互联网+煤层气元数据管理系统关键技术研究

李 梅,邹学森,毛善君,周恩波
(北京大学 遥感与地理信息系统研究所,北京 100871)

摘 要: 针对我国煤层气开发利用数据复杂多样、格式各异、分散存储以及海量数据无法充分共享等特点,在充分研究满足信息共享需求的基础上,设计了基于互联网的煤层气元数据管理系统。首先探讨了互联网环境下基于XML的煤层气开发利用元数据标准、元数据自动获取和互联网收割等技术,最后以某煤矿瓦斯地质图、某煤层气数据处理平台报告文档和煤层气利用数据库为例,给出了互联网环境下相应元数据自动生成的实例。本元数据管理系统贯穿了煤层气地质、开发和利用的全生命周期,通过元数据库信息的自动更新,能够实现煤层气信息的及时和有效共享,为煤层气开发利用的动态决策支持服务。

关键词: 互联网+; 煤层气; 元数据; 元数据抽取; 元数据收割
中图分类号: TP311.5; TD67 **文献标志码:** A **文章编号:** 0253-2336(2016)07-0080-06

Study on key technology of internet plus coalbed methane metadata management system

Li Mei, Zou Xuesen, Mao Shanjun, Zhou Enbo
(*Research Institute of Remote Sensing and Geographic Information System, Peking University, Beijing* 100871, *China*)

Abstract: According to the complicated and diversity, different format, separated storage of the coalbed methane development and utilization data in China and the massive data hard to fully sharing and other features, based on the full study to meet the requirements of the information sharing, the coalbed methane metadata management system was designed based on the internet. The paper firstly discussed the metadata standards of the coalbed methane development and utilization, metadata automatic extraction and internet reaping and other technology under the internet environment and based on the XML. In the end, based on mine gas geologic map, a report document of coalbed methane data processing platform and coalbed methane utilization database as the cases, the paper provided the relevant metadata automatic generation case under the internet environment. The metadata management system ran through the full life period of the coalbed methane geology, development and utilization. The automatic upgrading of the metadata database information could realize the coalbed methane information sharing timely and effectively and could provide the support services of the dynamic decision making for the coalbed methane development and utilization.

Key words: internet plus; coalbed methane; metadata; metadata extraction; metadata reaping

0 引 言

在我国煤层气的开发和利用领域中,形成了大量的与煤层气地质条件、开发与利用相关的数据[1-2],这些数据格式各异,包括图形数据、数据库数据和文档数据。其中,图形数据包括AutoCAD平台的dwg格式文件、LongruanGIS平台的lfmx格式文件、ArcGIS平台的shapefile格式文件和MapGIS平台的wt、wl、wp格式文件等;文档信息包括Microsoft Office Word格式和Adobe Acrobat PDF格式,而数据库又包含Microsoft SQL Server、Oracle以及Access等平台。同时,我国煤层气开发和利用数据

收稿日期: 2016-03-02; **责任编辑:** 赵 瑞 **DOI:** 10.13199/j.cnki.cst.2016.07.014
基金项目: 国家科技重大专项资助项目(2011ZX05040-005)
作者简介: 李 梅(1978—),女,陕西岐山人,副教授,博士。E-mail: mli@pku.edu.cn
引用格式: 李 梅,邹学森,毛善君,等.互联网+煤层气元数据管理系统关键技术研究[J].煤炭科学技术,2016,44(7):80-85.
Li Mei, Zou Xuesen, Mao Shanjun, *et al*. Study on key technology of internet plus coalbed methane metadata management system[J]. Coal Science and Technology, 2016, 44(7): 80-85.

均由生产单位自己保存和维护，各单位间缺乏有效的信息共享机制，导致煤层气开发和利用数据无法有效共享，“信息孤岛”现象严重; 企业间数据仍然采用传统的手工共享方式，导致信息共享效率低下。另一方面，在我国煤层气开发利用领域仍然缺少元数据的研究和成果，无法对煤层气开发和利用数据进行高效的管理[3]，这些都严重阻碍了煤层气信息的共享。

通过互联网技术，从各种数据源中进行元数据的主动抓取和信息的主动分发，以及数据库和信息平台的自助式更新已成为新的研究热点。中国科学院网络信息中心开发的地理空间云平台[4]，通过 8 年时间建立了覆盖周期长、覆盖范围广的地理信息数据资源库，完成 14 大类 109 个数据产品的汇聚，汇聚元数据 800 多万条。我国国家地质调查局就开发了基于 Internet 的元数据管理系统 CCOP-GIMS，通过该系统可直接向普通公众免费发布近 18 000 种、50 000 余件，总数据量高达 3.3 TB 的地质资料数据信息[5]。但是，在我国煤层气领域，尚缺乏一套行之有效的信息管理系统，来实现煤层气地质数据、开发和利用数据的有效共享。因此，为实现煤层气信息的充分共享，在充分研究我国煤层气数据格式特点和煤层气数据管理特点的基础上，笔者提出建立基于互联网的煤层气元数据管理系统。运用互联网技术，通过对煤层气信息主动抓取和元数据自动抽取，来实现数据库、元数据库和信息平台的自助式更新。同时，根据用户特点和权限，将数据库和相应元数据的更新信息主动发送给用户，从而克服了传统数据共享模式的缺点，实现我国煤层气地质信息、开发和利用信息的高效共享。

1　基于互联网的煤层气元数据管理系统设计

1.1　系统框架

针对我国煤层气领域数据的分散性、数据类型及格式的多样性和数据的海量性，以及通过互联网进行数据共享的技术特点，设计了基于互联网的煤层气元数据管理系统。系统采用三层体系架构，分为数据层、服务层和应用层，可以方便用户对煤层气领域多源异构数据的访问，实现数据的有效共享。系统的体系结构如图 1 所示。

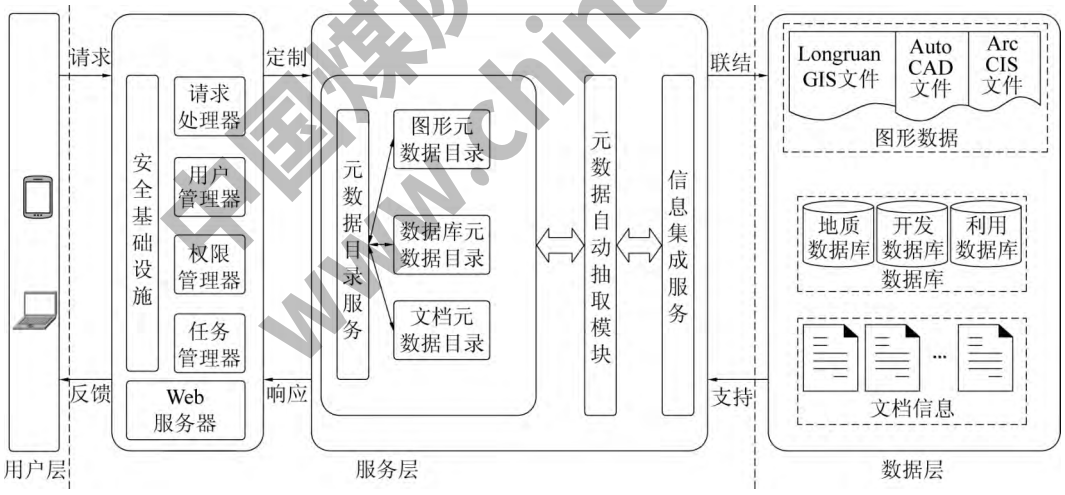

图 1　煤层气元数据管理系统的体系结构

Fig. 1　Structure of CBM metadata management system

1) 数据层。数据层包括煤层气地质、开发和利用的各类数据，这些数据以图形数据、数据库数据和文档数据 3 类形式分开存储。它们分布在全国不同地点和不同单位，通过 Internet 或各种无线通信设备实现物理连接。

2) 服务层。服务层提供了煤层气信息一体化管理与处理平台。当煤层气开发和利用领域的成果拥有者按照相关规定上传相应成果后，服务层检测到数据有变化，则会通过元数据自动抽取模块，自动解析相应数据的元数据，然后将集中的元数据成果主动推送给其他用户。用户通过推送的元数据信息，即可访问相对应的煤层气信息。

3) 应用层。应用层提供了面向应用领域的煤层气元数据集成环境，屏蔽了网络技术的复杂性，向用户提供了简洁便利的界面。用户通过 Web 页面或手机等便携设备即可实现对煤层气元数据的访

问、管理等操作，继而可实现对煤层气各类数据的高效访问。

1.2　系统的主要功能

系统功能部分包括元数据内容管理、结构管理和安全管理三大功能模块。元数据内容管理可细分为录入、编辑、查询、输出和发布 5 个模块。煤层气元数据管理信息系统的功能设计如图 2 所示。

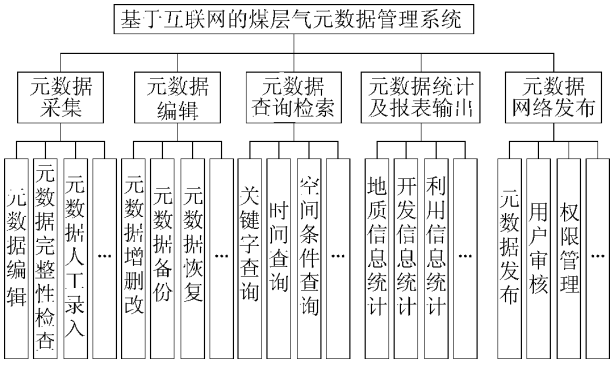

图 2　基于互联网的煤层气元数据管理系统功能设计

Fig. 2　Function design of Internet-based CBM metadata management system

1) 元数据自动抽取功能。从煤层气的图形数据、数据库数据和文档数据中自动抽取出元数据，形成元数据目录服务模块，对于无法自动生成的元数据，需要进行人工录入。

2) 元数据编辑功能。元数据编辑包括元数据记录的修改、添加、删除以及元数据备份等操作。

3) 查询检索功能。系统提供方便的菜单和按钮，使用户能通过手机等便携设备方便地浏览自己所关心的元数据信息。同时，系统也提供关键字查询、时间条件查询、空间条件查询等方式，帮助用户定位元数据信息，并能通过元数据定位和显示相应数据集。

4) 元数据统计及报表输出功能。系统为用户提供各种条件的统计和报表输出、打印功能，用户可以选择自己感兴趣的元数据信息，并按照自己要求的格式输出。

5) 元数据审核和网上发布功能。系统具备审核与网络发布功能，实现元数据以及相应的数据集网上发布，这是实现数据共享和交换的一项重要功能。

2　系统关键技术

2.1　煤层气信息元数据标准设计

煤层气信息复杂多样，格式各异，因此必须建立科学的煤层气信息元数据标准交换格式，方能实现煤层气信息的有效管理和充分共享。XML 是当前网络中常见的数据交换格式语言，目前大多数元数据标准都采用该语言作为文件格式[4]。笔者针对煤层气数据的 3 种类型，即文档数据、图形数据和数据库数据分别制定相应的元数据标准 XML，并对标准中的内容进行了初步规定。元数据以 XML 进行编码表示，后缀为 xml，以文件方式存储。还以 XML schema 定义了数据类型、命名空间等，记录 XML 的结构定义信息，可以根据 schema 对元数据的各个字段进行校验，使其符合元数据标准；针对不同专业可以指定不同的元数据模板。总的来说，煤层气元数据库的内容可以分为 3 类，如图 3 所示。

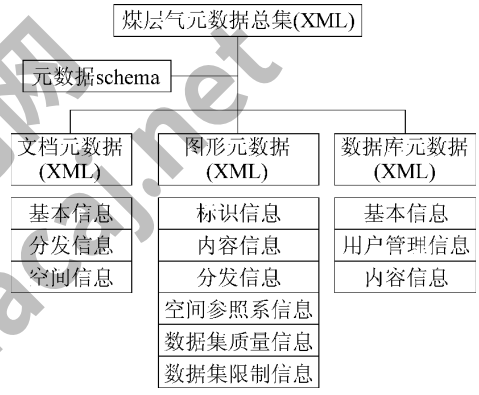

图 3　煤层气元数据库内容设计

Fig. 3　Design of CBM metadata database

1) 文档元数据库内容。①基本信息。包括基本类型、文档编号、文档名、编著者、形成单位、创建时间、修改历史、摘要、关键词、目录、附图、附件、参考文献、引用标准、审批人等；②分发信息。包括分发形式、内容格式、资料语种、密级、保护期、数据访问地址等；③空间信息。包括西边经度、东边经度、北边纬度、南边纬度、区域名称、区域代码、图幅号、图幅名等。

2) 图形元数据库内容。①标识信息。包括名称、内容描述、语种、空间分辨率等；②内容信息。包括要素类目说明、要素属性域说明等；③分发信息。包括分发介质、订购说明、分发联系方的详细信息等；④空间参照系信息。包括坐标系名称、坐标系类型、投影参数等；⑤数据集质量信息。包括验收说明、完整性说明等；⑥数据集限制信息。包括访问限制、安全限制分级等。

3) 数据库元数据库的内容。①基本信息。包括名称、内容描述、数据库类型、数据访问地址、数据

下载地址、发布时间、发布单位等; ②用户管理信息。包括用户口令登记表、用户工作分区和使用需求信息等; ③内容信息。包括数据表结构、数据内容、时间、地理位置和范围、数据质量、生产单位、获取方式等。

2.2　煤层气元数据自动抽取技术

上报的各类煤层气元数据可以通过 Web 页面进行提交,也可以通过条目的批量导入。提交或导入的条目在通过本地审核后都成为本地元数据库的条目,进而保存到元数据库中,供其他机构人员访问。无论是通过 Web 页面还是条目批量导入,都可以利用元数据批量自动抽取技术来生成相应的元数据。

1) 文档元数据自动抽取算法。煤层气文档元数据主要包括基本信息、分发信息和空间信息。从文档中可以抽取题名信息、作者信息、来源信息、关键词信息、摘要信息、参考文献信息、外部特征信息、最近保存时间等信息。文件格式主要是 Microsoft Office Word 格式和 Adobe Acrobat PDF 格式。现有的元数据自动抽取器主要有: 英国国家档案馆的 DROID 文件格式辨别工具、新西兰国家图书馆的 Metadata Extractor 软件和法国的 Metadata Miner Catalogue PRO 软件等[7]。这些抽取器大多针对英文设置,对中文无法做到有效提取。

针对上述问题,在该煤层气元数据管理系统中,对于 Word 格式文档,采用微软公司提供的 Word API[8] 提取 Word 文档元数据,使用的类包括 Microsoft.Office.Interop.Word 命名空间下的 Application 和 Document,其中 Application 类的实例对象表示运行的 Word 程序,通过 Application 类实例对象可以打开 Microsoft Office Word 格式文档,得到 Document 类的实例对象,获取 Document 类实例对象与有关文档元数据的属性予以保存即可。例如,提取文档的目录可以通过 doc.TablesOfContents 实现。标题、作者、关键词、摘要、参考文献作为文档中设置成为若干个顺序节点,类似 Web 文档的 DOM,约定每一个节点之间的相互线性关系。提取文档的这些节点信息,就可以实现相关信息的自动提取。

对于 PDF 格式文档,采用 PDFlib 提取文档中的元数据。PDFlib[9] 是 Adobe 提供的可移植文档格式(PDF) 文件的中间件,包含创建 PDF 输出(包括文本、矢量图形和图像以及超文本元素) 所必需的所有函数。PDFlib 为放置单行或多行文本、图像和创建表提供了强大的格式化功能。与 Word 提取方法类似,可以提取标题、作者、关键词等节点信息,并写入元数据库中。

2) 图形元数据自动抽取算法。煤层气相关图形数据来源多样,大多以 AutoCAD、MapGIS、ArcGIS 和 LongruanGIS 等进行数据制图和存储。煤层气各类设计图纸主要以 AutoCAD 的 dwg 格式存储,地理信息图形以 ArcGIS 的 shp 格式存储,瓦斯地质图等专题图以 LongruanGIS 的 lfmx 格式存储。针对不同文件生成的异构元数据文件,需要考虑不同的元数据抽取方法。

针对 ArcGIS 文件格式,可以直接采用.shp.xml[10] 文件。该文件是针对 shapefile 进行元数据浏览后生成的 xml 元数据文件,将该文件中提取的与标准元数据相关的字段,写入元数据库中。

针对 AutoCAD 文件格式,通过开放设计联盟 ODA(Open Design Alliance)[11] 平台开发进行元数据的抽取。ODA 开发用于技术图形应用程序的核心平台称为 Teigha。Teigha 支持 dwg、dgn、stl、pdf 之间的数据交换[12]。直接通过搭建平台可以访问 dwg 格式中的创建时间、文件头、图层、标题栏、图框内容等,然后写入元数据库中。

针对 LongruanGIS 格式[13],通过二次开发 COM 接口,读取文件头、图层、标题栏等信息搜索元数据。对于提供者信息,如提供者姓名、审核姓名等,可以通过 COM 接口提取瓦斯地质图的图框信息,自动抽取制图人、比例尺、审核人、单位、图名等信息,然后写入元数据库中。

针对 MapGIS[14] 数据格式,可以直接读取点、线、面文件,通过读取明码格式文件,获取相关元数据信息。

3) 数据库元数据自动抽取算法。数据库元数据是描述数据库的结构和建立方法的数据[15]。对于大型数据仓库,有描述元数据的数据库表,这些元数据表属于对数据仓库原始数据定义的语义层,可以帮助最终用户理解数据仓库中的数据[16]。而针对普通的数据库,除了数据库基本信息和外部管理信息等,主要是自动抽取数据库的表名和字段名等信息,并将其转化为普通用户可以识别的元数据语义。

按照数据库元数据的标准,需要抽取数据库内容信息元数据,其抽取步骤为: ①通过 ODBC、DAO、ADO、OLEDB 等数据库接口,读取每个数据库中的

数据表;②依次遍历每张表,读取每张数据表的列名,从而获得数据库的内容信息元数据;③所有信息读取完毕后,再根据规定的数据库元数据标准,将与标准相关的表名和字段名等写入元数据库中。④除此之外,将记录数据库的类型、数据访问地址、数据下载地址、发布时间、发布单位等写入元数据的基本信息中。

2.3　元数据收割技术

元数据收割技术是指在因特网上分布式检索不同的资源库,获取元数据在本地集中式建库的技术。开放的元数据协议一般有 HTML(直接将元数据字段显示在网页上)、OAI_ DC(封装为支持 OAI 收割协议 DC 元数据格式)、XML、Bib-Tex、Marc21(机读编目格式标准)、JSON、RDF、RIS(Endnote 等文献管理软件支持的元数据格式)、CSV、Excel 等[17]。笔者采用抽取普通 HTML 网页的策略,这部分元数据网页深度不一,网页深度有深有浅,可能达到 4 层或 4 层以上,为资源的自动采集带来了一定难度。

目前,在国家地质调查局等单位已经公开发布了一些关于煤层气的文档数据[18]。将这些网址指定为检索的资源库,爬取系统动态网页上的富文本元数据信息,深度为 2~3 级,通过解析文本获取标准元数据,如基本信息、分发信息、空间信息等,对这些元数据进行入库,从而实现海量多源、异构、多态数据的自动镜像和统一组织管理。

3　煤层气元数据管理系统实践应用成果

依据对煤层气元数据管理系统的设计和相应标准,开发了煤层气元数据管理信息系统。系统实现了元数据的录入、编辑、查看、检索和报表输出等功能,并在集成到全国重点煤矿区煤层气地质信息与开发利用决策支持平台后,能够为其提供数据支持。系统采集了晋城、松藻和黔西等煤矿的煤层气各类数据,建立了图形元数据库、文档元数据库和数据库元数据库,并实现数据的有效共享。当用户上传某煤矿瓦斯地质图时,系统根据前面的算法自动进行元数据的抽取。其中,图形元数据的自动生成如图 4 所示。系统自动抽取了其中的图名、创建日期、比例尺、制图人、制图单位、格式名称、格式版本等信息,人工填写了该瓦斯地质图的摘要信息、制图目的、语种、状况、更新频率、更新范围、关键词、关键词类型、数据表示类型等信息,从而依据图形元数据标准,完成瓦斯地质图元数据的标识信息部分。之后,根据自动抽取的其他信息和人工输入的信息,完成该瓦斯地质图元数据的数据集限制信息、数据集质量信息、空间参照系信息、分发信息、内容信息和元数据联系单位等信息,这样就制作完成了整幅瓦斯地质图的元数据。

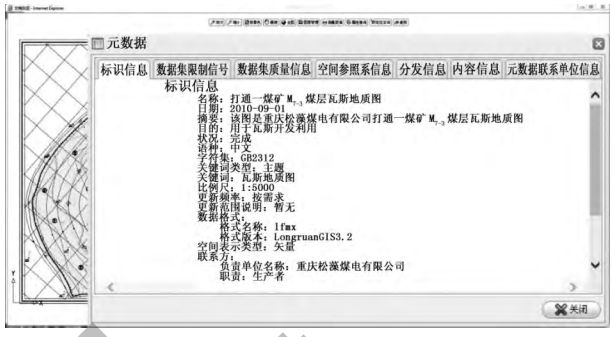

图 4　某煤矿瓦斯地质图元数据的自动生成

Fig. 4　Automatic generation of gas geological map metadata in a coal mine

数据库元数据和文档元数据均依照此方式进行制作,其中,文档元数据的自动生成如图 5 所示,生成了基本信息、分发信息和空间信息;数据库元数据的自动生成如图 6 所示,生成了基本信息、内容信息和用户管理信息。所有元数据制作完成后,自动存

图 5　某煤层气数据处理平台报告文档元数据的自动生成

Fig. 5　Automatic generation of metadata of a CBM data processing platform report document

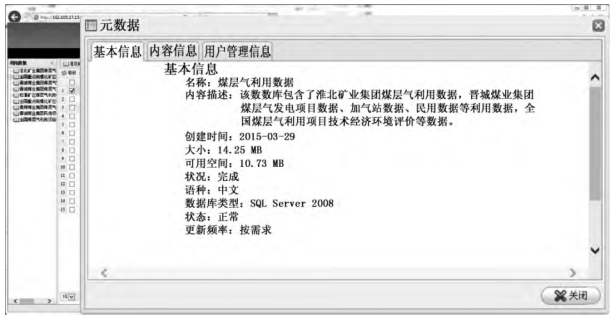

图 6　煤层气利用数据库元数据的自动生成

Fig. 6　Automatic generation of metadata for gas utilization database

入对应的元数据库中，元数据库建立完成后，用户通过此元数据管理系统，即可对煤层气图形数据、数据库数据和文档数据的元数据进行浏览、查询等操作，进而能够对整个煤层气地质信息、开发和利用信息进行有效的管理和充分共享。

4　结　　语

在煤层气项目开发过程中，国家、科研单位和相关企业形成了一批地质、开发与利用的图形、报告、数据库等，这些数据分散在各个地方，是国家宝贵的财富和社会重要资源，但是实现数据共享还很困难。互联网技术的发展，为解决这一难题提供了机遇。笔者从应用需求出发，结合煤层气信息元数据自动获取技术、元数据互联网收割技术等，在建立煤层气元数据标准的前提下，开发了煤层气元数据管理信息系统。该系统贯穿了煤层气地质、开发和利用的全生命周期，能够自动抽取煤层气文档、图形和数据库的元数据，实现元数据库和信息的自动更新，进而实现煤层气信息的及时、有效共享。本系统的研发为煤层气领域的数据共享做出了有益的探索。在此研究基础上，可以建立中石油、中海油、煤矿企业等多个煤层气开发主体的数据共享机制，实现面向国家、科研机构和企业的煤层气信息服务。

参考文献(References) :

[1] 樊振丽，申宝宏，胡炳南，等.中国煤矿区煤层气开发及其技术途径 [J] .煤炭科学技术，2014，42(1) : 44-49，75.
Fan Zhenli，Shen Baohong，Hu Bingnan，*et al*.Coalbed methane development and technical access in China coal mining area [J] .Coal Science and Technology，2014，42(1) : 44-49，75.

[2] 张子敏，吴　吟.中国煤矿瓦斯赋存构造逐级控制规律与分区划分 [J] .地学前缘，2013(2) : 237-245.
Zhang Zimin，Wu Yin.Tectonic-level-control rule and area-dividing of coalmine gas occurrence in China [J] .Earth Science Frontiers，2013(2) : 237-245.

[3] 毛善君.“高科技煤矿”信息化建设的战略思考及关键技术 [J] .煤炭学报，2014，39(8) : 1572-1583.
Mao Shanjun.Strategic thinking and key technology of information construction of high-tech coal mine [J] .Journal of China Coal Society，2014，39(8) : 1572-1583.

[4] 佚　名.地理空间云平台首页 [EB/OL] . [2016-02-10] .http: //www.gscloud.cn/.

[5] 颜世强，郭艳军，王　喆，等.基于 DSpace 的地质资料数字仓储系统建设与应用 [J] .地质通报，2013，32(7) : 1134-1140.
Yan Shiqiang，Guo Yanjun，Wang Zhe，*et al*.The construction and application of Institutional Repository System based on DSpace for digital geological data [J] .Geological Bullentin of China，2013，32(7) : 1134-1140.

[6] 黎建辉，吴　威，阎保平.一种基于 XML 的元数据映射与转换方法 [J] .微电子学与计算机，2008，25(1) : 34-38.
Li Jianhui，Wu Wei，Yan Baoping.An approach for metadata model mapping and metadata instances transfer based on XML [J] .Microelectronics and Computer，2008，25(1) : 34-38.

[7] 曾　苏，马建霞，张秀秀.元数据自动抽取研究新进展 [J] .现代图书情报技术，2008(4) : 7-11.
Zeng Su，Ma Jianxia，Zhang Xiuxiu.New development of automatic metadata extraction [J] .New Technology of Library and Information Service，2008 (4) : 7-11.

[8] 佚　名.Microsoft API 和参考目录 [EB/OL] . [2016-02-15] .https: //msdn.microsoft.com/zh-cn/library/ms123401.aspx.

[9] Lundin A，Bok C M，Aronsson L，*et al*.PDFlib plop: pdf linearization，optimization，protection [J] . Cellular Microbiology，2008，10 (5) : 1093-1103.

[10] ESRI. ESRI shapefile technical description an esri white paper [EB/OL] . [2016-02-17] .http: //www.esri.com/library/white-papers/pdfs/shapefile.pdf.html.

[11] Anon.Open design alliance home [EB/OL] . [2016-02-17] .https: //www.opendesign.com/.

[12] Anon. Teigha [EB/OL] . [2016-02-17] .https: //www.opendesign.com/the_oda_platform/Teigha.

[13] 董　平，毛善君.煤矿地测空间数据转换到 MapGIS 文件的技术方法研究 [J] .中国煤炭，2005，31(4) : 43-46.
Dong Ping，Mao Shanjun. Study on the technical method of the spatial data in coal mine converting to MapGIS file [J] . China Coal，2005，31(4) : 43-46.

[14] 钟世彬，郑贵洲.AutoCAD 和 MAPGIS 间的数据转换 [J] .测绘科学，2005，30(3) : 97-98.
Zhong Shibin，Zheng Guizhou. Data conversion between MAPGIS and AutoCAD [J] . Science of Surveying and Mapping，2005，30 (3) : 97-98.

[15] 徐　彬.网格环境下数据库系统的元数据服务 [D] .武汉: 华中科技大学，2004.

[16] 徐立臻，刘　安，董逸生.数据仓库系统中的元数据管理 [J] .计算机工程与应用，2002，38(24) : 193-196.
Xu Lizhen，Liu An，Dong Yisheng. Metadata management for data warehouse [J] . Computer Engineering and Application，2002，38 (24) : 193-196.

[17] 王思丽，马建玲，王　楠，等.开放知识资源的元数据自动采集策略研究 [J] .图书馆学研究，2013(12) : 47-51.
Wang Sili，Ma Jianling，Wang Nan，*et al*. Research on metadata automatic acquisition strategy of open knowledge resource [J] .Research on Library Science，2013(12) : 47-51.

[18] 佚　名.国家地质资料数据中心 [EB/OL] . [2016-02-20] .http: //www.ngac.org.cn/.