●郭少友

机构库建设的若干问题研究

摘 要 机构库建设过程中会遇到服务模式选择问题、法律问题、内容建设问题、资金问题等。其中服务模式有面向资源获取者的服务模式和面向资源提供者的服务模式。机构库建设者应选择适合于自己的模式。知识产权问题包括机构库软件本身的产权问题和研究成果的产权问题。内容建设主要是资源类型的选择和内容的质量控制,机构库存储哪些类型的内容,需要建设者有一个通盘考虑。参考文献 6。

关键词 机构库 机构库建设 服务模式 开放存取
分类号 G258

ABSTRACT In this paper, the author discusses some issues related to the development of institutional repositories, such as service patterns, legal issues, contents, financial issues. 6 refs.

KEY WORDS Institutional repository. Development of institutional repository. Service pattern. Open access.
CLASS NUMBER G258

SPARC 高级顾问、机构库权威专家 Raym Crow 认为,机构库(Institutional Repository,简称 IR)是学术机构为捕获并保存机构的智力成果而建立的数字资源仓库。它有两个重要使命:一是克服现有学术交流模式的弊端,实现研究成果的开放存取;二是长期保存机构的研究成果,并借此体现机构的学术声望、学术水平和社会价值[1]。

本文试图对机构库建设中可能遇到的问题进行分析,并提出一些建设性的意见,以期对国内一些正在进行或将要进行的机构库建设项目有所帮助。

1 服务模式选择问题

1.1 面向资源获取者的服务模式

促进机构的研究成果在研究人员之间免费、快速交流是机构库的一项重要功能,而这项功能是通过为资源获取者提供获取服务来实现的。从国外情况来看,机构库可提供 5 种服务方式,每种方式都允许资源获取者(本文将能从机构库中获得资源的团体和个人统称为资源获取者)获取机构库中的免费资源。第一种,外部引用服务。利用基于 DOI 的永久性保存与利用技术,为机构库中的每个数字对象分配一个永久性的、唯一的标识符,资源获取者通过数字对象的唯一标识符可在任何时候获得指定的信息对象,从而保证其创建的书签、个性化链接等信息永久有效[2]。第二种,通用搜索引擎爬行服务。建立链接页面指向机构库中的所有数据对象,并在此基础上提供一个入口页面,通用搜索引擎的爬行器通过链接页面执行爬行功能,抓取机构库中的所有数据,从而使资源获取者能通过搜索引擎来间接获取机构库中的资源。第三种,元数据收割服务。国外各种实用的机构库,几乎都遵循 OAI-PMH 协议,以 OAI 数据提供者的身份开放自己的数据库,对外提供基于 OAI-PMH 的元数据收割服务。第四种,功能模块层服务。随着 Web 服务技术的成熟和面向服务架构(SOA)的兴起,可以将机构库的各个业务逻辑和功能模块包装成 Web 服务(包括提供元数据收割和查询浏览服务的功能模块),并将这些 Web 服务发布到 UDDI 注册中心,以便其他系统来集成这些 Web 服务,从而实现对机构库中资源的获取。第五种,查询浏览服务。前面 4 种服务方式都是接口化的服务方式,需要借助于第三方系统(如 OAI 服务提供者、搜索引擎)或建立与机构库中信息对象之间的超链接才能使用相应的服务。而查询浏览服务方式是指机构库自身具备了检索功能,允许资源获取者直接进入机构库的检索界面来检索并获取机构库中的免费资源。

对于一个具体的机构库来说,这 5 种服务方式可以单独使用,也可以组合使用,从而形成多种不同的服务模式。欧美的机构库,元数据收割服务和查询浏览服务是两项最基本的服务,绝大部分机构库都提供这两种服务;功能模块层服务和通用搜索引擎爬行服务属于增强型的服务,提供这两种服务的机构库不多;外部引用服务不需要机构库提供专用的服务模

块，只要机构库中每个资源单位具有全球唯一的标识符，就可以被外部系统直接引用。机构库的建设者应根据本机构的实际情况和未来的发展战略，选定若干种服务方式，形成一个较为合理、实用的服务模式。

1.2 面向资源提供者的服务模式

机构库中的资源由机构内的各类人员（本文将能提供研究成果的各类人员统称为资源提供者）提供。如何为资源提供者服务，能将成果准确无误地提交到机构库中，是机构库建设者必须考虑的问题。根据国外机构库的实践，结合我国的实际情况，可有3种服务模式。第一种，分布式模式。这种模式完全由资源提供者上传并管理自己的研究成果[3]。资源提供者可从任意一台联网计算机进入机构库网站的提交页面，自行选择并输入一些关于研究成果的元数据，将元数据连同成果一起提交到机构库中。第二种，半分布式模式。这种模式由机构内的各个单位分别负责协助本单位的资源提供者上传并管理研究成果。一个机构往往包含若干个职能部门或分支机构，资源提供者将研究成果的原件提供给所在部门或分支机构，由部门或分支机构指定的专门人员为研究成果选择元数据，并提交到机构库。第三种，集中式模式。由机构图书馆或机构内指定的其他组织帮助资源提供者上传并管理研究成果。

上述3种模式中，分布式模式最方便快捷，资源提供者坐在家里就可以提交自己的研究成果。但由于机构库往往需要关于研究成果的元数据，而资源提供者给出的元数据往往不准确，从而导致资源获取者检索机构库时的查全率和查准率不高。半分布式模式由于采取部门集中审查、集中提交的方法，可以避免部分人不负责任地或恶意地提交，而且由于一个部门内的研究成果往往属于同一个学科，审查人员一般也具有该学科的专门知识，使得元数据质量和成果质量都有所保证。集中式模式由于机构图书馆的参与，能使机构库的质量得到进一步保证，但与半分布式模式类似，资源提供者对已入库成果的每一次修改，甚至是极小的改动，都要经过相关部门审查，对资源提供者很不方便。机构库建设者应根据机构的实际情况并结合各个模式的优缺点来选择适合自己的模式。

2 法律问题

机构库建设涉及的法律问题主要包括：知识产权问题和资源内容是否违反其他现有的法律法规。

2.1 知识产权问题

建设机构库时既要考虑机构库软件本身的产权问题，又要考虑研究成果的产权问题。

我国的各类组织在建设机构库时，首先需要确定是直接采用国外的现成机构库软件，还是自行开发。根据国际组织OSI（Open Society Institute）提供的一份指南，目前比较成熟的机构库软件主要有Archimede、ARNO、CDSware、Dspace、Eprints、Fedora、i-Tor、MyCoRe和OPUS等[4]，这些软件均可以免费下载、升级和重新分发，并且严格遵循最新的OAI元数据收割协议OAI-PMH2.0。采用这些免费机构库软件时，只要在网站首页的显著位置添加软件指定的、代表软件所有权的徽标，并加上指向该软件所有者网站的链接即可，不需要承担其他任何费用和责任，也不会引起相应的知识产权纠纷。如果采用一些只有购买才能使用的机构库软件或委托软件商开发机构库软件，必须与软件的提供者就知识产权问题达成协议，签署具有法律效力的文件，以免在使用过程中产生纠纷。

对我国的机构用户来说，最好的办法是直接采用国外比较成熟的开放源码软件系统，如Greenstone、CDSware、Eprints、DSpace等。一是因为这些软件的用户较多，已经比较稳定；二是可以直接在这些软件的源码基础上进行适当的、非商业化的修改，以符合自身需要。

机构库中研究成果所涉及的知识产权问题相对较复杂一些。虽然机构库的重要使命之一就是通过开放存取来促进学术交流，但并不排斥受限的存取，允许部分内容通过间接的方式来获取（这些内容的元数据可以免费访问，但全文则需要与机构和资源提供者进行沟通，经过许可后，通过电子邮件来获得机构或资源提供者发送过来的全文）。为此，机构库建设者至少应起草两个协议，即机构与资源获取者之间的协议、机构与资源提供者之间的协议。通过这两个协议来明确机构、资源获取者、资源提供者三者之间的权利与义务。至于协议的内容，则视具体情况而定。但从知识产权的角度来看，一般应至少包含下述内容：对于资源提供者而言，提交的内容不违反或侵犯其他人的版权、专利权和商标权，如果资源提供者在生成资源的时候受到某个组织的赞助，则由资源提供者负责履行相应的义务；对于机构而言，机构有权通过因特网发布资源提供者提交的内容，为了资源的长期保存和便于存取，有权存储、移值、复制或重新排列资源提供者提交的资源，如果发现资源存在违反版

权、商标权和专利权的情况，有权将其删除；对于资源获取者而言，可以免费复制、分发、显示从机构库获取的资源，可以在此基础上生成派生资源，获取的免费资源不能用于商业目的等。

2.2 研究成果是否违反知识产权法外的其他法律法规问题

资源提供者提交的研究成果，除了存在知识产权方面的问题，还可能违反现有的其他法律法规。如果采用半分布式模式或集中式模式，机构可以在上传之前对研究成果进行法律意义上的鉴定，过滤掉与现行法律法规相违背的内容。如果采用分布式模式，完全允许资源提供者自存档，则应该在机构与资源提供者之间的协议中，在关于知识产权条款的基础上进一步要求资源提供者遵守有关的法律法规，如《全国人大常委会关于维护互联网安全的决定》、《互联网信息服务管理办法》等，并保证不利用机构库从事非法活动，如传输非法的、骚扰性的、辱骂诽谤他人的、恐吓性的、伤害性的信息，传输庸俗淫秽、危害国家安全统一的信息等。并在协议中声明如果资源提供者利用机构库进行任何违法或侵权行为，由此导致的民事、行政和刑事责任将全部由资源提供者承担，机构库所有者将依法采取必要措施并向有关机关报告。

3 内容建设问题

机构库的内容建设涉及诸多方面，其中最为重要的是资源类型的选择和内容的质量控制。

3.1 资源类型的选择

机构库应该收录机构成员的研究成果，摈弃没有学术价值的普通资料。Raym Crow 认为，机构库是用来长期保存机构的研究成果的，机构成员的研究成果可以存储到机构库中，而行政记录、与机构的历史和活动有关的一些资料等则可以存储在机构的档案馆里[5]。当然，研究成果和非研究成果之间的界限不是绝对的，哪些内容存储到机构库，哪些内容存储到档案馆，需要机构库的建设者给出便于操作的指导性意见，以避免机构成员将一些没有学术价值的资料存储到机构库中，从而保证机构库的学术质量。

具体存储哪些类型的研究成果，需要机构库的建设者事先考虑。从文件格式看，研究成果的存在形式有图像文件、文本文件、视频文件、音频文件等多种类型，还可以是若干类型的混合体；从出版状态看，研究成果可分为已经在正式刊物上发表的和尚未发表的两种类型；从所隶属的学科门类看，研究成果可分为哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、管理学等多种类型；从学术研究的性质、目的和过程方面看，研究成果可分为基础研究、应用研究和开发研究 3 种类型。笔者通过分析 OAIster 网站（该网站列举了 405 个机构库，网址为 http://oaister. umdl. umich. edu）上列举的部分机构库的帮助文档，发现很多机构库都对所收录研究成果的类型进行了不同程度限定。我国的机构库建设者也应该根据机构的性质和机构库的建设目的，进行相应限定，如确定是否收录多媒体类型的文件，是否收录已经发表的论文，是否同时收录基础研究、应用研究和开发研究 3 个方面的成果，是否允许一些灰色文献（如机构的内部报告、调查报告、研究报告、技术报告、学术会议资料等）进入到机构库，专业性较强的研究机构是否允许专业领域之外的研究成果进入到机构库等。总之，可以在机构库中存储哪些类型的内容，需要机构库建设者有一个通盘考虑。

3.2 内容的质量控制

机构成员的研究成果在正式进入机构库之前，应该采取一定的措施对成果内容质量进行控制，进一步摈弃没有学术价值的成果。

内容的质量控制方法与面向资源提供者的服务模式有关。对于半分布式服务模式和集中式服务模式来说，由于研究成果在入库前有较为严格的审查机制，元数据也由专业人员提取并录入，对入库成果进行有效的质量控制是可以做到的。对于分布式服务模式来说，由于机构库往往采用资源提供者自存档的方式来接收新的研究成果，所以较难解决内容的质量控制问题，只能采取一些辅助手段来提高入库内容的质量。

在分布式服务模式下，内容的质量控制可分成两个级别：元数据级和内容级。

在元数据级，可通过机构库软件来自动控制元数据的质量。机构库建设者在选择现成的机构库软件并对其进行适当的改造或自行开发机构库软件时，应尽量选择通用的、简单的元数据格式，如都柏林核心集；尽量减少需要资源提供者录入的项目；对于一些必须由资源提供者录入的项目，尽量采用让其从列表框中选择数据的方法，如国别的选择、语言的选择等；至于主题词或关键词方面，如果机构的专业面不很宽的话，可以考虑由软件人员和专业人员配合，在列表框中列出主题词或关键词清单供资源提供者选择，同时还允许资源提供者自己键入合适的关键词。

在内容级,可分别由资源提供者、读者(读者是资源获取者的一种)和机构库管理人员对研究成果的质量进行评价,并由资源提供者或机构库管理人员进行必要的质量控制。如果机构库软件具有反馈机制的话,读者可以针对研究成果提出意见或建议并反馈给资源提供者;机构库管理人员也可以定期地、全面地对机构库中的研究成果进行扫描,对有问题的成果给出非强制性的修改意见并反馈给资源提供者。资源提供者可以根据反馈意见和自己的进一步研究,对已经入库的成果进行修改。通过研究现有的机构库软件,笔者发现大致有两种修改方式:一是资源提供者在本地对原有成果修改,然后上传到机构库中,并将原来的成果覆盖;二是资源提供者在本地修改,然后上传到机构库中,以同一成果的新版本形式存在,经过多次修改之后,同一成果将存在多种版本。一些机构库(如由香港科技大学图书馆负责建设的香港科技大学机构库)将第二种方式作为默认的方式,如果想删除以前的版本,必须另发电子邮件提出申请,由机构库管理人员删除[6]。

4 其他问题

4.1 机构之间的协作问题

词表的不统一,导致跨机构库进行检索的效率不高。读者检索某个机构库中的资源可以有两种途径:一是直接进入该机构库的主页,二是通过一些服务提供者(此处指OAI服务提供者)进行检索。第一种途径检索,查全率和查准率比较高,原因是一个机构库往往会要求所有资源提供者采用同一个词表来形成元数据。第二种途径检索,查全率和查准率较难得到保证,原因在于不同的机构库在形成元数据时采用的词表可能不同,同一个词在不同的词表中含义可能不一样,读者通过服务提供者的检索界面进行跨库检索时,查全率和查准率自然就会下降。

目前国外一般都以单个机构为对象进行机构库的创建、信息采集、组织和管理,对跨机构、跨区域分布式机构库的统一组织与控制技术重视不够,已经暴露出一些问题,有关组织正在商讨对策。我国的机构库建设者应该根据机构的性质和任务,与相关机构接洽协商,就未来分布式机构库的有关问题达成协议,共同为机构库之间的协作提供解决方案。

4.2 资金问题

建设机构库之前,机构库的所有者应有明确的资金支持计划,以保证机构库正常运行。有条件的机构可以通过募集的方式来获取建设和运行机构库所需要的资金,对于募集到的资金,机构还应该制定相应的管理办法。

4.3 组织管理问题

机构库的建设将会涉及机构内多方面的人员,主要包括图书馆员、信息技术人员、机构主管领导、档案管理人员、各类研究人员。应该制定合理的合作策略,建立完善的协作机制,以保证机构库的顺利建成和正常运行。

4.4 机构文化的培育问题

机构文化关系到机构库的成败。如何培养机构成员的集体荣誉感,在机构内营造良好的、自由的、开放的学术研究氛围,使各类研究人员愿意将自己的研究成果存储到机构库中,从而丰富机构库的内容,彰显机构的实力与水平,是建设机构库时应考虑的问题。尤其是目前传统出版模式的影响还非常大,更需要在培育机构文化的同时,制定合适的宣传策略,鼓励、吸引更多人员在机构库中存储研究成果。

参考文献

1,5 Raym Crow. The Case for Institutional Repositories: A SPARC Position Paper [EB/OL]. http://www.arl.org/sparc/IR/ir.html(2005-02-20查询)

2 李春旺. 网络环境下学术信息开放存取. 中国图书馆学报,2005(1)

3 ROY TENNANT. Institutional Repositories [EB/OL]. http://www.keepmedia.com/ShowItemDetails.do?item_id=246274(2005-02-20查询)

4 Raym Crow. A Guide to Institutional Repository Software [EB/OL]. http://www.soros.org/openaccess/software(2005-02-25查询)

6 香港科技大学图书馆. Institutional Repository[EB/OL]. http://repository.ust.hk/dspace/(2005-02-25查询)

郭少友 郑州大学信息管理系副教授,中国科学院文献情报中心博士生。通信地址:北京中关村北四环西路33号,中国科学院文献情报中心。邮编100080。

(来稿时间:2005-03-07)