中国科学院联合机构仓储系统的开发与建设*

祝忠明[1] 马建霞[1] 张智雄[2] 孙坦[2]
[1]（中国科学院国家科学图书馆兰州分馆，兰州，730000）
[2]（中国科学院国家科学图书馆，北京，100080）

[摘要]对中国科学院联合机构仓储系统的建设思路和实施过程中遇到的主要问题进行了讨论，包括如何选择合适的机构，进行试点和示范，以推动机构仓储的实施；如何进行中文机构仓储软件系统的定制和扩展开发；如何集成现有系统和组织机构仓储服务；如何建立联合仓储系统等。
[关键词] 中国科学院 国家科学图书馆 机构信息环境 联合机构仓储
[分类号] G250.73

Developing a Federated Institutional Repository in Chinese Academy of Sciences

Zhu Zhongming[1] Ma Jianxia[1] Zhang Zhixiong[2] Sun Tan[2]
[1] (Lanzhou Branch of National Sciences Library of CAS, Lanzhou, 730000, China)
[2](National Sciences Library of CAS, Beijing 100080, China)
[Abstract]This paper discusses the strategies and problems that are involved in the process of implementing a federated institutional repository in Chinese Academy of Sciences. They include how to select the institute and promote the implementation of the institutional repository to set up examples for other institutes to follow; how to customize and develop a suitable Chinese institutional repository system; how to integrate the existing system; how to set up a federated repository; and etc.
[Keywords]: Chinese Academy of Sciences, National Sciences Library, Institutional Information Environment, Federated Institutional Repository

1 前言

不断增长的研发投入为中国科学界带来了丰硕的成果，在全球的科学研究中，中国科学院的科学家们扮演着越来越重要的角色。但是，由于书刊价格的上涨及其他方面的因素，中国科学院的研究人员已经在国外发表的许多高水平论文却往往不能为本国的科学家所容易地查看或者获得。如何充分利用这些高水平的学术资源成为当前中国科学院的科学家、政策制定者和图书馆人面对的大问题。2005 年针对中国科学院内一些代表性研究所开展的调查显示[1]：中国科学院的科研机构迫切需要实现对数字资产的管理，以便进一步地提升机构的社会形象和扩大机构的学术影响，重用已有科研成果，实现成果的长期保存。

这种形势下，以中国科学院作为一个整体性的研究机构来规划和建设一个单一的和综合的机构仓储（Institutional Repository，以下简称 IR）系统，实现对全院数字科研资产的统一管理和利用，从表面看来，应是一种比较理想的选择。但根据中国科学院在组织结构和管理模式上的特点，这种单一性的方案将在可行性和实效性方面存在很大的疑问。中国科学院由分布于全国各地的 89 个研究所组成，各研究所以相对独立和自主的方式确定自己的研究方向和发展目标，开展有关的科研活动，并且都有自己的图书馆。因此，各研究所是全院范

* 本文系中国科学院国家科学图书馆“全院联合机构仓储体系建设”项目及国家社科基金项目“机构知识库建设研究与应用”(项目编号: 07BTQ019)的研究成果之一。

围内组织和开展具体科研活动的基本单元，也应该是对科研产出进行积累和管理的基本单元。而且，研究所已经开始认识和意识到通过对其数字科研产出进行管理，在促进其共享利用、扩大学术影响等方面的重要的意义和作用。据此，中国科学院国家科学图书馆提出了充分利用和调动研究所的积极性，构建全院联合机构仓储体系的构想。即：首先，以研究所为单元构建各所的 IR 系统；然后，依托分布于各研究所的 IR 系统，通过元数据开放获取和内容聚合的方式，建立起全院联合的机构仓储服务体系，从而发展和建立起：(1) 全院的知识中心，支持全院范围内数字资产的汇集和管理；(2) 全院研究成果有效分发、传播和共享利用的机制和平台；(3) 对正式出版渠道的补充，逐步形全院整体的 e-Scholarship 仓储和交流平台；(4) 全院知识产出的长期数字保存机制。

目前，按照上述构想，中国科学院国家科学图书馆已经启动了全院联合机构仓储体系的研究和实施计划，并取得了初步的成果。本文后续的部分将对建设过程的主要环节和涉及的主要问题进行讨论。

2 研究所 IR 系统的开发建设

研究所 IR 的建设是整个计划实施的第一阶段，也是最重要的基础环节，主要通过试点和示范的模式进行开发建设。目前，我们通过选择中国科学院力学研究所作为试点和合作单位，已经部署和建立起了该所的 IR 服务示范系统。

试点研究所选择的考虑

通过选择试点研究所，一方面，是为了能够准确地了解和把握研究所对 IR 的建设需求；另一方面，也希望通过在试点研究所进行 IR 的部署，为其他研究所树立榜样，发挥示范效应，从而带动和吸引其他研究所积极地加入全院机构仓储建设体系中来。为了达到这样的目的和效果，在试点研究所的选择过程中，我们主要考虑：

研究所对 IR 建设有较高的积极性。

首先，研究所的管理层特别是主管信息化工作的领导要对 IR 建设的作用和意义有明确的认识，对 IR 的建设持欢迎和积极的态度，已经打算或愿意支持 IR 的规划建设。某种程度上，这是决定是否能够启动 IR 规划建设的先决条件，也是 IR 建设所需要的各种政策环境、设备条件、人员等方面的条件得到持续有效支持的保障。

其次，由于研究所图书馆的领导和有关工作人员将是本所 IR 规划和建设的主要执行者，因此他们是否有足够的积极性和热情，对于 IR 的建设成败也有着重要的影响。

研究所能够保障 IR 实施的必要投入。

包括研究所是否已经具备 IR 实施所要求的基础软硬件环境，或者能够保证进行投入和从新建设；是否能保证 IR 建设的人员投入，特别是在计划实施的初期，应能够保证科研管理部门的有关人员和图书馆管理者有一定的时间和精力投入 IR 的规划和建设，并且能够指派专职的人员，专门负责 IR 的具体建设和实施；以及其它方面的经费投入的保证等。

研究所的信息环境有相对较为丰富的应用。

IR 做为研究所整体信息环境发展和构建的重要“组件”，要能够适应与其他相关系统进行互联互通的要求。如果研究所的信息环境过于简单或单一，将不利于识别、捕获和形成一个相对完善的功能需求集合，也不便于对 IR 进行实际地运行和测试，并可能影响到 IR 应用软件在全院范围内的通用性和适用性。

特殊功能和服务需求的识别

从 IR 的发展来看，采用成熟的开源软件已经成为一种主流的选择。因此，我们采取了基于开源软件 DSpace 建立原型系统，与试点研究所进行交流和沟通，以进一步捕获和明确需求的方法。事实证明，这是一种非常行之有效的方法。它不仅提供了一个对 IR 通用功能需求和服务分析的基础，而且也有助于快速地确定功能改进的需求、以及发现和捕获新的功能需求。按照这一过程和方法，我们发现除了通用功能和服务外，合作研究所提出了一些新

的服务需求，与新型机构信息环境中有关 IR 建设的情景描述不谋而合[2]。这些新的或特殊的功能和服务需求包括：

IR 是机构整体信息环境的一部分，需要与其它的系统有机地集成，能够从图书馆自动化系统、ARP 系统、以及其他类型的数据信息系统中自动地提取有关的信息，减少数据加工的重复和人工操作，提高系统之间的协同工作能力和效率。

尽可能减少由科研人员以自助提交方式进行内容提交和加工描述的要求，避免使科研人员对 IR 产生有使用繁琐、信息描述加工量大的印象，从而降低对系统使用的期望和使用频率。

IR 应能够提供对常见格式内容的自动分析和识别，从中抽取和形成有关的元数据描述信息，从而最大程度地减少人工描述和加工的要求。

IR 应该有灵活完善的访问控制策略，比如基于 IP 地址的全文内容访问控制、以及面向特定用户或用户组的全文内容访问和存取策略等。这虽然有违于 IR 实践开放存取的初衷，但确是机构的实际需求。

支持信息的多维组织、浏览、导航和检索，如基于研究室、内容类型、学科方向、主题、作者等的信息组织与浏览，提供全文检索、检索结果的链接检索、二次检索、相关检索等。

IR 不仅要能够与研究所内的各相关系统之间集成，还应该支持与与外部的应用系统之间的开放集成，方便系统之间的数据交换和共享。这就要求 IR 必须支持多种开放接口，如 RSS 内容聚合接口、OAI 元数据开放获取接口、SRW/U 标准检索接口、OpenURL 开放链接接口等。

IR 的开发建设

2.3.1 开发策略

由于 IR 的建设和发展在近年来得到了国内外特别是国外各著名大学和研究机构的普遍重视，支持 IR 构建的应用软件平台也开始逐步走向成熟，特别是以 DSpace、Eprints 等为代表的开源 IR 软件，基本上已经成为国际上 IR 建设和部署的主要技术选择。因此，我们也采取了通过选择和基于此类开源软件为基础进行定制和扩展开发的策略。

根据对一些得到广泛应用的开源 IR 软件的比较和分析，以及结合我院知识仓储系统建设的需求，我们认为 DSpace[4]具有系统结构比较合理、功能较为完善、支持任意类型的内容存储等特点，适合做为我院 IR 应用软件系统定制和扩展开发的基础。

2.3.2 基于 DSpace 的 IR 系统功能结构

总体上，研究所 IR 将不仅考虑作为研究所独立运行的知识资产管理系统，同时也将支持其作为研究所信息基础设施环境构建的重要组成部分，支持与其他相关信息系统之间的关联和集成。这种集成和交互关系，如图 1 所示。

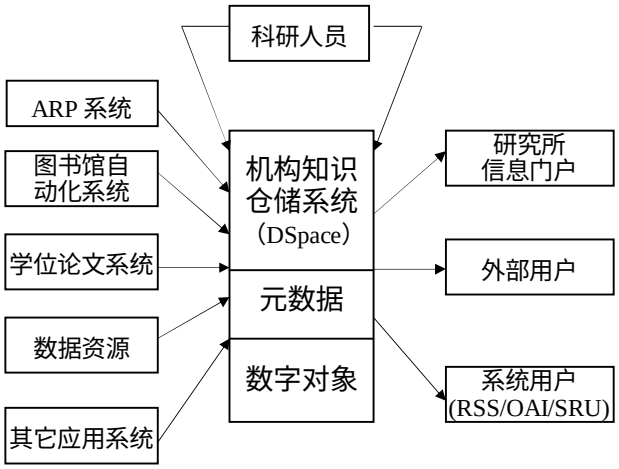

图 1　IR 与相关应用系统的交互关系示意图

结合前述对 IR 系统功能和服务需求的分析，这里给出我院基于 DSpace 的 IR 系统的功能结构图。如图 2 所示。

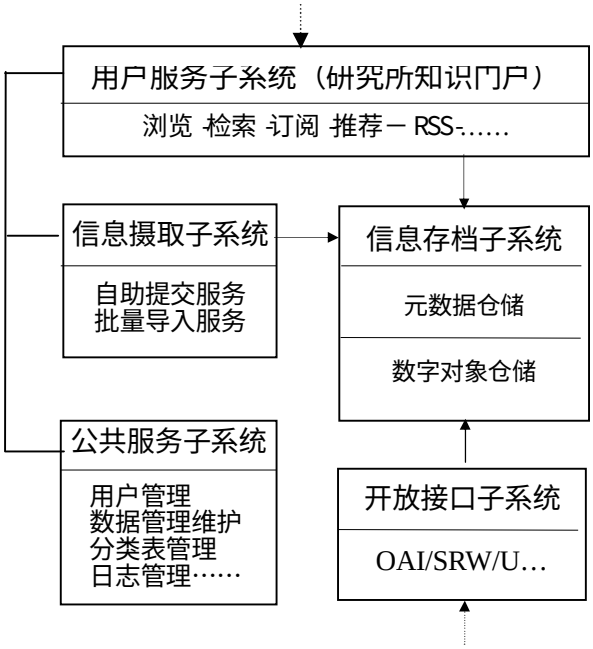

图 2　IR功能结构示意图

其中，用户服务子系统为用户提供多维的信息浏览和导航途径，提供简单检索、高级检索、全文检索等灵活多样的信息检索方式，提供内容订阅与推送服务，以及提供按照一定的条件组合，进行相关信息内容的分析与汇总、结果的可视化展现等扩展服务。

信息摄取子系统支持任何授权用户以自助提交和存档的方式，按照一定的工作流程和内容描述要求对提交内容进行描述和编辑，并提交到 IR 中。同时，也支持以批量导入的方式支持将符合系统数据格式要求的内容导入到系统中。

信息存档子系统由提交的数字对象（内容）及其元数据组成，支持对数字对象的长期保存、管理及存取利用服务。

开放接口子系统主要通过实现 OAI Data Provider，支持联合机构知识仓储系统实施元数据或数字对象的聚合，并根据与其他应用系统集成的要求，提供 SRW/SRU、OpenURL 等服务接口。

公共服务子系统，提供系统运行和服务过程中的各种公共服务功能和管理维护功能的支持，如分类表的管理、元数据及内容的维护与管理、用户管理、访问统计分析等。

2.3.3 主要的定制和扩展开发工作

包括 DSpace 的本地化，以及针对特定功能需求的定制和扩展开发。DSpace 的设计具有良好的分层架构，支持基于公共 API 进行功能和服务的扩展开发方式。在开发过程中，尽管我们尽量通过对其 API 的调用和扩展来进行功能的定制和开发，但在很多情况下，还是需要对其底层代码进行修改才能完成一些特定的功能开发需求。

2.3.3.1 DSpace 的本地化及界面定制[5][6]。

主要包括以下方面：

l 　DSpace 默认界面为英文界面，根据中文信息显示和处理的习惯，对各有关页面显示、处理过程、在线帮助等方面的显示信息进行了全面的中文化。

l 　通过对相关 jsp 文件及 css 文件的修改和调整，对各级页面从内容布局、色彩搭配等方面进行了美化设计和调整，提供显示主题和风格的定制和切换的功能扩展。并在首页增加了有关最新提交、RSS 订阅等方面的功能。

l 　作者姓名处理的本地化。由于 DSpace 在对作者姓名的处理上，将“ 姓 ”和“ 名 ”进行了分解和分别存储的处理，这对于西方作者来说是合适的，但并不符合中文姓名的处理和显示习惯。在不改变数据逻辑的基础上，通过对有关提交和显示过程的程序进行了修改，以适应中英文姓名能以比较习惯的方式得到显示和处理。

l 　中文排序的支持。DSpace 通过支持 Unicode 编码标准以实现国际化设计和支持，但这样带来的问题是无法支持按照汉语拼音顺序来对显示结果进行排序。通过集成有关的开源软件包，以及对结果显示处理的程序进行了修改，实现了按照汉语拼音顺序显示浏览和检索结果的功能。

2.3.3.2 元数据应用规范的扩展。

DSpace 支持以基于扩展的 DC 元数据标准为基础的元数据应用规范，我们主要从适应研究所提出的对多种类型的数字内容类型的描述和显示的要求基础上，通过元素修饰符扩展的方式进行了扩展，如根据学位论文、会议论文等的特殊描述和显示需求，做了 10 多项有关的扩展。

2.3.3.3 提交流程和界面的调整。

DSpace 默认的内容提交流程步骤较为繁琐，有关的提交界面也常常被分解到几个页面中。在保留默认提交流程及界面的情况下，提供了一套简化的流程及界面，以满足一般的内容提交和编辑习惯。

2.3.3.4 数据的批量导入功能。

主要进行了从研究所 ARP、图书馆自动化系统等系统中将有关数据进行导出和导入 IR 的专门工具的扩展开发。对有一定通用性的工具，将考虑集成到 DSpace 系统中，方便用户的使用。对于那些只是在系统初装过程中等场合使用的一次性数据导入导出工具，则主要作为外部的程序的方式提供使用。

2.3.3.5 存取控制的强化。

扩展了基于 IP 地址和用户组的数字对象访问控制功能，以满足研究所制定灵活的 IR 内容访问许可策略。

2.3.3.6 数字对象的访问统计功能。

扩展实现了基于信息条目级的访问利用统计功能，可以方便作者对发布在 IR 中的科研和学术成果的访问和下载情况进行及时了解和掌握，IR 管理者也可以据此对任一数字对象或所有数字对象的访问和利用情况进行统计和分析。

2.3.3.7 开放接口的定制和扩展。

首先，DSpace 已经提供了 OAI Data Provider 接口，这也是我们构建全院联合知识仓储服务系统必需的接口。同时，基于 OCLC 发布的 SRW/U[7]开源软件，我们也为 DSpace 扩展了 SRW/U 接口，可以方便地支持以标准的方式与各种检索应用服务系统的集成。

协助研究所进行 IR 的规划和实施

对于我院大部分的研究所来说，IR 的规划和实施还属于新生事物。针对这一点，我们编写了有关 IR 规划和实施的参考文档，拟提供给研究所使用，以促进 IR 概念、作用和职能的宣传和推介，帮助研究所进行 IR 的规划和实施，包括 IR 建设的投入分析、实施团队的组件、内容建设保障机制和政策、内容组织和提交的流程、内容的安全和长期保存机制等。

在具体的实施过程中，则主要通过远程或现场方式，为研究所进行 IR 应用系统的安装、部署，实现上线使用。

IR 与相关服务的集成

目前，研究所 IR 系统主要实现了与研究所范围内的图书馆自动化系统、ARP 系统之间的数据转换和集成。如在针对中科院力学所的 IR 实施过程中，已经实现了从其图书馆自动化系统中提取和导出学位论文的数据、以及从 ARP 系统中导出科研论文、会议论文、专著等产出物信息，并将这些数据转换和导入 IR 的处理。在与研究所图书馆网站、研究所门户网站的链接和集成方面，也提供了基于 IR 公共检索 API 及 SRW/U 标准的嵌入和集成机制。随着 IR 应用的逐步深入，将进一步对 IR 的 SRW/U 接口进行优化，并启动 OpenURL 的支持，逐步以标准的方式集成到有关的公共检索和服务系统中，如与中国科学院国家科学图书馆的集成检索服务平台的集成，使 IR 中有关的知识信息能在更大的范围内被检索、发现和利用。

3 全院联合的机构仓储服务系统建设

这是整个计划实施的第二阶段，将包括两方面的工作。

首先，在全院范围内开展 IR 的推广和部署。即，根据第一阶段 IR 的试点应用和部署，在形成功能完善和性能稳定的 IR 应用软件基础上，举办面向研究所科技管理人员、图书信息管理及应用人员的集中培训，并针对不同研究所的情况，通过现场安装、自助安装或者远程安装方式进行系统的安装、部署和上线应用。

其次，开发 OAI 元数据收割系统，对逐步部署和应用起来的研究所 IR 实施元数据收割和再组织，建立起全院联合的机构仓储服务系统。

在全院联合机构仓储服务系统的建设过程中，我们仍然以 DSpace 系统作为基础，通过扩展 OAI 元数据收割功能的支持，以形成全院联合的机构知识门户服务系统。在 OAI 元数据收割系统的扩展开发过程中，也沿用了基于开源软件以加快开发进度和节省开发成本的策略，选择了 OCLC 发布的开源软件 OAI Harvester 2.0[8]，并通过定制开发实现了与 DSpace 系统的集成。图 3 给出了全院联合机构仓储服务系统的总体功能结构图。

在这种集成的过程中涉及的关键问题包括[9]：(1) 元数据收割器的定制，其中涉及对目标仓储系统列表的维护，元数据获取的控制，增量更新获取的调度与控制，以及基于多线程的多目标仓储的并发搜寻与元数据获取等。(2) XML 数据解析器的实现，即从 OAIHarvester 获取的以 XML 格式的数据文件中将需要的元数据信息解析出来，进行必要的规范化和归并处理，为载入 DSpace 做好准备。(3) 数据批量导入接口的实现，即将解析并规范化处理后的元数据按照 DSpace 系统可以接受的格式，载入 DSpace 系统。

同时，联合机构仓储服务系统继续保持对 OAI、SRW/U、RSS 等开放接口的支持，保证其与全院层面上有关信息服务系统的方便集成。如与中国科学院国家科学图书馆的集成检索系统基于 SRW/U 的标准化检索集成等。

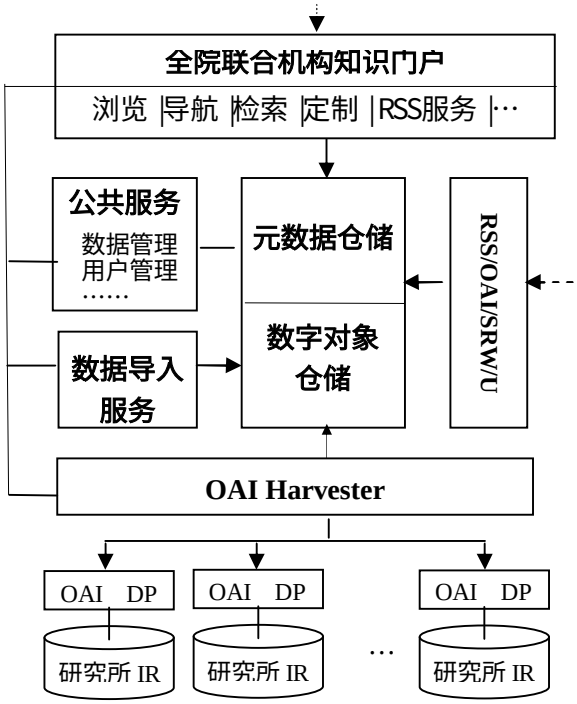

图 3 全院联合机构知识仓储服务系统

4 结语

到目前为止，我们已经完成了示范性 IR 的构建和部署，全院联合机构仓储服务实验系统也正在开发完善之中，而且很多研究所对 IR 的建设都响应积极，愿意早日实施本所的 IR。

在建设过程中，我们也吸取了许多的经验和教训。例如：

研究所的机构仓储建设，技术手段不是关键问题，了解研究所需求，按照研究所的相关机制定制和扩展相应的仓储系统才是重中之中。特别是在当前每一个中国科学院研究所图书馆的工作人员平均不到 3 人的情况下，研究所 IR 的提交和管理流程需要大力地简化。

研究所 IR 的建设必须考虑到研究所已有的信息系统。在当前研究所中存在着 ARP 系统、图书馆自动化系统等，IR 的建设必须与这些系统实现有机的集成和共享，一方面避免数据的重复录入，另一方面要避免构建一个个孤立的系统。

提前规划十分重要。在研究所 IR 的建设中，需要规划走在前面，提前为研究所规划。在我们的实践中，我们提出了包括政策、流程、机制、技术、管理各个环节在内一系列建设指南和最佳实践方案，供研究所参考。

研究所 IR 应当是一个开放系统。国家科学图书馆构建联合的机构仓储系统，目标在于提升全院的信息服务能力，为此我们为研究所 IR 提供了 SRU 等检索接口，使授权用户能够通过 SRU 规范检索仓储系统，提高仓储系统的集成能力。

尽管走了一些弯路，但是通过实践，我们也增强了信心，希望进一步了解和把握研究所的需求，争取他们的支持和配合，尽快在全院范围内逐步推广和部署，并最终形成有一定规模的联合机构仓储服务系统。

参考文献

1 张智雄，林颖,郭少友等.新型机构信息环境建设的思路与框架.现代图书情报技术,2006(3):1-5

2 张智雄,林颖.机构仓储及其在数字图书馆中的应用.2005 海峡两岸图书馆服务发展与创新

4 Dspace. [2007-10-30]. http://www.dspace.org.

5 祝忠明,马建霞,常宁等. 基于 DSpace 构建学科知识库系统的研究与实践,现代图书情报技术，2006(7):10-14

6 马建霞,祝忠明, 王渊命等. 基于 Dspace 构建甘青特有少数民族数字资源保存与服务系统,现代图书情报技术, 2007(1): 53-57

7 SRW/U. [2007-10-30]. http://www.oclc.org/research/software/srw/default.htm.

8 OAIHarvester2. [2007-10-30]. http://www.oclc.org/research/software/oai/harvester2.htm

9 刘勋,祝忠明. DSpace 系统元数据获取功能的实现.现代图书情报技术,2007(4): 17-20

[作者简介]祝忠明，男， 1968年生，研究员，发表论文 30余篇。马建霞， 1972年生，副研究馆员，发表论文 20余篇。张智雄， 1970年生，研究馆员 ,发表论文 60余篇。孙坦， 1970年生，研究馆员，发表论文 40余篇。