极地科学数据共享平台的设计与实现

程文芳[1], 王 伟[2], 张 洁[1], 夏明一[3], 杨 蕊[1], 朱建钢[1], 张北辰[1], 汪大立[1], 凌晓良[1], 李升贵[1]

(1. 中国极地研究中心, 上海 200136; 2. 国家海洋局信息中心, 天津 300171; 3. 上海橙智信息科技有限公司, 上海 201204)

摘要: 为了共享中国 30 年来获取的南北两极科学考察数据, 基于 Python 语言, 设计了具有数据检索、数据发布、数据申请审批、数据应用等功能的分布式共享信息系统。目前系统已对外服务, 成为我国唯一向国内外开放共享的极地元数据平台, 促进了更深层次的国际项目合作和数据共享。

关键词: Python; 南北极; 元数据; 数据检索; 数据共享

中图分类号: TP311.1 **文献标识码**: A **文章编号**: 1000-3096(2015)05-0075-10

doi: 10.11759/hykx20130510001

20 世纪 80 年代初, 欧美发达国家开始研究科学数据共享。1989 年全球气候变化主目录(Global Change Master Directory, GCMD)成立, 逐渐发展成全球最大的公共元数据库之一。1997 年南极数据管理委员会(Standing Committee on Antarctic Data Management, SCADM)成立, 用于协调管理南极数据合作及共享。1996 年澳大利亚南极数据中心(Australian Antarctic Data Centre, AADC)成立, 负责南北极科学考察数据的管理和发布, 并于 2011 年重构元数据库发布系统。作为南极成员国家, 德国、智利、英国等在 90 年代相继建立了自己的南极科学考察数据库。在亚洲地区, 日本于 1998 年成立数据中心, 于2007年建设数据共享平台; 韩国极地数据中心于 2010 年建立, 并于 2011 年建设了数据共享平台。

中国南北极数据中心作为南极科学委员会南极数据管理联合委员会的成员, 按照国际南极数据管理框架, 致力于建立一个基于因特网的中国极地科学数据共享平台, 试图为国际和国内从事极地科学研究的人员提供一个相对集中和完整的基础数据资源库[1]。1999年中国极地科学数据库系统(CHINARE, 项目编号: G99-A-02a)在中华人民共和国科技部资助下建成, 2003年该系统加入了国家科技部建立的“中国地球系统科学数据共享网”, 在《南极条约》和《中国极地科学考察数据管理办法》原则框架下面向国内外科学界和社会公众提供专业研究、管理决策和科普教育所需的极地科学数据、信息、研究成果等共享服务。2005年根据中华人民共和国科技部提出的建立“国家地球系统科学共享中心”的要求, 进行全面改版, 增加数据在线汇交系统、实现分布式的数据管理、平台升级与维护等功能。经过6年对外服务运行, 针对数据检索、数据预览的局限, 2012年对系统进行了重新构建, 对响应速度、数据集预览、数据检索效率, 数据应用和数据集成等功能进行了优化设计与实现。

1 CHINARE 系统设计

CHINARE 系统建设以实现极地南北极科学考察数据的检索、发布、申请和审批为目标。系统建设以 Python 为开发语言, 以开源插件为基础, 结合 Ajax(异步 JavaScript 和 XML)、JSON、Lucence(开源全文检索引擎的架构)等技术实现了极地科学数据的发布、浏览、检索、申请、统计分析等功能。Python 有高效率的高层数据结构, 简单而有效地实现面向对象编程, 适用于快速的应用程序开发[2]。由于极地科学数据共享平台共享了 30 GB 的数据集, 从数据增长速度以及系统访问速度因素考虑将 Web 应用服务器和文件服务器集中在一台硬件服务器中, 并配备 Oracle RAC 为数据库。

收稿日期: 2013-05-10; 修回日期: 2013-08-05

基金项目: 科技部基础条件平台——地球系统科学数据共享网(2005DKA32300); 国家海洋局青年创新基金(2012621); 中国极地科学战略研究基金项目(20120106)

作者简介: 程文芳(1979-), 女, 湖北黄石人, 工程师, 硕士, 主要从事数据共享、数据集成和数据挖掘, 电话: 021-58713642, E-mail: chengwenfang@pric.org.cn; 张洁(1970-), 通信作者: 男, 上海人, 高级工程师, 研究方向: 数据政策、数据管理、数据共享, E-mail: zhangjei@pric.org.cn

1.1 CHINARE 系统体系架构

系统架构分为四个层次: 物理数据层、逻辑数据层、业务层、用户层。物理数据层部署数据实体的存储。关系数据和日志存放于数据库, 数据集存放于磁盘, 快照数据存放于服务器缓存。逻辑数据层管理 CHINARE 平台的各类数据。业务层设计为系统业务功能区。见图 1。

来自互联网的 Web 请求通过 Apache 的 WSGI 转发至 Web 服务框架 Django, 核心应用业务由 Django 完成, 它主要负责处理业务请求事物, 根据不同的资源类型, 分别向 Oracle 数据库、Django 缓存(最近访问数据)、Web 服务器硬盘(数据集)请求资源。除了来自互联网的请求外, 系统定时需要进行的日志处理、数据分析、索引处理等其他任务通过 Web Services 方式与核心应用通信。

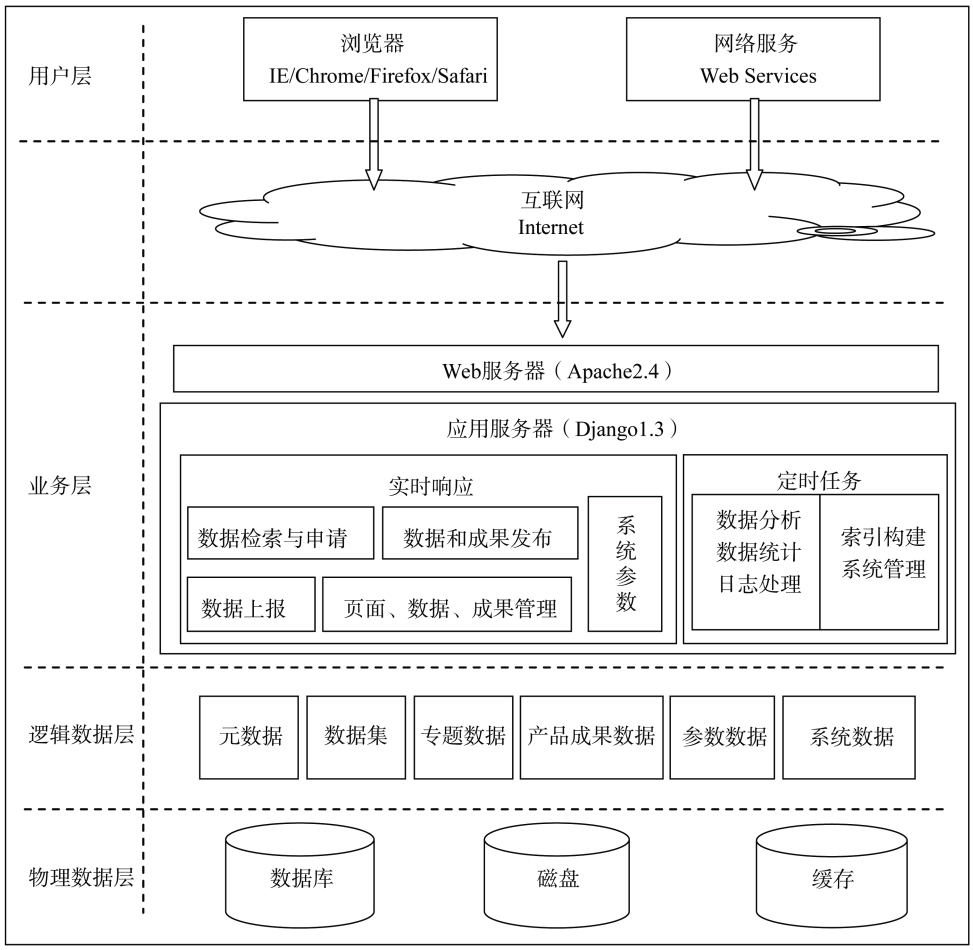

图 1 CHINARE 结构图

Fig.1 System structure of CHINARE

1.2 CHINARE 与其他系统之间的接口

极地科学数据共享平台与极地之门(极地考察信息门户)、极地各专题数据库、地球系统科学数据共享网总平台、国家科技基础条件平台、全球气候变化主目录 GCMD 之间存在多种依托和附属关系。CHINARE 平台系统向地球系统科学数据总平台汇交用户信息、元数据信息和日志信息; 向国家科技基础条件平台汇交资源数据和服务统计数据[4]; 与极地之门共享新闻模块, 实现单点登录、portlets 个性化定义、系统角色和权限分配, 实现公共基础数据(例如: 机构、人员、论文、项目等)共享; 向 GCMD 共享元数据信息。虽然 CHINARE 数据库与各专题库之间没有智能数据映射, 数据映射仅仅局限于专题数据发布时手工匹配元数据。 但是由于各专题库与极地之门实现了单点登陆, 实现了基础数据库(考察人员、机构、项目、队次、地名、论文、考察站、术语)共享, 因此 CHINARE 平台与各专题库之间数据互访及集成检索实现了一定的紧密度。CHINARE 平台系统体现架构如图 2。

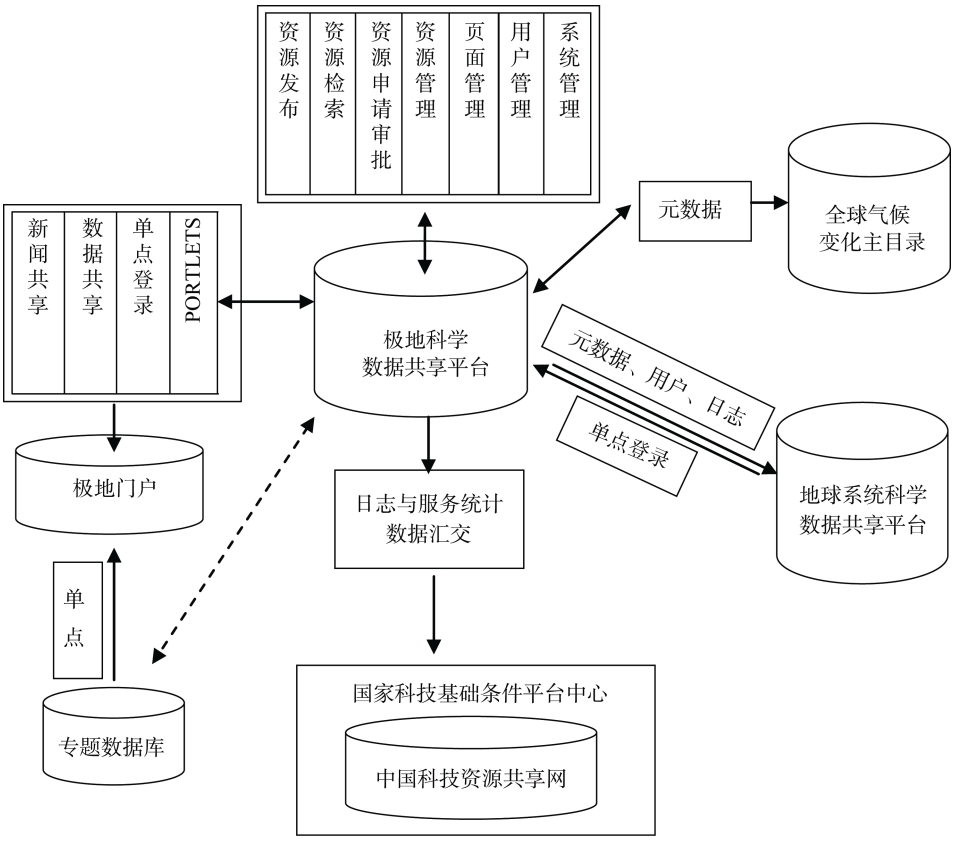

图 2　系统全局架构

Fig.2　The whole structure of the system

1.3　CHINARE 系统功能设计

除了元数据和数据集在线发布、数据检索的基础功能, CHINARE 系统设计了导航浏览、数据集预览、数据挖掘、数据申请审批、新闻统一发布、日志挖掘、科普宣传等功能(图 3)。

1.3.1　数据挖掘层

分析数据共享和使用情况, 提高专题数据质量。提供海冰监测数据、走航气象数据、走航温盐数据、走航 GPS 数据等专题数据; 提供考察家属和决策层考察船航行信息数据(依托雪龙在线系统); 提供数

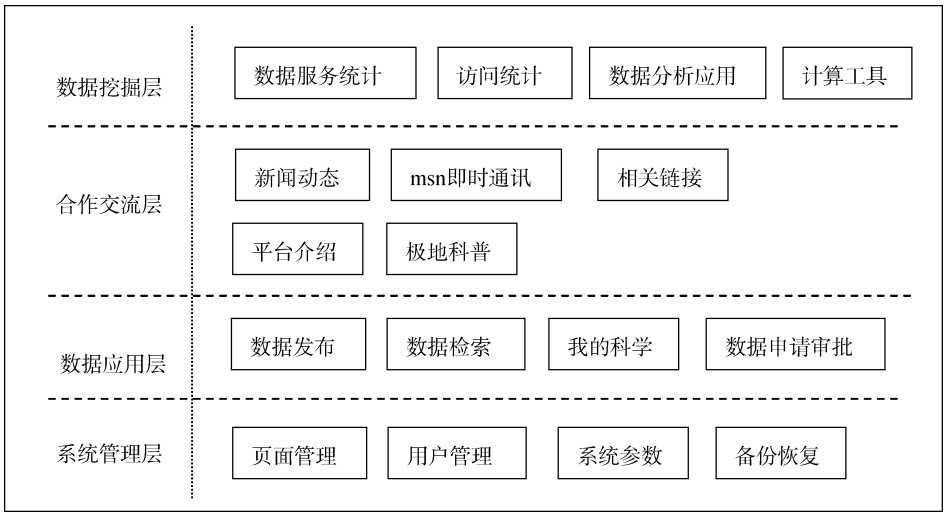

图 3　CHINARE 系统功能图

Fig. 3　System functions diagram of CHINARE

据分析常规工具(如时区转换工具、温度转换工具、速度转换工具、经纬度转换工具等)。根据数据共享情况、数据服务情况、数据质量情况统计分析数据的利用和效果。

1.3.2 合作交流层

对外提供极地科学知识普及、新闻宣传、相关国内外资源链接导航、数据使用即时通讯等功能[5]。

1.3.3 数据应用层

提供数据发布、数据检索、数据导航和资源申请审批等数据使用。利用“我的科学”数据空间记录数据使用情况，并推荐数据和相关考察人员动态。

1.3.4 系统管理层

对系统参数、元数据参数、数据集参数进行集中管理；提供管理员注册用户审核、元数据发布质量审核、数据集质量审核、元数据填报公共模板和角色克隆等系统管理功能；负责平台新闻公共内容编制；另外还支持系统备份和恢复。

1.4 数据库建模

数据库设计非常重要，一旦数据模型确认，整个系统底层通讯和架构就确定了。为了便于系统化标准化设计数据库，采用了 UML 语言和 Erwin 建模工具。Unified Modeling Language (UML)又称统一建模语言或标准建模语言，它是一个支持模型化和软件系统开发的图形化语言，为软件开发的所有阶段提供模型化和可视化支持，包括由需求分析到规格，到构造和配置[6]。可以在几个层次上显示系统如何工作，非常有利于用户和其他涉及项目人员之间的通信[7]。ERWin 是目前三大主流数据建模之一，是关系数据库应用开发的优秀 CASE 工具，主要用来建立数据库的概念模型和物理模型。它能用图形化的方式，描述出实体、联系及实体的属性[8]；可以方便地构造实体和联系，表达实体间的各种约束关系，并根据模板创建相应的存储过程、包、触发器、角色等。ERwin 可以实现将已建好的 ER 模型到数据库物理设计的转换，即可在多种数据库服务器(如 Oracle, Sql Server, Watcom 等)上自动生成库结构，并进行逆向工程、自动生成文档、支持与数据库同步，提高了数据库的开发效率。CHINARE 在 Oracle 数据库设计中，共设计了 8 个包：数据集(pDataset)、专题库(pSubjectDb)、元数据(pMetadata)、元数据要素(pMetadataElement)、统计(pStat)、系统参数(pSystemInterface)、极地科普(pOutreach)、系统信息管理(pSystemInfo)。系统 UML 图见图 4。

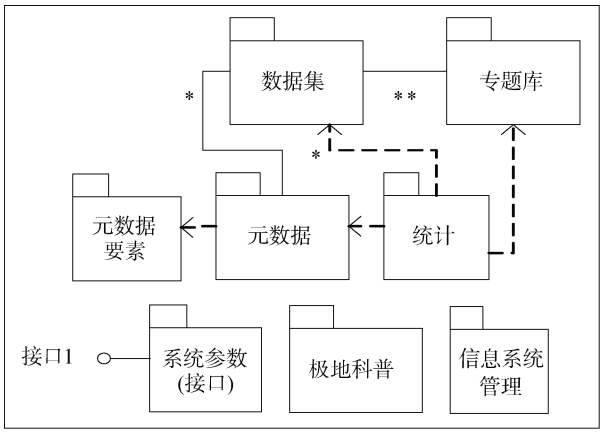

图 4 数据库 UML 静态图

Fig. 4 Statics UML figure of Database

以统计包为例，它包含资源统计(cDataStat)和访问统计(cVisit)两个类。资源统计类包含元数据统计和数据集统计两个子类，访问统计类包含服务天数、总点击数、总用户数、唯一 IP 总数、PV 总数、总下载量、用户来源国家数等 10 个子类。

1.5 元数据的处理流程

为了确保元数据检索速度，并提供多维检索模式，本文采用关系数据库存储元数据，并以 XML 格式汇交元数据(图 5)。当元数据处于编辑状态时，系统触发基于 Javascript 和 JSON 数据格式的 underscore 模板引擎将关系数据模式转换为 Json 格式；当元数据处于汇交状态时，系统设计将其转换为 XML 格式，并通过 XSD 验证[9]。

XSD(XML Schemas Definition)，描述了 XML 文档的结构。可以用一个指定的 XML Schema 来验证某个 XML 文档，以检查该 XML 文档是否符合其要求。文档设计者可以通过 XML Schema 指定一个 XML 文档所允许的结构和内容，并可据此检查一个 XML 文档是否是有效的。XML Schema 本身是一个 XML 文档，它符合 XML 语法结构。可以用通用的 XML 解析器解析它[10]。

JSON(JavaScript Object Notation)是一种轻量级的数据交换格式[11]。它基于 JavaScript(Standard ECMA-262 3rd Edition - December 1999)的一个子集。JSON 采用完全独立于语言的文本格式，但是也使用了类似于 C 语言家族的习惯(包括 C, C++, C#, Java, JavaScript, Perl, Python 等)。这些特性使 JSON 成为理想的数据交换语言。易于阅读和编写，同时也易于机器解析和生成。

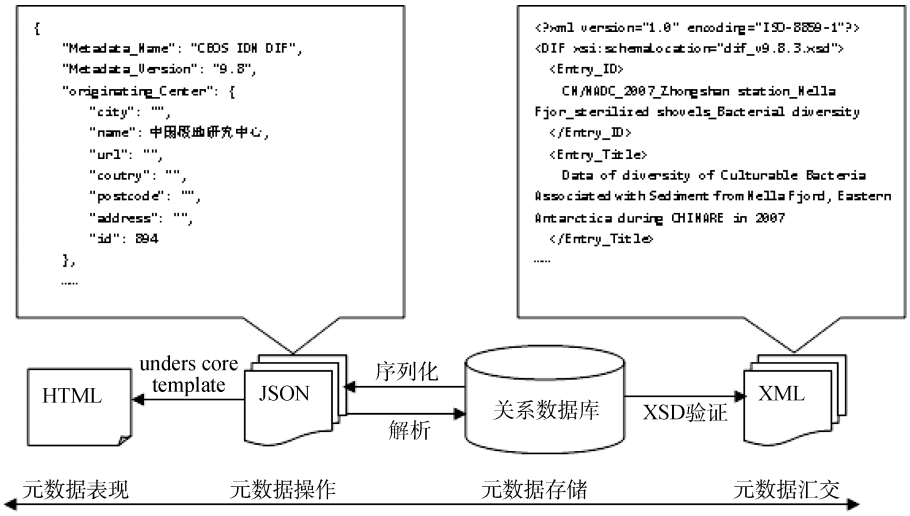

图 5　元数据存储汇交流程

Fig.5　Storage and collection of metadata

2　关键功能开发

系统的开发实现包括开发环境配置，数据浏览、检索、发布、元数据发布审批、查询结果处理、数据服务日志上报、系统管理等功能的开发。

2.1　元数据标准

国际上比较典型的地理空间元数据包括美国地理信息委员会 FGDC 标准，地理观测卫星委员会 CEOS IDN 标准，欧盟 CEN/TC211 标准，国际 ISO/TCC 211 等。不同的国家，不同的组织根据自身数据背景，数据特点、侧重点，数据管理原则而采用不同的元数据标准。例如，MMI 侧重于海洋监测，CODATA 侧重于交叉数据研究，OGC 采用 ISO 国标，美国航天宇航局 NASA 采用 FGDC 标准，美国雪冰数据中心 NSIDC 采用 CEOS IDN 标准。澳大利亚南北极数据中心扩展自己的标准，以适应于国际标准 ISO19115/19139。极地科学数据共享平台包括大量的遥感影像、地图、生态环境监测等考察数据，基于元数据交换的标准，在元数据标准的选择上采用 CEOS IDN 地理信息元数据标准 DIF V9.7，并根据国内数据的特性对 DIF 的元素属性进行调整，保留了 36 项要素，对要素的约束性进行了自定义，对子要素进行了简约、拆分、合并、预处理等处理。CHINARE 元数据实体信息见图 6。

元数据标准采用 UML 建模，再映射成 XML 实现数据存储与交换。为了保证元数据的质量和国际交换的有效性，在元数据的基础上建立了核心元数据，要素包括：元数据标识、元数据标题、参数、ISO 主题类别、摘要、关键词、学科分类、数据中心、

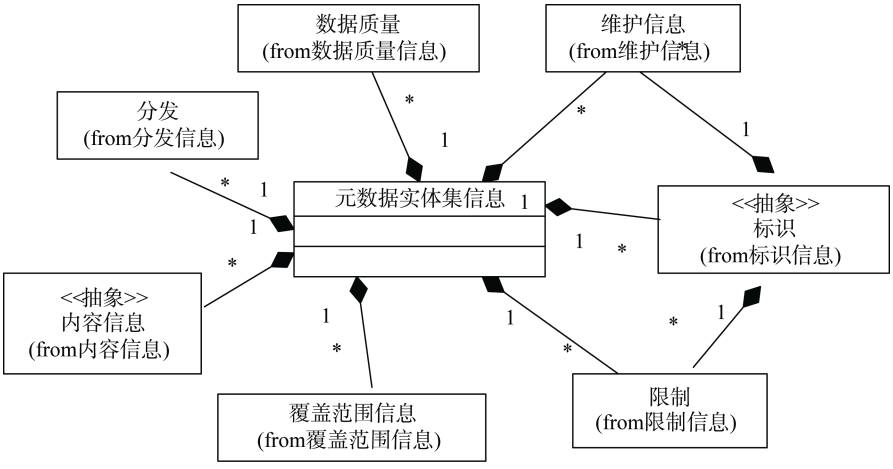

图 6　元数据库实体集信息

Fig. 6　Entity structure of metadata

元数据标准名称、元数据标准版本。元数据以二维关系数据格式存储在 Oracle 数据库中，并转换成 XML 文件存入 BLOB 大字段中便于向 GCMD 自动汇交。核心元数据的 XML 格式表达如下:

```
<xs: element name="DIF">
    <xs: complexType>
        <xs: sequence>
            <xs: element ref="Entry_ID" minOccurs="1" maxOccurs="1"/>
            <xs: element ref="Entry_Title" minOccurs="1" maxOccurs="1"/>
            <xs: element ref="Summary" minOccurs="1" maxOccurs="1"/>
            <xs: element ref="Discipline" minOccurs="0" maxOccurs="unbounded"/>
            <xs: element ref="Parameters" minOccurs="1" maxOccurs="unbounded"/>
            <xs: element ref="ISO_Topic_Category" minOccurs="0" maxOccurs="unbounded"/>
            <xs: element ref="Keyword" minOccurs="0" maxOccurs="unbounded"/>
            <xs: element ref="Data_Center" minOccurs="1" maxOccurs="unbounded"/>
            <xs: element ref="Metadata_Name" minOccurs="1" maxOccurs="1"/>
            <xs: element ref="Metadata_Version" minOccurs="1" maxOccurs="1"/>
        </xs: sequence>
    </xs: complexType>
</xs: element>
```

2.2 元数据发布

通过元数据标准可以保证元数据结构符合元数据标准的要求，但仍然无法保证元数据内容的正确性。因此，本文提出元数据的审查、发布流程，通过行业数据专家和平台管理员的参与，进一步检查元数据内容。其中元数据由数据采集人、元数据作者、数据支持者在线发布，尽可能保证元数据的完整性、正确性和客观性描述，更真实地反映目标数据；元数据在线共享审核则由平台数据管理员完成，通过元数据质量审核、元数据状态标识、数据期和保护期用户组管理实行元数据的在线发布共享。本文采用的元数据审核步骤有拒绝、通过和入库；元数据状态有状态不明、已完成、未完善、初始化和超期限。元数据发布流程[12]如图 7。

仅登录用户可在线发布元数据，且该用户为极地之门用户。用户登录后由门户进行身份识别和认证。极地之门采用 OAuth 协议来认证，它是一个开放标准，允许用户提供一个令牌，而不是用户名和密码来访问他们存放在特定服务提供者的数据。每一个令牌授权一个特定的应用系统在特定的时段内访问特定的资源。这样，OAuth 让用户可以授权第三方网站(如 CHINARE)访问他们存储在另外服务提供者的某些特定信息，而非所有内容[13]。

用户可通过公有模板和私有模板编辑元数据(图 8)，元数据填报采用“INPUT PROMPTING”技术有效地降低了用户输入成本，提升用户搜索体验，增加了网站的用户黏性，保证了元数据基础数据维护的一致性和权威性。CHINARE 的输入框自动提示功能的实现开发主要解决问题。

2.2.1 输入框的位置

为了在输入一些内容时能够显示提示框，在设计时需要建立一个 div，初始时不可见，当改变输入框内容时则显示出来，并且移动到正确的位置。另外，还需要确定一个元素的相对页面的绝对位置。本文采用绝对定位来计算输入框位置，通过 Offsetparent 取得父亲结点递归来计算该元素的相对页面的绝对位置。

```
function getabsposition(obj) {
        var r = {
              left: obj.offsetleft,
              top : obj.offsettop
  };
  r.left = obj.offsetleft;
  r.top = obj.offsettop;
if(obj.offsetparent) {
        var tmp = getabsposition(obj.offsetparent);
        r.left += tmp.left;
      r.top    += tmp.top;
  }
  return r;
}
```

2.2.2 输入框中内容侦测

假如采用 onchange 事件，只有当输入框失去焦点时才会触发，因此本文采用定时器的方式，每隔一个时间片检查一下输入框的值，假如值改变了，就执行更新过程。通过 Ajax 的回调函数将服务器传回的 json 数据解码并且以一定的方式显示在 div 中。

2.2.3 内容选择

在内容检测方面，本文采用了键盘和鼠标双向检测法。使用键盘的向上和向下的按键选择内容，用 Enter 键来确定内容；或者用鼠标上下滚动选择内容，用鼠标左键来确定内容。

2.2.4 输入框校正

解决浏览器自带的自动历史选择下拉框与 CHINARE 系统的智能输入提示框的冲突。本文通过

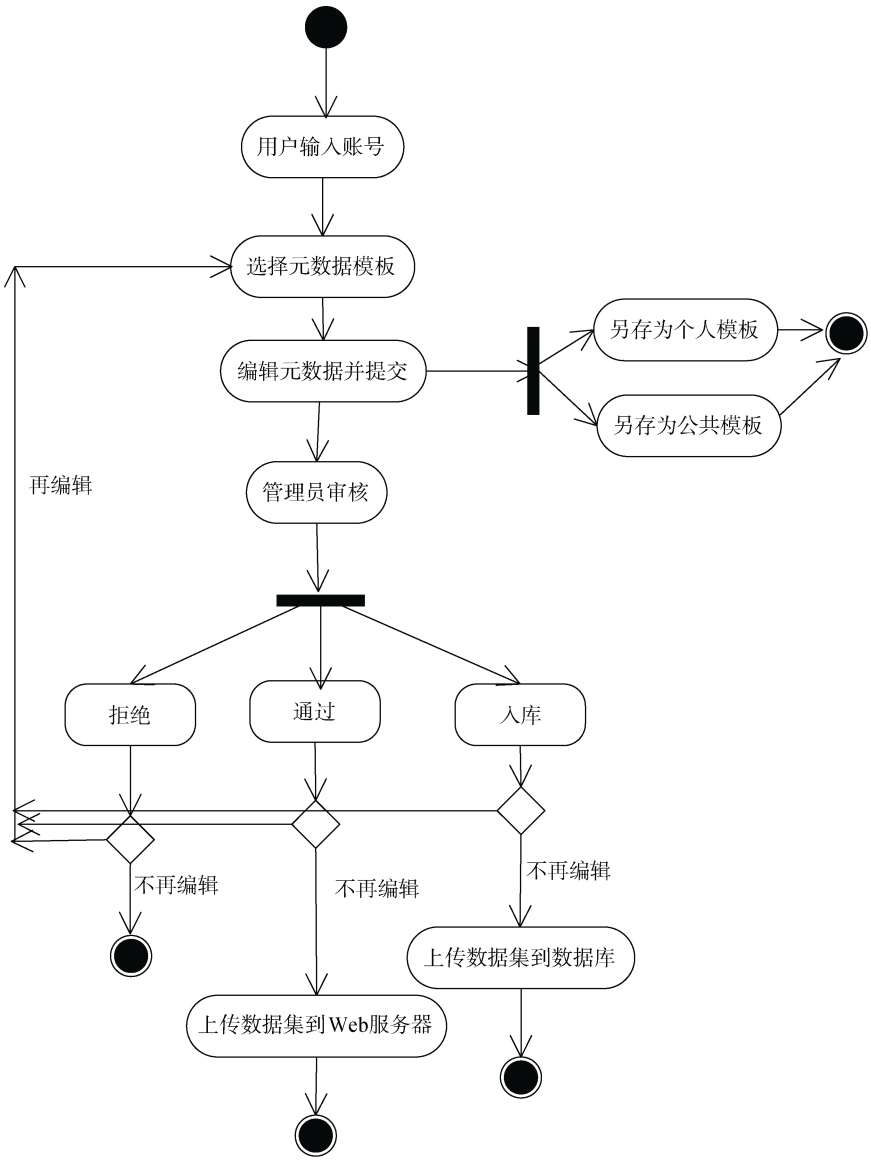

图 7 元数据发布活动图

Fig.7 Activity diagram of metadata publish

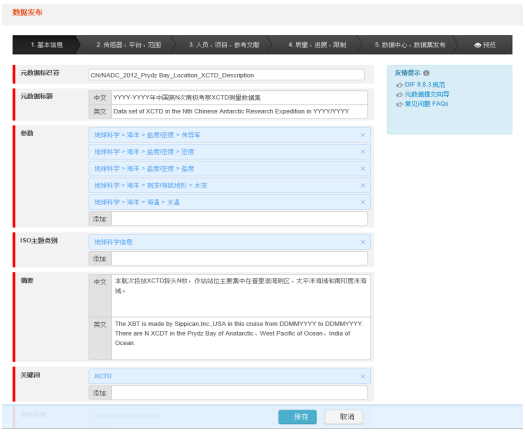

图 8 元数据的编辑页面

Fig. 8 Edit page of metadata

改变输入框的 oncomplete 属性实现。

常规下数据集的上传限制于网络带宽，针对带宽限制，采用 swfuploader 插件实现文件批量上传功能，且可实现 3 GB 数据的断点续传，解决了超大数据上传困难的瓶颈。但是，对处于代理模式下的 Apache，还需要处理超大文件的超时问题。

#配置说明：链接超时时间，单位为 s。

ProxyPass/ http://ip: port/ timeout=36000 Keepalive=On

同时，为了便于其他专题数据库中发布数据时能匹配到元数据，在系统设计中为每个用户发布的元数据提供了调用接口 API(图 9)。

2.3 元数据检索

系统实现中采用了元数据导航、元数据检索、元数据地图检索、数据集检索四种检索模式。元数据导航采用级联动态分类，系统类别由数据管理员从后台添加，数据与类别之间是多对多关系，数据可以归类于多个类别[14]。如果一个数据属于子类别，在父类别里也可见此数据。这种灵活的分类方式允许从不同的侧面对一数据进行归类，也为极地之门下的多数据平台共享一个元数据库奠定了基础。

元数据检索模式下，首先对学科分类、发布日期、传感器、搭载平台、地点、提交单位进行一级遍历，然后再对关键词和发布时间进行二级遍历。

```
<script type="text/javascript"src="http://beta.chinare.org.cn/js/jquery-1.7.1.min.js"></script>
<script type="text/javascript"src="http://beta.chinare.org.cn/js/chinare.api.1.0.js"></script>
<div id="container_mydif"></div>
<script>
        $("#container_mydif").GetMetadata({"creator":""});
</script>
```

图 9　元数据调用接口

Fig .9　Interface of metadata

$$Z = \mathrm{f}_{zi}(k_{zi}),\ 1 \leqslant i \leqslant 6$$

上述表达式，Z 为一级检索结果集，f 为检索方法，z_i 为检索因子，z_1 到 z_6 分别为：学科分类、发布日期、传感器、搭载平台、地点、提交单位；k_{zi} 为相应检索因子的关键词。

$$S=Z \mid X \cap Y \mid Z \cap X \cap Y$$

$$X=\mathrm{f}_{xj}(k),\ 1 \leq j \leq 16$$

$$Y = \mathrm{f}_{\mathrm{T}}(t_1, t_2)$$

其中，X 为对关键词检索结果集，Y 为对时间区间检索结果集；k 为输入关键字，x_j 为检索因子，x_1 到 x_{16} 分别为：标题、标识符、学科分类、提交者、关键词、摘要、传感器、搭载平台、地点、参数、源中心、作者、ISO 主题类别、数据中心、数据集进展、数据集语言；T 为元数据发布时间检索因子，t_1 为起始时间，t_2 为结束时间。检索结果集 S 可以是一级检索结果集 Z，或二级检索结果集 $X \cap Y$，也可以是在一级检索结果上再进行二级检索的结果集。

数据集检索模式下，首先对学科分类、文件格式进行一级遍历，然后再对关键词和发布时间进行二级遍历。

$$Z = \mathrm{f}_{zi}(k_{zi}),\ i = 1, 2$$

其中 z_1，z_2 分别指学科分类和文件格式。z_1 包含极地海洋学、极地地球物理学、极地大气科学、极地生物学、极地环境科学、极地地理学、极地地质学、极地工程、极地冰川学、南极天文学，z_2 包含 doc、xls、rar、txt、zip、csv、jpg、pdf、cdr、dat、xbt、raw、xlsx、bmp、docx。

$$S = Z \mid K \cap T \mid Z \cap K \cap T$$

$$K = \mathrm{f}_{\mathrm{zn}}(k) \cup \mathrm{f}'_{\mathrm{zn}}(k) \cup \mathrm{f}_{\mathrm{en}}(k) \cup \mathrm{f}'_{\mathrm{en}}(k)$$

$$T = \mathrm{f}_{\mathrm{T}}(t_1, t_2)$$

其中，f 为针对元数据标题检索方法，f'为针对数据集标题检索方法，zn 为中文标题检索因子，en 为英文标题检索因子(zn 和 en 为关键词系统互译)；T 为对时间区间检索结果集，t_1 为起始时间，t_2 为结束时间。检索结果集 S 可以是一级检索结果集 Z，或二级检索结果集 $K \cap T$，也可以是在一级检索结果上再进行二级检索的结果集。

数据集检索的结果，本文还通过数据预览功能提供用户下载前阅读。系统设计了 pdf、doc、jpg、bmp、xls、txt、 zip、rar 八种数据格式的数据预览(图 10)。压缩格式的数据列出数据目录，doc 格式的数据先转换成图片格式，再以省略图的方式预览，pdf 格式数据使用 Django 插件第三方 Reportlab 实现，xls 使用插件 Django 第三方 xlrd, xlwt 实现。

元数据地图检索使用 google map 加载数据实现可视化检索。采用常规地理位置信息标记方法无法

图 10　数据集预览

Fig. 10　Preview of dataset

快速加载数据, 因此系统改进算法, 使用 Cluster 聚类算法动态地对采集坐标点进行分组聚类, 提高了在地图中显示大量数据的响应速度。根据一定规则将 Marker 划分到不同的群中。根据地图的不同缩放级别显示不同的 marker 集群。当地图缩放的时候, 继续将更多的相关的 marker 集中在这个组中, 用这个组代替 marker 显示。当放大到一定级别再显示 marker 本身。目前主要有三种集群方式: 基于网格, 将地图分割成很多网格。根据地图不同的缩放级别的在同一个网格中的 marker 集群在一个分组中; 基于距离, 将距离近的 marker 显示在一个分组中; 基于分区, 将地图分成不同的分区, 比如根据省份将地图分成不同的分区, 将省份中的 marker 集中在一起显示。考虑到南北极地域性和南极条约限制, 本文采用了基于网格的聚类算法, 使用 Google 地图第三方类库 MarkerClusterer 实现数据检索。测试环境: HP Z400 Workstation, Intel Xeon CPU 3.07GHz, RAM 6GB, 64 Bit Windows 7, 加载 425 条元数据历经仅为 0.02 s。

2.4 敏感词过滤

基础数据库是极地考察信息门户中的一项重要服务内容, 它是一个开放的, 允许多人共同编辑的基础信息资料库。包括队次、人员、机构、论文、项目、术语、地名等多个信息库, 目前共收录 43902 条数据。其特点如下: (1)避免基础信息在各应用系统中的重复录入工作, 并保证了基础信息的一致性。(2)各应用系统可以以最小的程序修改代价, 便捷地引用基础公共数据。(3)提高各应用系统的数据与基础公共数据间的关联度。

为了有效地将数据关联起来, 扩展用户的延伸阅读。系统设计信息检索结果与极地之门中的资源词条关联在一起。为了识别极地考察信息门户网络文本的基础数据词典, 通过分析文本结构以及基础数据词典的特点, 建立二层过滤模型, 将整个基础数据词典划分为首字符过滤和与首字符关联过滤。数据接口为一个返回 json 数据格式的 web 服务, 可应用于 Ajax 请求。当向该接口发送一段文本后, 会将该文本中的关键字高亮显示, 并加上详细查看的跳转链接。

2.5 专题数据库

根据数据特点, 在数据挖掘层设计实施了 8 个专题数据库: 极地地磁数据库、极光数据库、极地海洋数据库、极地遥感与地理信息数据库、极地气象数据库、雪龙在线、科考航行数据库和走航卫星云图数据库。这些专题数据从空间上覆盖高空和海洋; 数据采集时间贯穿 28 个南极航次; 同时数据覆盖多学科最需要的地理数据、气象数据、海洋、GPS 数据和卫星云图等基础数据, 数据发布情况见表 1。

表 1 专题数据库数据发布情况
Tab.1 Data publish about subject databases

专题数据库	数据共享情况
极地遥感与地理信息数据库	全南极遥感影像图、海冰变化产品
极地地磁数据库	2004~2006 年, 2013 年地磁数据
极光数据库	1995~2010 年极光数据
极地海洋数据库	极地海洋物理化学基础数据: 南极第 13 航次~南极第 26 航次 北极第 1 航次~北极第 3 航次
极地气象数据库	1985~2013 年气象数据
雪龙在线	南极第 1 航次~南极第 31 航次(南极第 2, 4, 10, 17, 20, 23 航次数据缺失) 北极第 1 航次~北极第 6 航次
科考航行数据库	数据整理发布中
走航卫星云图	数据整理发布中

3 CHINARE 系统性能分析

经过反反复复的需求调研和揣摩, 详细的功能设计, 前沿的系统架构布局, 敏捷的开发流程, 人性化的界面分析使得基于 Python 开发语言的 CHINARE 系统大大节省了开发周期且更便于维护和升级。CHINARE 系统已经完成系统建设, 并且在系统响应速度和使用效率上都突显出极地科学数据共享的优势, 达到预期效果, 它支持 IE5.0, IE7.0, IE8.0, IE9.0, Chrome, Firefox 和 Safari 浏览器。平台运行以来首屏显示平均 0.361 s, 文件检索平均耗时 0.671s。已经发布 745 条元数据、2060 个数据集, 71.7 GB 数据, 日均点击数 1 288 次, 总点击数用户数 54.3 万个, 总下载量达 859.5 GB, 有来自 144 个国家的用户访问平台。

4 结论与展望

无论是系统功能设计, 还是页面布局, 极地科学数据共享平台已经受到美国、澳大利亚、韩国、日本等国家的关注, 同时得到数据管理人员和科学家的一致认可。

但是随着用户交互的快速增长, 系统在性能和功能上有待进一步完善:

1) 综合检索问题。本文的数据检索建立在关系型数据库和 xml 检索上，而且检索内容局限于元数据要素和数据集名称，无法对数据集全文检索，无法统一数据语义，也无法结合元数据和主题数据库查询。随着数据共享程度的提高，还需要通过对数据集成等技术手段实现数据检索，扩大数据共享范畴。

2) 资源数据链研究。目前 CHIANRE 系统已经完成元数据–项目–成果的数据链研究，但是与主题数据库(国内外)–数据产品–类似数据(关注数据)等关系网建设还在探讨中[15]。与此同时关系链系统自动匹配技术还需进一步调研分析。

3) 数据质量问题。元数据建立了审查机制，但是数据集的审查机制还在摸索中，还需要建立专家委员会等方式把握数据质量。

参考文献:

[1] 张侠，朱建钢，凌晓良，等. 中国极地科学考察数据管理与共享[J].海洋开发与管理, 2004, 5: 50-53.

[2] Swaroop C H，沈洁元译.简明 Python 教程[EB/OL]. [2005-02-17]. http://woodpecker.org.cn/abyteofpython_cn/chinese/.

[3] 屈展，李婵. JSON 在 Ajax 数据交换中的应用研究[J]. 西安石油大学学报(自然科学版), 2011, 26(1): 95-98.

[4] 诸云强，冯敏，宋佳，等.基于 SOA 的地球系统科学数据共享平台架构设计与实现[J].地球信息科学学报, 2009, 11(1): 1-9.

[5] 李新，南卓铜，吴立宗，等.中国西部环境与生态科学数据中心：面向西部[J].地球科学进展, 2008, 23(6): 628-637.

[6] 刘海宁，段斌华，李楠，等.设备状态智能诊断模型的自更新机制及其 UML[J].高技术通讯, 2011, 12: 1305-1311.

[7] 程文芳，张侠，何剑峰，等．极地生态环境监测与研究信息平台的设计与实现[J].极地研究, 2009, 21(4): 299-307．

[8] 杨国强. ERWIN 数据建模[M].北京：电子工业出版社, 2004.

[9] 冯敏，诸云强，王卷乐，等.分布式多标准地学元数据共享研究与实践[J].地理与地理信息科学, 2007, 23(6): 8-13.

[10] Hunter D, Rafter J, Fawcett J, et al. 吴文国译. XML 入门经典[M]. 清华大学出版社, 2009.

[11] Bray T. The JSON Data Interchange Format[DB/OL]. [2013-01-17].http://www. json. org/.

[12] 申利民，隋峰，白莲，等.基于扩展 UML 活动图的工作流模型研究[J].计算机应用研究, 2009, 26(2): 587-590.

[13] 诸云强.地球系统科学数据共享关键技术研究[D].北京：中国科学院地理科学与资源研究所, 2006.

[14] 南卓铜，李新，王亮绪，等.中国西部环境与生态科学数据中心在线共享平台的设计与实现[J].冰川冻土, 2010, 32(5): 970-975.

[15] 孙枢.地球数据是地球科学创新的重要源泉——从地球科学谈科学数据共享[J].地球科学进展, 2003, 18(3): 334-337.

System design and implementation of Data-sharing platform of polar science

CHENG Wen-fang[1], WANG Wei[2], ZHANG Jie[1], XIA Ming-yi[3], YANG Rui[1], ZHU Jian-gang[1], ZHANG Bei-chen[1], WANG Da-li[1], LING Xiao-liang[1], LI Sheng-gui[1]

(1. Polar Research Institute of China, Shanghai 200136, China; 2. National Marine Data & Information Service, Tianjin 300171, China; 3. Chengzhi Information Technology Co., Ltd., Shanghai 201204, China)

Received: May, 10, 2013

Key words: Python; polar; metadata; data retrieval; data sharing

Abstract: In order to share the data collected by Chinese Antarctic and Arctic expeditions, we designed a distributed system to effectively manage these polar data. The system has data retrieval, dissemination, application and thematic service functions, and it has been successfully applied as Chinese unique polar data sharing system, and it is also helpful for international projects cooperation and data sharing.

(本文编辑：刘珊珊)