# 元数据

## 1 元数据模型

### 1.1 原模型

#### 1.1.1 原模型概念

元模型是对元数据的定义，包括包、类、数据类型三种要素。包是将类、数据类型组织成语义上相关联的分组，每个类、数据类型只能属于一个包，包可以嵌套在其他包中，但只能嵌套在一个包内。元模型管理功能提供了浏览包、类、数据类型，增加、删除和修改包、类、数据类型信息的功能。

 **包**：是一个容器，它可以把元模型相关的类、数据类型按照特定的元数据来 源分组。

 **类**：定义了元数据对象的类型。例如，对于从 PowerCenter 这个 ETL 工具 中来的元数据，在 PowerCenter 包中，会定义一个 Source 类，用这个 类可以描述 PowerCenter工具中定义 Mapping 的源。类自身具有属性， 类之间具有关系，包括组合关系、依赖关系和继承关系。

**数据类型**：是用来定义属性的，例如 Source 类的“描述”属性，其数据类型是 文本型的，这样元数据系统就可以识别出该如何向用户展示这个属性。

#### 1.1.2 原模型结构

典型的元模型结构可以描述为：信息层、模型层、元模型层和元元模型层

**1．信息层（information layer）**

信息是由我们希望描述的数据组成，这些数据通常是一些用户数据(user data)，主要职责是描述信息领域中的详细信息。

**2.模型层（model layer）**

模型层是由元数据组成，元数据是描述信息层的数据，元数据的集合被称作为模型。模型层的主要职责是为描述信息层而定义的一种“抽象语言”(即没有具体语法或符号的语言)。信息层的数据，即用户数据，是模型层的一个实例。

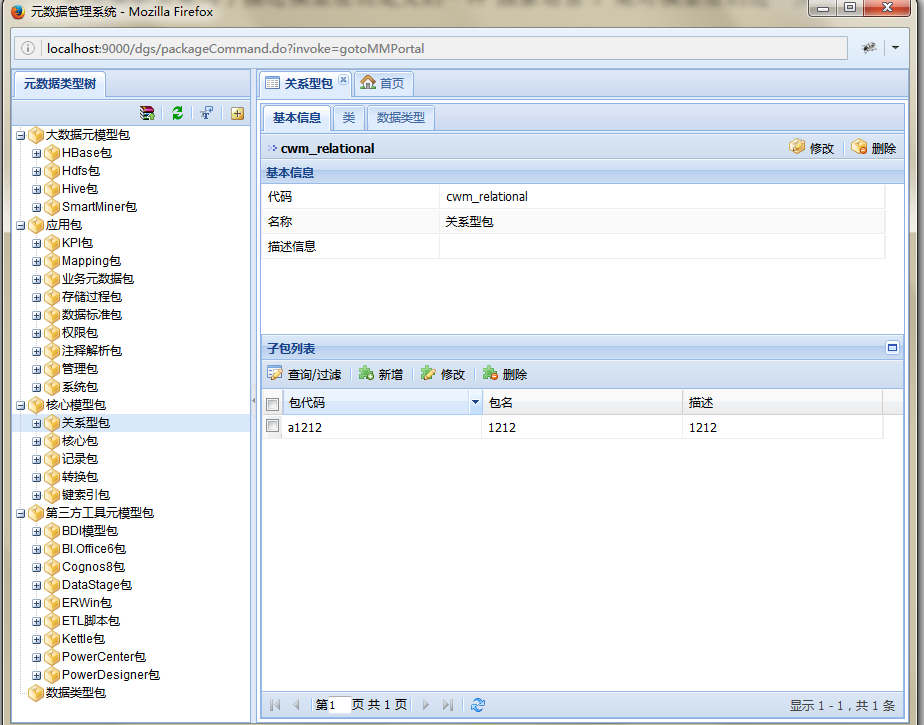
**3．元模型层（metamodel layer）**

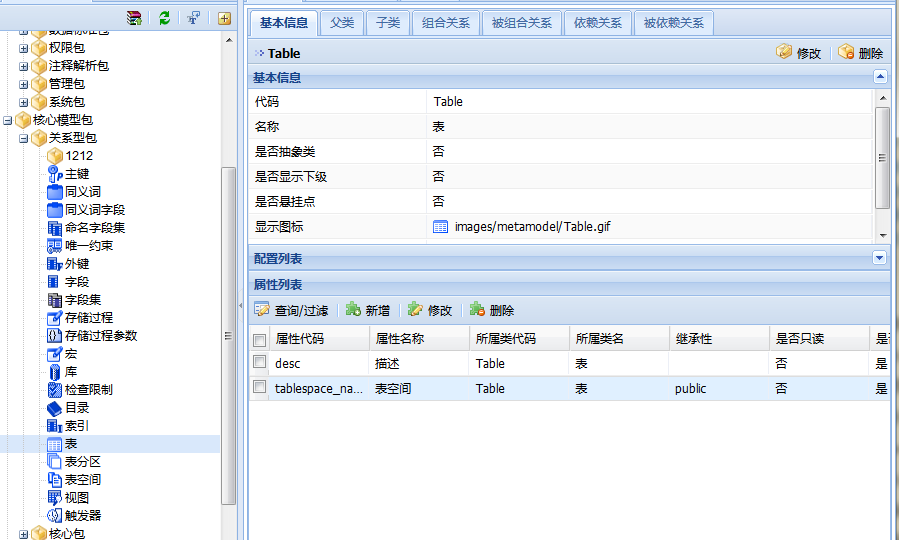
元模型层是由元一元数据组成，元一元数据 定义了元数据的结构和语义，元一元数据的集合被称作为元模型。元模型层的主要职责是为了描述模型层而定义的一种“抽象语言”，是对模型层的进一步抽象。也就是说，模型层描述的内容通常要比元模型层描述的内容丰富、详细。一个模型是元模型的一个实例。数据词典中的元数据是对[数据模型](http://baike.baidu.com/view/72142.htm)的描述。

**4．元元模型层（meta-metamodel layer）**

元元模型层是由元元数据的结构和语义的描述组成，这层的主要职责是为了描述元模型而定义的一种“抽象语言”。元元模型的定义要比元模型更加抽象、简洁。一个元元模型可以定义多个元模型，而每个元模型也可以与多个元元模型相关联。通常所说的相关联的元模型和元元模型共享同一个设计原理和构造，这也不是绝对的准则。每一层都需要维护自己设计的完整性。一个元模型是元元模型的一个实例。

#### 1.1.3 原模型显示





普元原模型管理中，根据功能的不同，分为大数据元模型包、应用包、核心模型包、第三方工具元模型包。每个功能包下都包含了具体的原模型包，而在具体的元模型包中包含了一组不同功能的原模型映射。如关系型包，包含关于数据库原模型的映射。

在原模型中，可以显示模型的基本信息、父类、子类、组合关系、被组合关系、依赖关系、被依赖关系。

### 1.2 元数据

#### 1.2.1 元数据定义

元数据（Metadata）是描述其它数据的数据（data about other data），或者说是用于提供某种资源的有关信息的结构数据（structured data）。元数据是描述信息资源或数据等对象的数据，其使用目的在于：识别资源；评价资源；追踪资源在使用过程中的变化；实现简单高效地管理大量网络化数据；实现信息资源的有效发现、查找、一体化组织和对使用资源的有效管理。

#### 1.2.2 元数据特点

1元数据一经建立，便可共享。元数据的结构和完整性依赖于信息资源的价值和使用环境；元数据的开发与利用环境往往是一个变化的分布式环境；任何一种格式都不可能完全满足不同团体的不同需要；

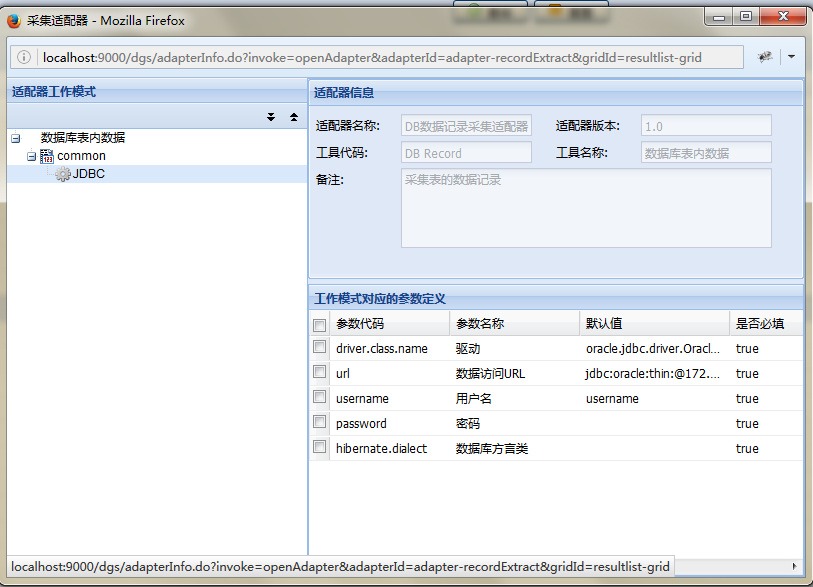
2元数据首先是一种编码体系。元数据是用来描述数字化信息资源，特别是网络信息资源的编码体系，这导致了元数据和传统数据编码体系的根本区别；元数据的最为重要的特征和功能是为数字化信息资源建立一种机器可理解框架。

元数据体系构建了电子政务的逻辑框架和基本模型,从而决定了电子政务的功能特征、运行模式和系统运行的总体性能。电子政务的运作都基于元数据来实现。其主要作用有：描述功能、整合功能、控制功能和代理功能。

### 1.3 适配器

#### 1.3.1 适配器概念

适配器是指针对不同的**元数据来源**，采用相应的采集方式**获取元数据的程序**。适配器包括对应的工具或产品的名称、版本，适配器工作的模式，工作参数列表。



#### 1.3.2 适配器的边界

目前普元数据采集系统支持的适配器有:



经过测试。有的适配器存在bug。如BI.office采集适配器，由于公司smartBi版本过高，程序无法解析而报错。

### 1.4 数据源

配置数据源是采集元数据的基础，它需要用户选择使用什么**适配器**，被采集的工具或产品的版本是什么，适配器需要按什么工作模式执行采集，以及采集所需要的连接配置参数，比如说采集数据库数据字典时，需要提供例如数据库的 IP 地址、端口、用户、密码等信息。数据源管理就是对采集配置信息的管理维护，您可通过界面来维护。而且根据选择的适配器挂载数据源，要保证**元数据模型要和适配器匹配**。

## 2 元数据采集

元数据采集是指获取数据源数据，对数据进行转换，然后将数据写入到元数据系统中的过程。整体上可分客户端与服务器端，客户端包括适配器，数据源，采集任务等的配置，服务器端则负责真正的采集数据、转换数据、入库落地等操作。元数据采集管理包括适配器管理、数据源管理、任务配置、手动采集、采集入库审核、采集日志查看。

### 2.1 DB数据字典采集

#### 2.1.1 采集说明

数据字典采集是采集数据库的数据结构及相关对象。包括表、视图、存储过程、存储函数等信息。这些采集模型是通过元数据模型来定义的。如果想要采集数据库其他信息，可以通过修改数据库的原模型。

DB数据字典采集适配器存在三种采集模式 1 JDBC自动采集，2文件上传手动采集，3 服务器文件定时采集。

#### 2.1.2 JDBC自动采集

**创建元数据**

创建元数据，选择数据库元模型（元数据类型树-->核心模型包-->关系型包-->目录）

**创建数据源**

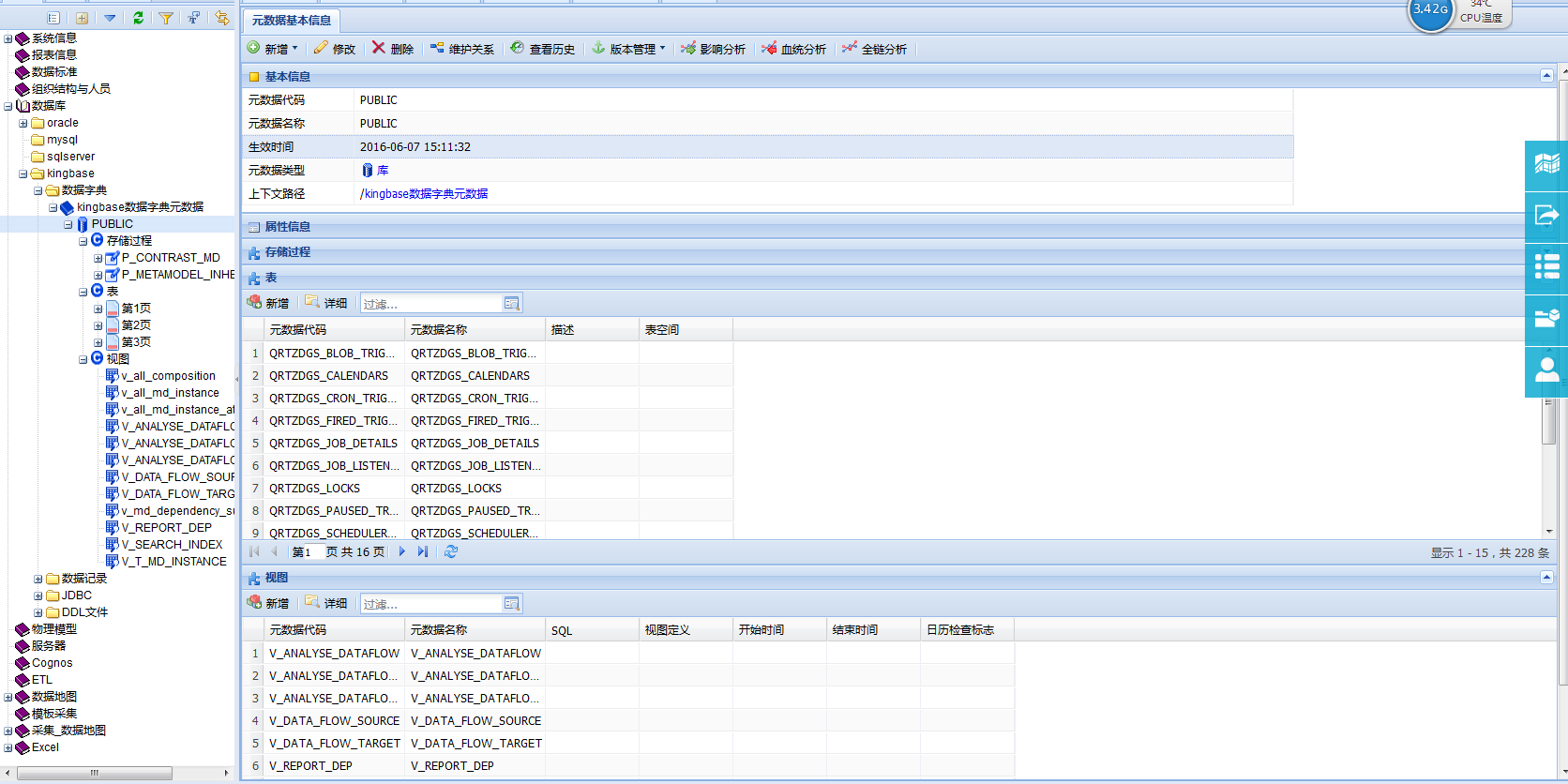
创建数据源，选择DB数据字典采集适配器(选择JDBC采集模式)。选择挂载点(上一步创建的元数据)。

**创建采集任务**

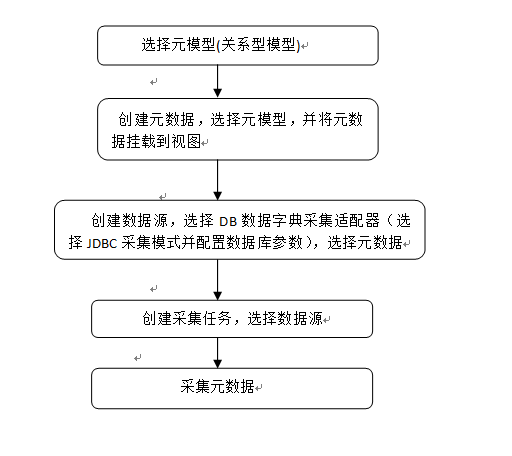
创建采集任务，选择上一步创建的数据源

**元数据采集**

任务保存完毕，可以直接运行或者定时运行（使用定时器）。采集过程中，可以通过采集日志来查看采集情况。采集成功完成后，元数据就可以通过视图显示出来。元数据采集结果如下:



采集流程图



#### 2.1.2文件上传手动采集

**创建元数据**

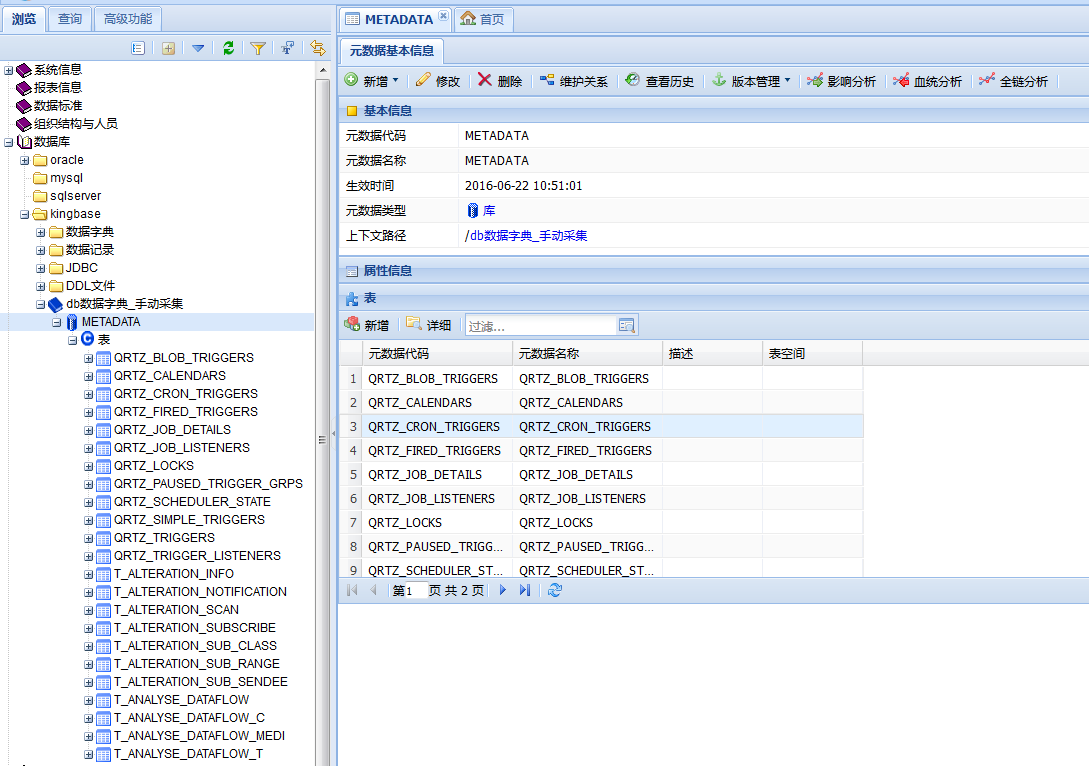
创建元数据，选择数据库元模型（元数据类型树-->核心模型包-->关系型包-->目录）

**创建数据源**

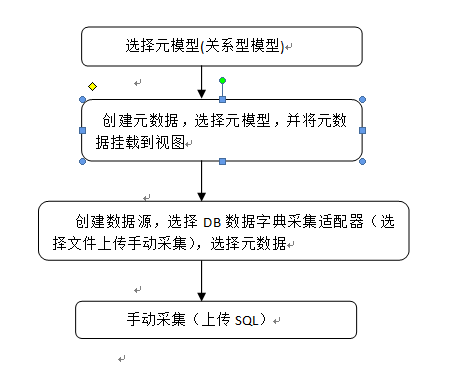
创建数据源，选择DB数据字典采集适配器(选择文件上传手动采集，普元手动采集DB数据字典只支持oracle的sql文件)。选择挂载点(上一步创建的元数据)。

**手动采集**

选择上一步创建的数据源，上传sql文件（支持sql，txt，zip三种文件格式）。然后点击保存，即可开始采集。采集过程中，可以通过采集日志查看采集过程。采集结果图如下：



**采集流程图**



#### 2.1.3服务器文件定时采集

略

### 2.2 DB数据记录采集

#### 2.1.1 采集说明

DB 数据记录采集是指直接采集数据库中表或视图的数据，这些数据可以是任意方面的元数据，例如可以是标准代码、指标定义、检核规则，也可以是结构化的数据字典等等。与 DB 数据字典采集类似，它也采用 JDBC 方式连接数据库，因此配置参数大致是一样的，如驱动类、数据访问 URL、用户名、密码。但是需要选择方言，因为要适应多版本的数据库。

#### 2.1.2 JDBC采集模式

**创建元数据**

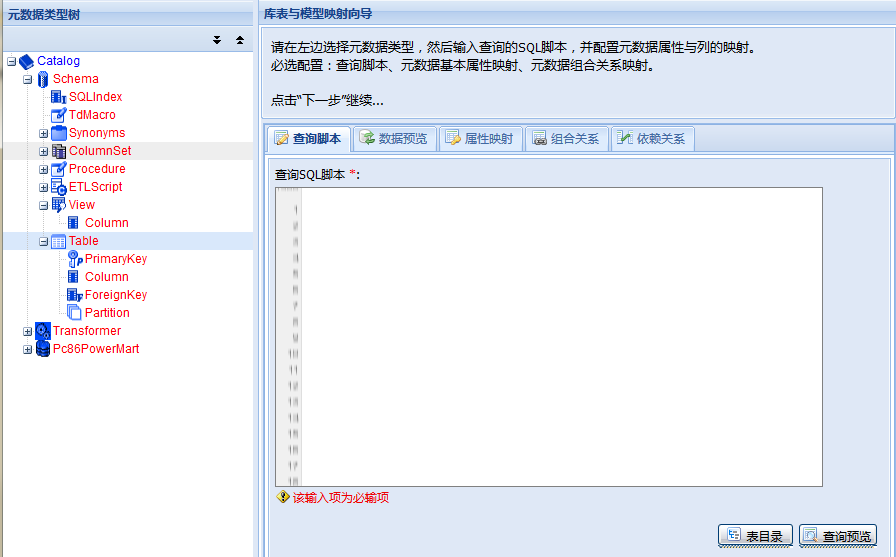
创建元数据，选择数据库元模型（元数据类型树-->核心模型包-->关系型包-->目录）

**创建数据源**

创建数据源，需要选择适配器（DB数据记录采集适配器），并且配置好数据库参数、方言，选择元数据(上一步创建的元数据)。

**DB数据采集配置**

创建DB数据采集配置，选择上一步创建的数据源。表、字段、主键等关系。



采集结果完全依赖于配置的关系。可以通过配置属性映射、组合关系、依赖关系来生成用户需求的数据。

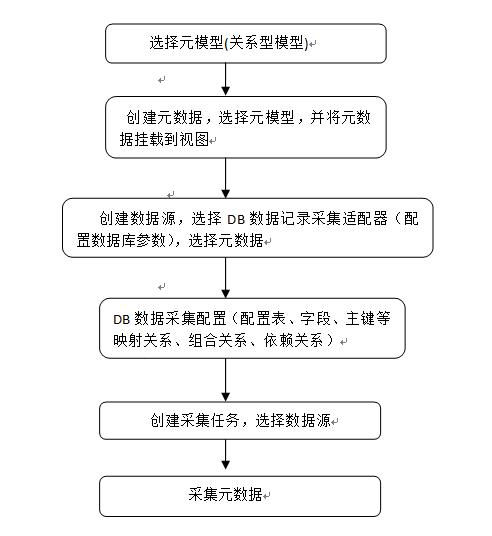
**创建采集任务**

创建采集任务，选择上一步创建的数据源

**元数据采集**

任务保存完毕，可以直接运行或者定时运行（使用定时器）。采集过程中，可以通过采集日志来查看采集情况。采集成功完成后，元数据就可以通过视图显示出来。

流程图



### 2.3 JDBC驱动采集

#### 3.1.1 采集说明

JDBC驱动采集类似于DB数据字典采集，是采集数据库的数据结构及相关对象。包括表、视图、存储过程、存储函数等信息。这些采集模型是通过元数据模型来定义的。如果想要采集数据库其他信息，可以通过修改数据库的原模型。

适用于采集不常见的数据库数据字典，只要该数据库实现了JDBC连接。需要注意的是，通过这种方式采集过来的数据字典，无法采集到数据字典之间的关系，如外键关联。

#### 3.1.2 JDBC自动采集

**创建元数据**

当创建元数据，选择数据库元模型（元数据类型树-->核心模型包-->关系型包-->目录）

**创建数据源**

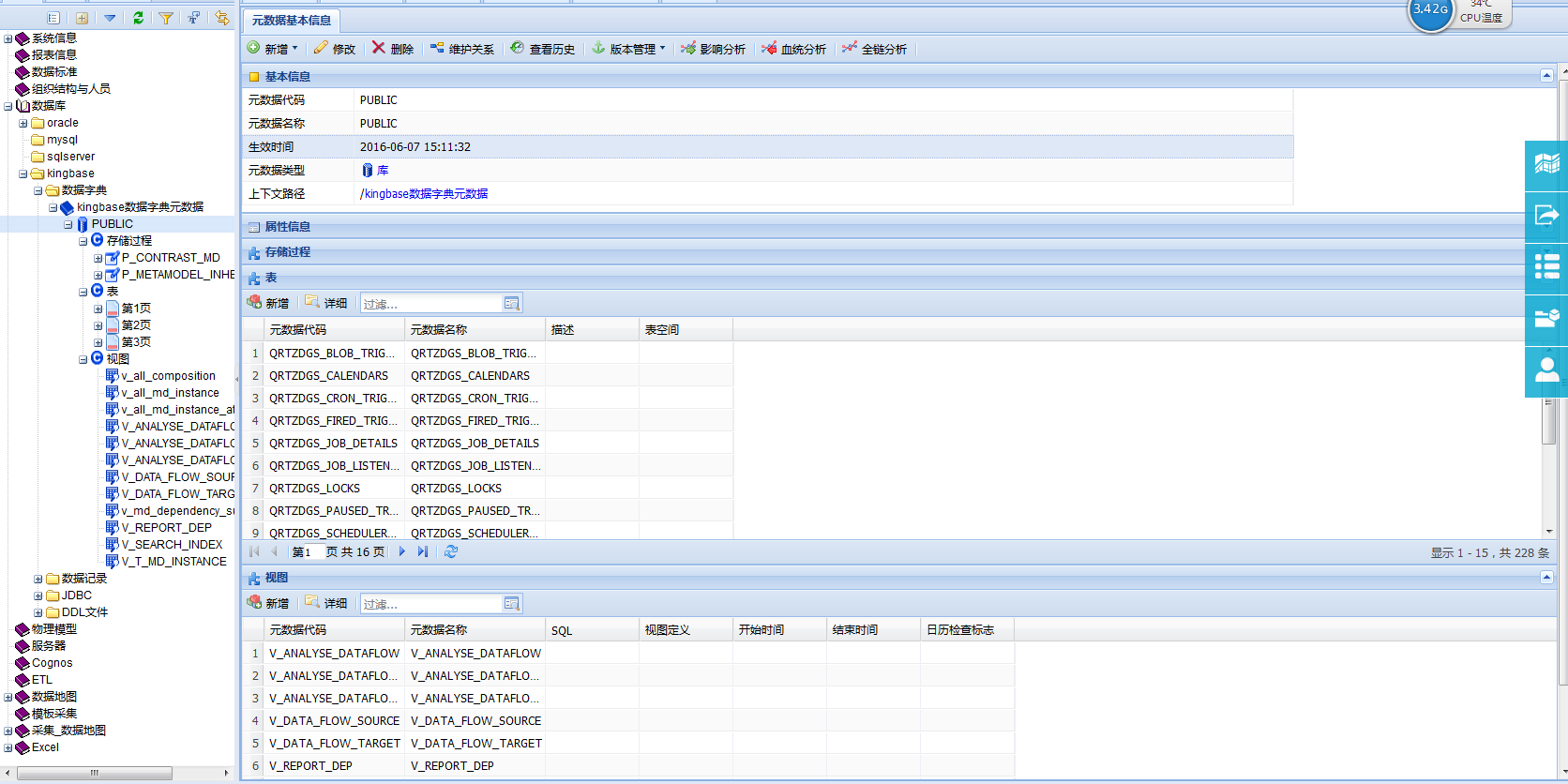
创建数据源，需要选择适配器（JDBC驱动采集适配器），并且配置好数据库参数，选择挂载点(上一步创建的元数据)。

**创建采集任务**

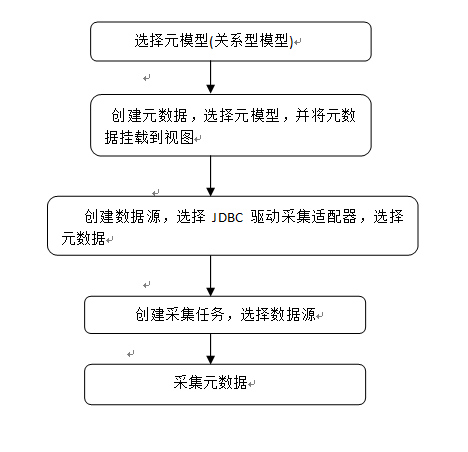
创建采集任务，选择上一步创建的数据源

**元数据采集**

任务保存完毕，可以直接运行或者定时运行（使用定时器）。采集过程中，可以通过采集日志来查看采集情况。采集成功完成后，元数据就可以通过视图显示出来。元数据采集结果如下:



流程



### 2.4 数据地图采集

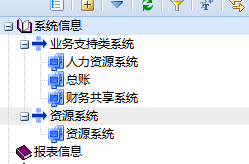
#### 2.4.1 数据地图简介

数据地图为您从宏观角度展示**系统以及系统之间数据流向关系**，并支持从系统到数据库的下钻及相关信息的展示。通过 WEB 方式编辑元数据或通过模板方式采集元数据都可进行元数据创建；对于创建好的初始数据地图，如果您不满意数据地图的布局，可通过数据地图位置调整及保存来调整数据地图；对于数据地图中单节点的信息可通过双击节点展示数据地图节点详细信息； 双击两个系统节点之间连线上的锚点 可下钻到这两个节点下的数据库之后。

#### 2.4.2 手动创建数据地图

**创建系统**

创建多个系统层次，并在每个系统层次中创建一个或者多个系统。

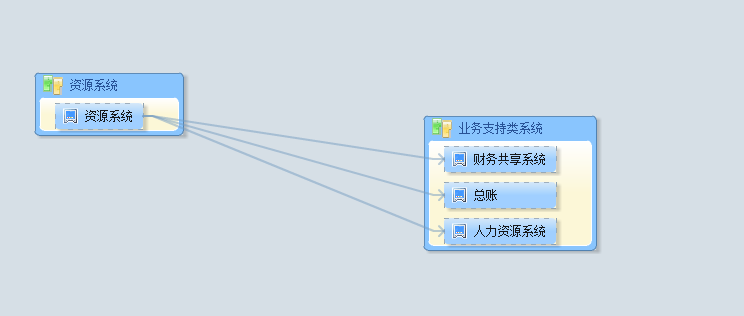


**添加系统之间的关系**

点击系统，选择维护关系。

**数据地图展示**

添加关系之后，就可以在数据地图中，查看简历的系统关系。数据地图如下



#### 2.4.3 模板采集创建数据地图

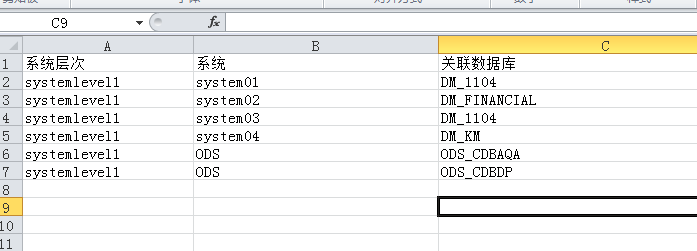
**创建元数据**

创建元数据，选择系统层次原模型(元数据类型树-->应用包-->系统包-->系统层次)

**创建数据源**

创建数据源，选择Template采集适配器

**建立模板映射**



**手动采集**

手动采集，上传模板映射文件。采集完成后，点击数据地图。即可展示采集的数据地图。

#### 2.4.4数据地图的下钻

为进一步查看两个系统之间所属数据库之间关系，您可通过双击两个系统节点之间连线上的锚点下钻到这两个节点下的数据库之间关系，可按以下步骤操作：

1. 登录产品，在首页中显示数据地图。

2. 鼠标悬到两个系统之间的连线上，系统会自动显示锚点。

3. 双击锚点，产品会展示两个系统所属数据库之间的关系。

4. 为进一步查看数据库节点的详细信息，您可通过双击该节点，产品会打开单独的 Tab 页面展示该节点详细信息。

5. 鼠标悬到两个库之间的连线上，系统会自动显示锚点。

6. 双击锚点会打开单独的 Tab 页面，显示这两个库之间表级间的关系以及他们之间的 ETL 作业。

7. 选择‘表间关系’的源 schema 分类，目标 schema 分类，源关键字，目标关键字，并点击‘查询’按钮，可以对记录进行过滤。

8. 点击‘表间关系’记录上的链接，可以单独打开该元数据的详细信息 Tab

9. 点击‘表间关系’记录上的‘下钻’图片上的连接，可以下钻到该源表和目标表之间的字段级关系。点击‘字段间关系’记录上的链接，可以单独打开该元数据的详细信息 Tab。

### 2.5 DB数据字典DDL采集

#### 2.5.1 采集说明

适用于采集DDL数据字典，当业务系统无法提供数据库，而只能提供DDL语句的时候。对DDL语句的语法有很高的要求。（普元目前只支持oracle数据库导出的sql文件）

#### 2.5.2 文件上传手动采集模式

**创建元数据**

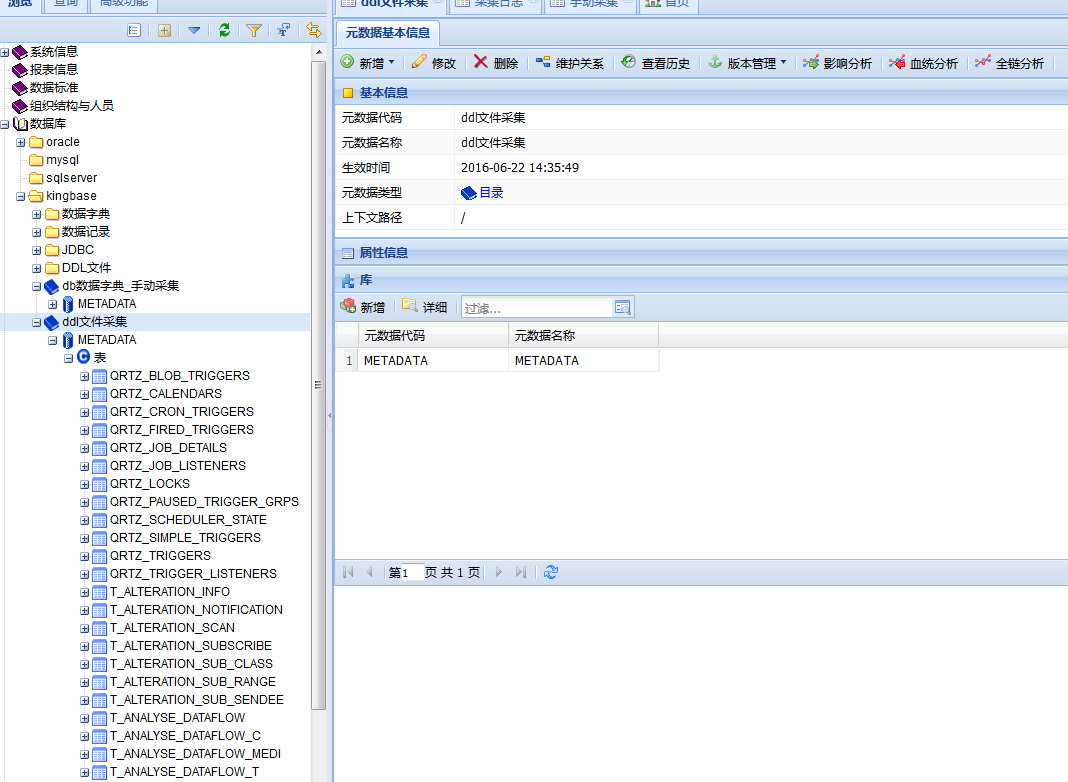
创建元数据，选择数据库元模型（元数据类型树-->核心模型包-->关系型包-->目录）

**创建数据源**

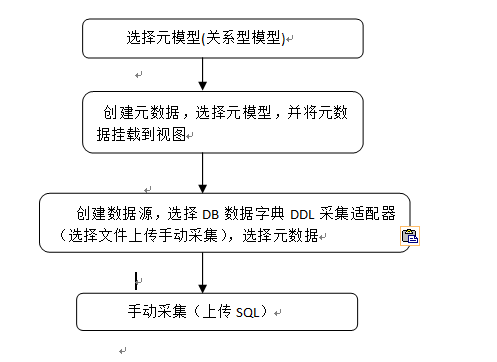
创建数据源，选择DB数据字典DDL采集适配器(选择文件上传手动采集，普元手动采集DB数据字典只支持oracle的sql文件)。选择挂载点(上一步创建的元数据)。

**手动采集**

选择上一步创建的数据源，上传sql文件（支持sql，txt，zip三种文件格式）。然后点击保存，即可开始采集。采集过程中，可以通过采集日志查看采集过程。采集结果图如下：



**采集流程**

****

#### 2.5.3 服务器文件定时采集模式

略

### 2.6 Bi报表采集

#### 2.6.1 采集说明

适用于采集Bi.office工具生成的报表。Bi.office是普元公司自己的一款报表工具。Bi.office报表采集只适合通过bi服务器接口连接来采集数据。无法通过bi报表来采集元数据。

#### 2.6.2 WebService采集模式

**创建元数据**

创建元数据，选择资源库（元数据类型树-->第三方工具原模型包-->Bi.Office6包-->资源库）

**创建数据源**

创建数据源，选择Bi.office采集适配器（配置Bi.office采集适配器的参数ip、port、username、password）。选择挂载点(上一步创建的元数据)。

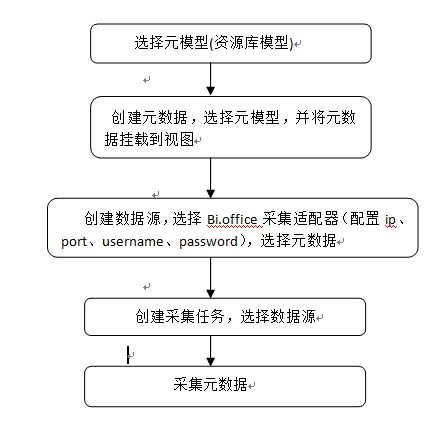
**创建采集任务**

创建采集任务，选择上一步创建的数据源。

**元数据采集**

任务保存完毕，就可以开始采集（立即采集或者定时采集）。采集过程中，可以通过采集日志查看采集结果。

**采集流程图**

****

### 2.7 Template采集适配器

#### 2.7.1 采集说明

适用于采集数据字典中文名补充，和使用excel整理出来的数据映射关系。对于一般类型的模板，元模型中每个“类”的元数据对应一个 Excel 中的 Sheet 工作表，元数据之间的组合关系通过被组合的元数据路径体现，元数据之间的依赖关系用专门的 Sheet 存放，这种类型的模板适用于系统所有元数据的导入导出，系统中的所有元模型（类）都可在本模板中体现。

#### 2.7.2 模板定制

**创建元数据**

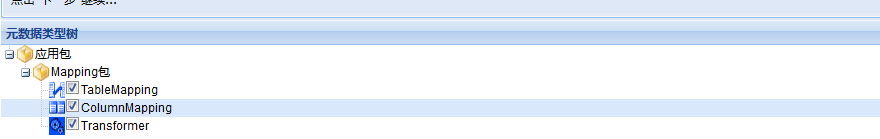
创建元数据，选择转换模型（元数据类型树-->应用包-->Mapping包-->转换）

**创建数据源**

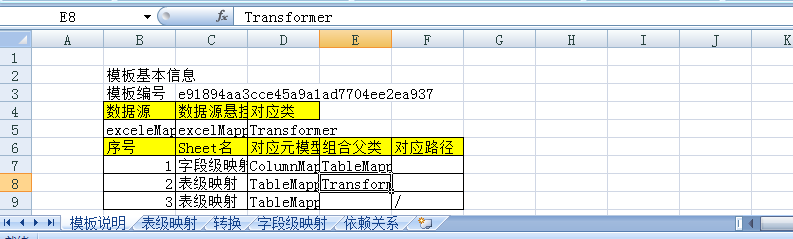
创建数据源，选择适配器，即可显示该适配器支持的元数据模型。选择挂载点(上一步创建的元数据)。

**创建模板**

选择创建的数据源，选择表级映射、字段级映射、转换。



配置 表级映射关系、字段级映射关系、转换关系即可创建模板。



#### 2.7.3 模板映射

**创建元数据**

创建元数据，选择转换模型（元数据类型树-->应用包-->Mapping包-->转换）

**创建数据源**

创建数据源，选择适配器，即可显示该适配器支持的元数据模型。选择挂载点(上一步创建的元数据)。

**上传模板**

选择数据源后，上传模板（Excel文件）

**模板映射配置**

#### 2.7.3 模板采集

参照 [数据地图采集](#_2.4_数据地图采集) [模板采集创建数据地图](#_2.4.3_模板采集创建数据地图)

### 2.8 PowerDesigner采集适配器

#### 2.8.1 采集说明

适用于采集PowerDesigner工具生成的数据模型文件，PowerDesigner是一款数据模型设计工具。

#### 2.8.2 文件上传采集模式

**创建元数据**

创建元数据，选择PowerDesigner模型（元数据类型树-->第三方工具原模型包-->PowerDesigner包-->PowerDesigner原模型）

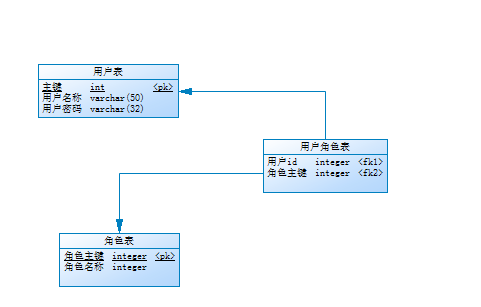
**创建数据源**

创建数据源，选择PowerDesigner采集适配器。选择挂载点(上一步创建的元数据)。

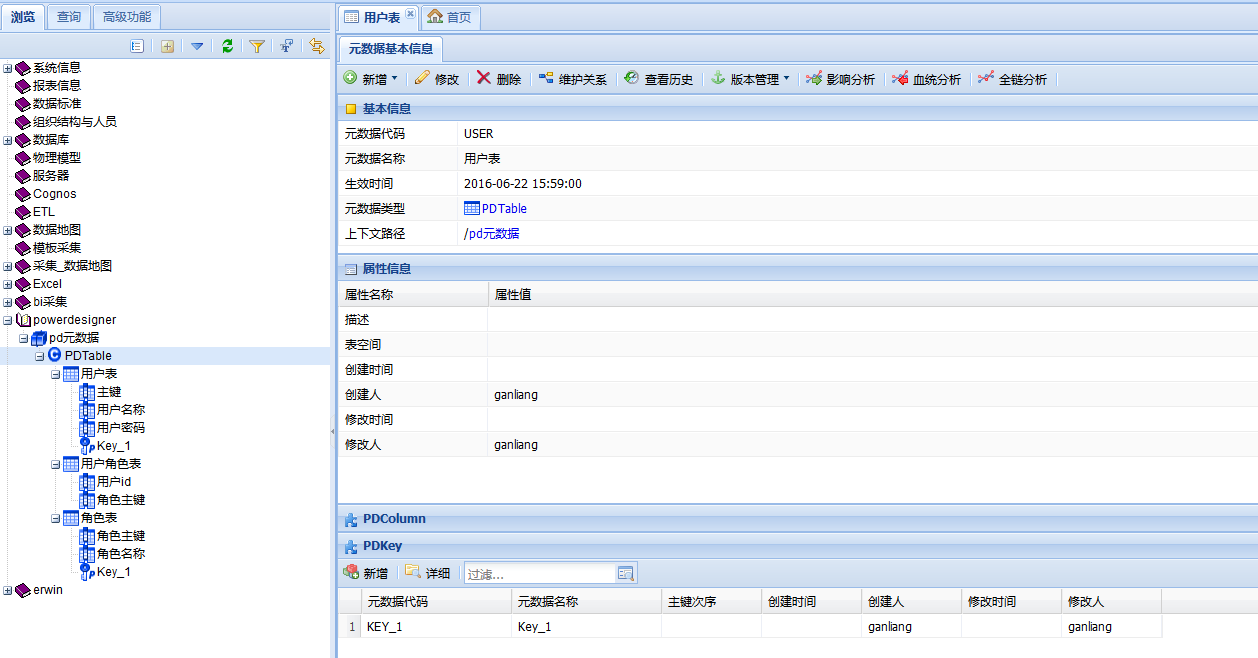
**手动采集**

选择上一步创建的数据源，上传文件（支持pdm，xml，zip三种文件格式）。然后点击保存，即可开始采集。采集过程中，可以通过采集日志查看采集过程。

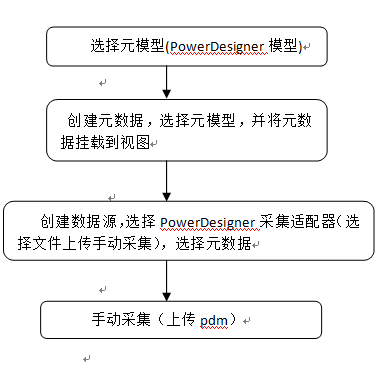
pdm如图



采集结果图如下：



**采集流程**



#### 2.8.3 元模型关系展现

通过PowerDesigner设计出模型关系。通过采集pmd生成元数据，是否可以通过元数据展示模型之间的关系。（如表和表之间的关系，是否存在外键等关系）

### 2.9 Erwin采集

#### 2.9.1 采集说明

适用于采集Erwin工具生成的数据模型文件，Erwin是一款数据模型设计工具。

#### 2.9.2 模型采集

**创建元数据**

创建元数据，选择ER模型库（元数据类型树-->第三方工具原模型包-->ERWin包-->ER模型库）

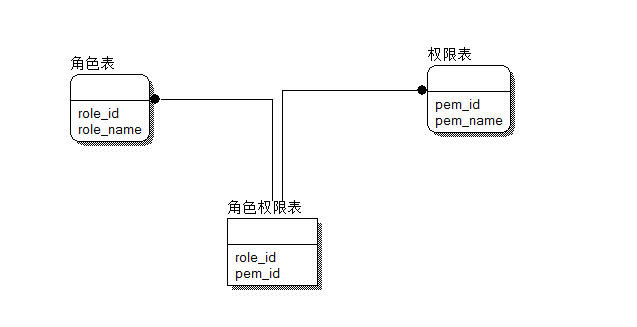
**创建数据源**

创建数据源，选择Erwin采集适配器(配置参数extractContent选择采集的数据内容物理模型和逻辑模型或者逻辑模型)。选择挂载点(上一步创建的元数据)。

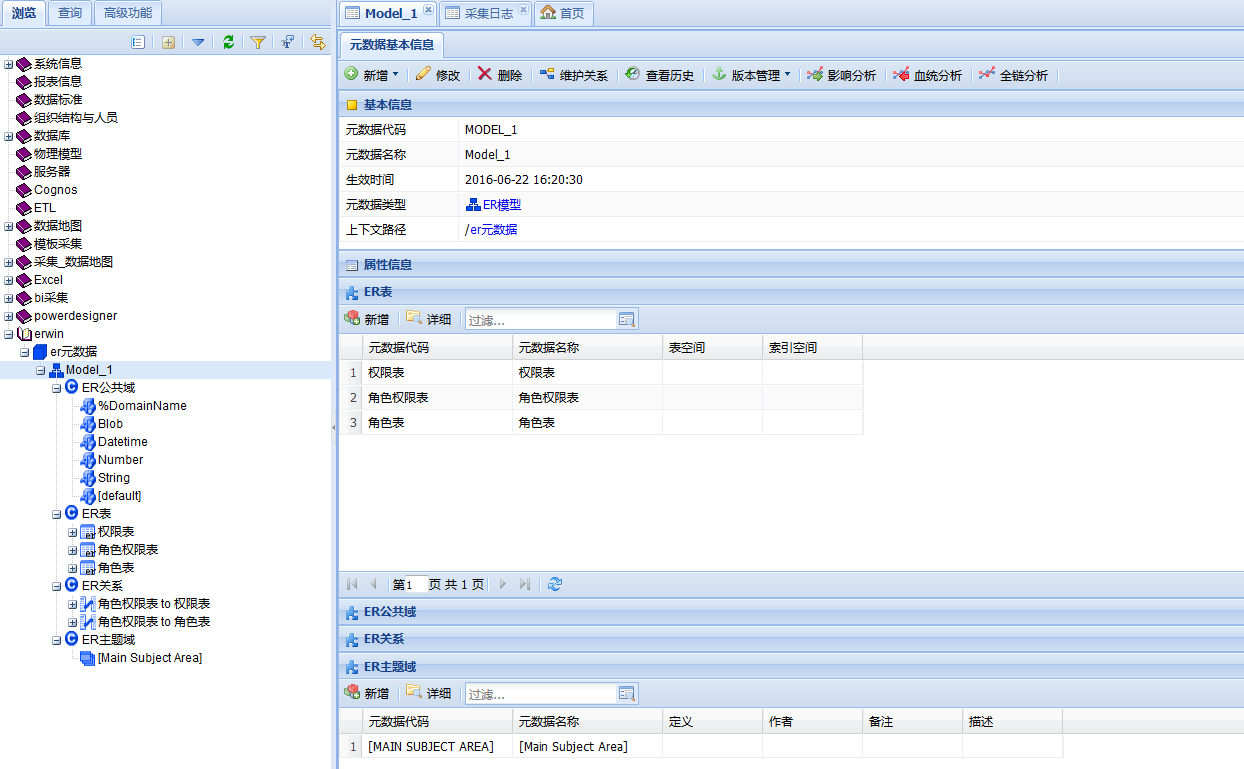
**手动采集**

选择上一步创建的数据源，上传文件（支持ERWin导出来的xml，zip两种文件格式）。然后点击保存，即可开始采集。采集过程中，可以通过采集日志查看采集过程。

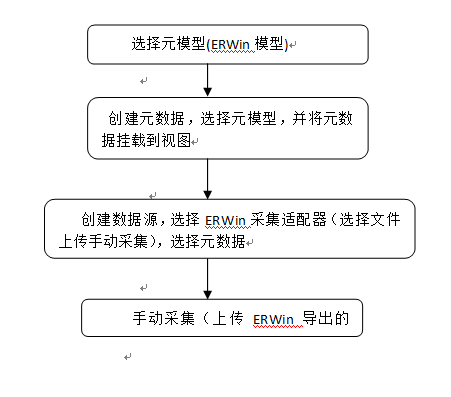
ERWin关系图如下：



采集结果图如下：



**采集流程**



### 2.10 Kettle采集适配器

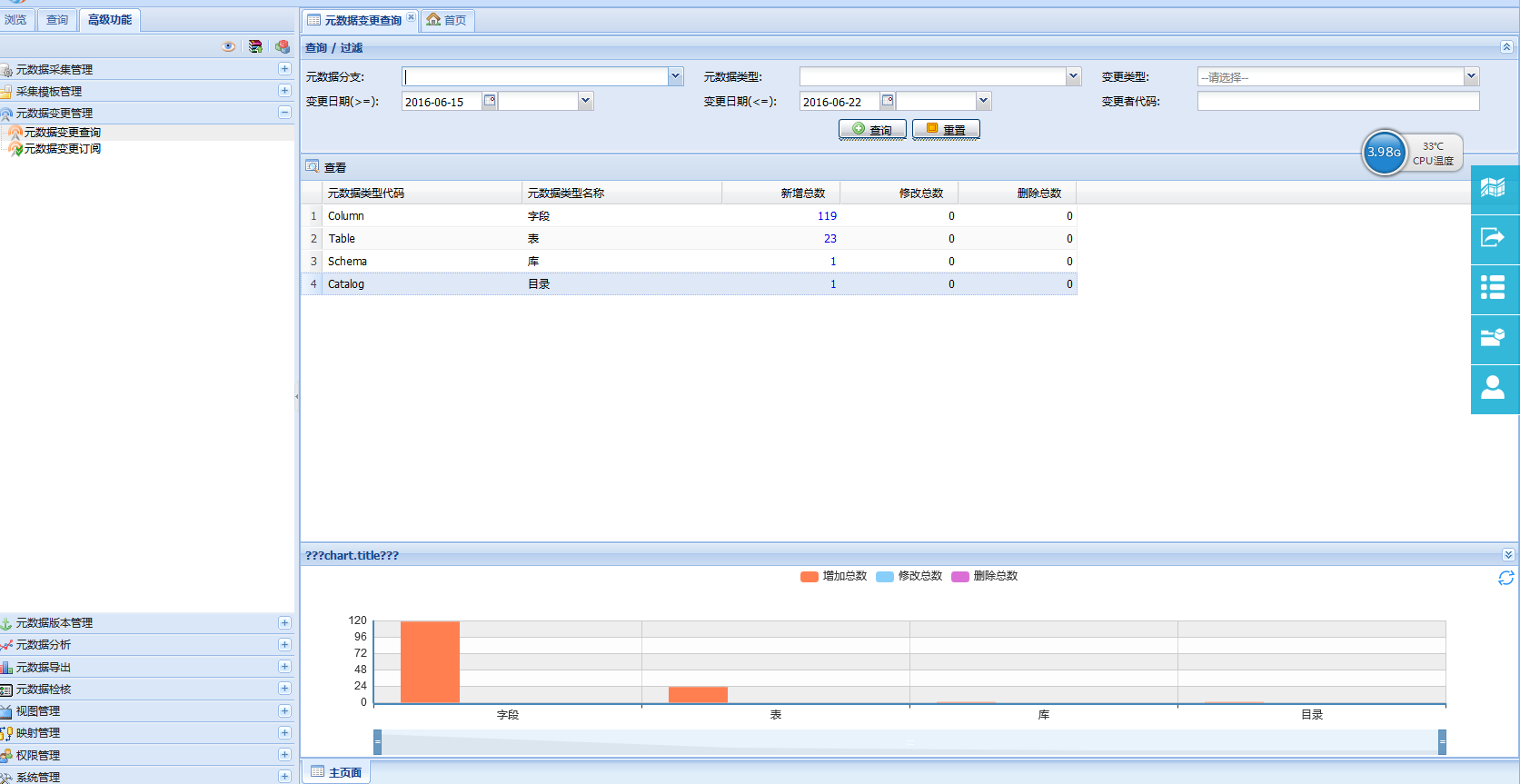
## 3 元数据管理

### 3.1 元数据变更管理

元数据变更管理也就是，当元数据发生变化的时候，通知给对元数据感兴趣的用户（通过订阅的方式实现）。每一次元数据发布的时候，普元后台都会自动检测元数据是否发生变更，如果发生变更，就查询订阅了该元数据的用户，然后对用户发送Email通知元数据的变更。

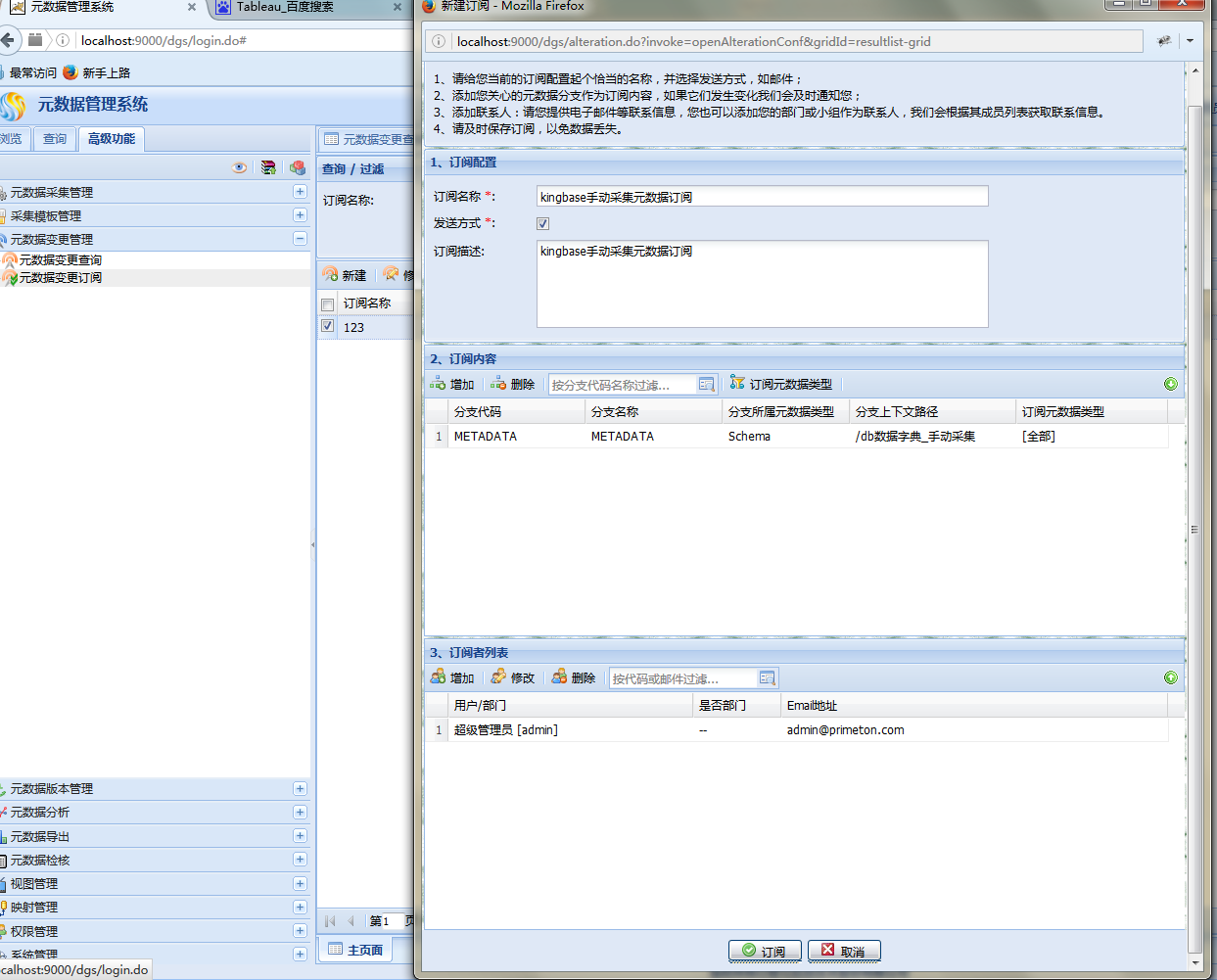
**元数据变更查询**

元数据变更管理-->元数据变更查询

****

**元数据变更订阅**

点击添加订阅。在订阅列表选择元数据，选择订阅者。即可完成订阅。



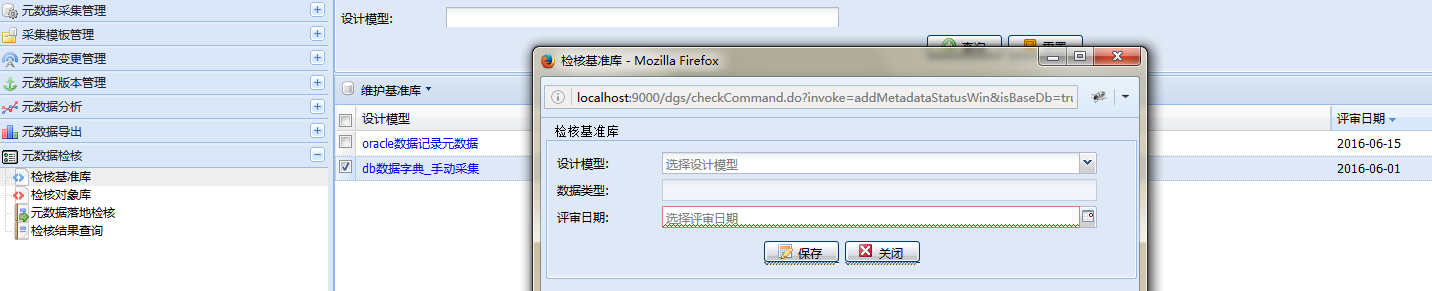
### 3.2 元数据检核

元数据检核的主要功能就是检验当前的元数据库是否和标准的元数据库之间的差别。以防元数据库发生太大的变化。有利于元数据库的维护。

元数据检核的元数据必须先要定版，才能够检核。因为检核的一般都是同一元数据的不同版本。从而检核出不同版本之间的差异。

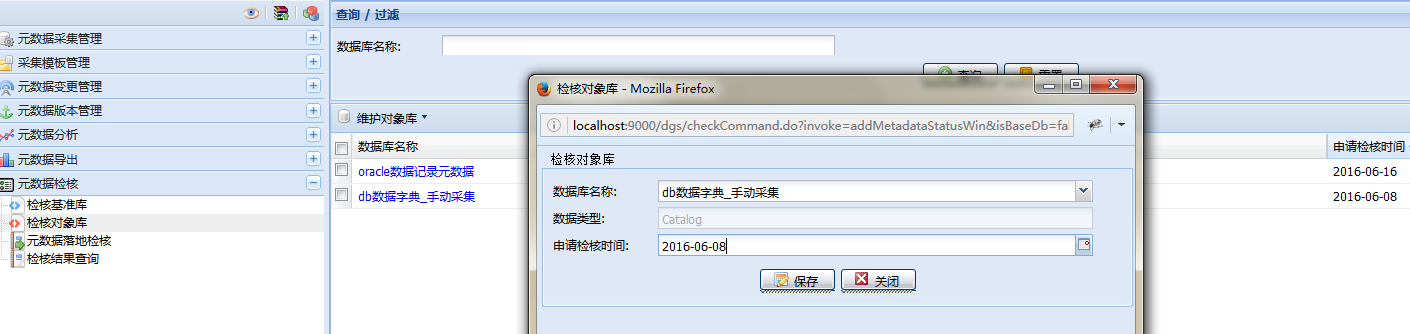
**校核基准库**

元数据检核-->检核基准库。添加元数据（基准）、评审日期即可保存基准库。



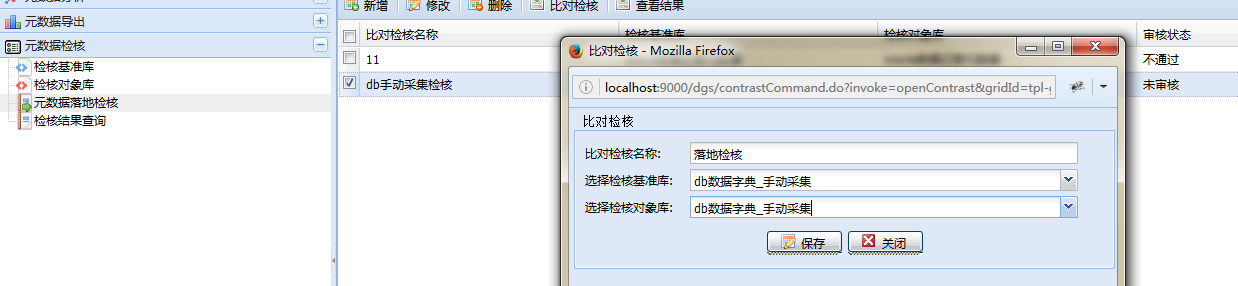
**检核对象库**

元数据检核-->检核对象库 添加元数据（对象库）、评审日期即可保存基准库。



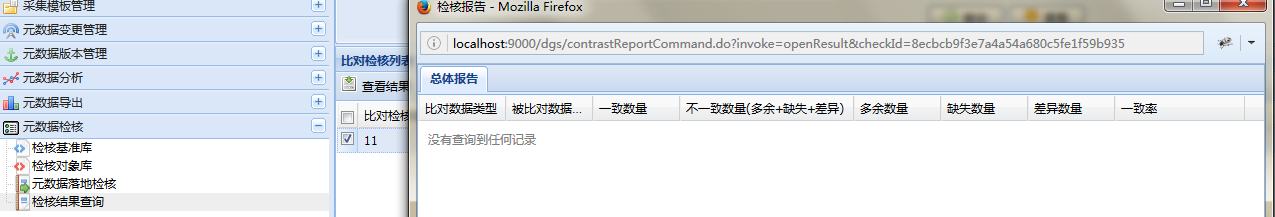
**元数据落地检核**

元数据检核-->元数据落地检核 选择检核基准库、选择检核对象库。然后保存即可比对检核。比对检核后，可以从查看结果中查询检核结果。



**检核结果查询**

元数据检核-->检核结果查询 选择检核结果，查看结果。可以展示基准数据库和对象库之间的差异。

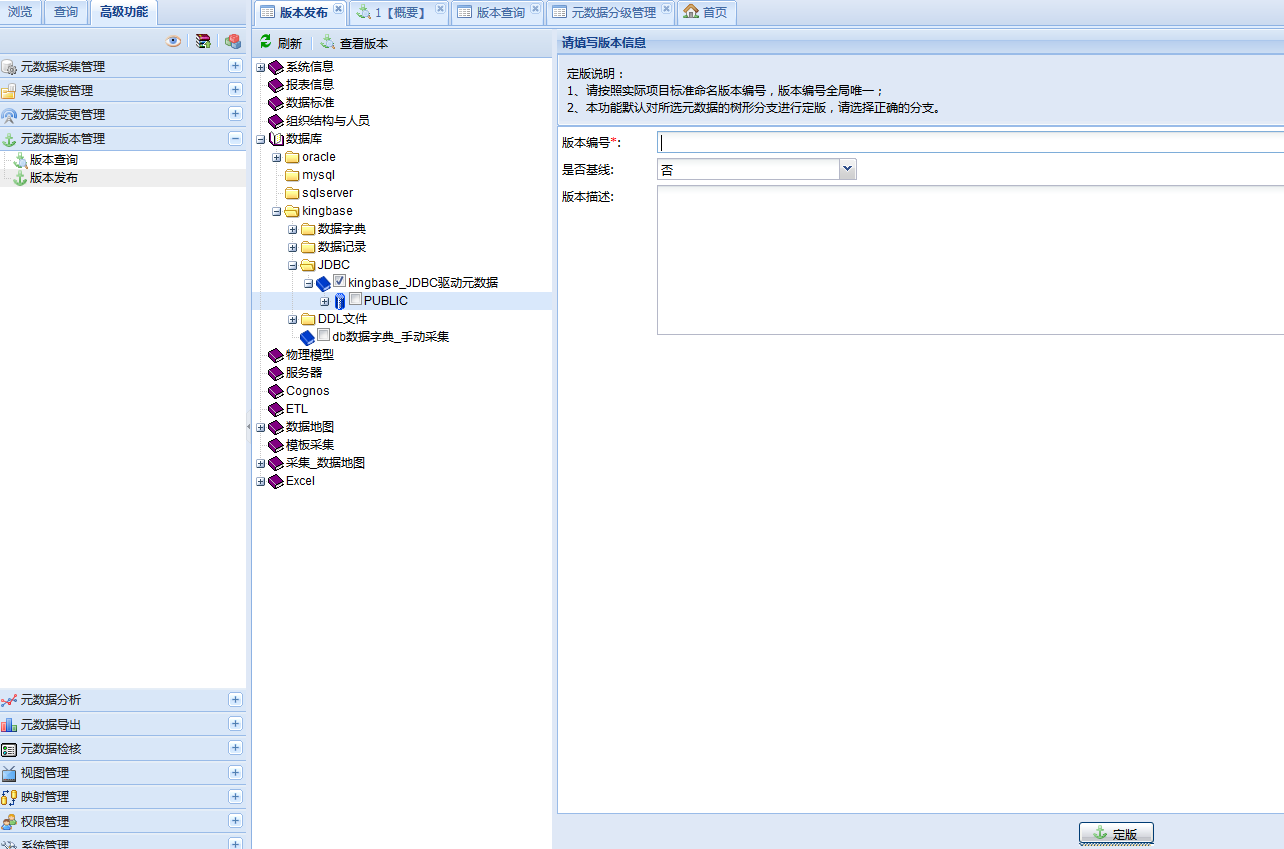
****

### 3.3版本管理

元数据针对不同时期，会有变化，为了管理这个变化，需要建立元数据的版本管理，用于记录版本和变化轨迹。

元数据版本管理-->版本查询 可以查询到当前已经定版的元数据版本。双击版本，可以展示元数据的详细信息。

元数据版本管理-->版本发布 选择左侧元数据，然后点击定版就可以完成版本的定版。

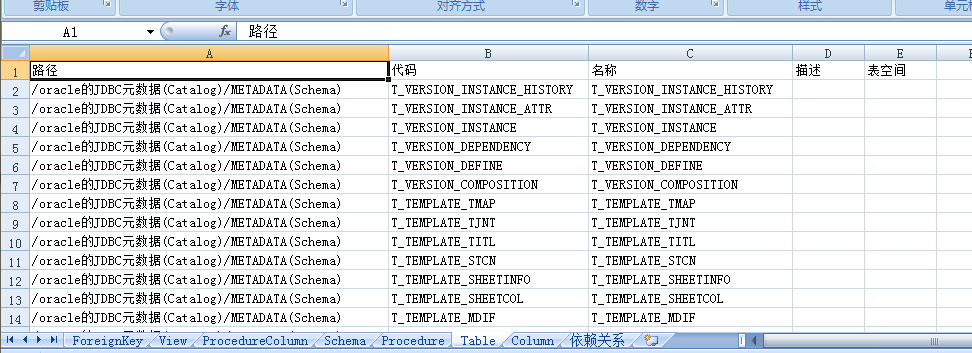


### 3.4元数据导出

**元数据导出（Excel）**

元数据导出-->元数据导出 左侧选择元数据，点击导出Excel。下载Excel即可显示Excel详细信息。

元数据导出会将 关于元数据的元数据模型导出到Excel。每个原模型都包含很多元数据对象类型。例如，关系型元数据模型包含外键、视图、存储过程列、方案、存储过程、表、依赖关系等元数据对象类型。导出Excel，如下图:



**视图导出（Zip）**

元数据导出-->元数据分级管理 左侧选择视图名称，可以将整个视图导出到Zip包。同时也可以将元数据导入到视图下。

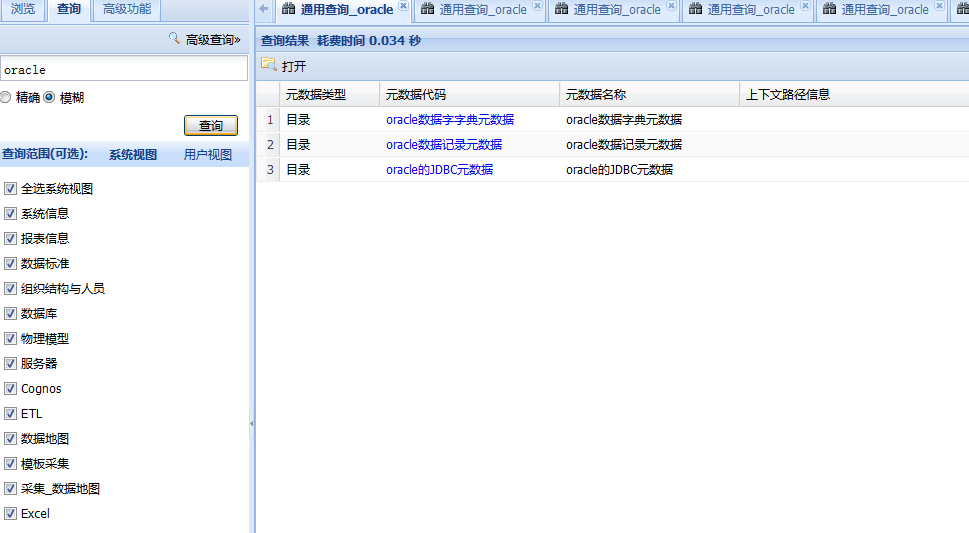
普元元数据管理系统中，是把元数据挂载在视图下的。因此一个视图下面可以包含多个元数据。当导出视图的时候，后台程序元数据打成zip包来下载。

## 4 元数据预览

### 4.1 元数据检索

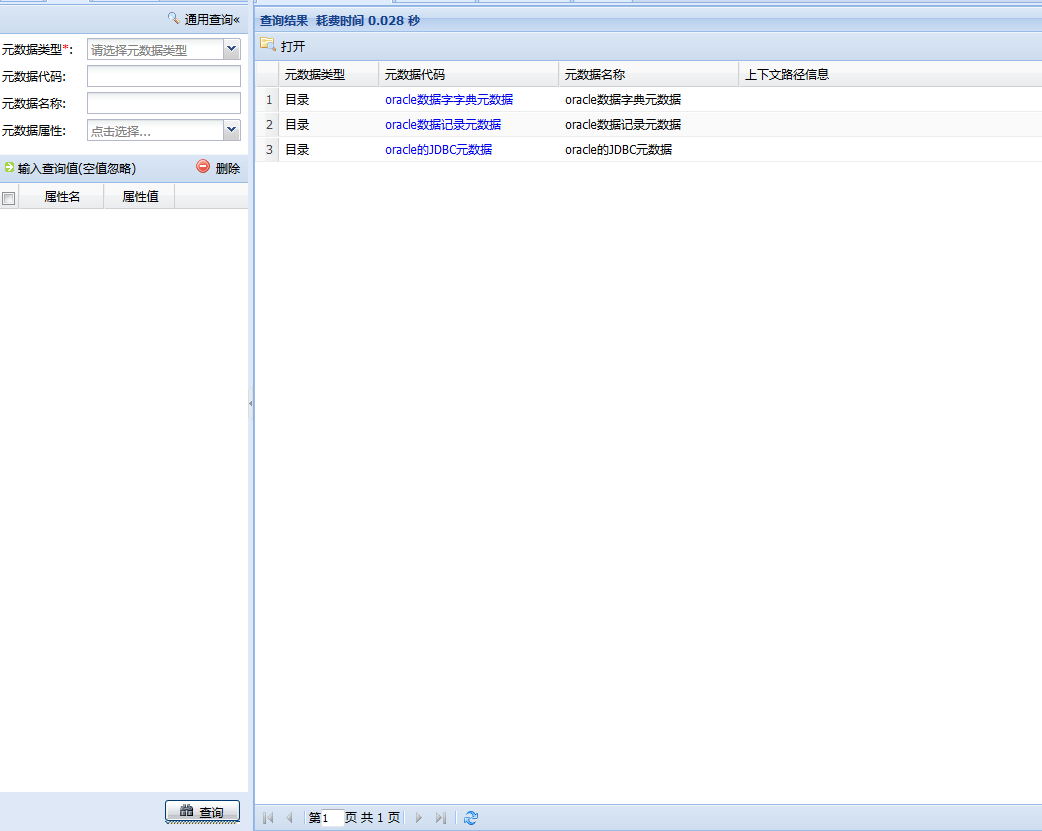
**通用查询**

在通用查询Tab中，选择精确查询或者模糊查询，选择系统视图或者用户视图。输入关键字，点击查询，右侧Tab展示查询出来的元数据。



**高级查询**

先选择元数据类型，输入元数据代码、元数据名称，选择元数据属性等来检索元数据。

****

### 4.2元数据分析

**影响分析**

影响分析采用图形方式展示了以某个元数据为起始节点，其后与其有关系的所有元数据，反应数据的流向与加工过程，使用影响分析可分析数据流向和数据转换中错误的定位。

**血统分析**

血统分析采用图形方式展示了以某个元数据为终止节点，其前与其有关系的所有元数据，反应数据的来源与加工过程，使用血统分析可分析数据来源和数据质量问题的定位。

**全链分析**

全链分析采用图形方式展示了以某个元数据为目标节点，其前后与其有关系的所有元数据，

反应数据的来源与加工过程，使用全链分析可分析数据来源和数据质量问题的定位。

**表关联度分析**

表关联程度分析描述了数据库表元数据与其他元数据的关系出现次数情况，如：表与 ETL 程序、表与 OLAP、表与指标等关系次数情况，产品通过分析数据库表关系次数的多少，来体现表关联程度的高低。

**属性值差异分析**

产品提供同类元数据进行元数据属性值差异分析功能，您可通过属性值差异分析功能直观获知两个同类元数据之间的相同和不同之处。