·业务研究·

机构知识库内容快速建设方法

徐以鸿 朱 涛
(中国科学院力学研究所图书信息中心, 北京 100190)

〔摘 要〕本文指出目前国内机构知识库内容建设遇到的问题, 探讨了一种可以加快机构知识库内容建设的方法。并以 WoS 和 CNKI 两大国内外网络著名数据库为例, 对该方法进行了详细说明。

〔关键词〕机构知识库; 内容建设; 批量导出; 批量导入
DOI: 10.3969/j.issn.1008-0821.2011.04.040
〔中图分类号〕G250.74 〔文献标识码〕B 〔文章编号〕1008-0821(2011)04-0148-04

Establishment of Institutional Repository's Content Development

Xu Yihong Zhu Tao
(Library of Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China)

〔**Abstract**〕This article pointed out the existing problems of institutional repository's content development, and put forward a kind of method of accelerating institutional repository's content constructation. Finally, the paper gave two examples especially illustrating the method.

〔**Key words**〕institutional repository; content development; batching export; batching import

机构知识库(Institutional Repository, 简称 IR)是科研教育机构对本单位员工所创造的各类有价值的知识产出进行统一收集、集中管理、长期保存并提供检索利用等增值服务的知识资产管理系统。建设机构知识库可以实现科研教育机构知识产出的系统积累、长期保存和统一管理, 集中展现和反映机构的整体研究实力和水平, 利于知识共享, 促进知识更新, 提高学术成果被发现和引用的几率, 扩大科研人员及机构的学术影响和声望。目前, 国内外许多高校和研究机构在着手建设或正在建设本单位的机构知识库。国外已建成的IR 就有1 800多家[1], 国内已建成的有香港科技大学、香港大学、香港中文大学、香港城市大学、香港教育学院、澳门大学、清华大学、厦门大学、浙江大学、北京理工大学、福建师范大学、中国农业科学院、中国国家科学图书馆及中国科学院下属的 60 多家研究所, 还有台湾机构典藏等机构知识库。

机构知识库的建设主要有三部分: 平台建设、内容建设和政策机制建设。平台建设大多是利用现有的开源软件如 Dspace, DSpace 是美国麻省理工大学(MIT)和惠普(HP)公司共同开发的开放获取软件, 是最早应用的机构知识库的软件系统, 也是发展最快的机构知识库系统; Eprints 是最早的 IR 构建工具, 也是第一个遵循 OAI 协议的 IR 软件, 由英国南安普顿大学开发。政策建设主要是制定一些政策法规以及机构知识库运行管理办法, 以确保机构知识库建设的顺利开展和长期稳定运行。机构知识库的内容建设大多是通过机构的科研人员自行提交。由于各种原因, 机构人员并没有把数据提交到本机构库中。目前有不少作者撰文, 提出了许多很好的观点和应对措施。如郎庆华提出了应积极开展机构知识库自存储的宣传和推广工作, 建立有效评价机制、加强版权保护以及提供自存储资源的提交服务[2]; 赖辉荣就目前机构知识库“有站无车、有车无人”的局面提出了一些策略: 如加强宣传、提供利用率推送服务、简化提交步骤、制定激励和强制性自存储政策、提供知识产权保护以及建立质量控制机制[3]。

这些文章是从机构知识库平台搭建后出发, 阐述了保

收稿日期: 2011-03-03
基金项目: 本文受力学所机构综合数字知识管理特色分馆项目支持。
作者简介: 徐以鸿(1969-), 男, 副研究馆员, 研究方向: 机构知识库建设和文献传递, 发表论文 10 篇。

障科研机构正在产生或将来的产出物的自行提交论点，而鲜有文献就 IR 如何收集机构已有的知识产出物进行论述。由于 IR 在国内起步较晚，科研人员对其还很不了解，用户自行提交内容基本都是当年公开发表的期刊论文、会议论文，报告以及学位论文等，很少会提交历史数据。要用户提交以前的知识内容一来不现实，因为科研人员都很忙，没时间、没精力把历史数据提交到 IR 库；二来科研人员很难收集全自己公开发表的论文、报告等知识内容，所以机构以往的知识产出提交成了 IR 内容建设的一大难题。而研究机构或大学的图书馆作为机构的知识产出物收集、管理部门，应该当仁不让的 IR 库的主要内容提供者。但如果让图书馆部门的工作人员也按自行提交模式一条一条提交数据既费时也费力。同理，图书馆的人员很多都是一人干着几项工作，要想短时间把机构以前的知识产出物逐条提交到 IR 库，几乎是不可能完成的工作。另外，很多机构都已经建有较为全面的产出物保存系统，如：期刊论文数据库、研究生学位论文数据库、成果奖励专利、国际会议论文统计系统等数据库，以及中国科学院 ARP 系统中的论文产出库，许多科研院所和大学购买的各种数据库。这些数据库大多都提供批量输出检索到的文本格式数据或 EXCEL 电子表格数据。如何利用从这些现有数据库抽取本机构的论文数据，再批量提交到 IR 库？马建霞撰文利用自行开发程序嵌入到 IR 系统，实现了批量提交数据到 IR 库[4]。

下面以力学所机构知识库的内容建设为例，提出一种 IR 内容建设关于历史知识产物的收集方法：IR 数据批量导入建库过程。

1 力学所机构知识库的简要介绍

力学所机构知识库（简称 IMECH - IR）作为“支持综合知识内容管理”特色分馆建设项目于 2008 年正式启动，于 2009 年 6 月中旬开通。IMECH - IR 按力学所现有的研究部门和 2008 年前的知识产出建立了 9 个社群，分别是力学所知识产出（1956 - 2008）、非线性力学重点实验室、高温气体动力学重点实验室、国家微重力实验室、水动力与海洋工程重点实验室、环境力学重点实验室、先进制造工艺力学重点实验室、等离子体与燃烧中心以及职能与支撑部门，每个社群下按内容类型分若干个研究专题。内容类型主要分：期刊论文、会议论文、学位论文、专著和会议文集、专利、研究报告、演示报告、成果以及其他。力学所知识产出（1996 - 2008）社群下有数据 8 700 多条，这些数据都是从其它数据库或网站导出再导入到 IR 库的。其中期刊论文、会议论文专利、成果等数据大多是从 ARP 系统导出成 EXCEL 表格格式数据，学位论文数据是从力学所图书馆公共检索系统中导出，有些从中国知网和 Web of Science 导出的。2009 年按研究部门建立的研究社群下的数据多数是注册用户自行提交的。

鉴于国内多数高校和科研院所都购买了 Web of Science 和中国知网 CNKI 两大数据库。本文就以这两个具有代表性的数据库为例说明 IR 数据的批量导入建库。

2 IR 数据批量建库前的数据准备

对于 SCI 收录的论文全部可以从 Web of Science 收集到，而发表在中文期刊上的论文大多可以从中国知网 CNKI 上下载题录数据。

2.1 从 Web of Science (WoS) 数据库批量采集数据

ISI Web of Science[5] 是 Thomson Scientific 建设的三大引文数据库的 Web 版，由 3 个独立的数据库组成（既可以分库检索、也可以多库联合检索）分别是 Science Citation Index Expanded（简称 SCI Expanded）、Social Sciences Citation Index（简称 SSCI）和 Arts & Humanities Citation Index（简称 A&HCI）。内容涵盖自然科学、工程技术、社会科学、艺术与人文等诸多领域内的8 500多种学术期刊。其中的 SCIE 数据库——《科学引文索引》网络版收录5 900余种期刊文摘和引文，内容涉及自然科学、工程技术的各个领域。

从 WoS 采集数据过程如下：先登录 ISI Web of Knowledge 网站，选取数据库 Web of Science，在作者地址栏输入机构名称的各种英文拼法，用 or 连接，设定检索年限；检索出结果，标记结果，按 IR 库元数据字段要求设定字段输出纯文本格式数据文件，见图 1。Web of Science 一次输出结果不超过 500 条，如检索结果超过 500 条，多次输出即可。

图 1 Web of Science 标记记录输出页面截图

2.2 从 CNKI 数据库批量采集数据

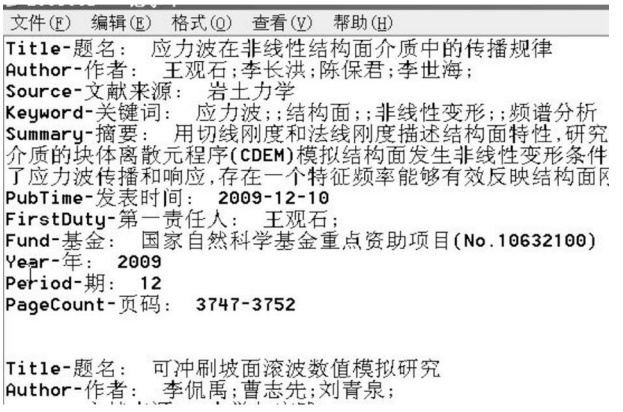
《中国期刊全文数据库》CNKI 期刊数据库[6] 是目前世界上最大的连续动态更新的中国期刊全文数据库，收录国

内8 200多种重要期刊，以学术、技术、政策指导、高等教育及科普为主，内容覆盖自然科学、工程技术、农业、哲学、医学、人文社科等各个领域；核心期刊收全率达到99%，内容收录完整率在 99%，出版时间不迟于纸本出版后2 个月。从 CNKI 网站导出题录数据与从 WoS 类似，只是 CNKI 一次输出最多 50 条，导出步骤见图 2。

图 2 CNKI 选中的文献记录输出页面截图

3 数据格式转换

对于从WoS数据库导出的纯文本格式数据，根据 Excel 电子表格中的自外部导入数据功能即可生成表格。而由 CNKI 数据采集到的文本数据转换成表则要复杂些。CNKI 导出的数据文件见图 3。



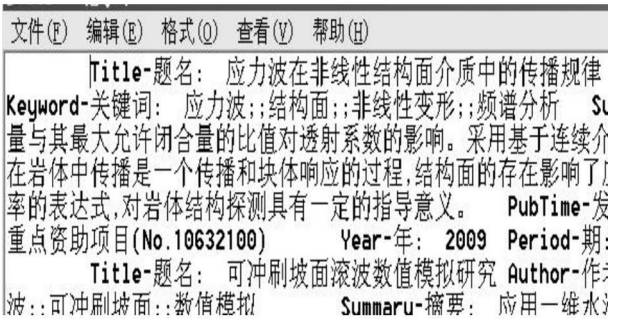
```
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
Title-题名: 应力波在非线性结构面介质中的传播规律
Author-作者: 王观石;李长洪;陈保君;李世海;
Source-文献来源: 岩土力学
Keyword-关键词: 应力波;;结构面;;非线性变形;;频谱分析
Summary-摘要: 用切线刚度和法线刚度描述结构面特性,研究
介质的块体离散元程序(CDEM)模拟结构面发生非线性变形条件
了应力波传播和响应,存在一个特征频率能够有效反映结构面
PubTime-发表时间: 2009-12-10
FirstDuty-第一责任人: 王观石;
Fund-基金: 国家自然科学基金重点资助项目(No.10632100)
Year-年: 2009
Period-期: 12
PageCount-页码: 3747-3752

Title-题名: 可冲刷坡面滚波数值模拟研究
Author-作者: 李侃禹;曹志先;刘青泉;
```

图 3 CNKI 导出的文本格式数据文件截图

格数据截图从图 3 可知，CNKI 的导出数据与 WoS 数据库导出数据格式不一样，每条数据中的每个数据项（或称为字段）是一行，数据与数据是两个回车符（段落符号），没有制表符。如直接导入到 Excel 表，生成的表格数据如图 4，只有一列数据，一行是一个字段。要生成行对应一条数据，列对应一条数据中的数据项形式规格的电子表格数据，必须对原 CNKI 导出的文本数据进行转换。通过 Ultra Edit 编辑器可以转换，在 Word 中也可置换。两个回车符替换为一个回车符，并在文件首添加一制表符，再另存为纯文本文件即可，转换后纯文本数据见图 5。在生成 Excel 表格数据时要注意：因 CNKI 导出的每条数据的数据项没数据时并不默认为空单元格无数据内容，而是没有该数据单元格，这就导致了生成后的表格数据列会错位。通过数据排序把没有数据项的插入单元格，以使表格格式规范，规范后的表格数据见图 6。

1	Title-题名: 应力波在非线性结构面介质中的传播规律
2	Author-作者: 王观石;李长洪;陈保君;李世海;
3	Source-文献来源: 岩土力学
4	Keyword-关键词: 应力波;;结构面;;非线性变形;;频谱分析
5	Summary-摘要: 用切线刚度和法线刚度描述结构面特性,研究结构面初始
6	PubTime-发表时间: 2009-12-10
7	FirstDuty-第一责任人: 王观石;
8	Fund-基金: 国家自然科学基金重点资助项目(No.10632100)
9	Year-年: 2009
10	Period-期: 12
11	PageCount-页码: 3747-3752
12	
13	
14	Title-题名: 可冲刷坡面滚波数值模拟研究
15	Author-作者: 李侃禹;曹志先;刘青泉;
16	Source-文献来源: 力学与实践

图 4 CNKI 导出直接导入 EXCEL 生成的表



```
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
Title-题名: 应力波在非线性结构面介质中的传播规律
Keyword-关键词: 应力波;;结构面;;非线性变形;;频谱分析 Su
量与其最大允许闭合量的比值对透射系数的影响。采用基于连续介
在岩体中传播是一个传播和块体响应的过程,结构面的存在影响了
率的表达式,对岩体结构探测具有一定的指导意义。 PubTime-发
重点资助项目(No.10632100) Year-年: 2009 Period-期:
Title-题名: 可冲刷坡面滚波数值模拟研究 Author-作
波;;可冲刷坡面;;数值模拟 Summaru-摘要: 应用一维水
```

图 5 CNKI 导出数据经转换后的文本数据截图

Ti	Au	Sc	Kd	Ab	Py	RpAu	Is	Page	Email	Vol	Dep	Cite	FullTxt	Fund
参数模型	白玉湖	石油钻探水合物,	#######		2009-11-25	白玉湖	6	11-17			h	石油钻探	J2009/c(	国家高
考虑介质	李小波	力学学报乳状液,	#######		2009-5-18	李小波	3	313-317		41	h	力学学报	J2009/c(	国家基
乳化和游	李小波	石油学报强化采油利用耗			2009-3-15	李小波	2	259-262+266			h	石油学报	J2009/c(	国家重
压电驱动	李敏;	航空学报压电驱动	#######		2009-12-25	李敏	12	2301-2310		30	h	航空学报	J2009/c(	国家自
节流器内	李国君	石油学报水力加砂建立了考			2009-1-15	李国君	1	145-148			h	石油学报	J2009/c(	国家自
三峡库区	尹小玲	水力发电水环境,	三峡水库		2009-12-25	尹小玲	6	43-48		28	h	水力发电	J2009/c(	国家自
让舰艇经	李家春	科技导报抗冲击能(正)200			2009-7-28	李家春	14	3	jcli05i		h	科技导报	J2009/c061.pdf	
高压深层	刘曰武	中国科学渗流, 不	#######		2009-4-15	刘曰武	4	606-616	lywu@i	39	h	中国科学	J2009/c062.pdf	
钱老的科	李家春	力学进展力学研究(正)"中			2009-11-25	李家春	6	654-655	jcli05i	39	h	力学进展	J2009/c063.pdf	

图 6 规范后的 CNKI 表格数据截图

4 编写 XML 格式的转换文档

根据文献 [4] 提供的 xml 文档示例，编写相应的转换 Excel 表格数据的 XML 格式文档，把编写好的文档上传至机构知识库系统服务器相应的文件夹。以管理员省份登录机构知识库网站，进入管理控制区的数据导入项。转换示例见图 7。

图 7 EXCEL 表格数据导入到 IR 库网页截图

5 结 语

目前国内机构知识库的建设正如火如荼的开展，但同时很多机构或大学都或多或少的遭遇了机构知识库平台建好了，相应的政策也制定了，也在本机构的各种场合进行现场的、虚拟的宣传，却叫好不叫座，缘由是内容建设跟不上。为了走出目前这种尴尬局面，本文试图从批量采集数据再批量导入到机构知识库系统中，从而加快 IR 的内容建设。文章指出，利用各机构已有的各种数据库，包括自建的、购买的或能免费获得的，从这些数据库采集本机构的知识产出物，生成规范的电子表格数据，再利用文献 3 提供的批量导入程序把数据导入到 IR 系统中，从而加快 IR 内容建设。文中以 WoS 和 CNKI 为例，详细说明了数据采集、整理、导出成 EXCEL 表格数据，再批量导入 IR 库的全过程。

参 考 文 献

[1] Search or Browse for Repositories (Open DOAR) [OL]. http: // www. opendoar. org/find. php , 2010- 12- 15.

[2] 郎庆华. 机构知识库长期保存的策略分析 [J]. 情报理论与实践, 2010, (5): 47- 51, 62.

[3] 赖释荣. 破解机构知识库建设中资源收集难题之策略 [J]. 国家图书馆学刊, 2009, (03): 59- 61.

[4] 马建霞, 祝忠明, 唐润寰, 等. 机构知识库与科研管理信息化环境集成的尝试 [J]. 现代图书情报技术, 2008, (2): 14- 18.

[5] http: // apps. isiknowledge. com [OL]. 2010- 12- 15.

[6] http: // epub. cnki. net/ grid2008/ index. htm [OL]. 2010- 12- 15.

(上接第 147 页)

育，以培养员工在全球化市场条件下从事信息化管理的能力。要加大培训的投入，实施人才本地化策略。

4.5 企业要加快实现管理模式创新

对于中国传统的工业企业而言，要实施企业信息化，首先要进行业务重组，改变传统的企业组织及管理模式。但是我国企业传统的管理机制、管理思想、管理方法与先进的市场经济管理模式有很大的差距，所以企业的信息化首先是管理模式的创新。企业信息化不仅仅是电子信息技术在企业生产及管理领域中的简单应用，它对企业所产生的影响是全方位的，只有把企业的技术创新与制度创新结合起来，使企业的信息化与管理的现代化相结合，企业的信息化工作才会成功。

装备制造业管理信息化是一个通过将信息技术和其他高新技术与制造技术不断融合，从而不断改善企业生产、经营、管理和产品开发行为，其潜力是巨大的。只要我们抓住重点，科学决策，装备制造业的管理信息化就一定能够在提高企业的经济效益和竞争能力的同时，推动并逐步实现装备制造业的产业转型升级。

参 考 文 献

[1] 中华人民共和国国民经济和社会发展第十二个五年规划纲要 [EB]. http: // www. ce. cn/ xwzx/ gnsz/ gdkw/ 201103/ 16/ t20110316_22305305. shtml

[2] 黄谊江. 漫谈企业信息化 [EB]. http: // www. manaren. com , 2010- 08- 02.

[3] 崔新升. 解读装备制造业信息化的现状、重难点和对策 [EB]. http: // bbs. it168. com , 2009- 07- 10.

[4] 中央企业信息化“不差钱”却难在意识 [EB]. http: // www. mie168. com/ read. aspx , 2010- 04- 04.

[5] 中华人民共和国国家统计局, 第三次全国工业普查办公室. 关于第三次全国工业普查主要数据的公报 [EB]. http: // c. ccfv. cn/ was40/ gjtjj_ detail. jsp?channelid= 47417&record= 2 , 2001- 09- 01.

[6] 安筱鹏. 推进中国企业信息化进程的发展战略 [EB]. http: // www. echinagov. com/ gov/ index. shtml , 2008- 10- 16.