机构知识库与科研管理信息化环境集成的尝试*

马建霞 祝忠明 唐润寰 李富强 王渊命
（中国科学院国家科学图书馆兰州分馆 兰州 730000）

[摘要]本文介绍了我们基于 Dspace 建设的机构知识库与中国科学院资源规划(ARP)集成的尝试。包括一个 ARP 中不同类型科研产出数据导出工具的开发，一个基于 B/S 架构的 EXCEL 数据导入 DSpace 工具开发，并介绍了如何针对不同类型的科研产出数据定制 Dspace 元数据输入和显示界面。

[关键词]机构知识库 ARP 集成 DSpace 科研信息环境 元数据定制

[分类号] **G250**

A Try to Integrate Institutional Repository With Information Environment of Management of Scientific Research

Ma Jianxia, Zhu Zhongming, Tang Runhuan, Li Fuqiang Wang Yuanming
（Lanzhou branch of National Scientific Library,CAS 730000）

[abstract]In this paper we introduce how we integrated the information of ARP in Chinese Academy of Sciences with institutional repository based on Dspace. It includes an exporter tool which export different types of data such as journal paper, conference paper, books, thesis and so on, from ARP into EXCEL, a web tool based on Browser/Server which import EXCEL file into Dspace, and introduce how we customize the input form and display form according to different type of documentation.

[key words]Institutional repository, Dspace, Scientific information environment,ARP

1.背景

机构知识库（Institutional Repository 简称 IR）是研究所科研知识产出长期保存的有效手段。中国科学院国家科学图书馆计划实施面向科研院所的机构知识仓储建设，兰州分馆信息技术部承担了技术开发和支持任务，并选定 DSpace 为软件平台，以中国科学院力学研究所为首批建设实施单位，在实施中，力学所提出为了减轻科研人员提交数据的负担，要将 ARP 系统中的数据集成到 IR 中，同时由力学所图书馆人员审核编辑元数据，并提交论文全

*基金项目:本论文受到国家社科基金项目 机构知识库的研究与建设（项目编号 07BTQ019） 及中国科学院国家科学图书馆全院联合机构知识仓储体系建设项目支持。

文。

根据力学所的需求，我们分析了 ARP 系统中的数据。中国科学院资源规划(ARP)项目[1]（以下简称 ARP），就是院所的资源计划管理，实际上是中国科学院科研管理信息化的平台，ARP 应用系统包括科研计划与执行管理系统、人力资源管理系统、综合财务管理和监督系统、科研条件管理系统、基本建设管理系统、电子政务系统、教育资源管理系统、评估评价系统。其数据库管理系统为 ORACAL9i，操作系统为 linux，为了安全起见，研究所人员访问 ARP 系统数据往往要通过 VPN。其中对于科研人员的知识产出，包含十种类型的数据：专著、会议报告、期刊论文、标准、成果鉴定、奖项、软件、药物、专利、咨询报告等，每种类型的数据有共有字段：题名、责任者等，但更多的是每种类型数据信息特有的字段。

ARP 设计之初，由于主要面向科研院所的管理信息化的需求，在其收录的科研人员知识产出的相关数据库表中，包括了部分元数据字段，但是没有全文信息。这将不利于对科研人员知识产出的有效管理和长期保存。

而我们目前实施的机构知识库旨在集中揭示和统一管理机构内分散的知识产出，实现对机构知识产出的长期保存，促进机构知识产出的共享利用，并通过机构知识库的服务，扩大和提升机构及科学家个人的学术影响和声望。

将 DSpace 和 ARP 相结合,获得 ARP 中已有的知识产出元数据，并定期获得更新的知识产出元数据，在 DSpace 中保存便于长期保存的知识产出的元数据和全文等数字对象，将达到有效利用已有系统中的信息，并集成到旨在长期保存和开放获取的机构知识库中的目的。

经过分析我们认为有 4 个问题需要解决：

1. 定期从 ARP 中导出数据;
2. 将 ARP 导出的数据导入到 Dspace;
3. 根据不同类型的文献设定录入界面;
4. 根据不同类型的文献定制显示界面。

2.将 IR 与科研管理信息化环境集成的尝试

2.1 开发数据导出工具

为了兼顾 ARP 系统本身的安全性和 Dspace 系统数据维护人员工作的便利，我们采用 Microsoft Windows.net，access 开发了运行于 Windows 环境的专门的数据导出工具。

该工具的运行需要安装.net framework, Oracle9 客户端，VPN 客户端。将 ARP 中的文献数据和人员信息导出到 Access 临时库，并进行数据查重，保证导出的数据是最新进入 ARP 系统的数据，为保证在不同研究所导出导入数据的通用性，将数据转换成 excel 格式。

2.2 修改 Metadata-Schema

由于 ARP 中导出的数据包括十种不同的文献类型，为了能够清全面的反映文献信息，我们在 Dublin-core 的基础上进行了字段扩展。进入 Dspace 管理员界面，再进入元数据模式注册页面，可以看到命名空间，在本系统中我们仍然采用 Dublin Core 的命名空间，并在此基础上进行扩展，点击 http://dublincore.org/documents/dcmi-terms/，在这里可以新增或者修改元数据字段。比如新增一个字段索书号,我们可以新增元素 identifier, 其修饰词为 callnum，并添加使用范围注释为 callnum of the thesis，这样我们就新增了一个

字段。

在此界面新增后，可以看到在源码的/config/registries/Dublin-core-types.xml 文件中，新增了一段：

```
<dc-type>
    <schema>dc</schema>
    <element>identifier</element>
    <qualifier>callnum</qualifier>
    <scope_note>callnum of the thesis</scope_note>
</dc-type>
```

2.3 将 ARP 数据导入到 DSpace

由于 Dspace 自带的数据导入导出工具[2]必须使用命令行来执行，而且需要数据完全满足 Dspace 要求的格式：即每个条目实际是一个文件夹，文件夹中包括 content、dublin_core.xml、handle，license.txt 以及全文（或图片文件）等文件。而根据需求，我们认为利用浏览器页面，直接调用导入工具，并导入 excel 格式的文件更为方便数据维护人员使用。

（1）Dspace 数据导入工具设计思路

参照 DSpace 自带的 itemexporter 工具导出数据的功能，每一个条目的元数据对应一个 xml 格式的文件，参照这个 xml 配置文件，可以实现 excel 文件到 Dspace 中条目元数据的对应。

导入数据时,将配置文件作为 dublin_core 的模版,然后将 excel 每一行的每个值填充到模版里,构成一个完整的 xml 文档,最后将该文档转换为 DSpace item ,插入数据库。

（2）类说明

net.teamhot.dspace.tools.ConfigReader 配置文件读取类,用 dom4j 作为 xml 解析工具.[3]

net.teamhot.dspace.tools.ExcelReader Excel 文件读取类,用 apache 的开源库 poi 作为[4]excel 文件解析工具.

net.teamhot.dspace.tools.ExcelRow 对 Excel 文件的一行数据的包装.

net.teamhot.dspace.tools.Importer 数据导入主类.

net.teamhot.dspace.tools.ImportServlet 数据导入程序的 http 请求响应类.

（3）配置文件说明

导入数据的配置文件是一个 xml 文件。其结构如下：

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<config>
 <dublin_core schema="dc">
    <dcvalue element="contributor" qualifier="author">
        <col checktype="emptyIgnoreValue" isMulti="true" >C</col>
    </dcvalue>
```

… ….
</dublin_core>
</config>

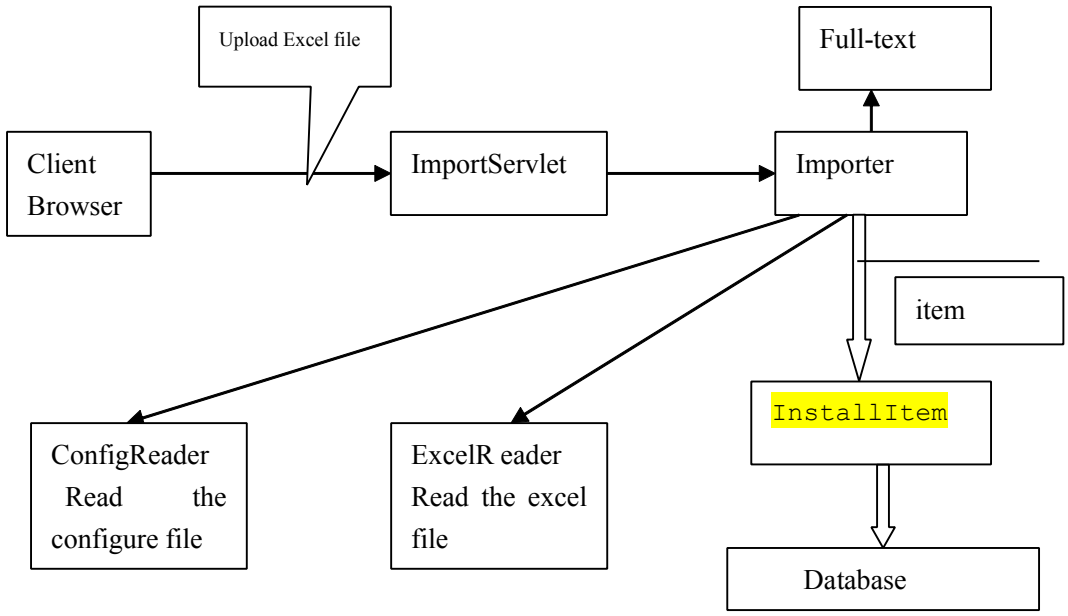
在这个文件里定义了 EXCEL 文件中字段与 DSpace 中 DC 元素的对应关系。比如：

<dcvalue element="contributor" qualifier="advisor">
<col checktype="emptyIgnoreValue" isMulti="true" >E</col>
</dcvalue>

表示：EXCLE 表中的 E 列，对应 Dspace 中的 contributor.advisor 元数据，该字段是可重复的，如果该值为空则忽略。

我们为不同的文献类型定义了不同的导入程序配置文件。

（4）流程图



导入时，打开浏览器，利用 Dspace 管理员工具中 ARP 数据导入按钮，选定用 ARP 数据导出工具导出的 EXCEL 格式文献，此时，程序执行上传 EXCEL 文件的动作，并调用 ImportServlet 类和 Importer，利用 ConfigReader 读取配置文件，ExcelReader 读取 EXCEL 文件，根据配置文件进行字段的匹配，并用 InstallIte 装入 Dspace 数据库。

2.4 根据不同类型的文献定制录入界面的元数据[5]

在Dspace中，input-form.xml控制提交页面元数据。通过修改input-form.xml定制元数据录入页面表单的页数，每一页上显示的元数据字段，他们的顺序，字段的标签、注释信息，每个菜单驱动的字段的可选选项等。[6]

（1）input-forms.xml的结构

①元素结构

这个文件有一个顶层元素input-forms,它按顺序包括了三个元素：form-map,form-definitions,form-value-pairs。其中form-map中定义了专题与欲显示表单元数据的对应关系，form-definition中定义了定制的表单元数据，form-value-pair定义了一些下拉菜单显示的值。

②字段的组成及含义

<form-definitions>中每个<field>字段按照顺序包括一下元素，有些元素是必备的。

dc-schema	指所采用元数据模式的名称
dc-element (Required)	DC元素的名字，比如：contributor
dc-qualifier	DC元素的限定词，当这个元素的值是contributor.advisor，就意味着这个元素是advisor.如果为空的话，意味着输入了一个没有限定词的dc元素。
repeatable	true代表该值可重复，false代表不可重复。如果是可重复的，用户界面上就会多一个控件，允许用户录入更多的该字段，比如关键词，作者等。
Label (Required)	本字段的文字标识，说明您要录入的字段的含义
input-type (Required)	定义了表单中不同dc字段的交互控件的类型。其内容必须是下述值之一： onebox -- 单一录入框 twobox -- 一对文本录入框。常用在可重复的值，比如dc主题词 textarea -- 文本输入区，可录入多行文本。比如文摘。 name -- 人名 date -- 日期 dropdown --下拉菜单 注意：必须包括每个菜单选项的名值对。以此来限定录入选项的值。 qualdrop_value -- 包括一个下拉菜单和一个自由文本框。可输入可选的标识符代码。 注意：必须包括每个菜单选项的名称和值。以此来限定录入选项的值。
hint (Required)	录入说明
Required	说明该元素是必备的。当有这个词时，而用户没有输入信息，将会出现提示:比如：， <required>题名字段是必填项。</required>

在input-form.xml中可以定义表单和页码。这个表单的内容是一系列元素，一个表单可以包括1-6页，每个页码元素必须包括一个number属性，比如：

<page number="1">

（2）定制录入界面的详细步骤

由于会议论文的有些元数据字段在DSpace的缺省录入界面里没有包括，这里，以会议论文为例说明不同类型文献定制录入界面的步骤，其他类型的文献可以参照这个步骤来实现录入界面的定制。

①增加一个form map

在form-map中的每个name-map元素都用一个form的名字和一个专题相关，专题用本专题的handle代表，其form-name属性是该表单的名字，它必须与一个form元素的的属性相匹配。

例如：下面的片段表示了以handle为“12345.6789/42”的专题与“conf”表单相关联。

```
<form-map>
    <name-map collection-handle="12345.6789/42" form-name="conf" />
    ...
  </form-map>
  <form-definitions>
    <form name="conf">
    ...
  </form-definitions>
```

最好保留input-forms.xml中的缺省name-map的定义，这样对那些未定制的专题就可以采用缺省的录入格式。要获得一个专题的Handle，可以通过浏览“社群和专题”，点击特定的专题，可以获得的地址形如: http://myhost.my.edu/dspace/handle/12345.6789/42 ，其中下划线部分就是改专题的handle。

②增加一个表单：

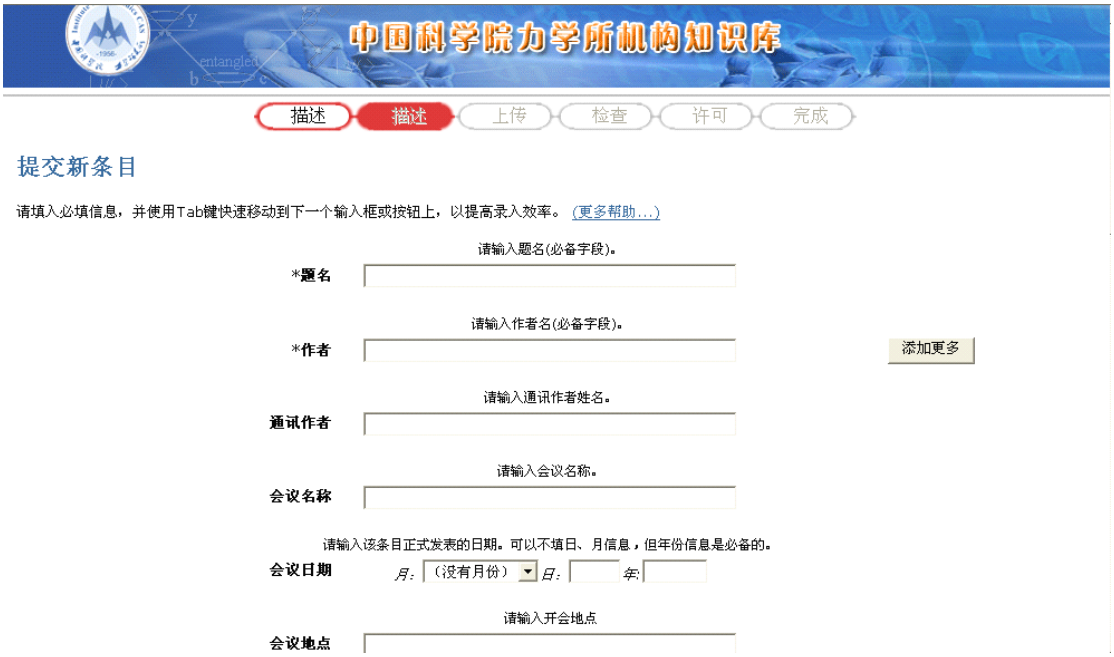
然后通过在form-definitions中创建新的表单元素来创建新的表单。它有一个属性：名称，这个名称要与name-map中定义的值保持一致。

```
〈form-definitions〉
 <form name="thesis">
      <page number="1">
       … …
 //以下可加入定制的元数据字段
<field>
            <dc-schema>dc</dc-schema>
            <dc-element>publisher</dc-element>
            <dc-qualifier>nameConference</dc-qualifier>
            <repeatable>false</repeatable>
            <label>会议名称</label>
            <hint>请输入会议名称。</hint>
            <required></required>
            <input-type>onebox</input-type>
       </field>… …
</form-definitions>
```

之后，还要在form-value-pair定义下拉菜单显示的值。

部署定制的表单，必须要重新启动tomcat,如果该文件有错误，将会导致用户界面的错误，这些错误详细记录在dspace.log文件里。

这样，会议论文专题录入界面就按照input-forms.xml中定制的，新增了会议名称、会议日期、会议地点等字段。会议论文录入界面截图如下：



2.5 定制不同类型文献的元数据显示页面[1]

在数据库中 DC 字段扩展完毕后，还要相应地修改数据显示页面，以便根据不同类型的数据类型显示对应的元数据。

在这里我们采取修改 DSpace 配置文件，并修改 DSpace 源码负责元数据页面显示的程序来解决这个问题。

即在DSpace.cfg中添加webui.itemdisplay.full，并列出不同类型文献的元数据字段名称和修饰词。比如webui.itemdisplay.full = dc.title, \

```
dc.title.alternative, \
dc.contributor, \
dc.contributor.author, \
dc.contributor.correspondent, \
dc.publisher.nameConference, \
.....
dc.identifier.callnum, \
... ... .....
```

然后对 srn/org/dspace/app/webui/jsptag/ItemTag.java,进行适当的修改。这样根据配置文件和 ItemTag.java，就可以在显示页面根据我们针对不同文献定制的元数据来显示不同的元数据字段了。

这样会议论文的显示界面如下：

在此界面中，我们除列出会议名称外，还将会议的名称、时间、地点组合显示在出处字段。

3 结语

目前，本项目利用上述方法基本实现了 ARP 系统中不同类型文献 excel 格式元数据的导入功能，不同类型文献录入界面的定制、不同类型文献显示界面的定制，进行了将 IR 与科研管理信息环境集成的初步尝试，将在以后的工作中，尝试减轻科研人员和图书馆人员提交条目的负担，将 ARP 数据导出工具集成到 B/S 架构下、在对标准 pdf 文档的元数据抽取、个人知识管理工具如 endnote 中数据与 IR 的整合、IR 与其他信息系统关联整合等方面进行进一步研究和实验，达到将 IR 融入科研信息环境的目标。

参考文献：

[1] http://www.arp.cn/ [2007-04-15]

[2]DSpace System Documentation: Contents
http://www.dspace.org/index.php?option=com_content&task=view&id=151&Itemid=116[2007-03-12]

[3] dom4j: flexible XML framework for Javahttp://sourceforge.net/projects/dom4j/ [2007-03-10]

[4] Apache POI - Java API To Access Microsoft Format Files.http://poi.apache.org/ [2007-03-15]

[5]http://www.dspace.org/index.php?option=com_content&task=view&id=155 [2002-07-15]

[6]Custom Metadata-entry Pages for Submission.
http://www.dspace.org/index.php?option=com_content&task=view&id=155#crosswalk[2007-03-12]

致谢

本文的研究与实施受到国科图信息技术部张智雄老师、学科咨询部张冬荣老师及中科院力学研究所周涛老师和徐以鸿老师的大力支持，在此表示感谢。

(作者 E-mail:majx@lzb.ac.cn)