基于 Dspace 构建高校科学数据管理平台*

——以蝎物种与毒素数据库为例

洪正国 项英

[摘 要] 以武汉大学蝎物种与毒素数据库为例,介绍如何利用 Dspace 构建“蝎物种与毒素数据管理平台”。首先对系统的平台选择、系统目标、数据类型进行分析,继而描述利用 Dspace 构建“蝎物种与毒素数据管理平台”的本地化实施与二次开发过程,最后提出利用 Dspace 构建科学数据管理平台的一些思考和建议。

[关键词] 科学数据管理 开源软件 Dspace 数字资源管理 高校
[分类号] G250

1 引 言

随着科学技术和实验设备的发展进步,科研和教育环境正逐渐走向数据密集型,可能需要花费更多精力来组织、管理、保存和共享数据,传统的数据管理方式已不能适应科学研究的需要,迫切需要功能完备的管理平台进行数据管理,提高数据资源的利用价值。

武汉大学生命与科学学院“中国蝎类及其毒素基因资源的调查与鉴定”项目(简称“蝎资源项目”)为 2008 年国家科技部重点平台项目,主要进行中国地域内的蝎子物种发现、蝎子毒素蛋白、毒素核酸的测序工作。该课题研究过程中会产生比较典型的科学数据,包括蛋白质序列以及核苷酸序列数据、蝎物种特征多样性及其图片数据、相关研究论文与报告。其中序列数据必须符合 National Center for Biotechnology Information(NCBI)基因数据库的规范,为了科学管理这些数据需要开发一套数据发布与管理系统。根据需求,笔者采用 Dspace 构建了“蝎物种与毒素数据管理平台”。

2 系统需求与分析

2.1 软件平台分析与选择

通过调研发现,国内外目前主要的科学数据管理平台构建方式有以下三种:

2.1.1 专业数据管理平台 主要是一些大型数据机构或科研机构为某个领域研究开发的系统,如气象、地球物理、生物领域等。这些系统往往面向特定学科的数据处理需求,不能应用于其他学科,如 NuGenesis 科学数据管理系统[1](主要用于医院、生物技术方面)、Nesstar 系统[2](主要用于社会调查领域)等。

2.1.2 自开发系统 主要是利用 asp、java、php 等语言开发的适合本单位应用的数据管理系统。如中国社会调查开放数据库[3](Chinese Social Survey Open Database,CSSOD)就是以 Linux + Apache + MySQL 构建的社会调查数据管理平台。

2.1.3 利用开源的数字资源管理软件构建的平台 主要是利用开源软件构建的数据管理平台,这在国内外的高校中应用比较普遍,如美国康奈尔大学图书馆的 DataStaR 项目[4]和美国约翰霍普金斯大学的 Data Conservancy 项目[5]均采用 Fedora 构建,香港科技大学机构仓储库项目[6]采用 Dspace 构建。

经广泛调研和比较分析,Dspace 系统进入了我们的视野。Dspace 最初由美国麻省理工学院(MIT)和美国惠普公司合作开发,是以内容管理发布为目标的数字资源存储系统,可实现对各种格式数字资源的收集、存储、索引和发布。Dspace 具有完善的用户界面,可定制性强,易于实施;同时也具有较好的扩展性,提供了二次开发的可能。目前 Dspace 已广泛地应用于全球各地的数字资源系统,拥有众多用户和成功案例。

在综合比较了各种开源软件在系统结构、用户界面、二次开发等方面的特点,后鉴于 Dspace 具有用户界面成熟、用户多、易于二次开发等优势,故笔者决定利用 Dspace 来构建蝎物种与毒素数据管理平台。

* 本文系 CALIS 三期预研项目“高校科学数据管理机制及管理平台研究”(项目编号:03 - 3043)研究成果之一。
[作者简介] 洪正国,武汉大学图书馆馆员,硕士,E-mail:zghong@lib.whu.edu.cn;项英,武汉大学图书馆馆员,硕士。
收稿日期:2013 - 01 - 08 修回日期:2013 - 03 - 10 本文起止页码:39 - 42,84 本文责任编辑:高丹

2.2　系统目标与功能

本系统开发的目标是确保蝎资源项目组成员可以随时向数据库添加相关数据及相关文献，并发布和管理蝎子毒素领域的最新研究成果；通过权限控制，使中国蝎子毒素研究人员获取相关信息，实现数据共享。

蝎物种与毒素数据管理平台下建有 4 个科学数据库，分别是：蝎物种资源数据库、蝎遗传基因核酸数据库、蝎遗传蛋白数据库、蝎资源文献数据库。系统的主要用户为蝎资源项目组成员及业界同行。系统界面力求简洁并符合生命科学研究者的阅读习惯；页面语言以英文为主，对适用于科学普及的蝎物种数据库，则设置中英文两种语言界面；数据库需提供检索和浏览功能，且能嵌入序列数据比较工具（BLAST），实现序列对比功能。

2.3　数据分析

2.3.1　*数据类型分析*　根据需求，平台需要管理的数据包括文献数据（蝎资源相关文献及项目研究成果）、物种数据（蝎物种的图片采集资料）和序列数据（蝎蛋白及蝎核酸序列测定数据）三种。文献数据，主要指与课题相关的文献资源，如结题报告、论文、专著等。蝎物种数据，主要是描述于蝎物种特征及其图片。该数据由两部分构成：①物种图片，格式包括 bmp、jpg、gif、png、tif 等；②与该物种有关的元数据信息。序列数据包括核酸和蛋白质两类，以核苷酸碱基顺序或氨基酸残基顺序为基本内容，并附有注释信息。在蝎资源项目中主要是蝎物种遗传基因和蝎物种遗传蛋白测序数据。蝎物种遗传基因资源数据形式如图 1 所示：

```
>BmP05
atgaagttcc tctacggaat cgttttcatt gcacttttc taactgtaat gttcggtaag
tgattgccaa tatttatgtt aaagaattta aaatcaataa tatgaaatta atttttattt
cgtaataaca tattattttc tttctgtagc aactcaaact gatggatgtg ggccttgctt
tacaacggat gctaatatgg caaggaaatg tagggaatgt tgcggaggta ttggaaaatg
ctttggccca caatgtctgt gtaaccgtat atga
```

图 1　蝎物种遗传基因资源的数据形式

2.3.2　*数据关系分析*　数据关系指的是数据间发生引用、包含、被包含、映射等关系，它是提供数据关联检索、构建知识图谱的基础。在科学数据管理中，揭示数据间的关系是其重要功能之一。在本项目中，三种数据类型事实上包含 4 种数据，即文献数据、物种数据、基因测序数据、蛋白测序数据。其中后三种数据存在着包含、相互映射的关系，即物种数据中可能涉及多个基因测序数据，基因测序数据也可能包含多个蛋白测序数据；文献数据则通过引用的方式与这三种数据产生联系。数据间的关系如图 2 所示：

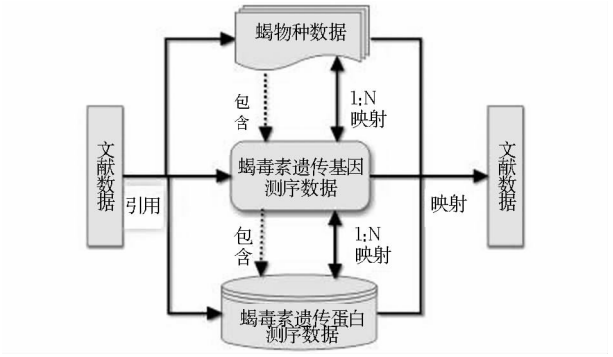

图 2　蝎资源项目数据关系

2.4　系统功能设计与分析

系统主要功能是为了实现蝎资源项目组成员日常实验观察中 4 种类型数据的管理，包括数据著录与提交、数据检索、检索结果显示、权限控制、用户界面等。

通过分析，利用 Dspace 构建数据管理平台主要解决以下几方面问题：页面汉化与设计、非 DC 类型的元数据结构处理、检索功能的实现（简单/高级/跨库）、检索结果的显示、不同数据库间数据的参照、序列数据的比较（Basic Local Alignment Search Tool，BLAST）、用户管理（与图书馆用户集成/权限控制）。这些功能有些可以通过 Dspace 实现，有些需通过系统参数配置实现，有些则需要通过二次开发实现。本系统主要功能、Dspace 对应的功能模块以及对应的解决方式如表 1 所示：

表 1　系统主要功能模块及实现方式

功能模块	Dspace 对应的功能模块	实现方式
元数据管理	元数据注册与管理	自有功能实施
数据提交	数据提交	参数设置/二次开发
检索	检索	参数设置/二次开发
检索结果显示	检索结果显示	参数设置/二次开发
权限控制	权限管理	自有功能实施
页面元素与布局	页面布局	二次开发
用户管理	用户管理	二次开发
整合数据分析工具（BLAST）功能	无	二次开发

2.4.1　*元数据管理*　Dspace 系统缺省采用的是 DC 元数据，而蝎资源库包含有多种非 DC 类型的元数据类型，因此需要对 Dspace 系统的元数据注册与管理模块进行配置和改进。

2.4.2　*数据提交*　包括元数据著录和对象数据提交，根据元数据类型和数据类型的不同，字段内容、提交的步骤等方面都会有些不同，因此系统应具有不同的提交界面。文献资源数据库提交的主要是文本类型的数据，可以利用 Dspace 的数据提交功能实现；物种和序

列数据由于和 DC 差别较大,同时输入方式需要符合蝎资源项目成员的习惯,故需要对 Dspace 进行二次开发,并重写 SubmitStep、SubmitDescript 等相关类来实现。

2.4.3 检索与检索结果显示 提供简单检索和高级检索,检索字段可以根据不同的数据类型定义,同时提供跨多个数据库的检索。检索结果概览和细览显示根据数据类型的不同而显示不同的字段。实现这些功能需要对 Dspace 检索与显示模块进行二次开发。

2.4.4 用户管理与权限控制 用户包括通过邮件注册的用户和使用图书馆集成系统的用户。系统可以对用户访问数据进行多级控制——可以控制到数据库一级,也可以控制到记录一级,还可以控制到对象数据一级。用户管理功能的实现需要对 Dspace 的用户管理模块进行二次开发;权限控制功能利用 Dspace 的权限管理功能基本可以实现。

2.4.5 页面元素与布局 主要涉及页面布局、美工等用户界面。可以通过修改 Dspace 的模板 jsp 文件实现。

2.4.6 序列数据分析工具 BLAST 整合 在序列数据的使用中,经常会用到两个序列数据进行比较,这就是 NCBI 提供的 BLAST 工具。BLAST 工具源代码是公开的,可以通过下载 NCBI 上的 BLAST 代码并进行部分修改,实现本系统的序列数据比较功能。

3 系统实现

Dspace 中二次开发主要有两种:①修改模板文件,这些主要是 jsp 文件,修改后立即生效;②修改 java 源代码文件,这种文件的修改必须经过编译发布才能生效。

Dspace 系统采用多层构架,分为表示层、业务层和存储层,下层提供接口供上层调用。考虑到系统的完整性及升级的方便,源代码修改主要在表示层完成。

Dspace 提供了两种界面:jspui 界面和 manakin 界面。通过分析,笔者采用了更成熟、二次开发更简单的 jspui 界面。

3.1 本地化配置

结合平台的需求,部分功能可以通过参数配置完成,其中涉及的主要参数配置方法如下:

3.1.1 中文语种支持 通过增加 messages_zh_CN. properties 文件,实现按钮和标签中文化。

先编写 messages. cn:

metadata. dc. contributor. author = 作者

然后利用 java 的 native2ascii 将上述文件转换成 unicode 格式:

native2ascii -encoding GBK messages. cn messages_zh_CN. properties

metadata. dc. contributor. author = \u4f5c\u8005

3.1.2 数据提交界面和流程 通过修改 input - forms. xml 和 item - submission. xml 可定制元数据项的输入项和加工流程。其中 item - submission. xml 定义了一个数据集(collection)的数据提交流程;input - forms. xml 定义了一个数据集在每个提交阶段中的元数据输入界面,包括提示、是否必备。

3.1.3 中文检索支持 为了支持中文检索,需要修改 Dspace. cfg 文件中的 lucene 配置部分,指定中文分词程序:

search. analyzer = org. apache. lucene. analysis. cn. ChineseAnalyzer(标准分词系统)或

search. analyzer = net. paoding. analysis. analyzer. PaodingAnalyzer(庖丁解牛分词系统)

在本系统中笔者采用了免费的庖丁解牛分词系统[7]。

3.1.4 元数据管理 Dspace 缺省采用的是 DC 元数据,本项目除了文献资源可以用 DC 描述外,其他三种元数据都无法用 DC 来描述,为了使得系统能处理其他类型的元数据,必须在系统中进行元数据注册。可通过 Dspace 的元数据管理模块完成:先注册元数据名称和 URI 地址,然后逐项添加物种数据、基因数据和蛋白数据等各元数据项及其说明。

3.1.5 Dspace. cfg 文件中其他主要的配置信息

- 索引字段配置:比如核酸库的 Definition 字段需要参与检索,可以在索引配置项增加以下定义:

search. index. 14 = Definition:nucl. Definition

- 检索结果概览字段:以下定义实现了 title、LOCUS、Source、Lineage、contributor 字段在检索概览页面的显示。

webui. itemlist. columns = dc. title, nucl. LOCUS, nucl. Source, ss. Lineage, dc. contributor. *

- 检索结果细览字段:以下设置定义某个库检索细览结果显示的字段。

webui. itemdisplay. nucl = dc. title, nucl. LOCUS, nucl. Definition, nucl. Accesstion, \

nucl. Version, nucl. Keywords, nucl. Source

webui. itemdisplay. nucl. collections = 123456789/2, 123456789/7

3.2　二次开发

3.2.1　界面元素　由于 Dspace 的界面比较简单，为了符合蝎资源科学数据的风格，需要对 Dspace 的图片、按钮、页面布局等做修改。页面元素的修改主要是通过修改 Dspace 系统中 layout 目录下相关的 jsp 文件来完成。

3.2.2　数据提交　在蝎资源科学数据管理系统中，有些数据具有特定的意义。比如描述序列数据的 LOCUS 字段的内容为 AF242736 391 bp mRNA linear INV 02 - APR - 2001，其内容与其他字段是相关的：AF242736 是 Accesstion 字段的内容；391 是这个序列的长度；mRNA 是序列的类型。为了保证数据的准确性，在数据著录时必须考虑各字段间的关系，自动生成某些字段以及数据的校验。为了实现自动生成某些字段的功能，笔者在 input-forms. xml 中扩展了 input - type 的类型，增加了一种自动从某些字段中提取内容的输入方式，保证了数据的一致性和完整性。

数据提交涉及到的二次开发主要是通过修改或重写 SubmissionController、JSPStep、org. dspace. app. webui. submit. step 包中的大部分类来实现。

3.2.3　数据检索及检索结果显示　Dspace 系统主要是为数字化文献资源进行设计的，其检索界面和结果的显示更符合文献检索系统。比如检索字段主要是题名、著者、出版日期等，结果是以表格形式显示的。由于蝎资源数据涉及到多种类型的数据，每种类型的数据检索字段各不相同，显示结果也各不相同。

为了实现不同数据类型提供不同的检索字段，需要在 java 的 session 对象中记录当前用户所选择的数据库，来调用不同的 jsp 包含文件。

同时，Dspace 原系统中检索结果概览页面是以表格方式显示的，这样显示字段的数量必定会受到限制，而且没有内容的元数据项也会占据表格空间，故这种形式不利于检索结果概览页面信息的展示。为了实现更多信息的显示和页面的美观，笔者采用了每个元数据项一行的方式；如果该元数据项为空就不显示，以避免出现空行。这样的设计可以方便灵活地显示各种复杂的数据，从而避免了 Dspace 原系统表格方式不能显示太多元数据项的弊端。

数据检索主要涉及 DSQuery、QueryResults；检索结果显示主要涉及 ItemListTag 和 ItemTag 两个类。其二次开发主要是通过修改或重写以上 4 个 java 类来完成。

3.2.4　用户管理　Dspace 系统使用电子邮件账号作为用户标识，用户可以自己注册成为系统的用户。为了方便学校用户的使用，笔者考虑通过图书馆集成系统账号认证的方式，这样用户不需要注册而直接使用图书馆的账号即可以使用本系统；同时结合武汉大学图书馆的统一认证系统实现统一认证和单点登录。这需要修改 Dspace 源代码中的 MyDspaceServlet 和 Authenticate 相关 java 类来实现。

3.3　第三方数据分析工具(BLAST)集成

BLAST 是一种基于成对局部序列对比的数据库相似性搜索工具。在 NCBI 的网站上提供了 BLAST 源代码下载，通过下载源代码，并参照相关说明修改 wwwblast. cpp 文件，最后编译成 cgi 程序；同时将本系统中的序列数据转化成符合 NCBI blast 规范的数据作为基础数据，在系统中发布以供其他系统调用进行相似性搜索和比较。

为了实现在核酸和蛋白质数据检索细览页面提供 BLAST 工具与本系统中或 NCBI 相关数据库中的序列进行相似性搜索，在系统的检索结果细览页面中，通过提取系列的位置、序列值等数据调用 blast cgi 相关程序，达到核酸或蛋白质序列数据的相似性检索的功能。

4　思考和建议

通过研究和实践，笔者认为 Dspace 设计合理、功能强大，是科研机构和高校在数字资源管理方面的首选系统。虽然科学数据也是数字资源的组成部分，但毕竟 Dspace 系统主要是为了管理电子文献资源而设计的，存在对非 DC 类型数据处理不便的问题，因此将其应用于科学数据的管理，须事先确定其是否能够满足所管理数据对象的需求。

Dsapce 处理大量数据的能力和效率还有待提高，利用 Dspace 构建的数据管理系统比较适合于数据量不大、学科多的高校。如果对特定学科的大量数据进行管理，则可以考虑更专业的科学数据管理平台。

利用 Dspace 构建的系统能够满足科学数据的提交、发布、存储和检索等一般性需求。但如果涉及数据本身的计算、分析等更高要求，Dspace 处理起来就相对比较困难。因此在构建科学数据管理平台时，须先行确定平台的目标和功能——有的放矢方能选择合适的软件系统。

针对 Dspace 系统的扩展和二次开发，建议尽量采用 Dspace 开发文档上推荐的方法，保证系统的三层结构，尽量针对表现层进行修改并调用 Dspace 核心层的 API 实现，以保证版本的升级和系统的兼容。

（下转第 84 页）

[18] 国家“十一五”时期文化发展规划纲要[EB/OL].[2012-12-14]. http://news.xinhuanet.com/politics/2006-09/13/content_5087533.htm.
[19] 国家“十二五”时期文化发展规划纲要[EB/OL].[2012-12-14]. http://www.ce.cn/culture/gd/201202/16/t20120216_23076264.shtml.
[20] 中国图书馆学会. 图书馆服务宣言[J]. 中国图书馆学报,2008,34(6):5.
[21] 联合国教科文组织. 公共图书馆宣言[EB/OL].[2012-12-14]. http://www.ifla.org/VII/s8/unesco/chine.pdf.

Status Investigation and Thought of Chinese Public Library Services for Vulnerable Groups Based on Empirical Survey

Wang Ping　Wang Zhenmeng　Huang Shang
Information Management Department, Zhengzhou Universdy, Zhengzhou,450001

[Abstract] Services for vulnerable groups have been universally carried out in Chinese public libraries. Based on the content analysis of the public library website, this paper finds that there is no clear definition about the object types, the essence and the necessity of vulnerable groups' services; the youth and children, the aged and the disabled are the main service objectives, but the current services stay the surface level; the services of vulnerable groups are positively related with the development of economics in certain degree, but it can not exclude the different situation of services in the same and different conditions. This paper points out that Chinese public library vulnerable group services need clearly definite the basic motivations, the internal essence, the security mechanism, the influencing reasons and the evaluation criterion.

[Keywords] public library　vulnerable group　status　empirical survey

(上接第 42 页)

利用 Dspace 构建的“蝎物种与毒素数据管理系统”基本达到了项目组数据管理的要求,在页面设计、系统功能、数据检索与显示等方面得到了生命科学学院蝎资源项目专家和老师的肯定,并希望该系统能成为“蝎资源 NCBI”(如试用请访问 http://sdm.lib.whu.edu.cn/jspui/handle/123456789/1?locale=en)。

当然本系统还存在需要完善之处,如大数据流处理问题,尤其是对大文件(GB 以上)的处理和大容量数据(亿级)的处理以及系统内数据的分析和挖掘问题。这些都是科学数据管理中非常重要的内容,有待于后续研究来使之逐渐完善。

参考文献:

[1] 沃特世. NuGenesis[EB/OL].[2012-07-12]. http://nugenesis-sdms.waters.com/
[2] 蒋颖. 欧洲社会科学数据的服务与共享[J]. 国外社会科学,2008(5):84-89.
[3] 中国社会调查开放数据库[EB/OL].[2012-08-10]. http://www.cssod.org/index.php.
[4] DataStaR[EB/OL].[2012-08-10]. http://datastar.mannlib.cornell.edu/.
[5] Data Conservancy [EB/OL]. [2012-08-10]. http://dataconservancy.org/.
[6] HKUST IR[EB/OL].[2012-08-10]. http://repository.ust.hk/dspace/.
[7] 庖丁解牛分词系统[EB/OL].[2012-08-10]. http://baike.baidu.com/view/7324777.htm.

Construction of University Scientific Data Management Platform Based on Dspace ——A Case Study of Scorpion Species and Toxins Database

Hong Zhengguo　Xiang Ying
Wuhan University Library, Wuhan 430072

[Abstract] This paper takes Scorpion Species and Toxins Database of Wuhan University library for example, and introduces the construction of scorpion species and toxins data management platform based on Dspace. It analyzes the platform selection, system objects and data types, and describes its localization implementation and secondary development. Finally, it provides some thoughts and suggestions on constructing Dspace-based scientific data management platform.

[Keywords] scientific data management　open source software　Dspace　digital resource management　university