

基于元数据的海量试验数据管理系统研究与设计

刘 斌

(西安邮电学院, 陕西 西安 710121)

摘 要: 分析了试验数据管理发展现状及存在的不足, 提出了基于 RDF 元数据模型的试验数据描述语言 (Testdata Description Language, TDL), 给出面向海量试验数据管理的系统平台设计思路和架构、以及具体试验数据解析算法. 通过对设计架构的分析, 表明该设计架构能够有效满足试验数据管理的系统平台化需求, 提高了科研产品研发效率, 降低了试验成本, 同时为数据管理平台化建设积累了经验.

关键词: 试验数据管理, 元数据, 全文索引, TDL

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2012)04-0158-04

Metadata-Based Mass Test Data Storage System Study and Design

LIU Bin

(Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract: With the wide usage of test technology, test has becomes one of the most important methods for product development. The existing test data management system can't solve the test data management problems. This paper based on the background of mass test data management, puts forward test data description language (TDL) that based on Resource Description Framework (RDF), and supplies mass test data management oriented system design structure, and also presents algorithm of test data analysis. The test data management system meet the demand of openly platform.

Key words: mass data management; metadata; full-text indexing; Test Data Description Language

1 引言

随着试验技术在各行各业产品研发过程中的广泛应用, 试验数据格式多样, 且以文件的形式存储较多, 数据管理困难, 对试验数据管理系统在管理效率、平台化、精确性以及安全性等方面提出了越来越高的要求^[1].

目前, 在试验数据管理系统^[2]的研究上, 比较成熟的系统都集中在针对某一种或者几种试验的数据管理上. 其主要的研究方向集中在如下几个方面: (1) 通过手工建库及手工创建数据导入脚本的方式来实现试验数据导入的工作; (2) 通过实现多种接口, 在遇到试验仿真设备或者数据格式发生变化时, 进行有针对性地二次开发; (3) 改变试验数据的存储

格式, 以 XML 文件的方式保存试验数据, 并通过试验数据管理系统保存到数据库^[3].

试验数据管理面临的问题如数据缺乏统一有效的保存和管理、数据处理效率低下、准确性难以保障、综合利用率低下、不具备平台化等.

2 相关主要技术

试验数据管理相关技术较多, 如 Oracle 数据库导入工具、元数据、全文索引、Hadoop 等.

元数据^[4]可以很好的解决平台化需求, 对试验数据进行描述, 数据管理系统只要通过相关的元数据来识别试验数据, 不再需要关注仿真设备.

元数据最本质的定义为 Data about Data. 在不同的领域, 元数据有不同的定义和应用.

由于元数据的广泛存在,元数据格式也呈现出了多样性.下面是几种不同格式的元数据^[5].

(1) 编码文档描述(EAD)

EAD 主要用于描述档案和手稿资源,并利用网络检索来获取档案手稿类信息资源.其顶层元素主要有有一个头标和一个档案描述.每一层元素又包括多个小项以及若干细项.

(2) 机读目录(MARC)

机读目录(Machine-Readable Cataloging)被广泛用于图书馆书目记录数据,是目前图书馆描述、存储、交换、处理以及检索信息的基础. MARC 记录的总体结构有以下特点:可变格式和可变长字段的记录格式.每条 MARC 记录分三个区(头标区、目次区、数据区).

(3) 都柏林核心集(DC)

DC 元数据基于 XML 格式,是在充分吸收了图书情报界所具有的编目、分类及计算机、网络的自动搜索、编目、索引、检索等研究成果的基础之上发展起来.它是描述、支持、发现、管理和检索网络资源的信息组织方式,其最大的特点是数据结构简单,信息提供者可直接编码.

(4) 资源描述框架(RDF)

资源描述框架(Resource Description Framework)是一个能对结构化元数据进行编码,交换及再利用的体系框架. RDF 采用 XML 作为交换和处理元数据的通用语法结构体系,提供了一种强有力的句法转换模式.

3 试验数据描述语言 TDL

根据 TDL(Test Data Description Language)描述试验数据功能的不同,试验数据描述分为基本信息描述和存储信息描述.

3.1 试验数据基本信息描述

试验数据的基本属性包括:试验整体信息,试验名称,仿真设备,数据输出格式,数据输出频率,数据输出频率和数据量,以及每列数据的:数据编码,数据类型,长度,默认值,是否允许为空^[6].图1是 TDL 描述试验数据的基本信息示意图.

3.2 试验数据存储信息描述

试验数据的存储信息包括:存储试验数据的数据表,创建数据库及数据表的方法,试验数据的列和存储表之间的对应关系,试验数据的解析方法.图2显示试验数据存储部分的元数据描述内容.

3.3 TDL 数据接口

TDL 是试验数据管理系统和仿真试验的接口,

图3中描述了 TDL 元数据管理模块和仿真试验及试验数据管理系统之间的关系. TDL 降低了试验数据管理系统和仿真设备的耦合性,具有极大的开放性.

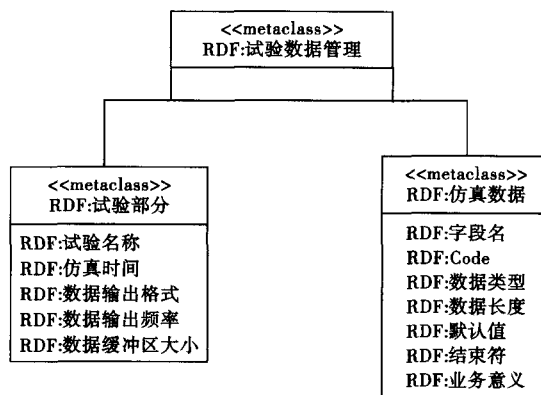


图1 元数据模型

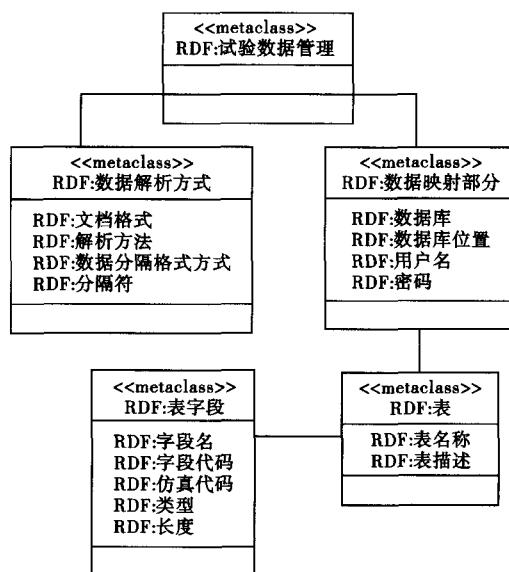


图2 试验数据存储部分元数据表述

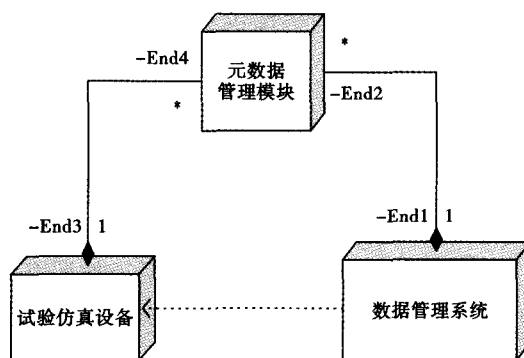


图3 TDL 元数据管理模块关系

3.4 试验数据解析算法

仿真设备在发送完 TDL 文件后,试验数据管理系统进行相关的检查和设置.设置完成后,仿真设备发送对应的试验数据文档.试验数据管理系统收到试验数据或者试验数据文档后,按照 TDL 中对试验数据的描述,对试验数据进行解析,并将解析后的试验数据存储到试验数据管理系统中.算法描述如下:

输入值:输入值为两个数据文件,一是系统定义的 TDL 文件,二是需要解析的试验数据文件.

输出值:无,系统解析每一行试验数据.并逐把分析的结果保存到指定的数据库中.

算法流程:图 4 为试验数据解析流程图.

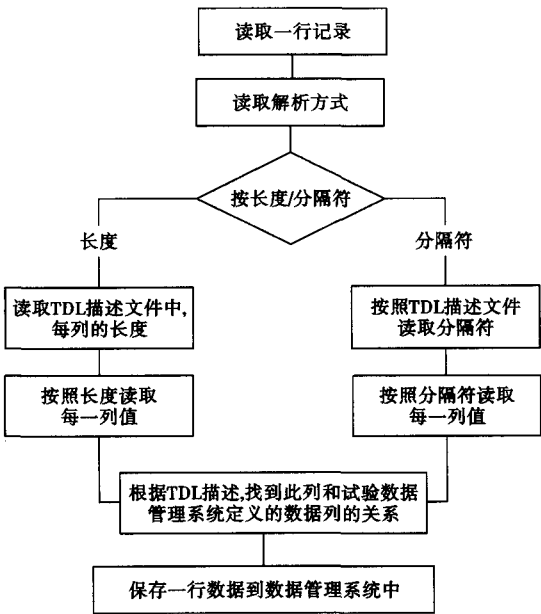


图 4 试验数据解析流程

(1)试验数据管理系统从试验数据文档中读取一行记录.

(2)读取 TDL 中对解析方式的定义,按照分隔符划分列和按照每列长度进行解析,并读取数据分隔符.

(3)如果是按照分隔符划分列,系统按照分隔符对读取到的一行记录进行解析.如果是按照长度进行解析,系统循环从 TDL 中读取每列的长度,并截取对应的列.

(4)一列的数据解析完成后,系统按照 TDL 定义的对对应关系,将试验数据保存到试验数据管理系统中.

4 试验数据管理平台设计

在设计试验数据管理系统时,平台化是系统设计的目标.试验数据管理平台利用基于 RDF 元数据框架的试验数据描述语言(TDL),实现系统开放性平台的设计.试验数据管理系统主要包括元数据管理模块,数据解析和导入模块,试验项目管理模块,试验文档检索模块及其他辅助模块.

4.1 设计架构图

图 5 是试验数据管理系统结构图.

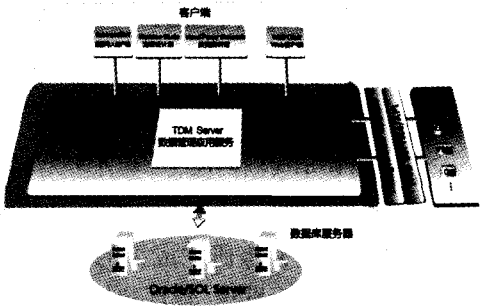


图 5 系统功能框架图

4.2 试验项目管理

面向航空航天领域的试验管理系统以试验项目为单位进行管理,其中包括试验项目基础信息管理,项目人员管理,项目设备管理,项目权限管理.图 6 是试验项目管理模块的用例图.

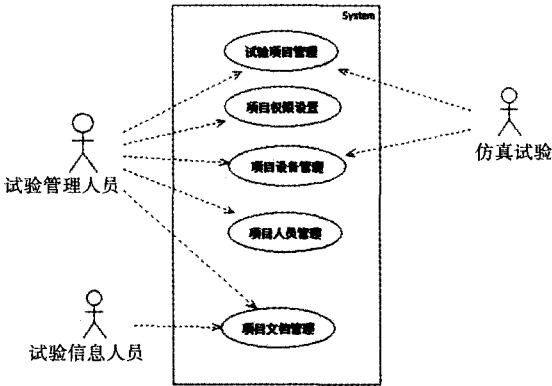


图 6 试验项目管理用例图

4.3 元数据定义

为了方便试验管理人员定义对应试验的 TDL 实例,系统提供专业的用户界面来帮助用户定义 TDL 实例,界面设置完成后,系统自动生成相对应的 XML 格式的 TDL 实例.

从功能上区分,此部分主要分为三个子模块,仿真数据定义模块,试验数据存储定义模块和试验数

据解析模块。

首先,试验管理人员从试验项目列表中选择试验项目,选择此试验对应的仿真设备输出的试验数据文档的格式及解析方式。

下一步需要定义试验数据。试验数据的定义包括数据名称,数据类型,长度,区分此字段时是通过固定长度区分或者是通过分隔字符串,数据意义。表 1 中显示了定义试验数据需要保存的内容。

表 1 试验数据定义内容

试验数据定义	
数据名称	试验数据中,某一列数据的名称
数据类型	此列数据的类型
数据长度	数据的最长长度
分隔方式	按照固定长度分隔或者是按照固定的分隔符
业务意义	描述此列数据业务意义
默认值	

其次要定义试验数据在系统中存储的方式和格式。例如,数据库名称,数据库 ID,数据库存储位置,创建数据库的脚本,访问数据库的用户名称和密码。

数据库定义完成后,需要定义存储数据的数据表格式。其中包括数据类型,数据属性,校验信息及关联关系。

创建存储试验数据的数据表需要定义表格名称,表格对应的数据列,要实现上述功能,需要设计保存对应元数据的数据表。表结构如表 2 所示。

表 2 数据表元数据设计

Table Name: MD_Table			
Description: 元数据管理表格管理			
Column:	Type	Length	Nullable
TableID	int	4	FALSE
DatabaseID	int	4	FALSE
TableName	Varchar	30	FALSE
Table Description	Varchar	200	TRUE
Table Name: MD_Column			
Description: 元数据管理表格列管理			
Column:	Tpe	Length	Nullable
ColumnID	int	4	FALSE
TableID	int	4	FALSE
ColumnName	Varchar	30	FALSE
ColumnTpe	SmallInt	2	FALSE
Length	int	4	FALSE
Nullable	BIT	1	FALSE
Remark	Varchar	200	TRUE

实现的功能为:(1)输入需要创建的表名;(2)表功能描述及说明;(3)输入表格每列的列名,类型,长度,是否允许为 Null,每列说明。

系统提供了多种试验数据解析方法。试验数据存储位置定义完成后,紧接着就需要定义试验数据的解析方法。根据第一步定义的试验数据保存格式,选择对应的数据解析方法,并设置数据分隔符,数据缓冲区等。

5 结束语

本文利用试验数据描述语言(TDL),实现了面向海量试验数据管理的平台设计。该设计架构降低了试验数据管理系统和仿真设备之间的依赖性,使系统成为了一个开放的平台,满足了数据数据管理系统开放性的需求,提高了试验数据管理效率,降低了成本。该系统目前已经在某航空科研单位试用,并取得了初步的认可,相信随着用户的使用和不断反馈,系统将会进一步完善。

参考文献:

[1] Yongxing Luo, Lijun Jiang, Tian-ge Zhuang. Data management for grid environments high-performance computing[J]. Computer Science, 2001: 151-160.

[2] 李沛然, 邓兆祥. 汽车 NVH 试验数据管理系统关键技术研究[J]. China Academic Journal Electronic publishing house 2010(29):163-166.

[3] Daniel J. Abadi data management in the cloud: limitations and opportunities[J]. IEEE Computer Society Data Engineering. 2009,32(1):3-12.

[4] Xiu An, Jian Huang, Wei Yu. Advanced research on information science[C]// International Conference on information science, Automation and Material System Zhenzhou; IEEE,2011:219-220.

[5] Grahm R W. The global reach ombined test force enterprise test data management system (ETDMS) [C] // American Nashville, Institute of Aeronautics and As2. TN, United States; IEEE,2011.

[6] James Hollenbach, Joe Presbrey, Tim Berners-Lee. Using RDF metadata to enable access control on the social semantic web[D]. USA: Decentralized Information Group, 2010.

作者简介:

刘 斌 男,(1968—),硕士,工程师。研究方向为软件工程。