数据转换成一致的数据。在分析数据的特性、研究数据清洗的必要性及数据清洗的基本原理的基础上，设计了数据清洗的 XMLSchema。

2 数据清洗的 XML Schema 的设计

2.1 数据清洗需求分析

如图 1 所示，简单地示出数据来自各个不同的数据源（来自物理上不同的数据采集点），以信息发布的方式传送到数据处理中心的过程。

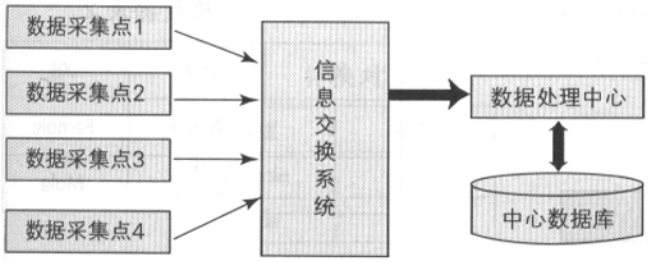

图 1　数据集成过程

在实际的数据环境中，上图就显得过于简单了，随着数据采集点的增多，数据的复杂性就成指数上升，数据的不一致性就越严重。假设银行系统里，由于数据分散在各个部门，使用自己定义的数据结构，就会出现这样的几种情况：（1）相同的数据，不同的名字；（2）不同的数据，相同的名字；（3）不同的关键字，相同的数据；（4）在一个出方出现的数据，在其它地方找不到。如图 2 所示。这些情况说明了数据缺乏一致性。

为了取出数据，必须对数据进行一致性处理。如下图 3 所示，有四个应用实例 A、B、C、D，第一种情况是它们的字段值不一样，但都表示了同一个意思：性别。为了统一数据，就要作数据值的转换，以 A 应用的文字表达为标准，其它应用都转换成 A 应用的数据格式。这样，数据就达到了一致性。第二种情况是它们的度量单位不一致，为了统一数据，就要作单位的换算。也以 A 应用的度量单位为标准，其它应用分别用相应的单位转换公式作变换。同样，数据也达到了一致性。第三种情况是定义数据表的结构时用了不同的字段名称，同样以 A 应用为标准，其它应用的字段名改成 A 应用的字段名。最后，数据达到了一致性。通过各种的变换（清洗）之后，数据的一致性得到改善。另外，数据清洗中对随机数据按正态分布的原则过滤掉边角的数据，为数据设置阀值等等。这些都说明了数据清洗的重要性。

2.2 数据清洗的过程

2.2.1 数据清洗过程分析

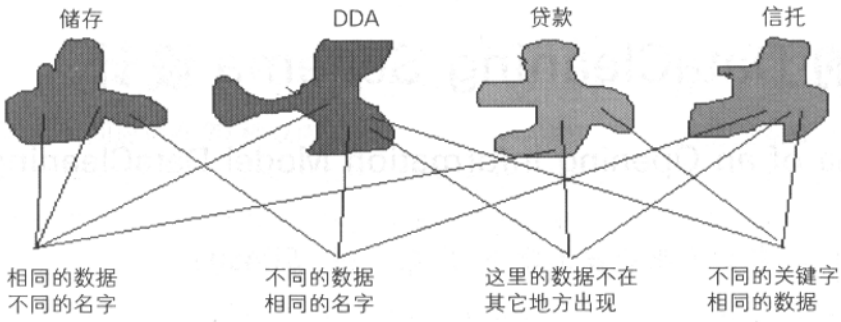

图 2　不同数据源的数据的比较

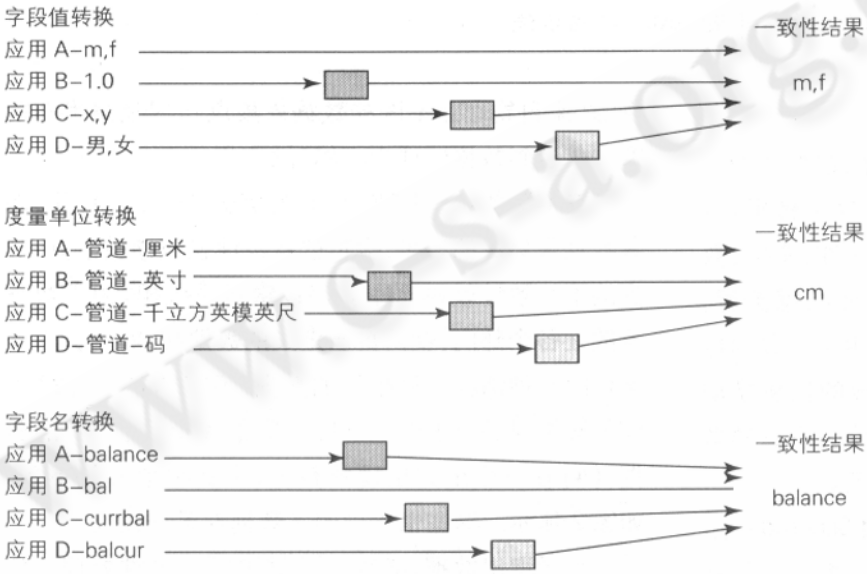

图 3　数据的一致性转换

为将现有系统环境中的数据正确移到中心数据库，必须进行集成。因此，在把数据导入之前，必须有一个对数据进行过滤清洗的过程。数据清洗的过程图，如图 4 所示。

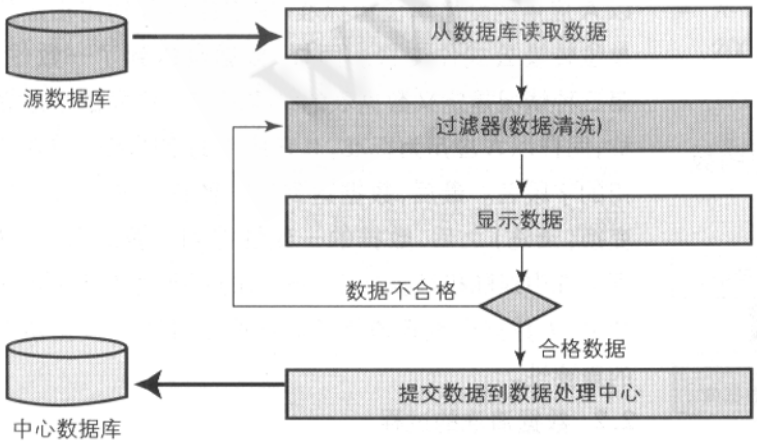

图 4　数据清洗过程

数据清洗首先从源数据库里读取数据，然后把数据送进数据过滤器进行清洗，数据清洗完之后，可以抽样显示数据，如果发现数据不合格，再进行清洗，如此循环，直至数据满意，最后把合格数据提交到中心数据库。

2.2.2　数据过滤器的功能

按照“数据库 - > 表 - > 字段名 - > 字段类型 - > 字段值”这一顺序进行分析，数据过滤器实现四个功能：段名转换；字段值的转换；字段度量单位的转换；指定缺省值。

(1) 字段名和字段值的转换

假设两个表，有相同的表结构，记录相同类型的数据，但表的字段名不同，字段的值也用了不同文字表达了同一意义，因而需要一个统一的标准。比较表 1、表 2 和表 3，首先，新数据表把字段名称统一起来，用明确意义的中文表示；其次，在性别这一字段里，字段的值也用了统一的中文表示；第三，两个原数据表的数据在无损的情况下转换到新数据表。

表 1　原数据表 1

SID	Name	Sex
1	李明志	Man
2	王清芳	Woman

表 2　原数据表 2

XH	XM	XB
1	童青桐	Female
2	张东方	Male

表 3　转换得到的新数据表

学号	姓名	性别
1	李明志	男
2	王清芳	女
3	童青桐	女
4	张东方	男

通过上面的分析，我们得出如下的四个转换对照表。其中表 4 和表 5 分别对表 1 和表 2 的字段名称进行转换，这两个转换比较容易实现。字段名称转换的实现，只要在查询语句中加入字段别名就可以。

表 4　转换对照表 1

原值	新值
SID	学号
Name	姓名
Sex	性别

表 5　转换对照表 2

原值	新值
XH	学号
XM	姓名
XB	性别

表 6 和表 7 分别对应了表 1 和表 2 的字段值，也就是实际的数据值。表 6 和表 7 将表 1 和表 2 的某一列可能的值罗列出来，通过用户界面让用户自行定义转换策略，最后提交执行。

表 6　转换对照表 3

原值	新值
Man	男
Woman	女

表 7　转换对照表 4

原值	新值
Female	女
Male	男

(2) 字段度量单位的转换和指定缺省值。

2.3 数据清洗的算法

数据清洗的算法流程，如图 5 所示。首先连接数据源，取得要清洗的数据表，然后读取数据表的结构信息。接着进入数据清洗处理流程。数据清洗决策中心的作用是：根据实际数据的情况进行分析，为数据的清洗制定清洗的规则方法，然后把数据的转换策略送给相对应的清洗处理流程进行处理。例如，当字段名的转换处理完毕后，把结果送到字段值的转换处理流程，处理完毕后再送到字段缺省值和字段度量单位转换处理流程处理。接着，把数据清洗结果送到检验中心进行检验，如果数据不是理想数据，则重复上面的流程进行处理，直到得出满意的数据为止。最后，把整个数据清洗的策略保存起来。整个清洗流程结束。

2.4 数据清洗的 XML Schema 设计

根据数据清洗的算法，“清洗”并不是真正意义上

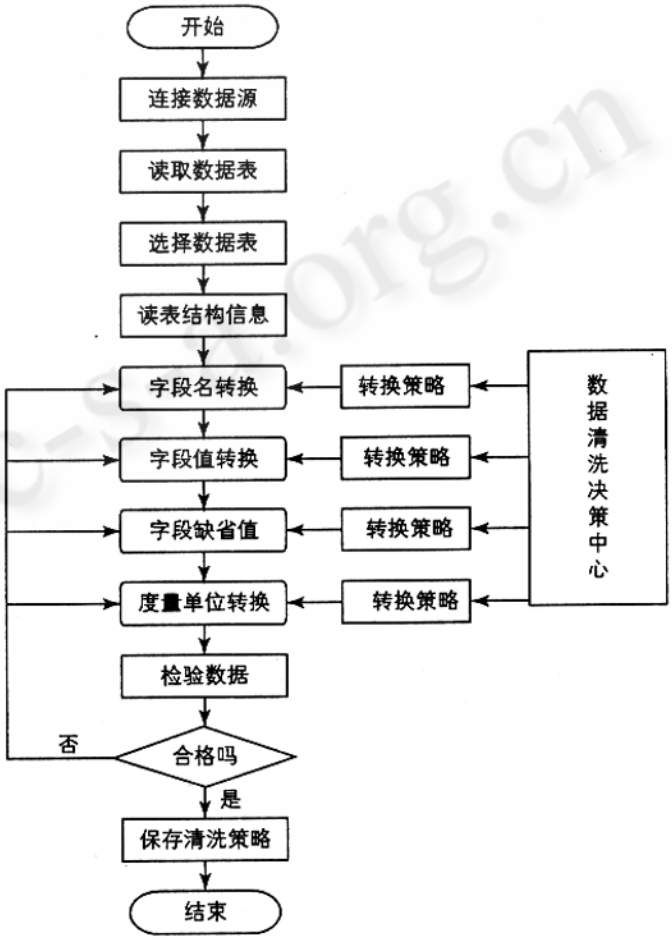

图 5　数据清洗算法流程

的清洗，只是通过对数据的分析，为数据制定一个清洗的策略。所以，清洗的最终结果不是得到数据，而是得到数据清洗的策略，我们需要把这个策略保存下来，以便日后执行真正的“清洗”。

为了得到开放统一的数据格式，采用了 XML 格式来存放这些策略，实现数据清洗的 XML Schema 的设计。

设计数据清洗的 XML Schema 时，先定义一个全局的 MapType 类型组件，如图 6 所示，这个组件有两个子元素，一个称为“源”，另一个称为“目标”；“源”又包含两个子元素：源表和源字段，“目标”也包含两个子元素：目标表和目标字段。因为在一种类型转换里，可能有多对 <Source，Object>，所以，整个类型的基数是 1..∞。

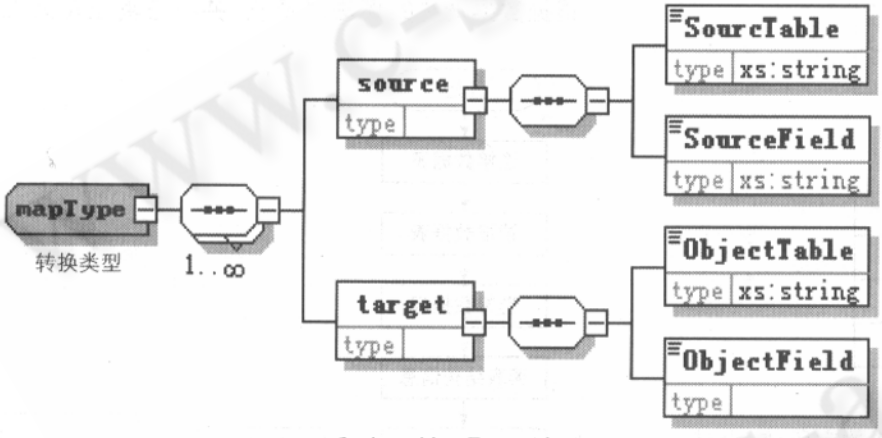

图 6　MapType 模型组

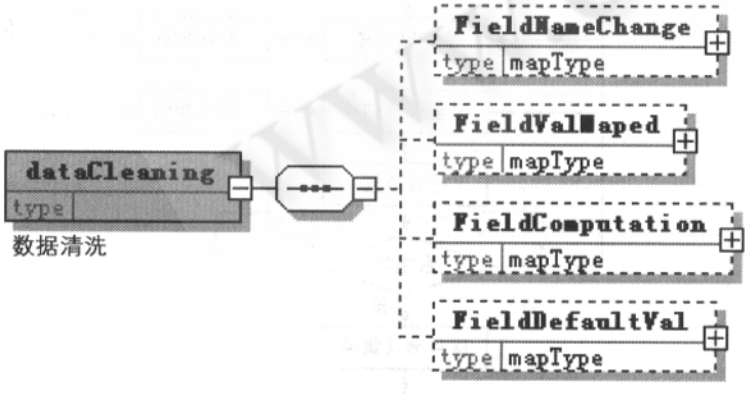

图 7　DataCleaning 模型组

有了 MapType 类型组后，定义整个数据清洗的模式，如图 7 所示，根元素 dataCleaning 下有四个子元素，分别描述了字段名转换、字段值转换、字段计算（字段度量单位转换）、字段缺省值设置。因为每种转换操作都是可选的，所以各元素的最小基数是 0。至此，数据清洗的模型完成。

3　结论

在数据交换系统里，信息的生命周期包括信息采集、信息定义、信息发布、信息消费。数据采集是信息生命周期的开始，在整个系统里面起着举足轻重的作用，而数据清洗又是信息采集的一个重要环节。为了实现数据的无缝交换，定义一个统一的格式来描述数据是很有必要的。文章对数据作了详细分析，然后设计出数据清洗的算法流程，最后实现数据清洗的 XML 模式设计，为整个信息交换系统的信息模型的实现打下了坚实的基础。

参考文献

1 XML Schema Part 0：Primer W3c Recommendation 2 May 2001. http://www.w3.org/TR/2001/REC-xmlschema-0-20010502

2 Patrick O'Neil Elizabeth O'Neil Database Principles，Programming，and Performance（第二版）.1998.

3 MSxml4.0 SDK Document Microsoft XML Code Services.

4 [美] Peter G. Aitken. 微软 XML 技术指南北京，中国电力出版社 2000：105～155.