支持科学数据专业类型的统一检索框架①

史晓磊　沈志宏　黎建辉　(中国科学院 计算机网络信息中心 北京 100190)

摘　要： 中科院科学数据库在“十一五”信息化建设中将建成一个由 51 家参建单位组成的庞大数据应用环境。为方便科研人员从这些海量科学数据中得到感兴趣的内容，有必要设计一个统一检索工具。然而传统的统一检索技术不支持科学数据专业数据类型，这带来了数据表达与展示两方面的问题。对此，提出一种支持科学数据专业类型的统一检索框架，该框架使用数据建模中间件实现专业类型数据的统一格式表达与发布，运用模板技术为数据提供灵活的展示方式，并以一种可扩展的插件方式管理这些科学数据专业类型。文章最后还介绍了基于此框架的统一检索系统 Voovle 的应用现状。

关键词： 统一检索；科学数据；专业类型；插件；概念模型；中间件

A Cross-Search Framework Supporting the Professional Type of Scientific Data

SHI Xiao-Lei, SHEN Zhi-Hong, LI Jian-Hui

(Computer Network & Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The CAS 11th Five-year Informationization Construction Program will build a huge data application environment, which is composed of 51 institutions. It is necessary to build a cross-search tool to help people get useful information from those massive data. The traditional cross-search technology doesn't support the professional type of scientific data, which brings difficulty in data formatting and data presentation. This paper proposes a cross-search framework which supports these professional types. This framework uses data modeling middleware to publish scientific data and metadata on the basis of the universal data exchange format of the profession type, and uses template technology to display these data flexibly. The framework also manages professional type as a plug-in, which has good extensibility. This paper also introduces the application of the cross-search system Voovle, based on this framework.

Keywords: cross-search; scientific data; professional type; plug-in; conceptual model; middleware

中国科学院“十五”期间科学数据库建库单位有 45 家，专业子库数量 503 个，总数据量 16.6TB，其中可共享的数据量达到 9.48TB。“十一五”期间，科学数据库将继续建设一个由 51 家参建单位组成的包含 2 个参考型数据库、8 个主题数据库、4 个专题数据库、37 个专业数据库的更加庞大的数据应用环境。面对如此浩瀚的数据环境，如何设计一个科学数据统一检索工具，帮助用户快速得到感兴趣的数据，是科学数据共享进程中必须解决的一个问题。

1　统一检索技术研究

统一检索又称整合检索、联邦检索，它为用户提供统一的检索界面，一次同时对多个分布的、异构的数据库进行检索，将返回结果整合后统一展现给用户。

① 基金项目:中国科学院信息化专项项目 (INFO-115-C01)
收稿时间:2010-04-02;收到修改稿时间:2010-05-12

统一检索应用最多的是数字图书馆领域，在企业门户信息系统也有应用。其成功应用有清华同方异构数据库统一检索平台 USP、国家科学数字图书馆跨库集成检索系统,以及国外的 Metalib 系统、Web Feat Prism 系统等。目前主流的统一检索方式有以下四种[1]：元数据整合方式、中间件方式、网页搜索代理方式、依附方式。在这四种方式中，中间件方式需要设计复杂的查询分解算法，网页搜索代理方式的前提是分布数据源都发布 web 网站，依附方式的前提是分布数据源自发互联，这些条件在科学数据建库单位难以满足。相比较而言，元数据整合方式更加轻量、适应性更强，因此我们选择使用元数据整合方式为科学数据提供统一检索。

OAI-PMH(OAI Protocol for Metadata Harvesting)协议[2,3] 是元数据整合方式的标准协议，它规范了服务提供方与数据提供方的职责，使用 6 个动词规范元数据收割的通信过程。基于该协议的开源软件有 OCLC 开发的 OAIHavester2 与 Old Dominion University 开发的 Arc[4]。

2 专业类型科学数据统一检索的问题

OAI-PMH 协议是从数字图书馆领域提出的，已在该领域广泛使用，在企业应用数据领域也有成功案例[5]，然而它却并不完全适用于科学数据领域。这主要是由科学数据的专业类型引起的。具体地，数字图书馆数据和企业数据大致可以归结为数值、字符、时间等少数几类基本类型，而科学数据却有学科领域的专业数据类型。传统的统一检索协议无法支持科学数据专业类型，它带来的困难主要表现在：

2.1 科学数据与元数据统一格式表达问题

科学数据在专业学科领域往往有一套统一的专业表达格式，它在学科内起到数据交换标准的作用。例如数学数据的 MathML、地理信息数据的 GML、化学分子式数据的 CML 等。专业格式以 XML 形式表达，这些 XML 的组织方式蕴含了复杂类型各子元素之间的逻辑关系信息。然而多数 RDBMS 不支持这些科学数据专业类型，复杂的专业类型数据往往被退化成关系型数据存储。易见，在科学数据统一检索的元数据收割与数据收割过程中，应该按照专业类型的统一格式表达与发布，而不是简单按照存储结构发布关系型数据，这样才能使之保持各子项的内部逻辑，符合学科内数据交换标准。这是我们要解决的第一个问题。

2.2 科学数据专业类型的展示问题

统一检索的目的在于共享数据，而共享价值得以体现的前提在于数据能被理解。科学数据专业格式表达解决了“机器理解”的问题，然而要把统一检索的结果以 web 页面的形式展现给用户,解决“人类理解”的问题。专业类型的科学数据往往有自己独特的展现方式。例如，化学结构式数据通常会以一个可视化的 ActiveX 控件(或 Java Applet、IFRAME 等)显示，而 GPS 数据往往会通过调用 Google Map 的接口，以图片的形式嵌入 web 页面。因此,如何提供一种灵活的、适应性强的方式展示检索到的专业类型科学数据，是科学数据统一检索的另一个问题。

综上所述，科学数据复杂专业类型为科学数据统一检索带来统一表达与灵活展示两方面的问题，这使得在 OAI-PMH 等协议不能完全适应科学数据的检索。对此，我们需要在“元数据收割-查询定位-数据收割”思想的基础上，在数据表达与数据展示方面做若干改进,以适应专业类型科学数据统一检索的需求。

3 支持专业类型的统一检索关键技术

3.1 支持专业类型统一表达的科学数据建模与发布中间件

科学数据/元数据专业类型统一格式表达的问题，实质上是专业类型 XML 概念结构向异构关系型存储结构映射的问题。为此我们提出一种支持科学数据专业类型统一格式的概念模型。该概念模型中数据集包含若干个类(Class)。类是视图的概念，即对一张或多张数据表(Table)施以连接(表关联)、选择(发布部分数据)及投影(自定义检索、详览、细览字段)运算的结果。类有若干属性(Property)，每个属性绑定一种专业类型。通过这种绑定，我们约束了属性必须符合专业类型统一格式。这样我们即可通过由概念模型的属性向数据库中字段(Field)的映射，提供对专业类型的统一格式表达的支持。

为方便映射，我们提供一套支持专业类型统一表达格式的科学数据建模与发布中间件 SD-PUB。该中间件分为数据透明访问中间件 JDBC-X、模型映射中间件 VDBCatalog 和数据发布中间件 SDSP(Scientific Data Service Publisher)。

JDBC-X 子中间件定义了数据库操作的抽象接口，并针对不同 RDBMS 分别提供实现。这样，上层应用可以以统一接口对底层数据库透明访问，屏蔽了异构

数据库的差异。

VDBCatalog 是实现专业类型统一表达的核心部件，它为属性与字段的映射提供可视化界面。具体地，VDBCatalog 将识别绑定在属性上的科学数据专业类型，解析该类型统一表达格式 XML，将该表达格式分解成若干子项，并为每个子项提供下拉菜单，以便用户选择绑定的逻辑字段。用户通过选择绑定，完成专业类型统一格式向异构数据源逻辑结构的映射。所有的映射规则将存储到一个建模配置文件中。

由 VDBCatalog 包装的数据在模型上符合专业类型统一表达，在此之上我们提供数据发布子中间件 SDSP，按专业类型统一表达格式发布 XML 形式的科学数据与元数据，供科学数据统一检索工具收割。

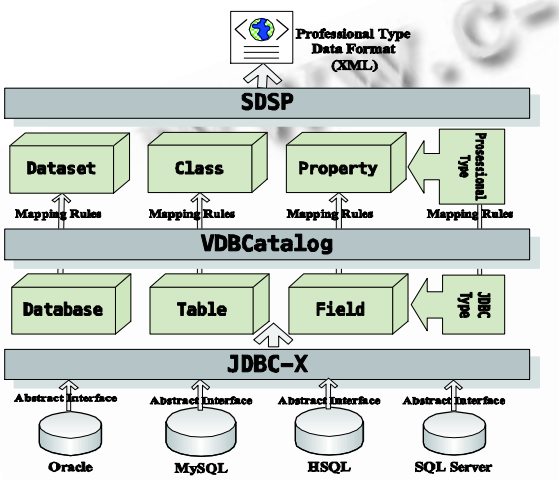

图 1　支持专业类型统一表达的科学数据建模发布中间件

3.2 基于 Velocity 模板技术的科学数据专业类型展示方式

一方面，基于统一表达格式的科学数据具有确定的组织结构，我们很容易将收割到的 XML 格式的科学数据转换成具有确定结构的 JavaBean；另一方面，每种专业类型的展示方法也是确定的，同种类型的科学数据有同一种显示方法。基于科学数据专业类型以上两方面特点，我们可以使用模板技术，为科学数据专业类型提供的灵活展示方式。Velocity[6]是一个基于 Java 的模板引擎，我们为每一种科学类型提供一个 Velocity 展示模板，在模板中定义该类型的展现方式。这样，Velocity 模板引擎将按照定义好的格式，把放入上下文(Context)中的科学数据渲染成 HTML 页面。通过这种模板技术进行的数据展示具有很强的适应性与灵活性。

3.3 科学数据专业类型的插件式管理

科学数据专业类型种类繁多，我们无法对所有专业类型一一提供支持；但我们可以提供一种插件式的类型管理方法，让用户以对自身专业领域的数据类型进行扩充。

插件式类型管理通过以下两种技术实现：(1) 业务逻辑层多态技术：我们从专业类型数据的业务逻辑中抽象出类型绑定接口 AbstractType 和类型操作接口 AbstractTypeDriver，每种专业类型只需实现以上两个接口，即可以多态的形式“插入”业务逻辑层。(2) 显示层模板技术：我们为每种专业类型的 VDBCatalog 映射配置页面提供模板 option.html，为数据检索统一展示提供模板 view.html。当需要对专业类型进行相关展示时模板引擎将自动加载对应的模板。这样，我们可以通过模板方式将专业类型“插入”显示层。

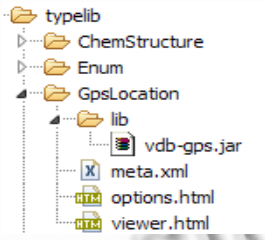

图 2　科学数据专业类型的插件式管理

4　支持科学数据专业类型的统一检索系统

4.1 科学数据统一检索系统整体架构

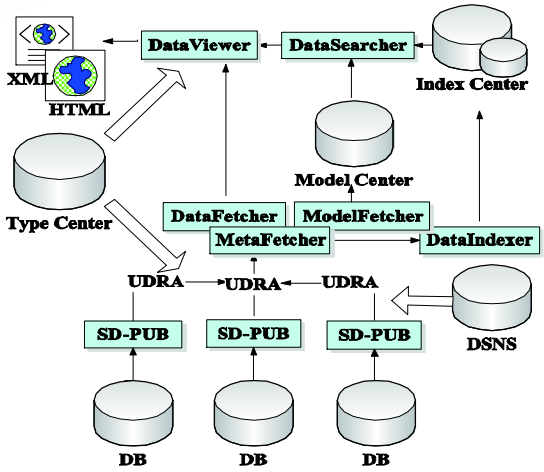

图 3　支持专业类型的科学数据统一检索系统整体架构

基于以上三点关键技术，我们可以设计一套支持科学数据专业类型的统一检索系统 Voovle。该系统整体架构如图 3。

该架构主要以下六个模块：

(1) 科学数据建模与发布中间件 SD-PUB

SD-PUB(Scientific Data Publisher)部署在各科学数据站点，它按照专业类型统一表达格式对异构关系数据进行建模，并按专业 XML 格式发布数据与元数据。详见 3.1 节。

(2) 专业类型标准插件中心 Type Center

新开发的类型插件需要通过审核才能成为该类型的标准插件。Type Center 是一个专业类型标准插件库，建库单位从该中心下载所需类型的标准插件加载到本地 VDBCatalog，以实现科学数据/元数据的专业类型统一表达格式发布。此外标准插件也为相应类型数据的检索结果展示提供支持。

(3) 数据集名字服务器 DSNS

DSNS(Data Set Name Server)提供数据集 URI 向域名的解析。所有参与统一检索的异构数据集都要在此注册。

(4)元数据抓取模块 MetaFetcher、数据抓取模块 DataFetcher 与模型抓取模块 ModelFetcher

MetaFetcher 按一定调度策略抓取异构数据源 SD-PUB 发布的元数据；DataFetcher 在检索到某条记录后根据定位标识抓取专业类型表达格式的数据；ModelFetcher 抓取数据集概念模型并存储到模型中心 Model Center。

(5) 索引模块 DataIndexer 与搜索模块 DataSearcher

DataIndexer 为 MetaFetcher 抽取的元数据建立倒排索引以备搜索；DataSearcher 从索引中检索符合条件的元数据，并按照一定的 Data Rank 排序算法返回检索结果。

(6) 数据展示模块 DataViewer

DataViewer 分析检索结果的数据的专业类型，并根据该类型插件定义的展示方式为 XML 专业数据生成 web 页面。

我们已基于以上架构完成科学数据统一检索系统 Voovle 的开发，并在青海湖基础数据平台、纳米科技专业数据库及生态专题数据库等科学数据参建站点完成部署。针对这几个专题库的统一检索测试效果良好。

图 4 科学数据统一检索系统检索效果演示

4.2 科学数据专业类型插件示例——GPS 类型插件

GPS 类型是一种常见的科学数据专业类型，它在青海湖基础数据平台等应用中广泛使用[7]。GPX(GPS eXchange Format)是 GPS 数据专业类型的统一表达格式[8]，用以学科间 GPS 数据交换。GPX 规定 GPS 数据至少经度、纬度两个维度，此外还包括海拔等可选维度。本节将以该类型为例，展示科学数据专业类型插件对统一检索的支持。

我们在 VDBCatalog 中加载 GPS 类型插件，得到如下可视化配置界面：

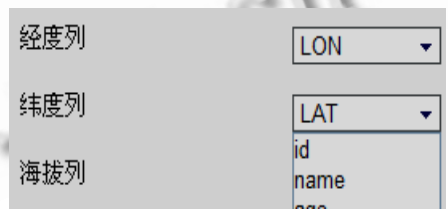

图 5 GPS 专业类型数据建模的可视化配置界面

通过配置我们实现了 GPS 类型统一表达格式向异构数据源的映射。在青海湖基础数据平台中映射规则如下：

```
<property>
    <uri>cn.csdb.qhh.pda.location</uri>
    <latColumn>LAT</latColumn>
    <altColumn>ALT</altColumn>
    <lonColumn>LON</lonColumn>
</property>
```

在展示方面我们定义 GPS 类型展示模板如下：

<IFRAME
src=http://visualdata.csdb.cn/google-map/view.jsp?
param=[{"lon":$bean.get('$prop.name').lon,"lat":$bean.get('$prop.name').lat,"grade":0}] />

其中 http://visualdata.csdb.cn 是一个第三方数据展示平台,经它中转可以调用 Google Map 接口,为 GPS 数据产生图片嵌入 IFRAME。基于该类型插件的数据展示如下:

图 6　GPS 专业类型科学数据展示效果

由此可见,GPS 类型插件为 GPS 数据的统一格式表达与显示提供了良好的支持。

5　结束语

本文在借鉴数字图书馆统一检索技术的基础上,提出一种支持科学数据专业类型的统一检索框架。该框架有一下特点:(1)支持科学数据专业类型统一表达格式;(2)提供灵活的专业类型数据展现方式;(3)对科学数据专业类型提供可扩展的插件式支持。目前,我们已在该框架基础上为相关学科提供了 GPS 类型插件、化学结构式类型插件等多个学科领域的类型插件,并实现了针对生态专题数据库、纳米科技专业数据库等多个分布异构科学数据专题数据库的统一检索。

参考文献

1 任瑞娟,米佳.基于网格技术的跨库检索研究.河北科技图苑, 2008,(7):15－17.

2 The Open Archives Initiative Protocol for Metadata Harvesting.[2007-09-02].http://www.openarchives.org.

3 曾婷,张成昱.基于 OAI-PMH 和复杂对象格式的资源收割机制探讨. 现代图书情报技术, 2005(11):14－18.

4 Liu XM, Maly K, Zubair M, Nelson ML. Arc — An OAI Service Provider for Cross-Archive Searching. The 1st ACM/IEEE-CS joint conference on Digital libraries, Jan. 2001.

5 陈硕,陈真勇,熊璋.企业门户信息系统统一检索平台的研究.微计算机信息, 2008,24(3):4－6.

6 GPX. [2010-5-5].http://velocity.apache.org.

7 黄志一,周园春,常青玲,沈志宏,侯元生,阎保平.可定制移动数据采集系统的研究和实现.计算机系统应用, 2009,18(11):11－15.

8 The Apache Velocity Project. [2010-5-5]. http://en.wikipedia.org/wiki/GPX.