开放会议文献资源的采集和跟踪*

朱 江 张春玲 姜恩波 刘春江 杨 宁

【摘 要】为实现对开放会议文献资源的有效采集和跟踪，调研开放会议文献资源现状，分别从资源类型特点、发布主体、资源发布网站特点、组织方式、资源格式、版权要求等方面对开放会议文献资源情况进行系统分析；根据会议文献资源类型和来源的不同，拟定资源采集的方案，包括采集原则、采集对象、采集参数、采集途径等；最后根据资源出版方式、主体和周期的不同，提出针对性的资源跟踪和发现的方法。

【关键词】开放会议文献 采集方法 跟踪方法 资源发现

Abstract: To efficiently collect and track the open conference papers, the detailed understanding of the status of open conference papers is required. This paper systematically analyzes the open conference papers from the aspect of resource type, publishing subject, site feature, organization method, resource format, the copyright requirements, etc. And then it puts forward the collecting programs of the open conference papers according to the type and source of resources, which includes the collecting principles, objects, parameters and ways. At last, it presents the specific methods of tracking and discovery the open resources in terms of the publishing way, subject and period.

Key words: open conference papers collecting method tracking method resource dsiscovery

1 引言

在学术交流和出版发生巨大变化的今天，部分学术会议上宣读或发表的论文不再结集出版，而是以多种形式直接在互联网上发布，供读者免费使用。笔者将这类资源统称为“开放会议文献资源”。随着这类资源数量的不断增加，学科范围的不断扩展，其学术影响力也在快速提高，已引起图书馆、学术机构和研究人员的广泛关注，部分重要的出版机构也开始涉足开放会议文献出版领域。

由于开放会议文献资源中的大多数资源的生命周期比较短暂，如不及时采集、保存，很可能“灭失”。为系统采集、保存这类文献资源，中国科学院启动了“重要会议开放资源采集与服务系统的建设”项目，在对开放会议文献资源调研、遴选的基础上，实现了对重要的开放会议文献资源的系统采集和保存，并向读者提供免费的检索和全文链接服务。

本文将着重分析开放会议文献的类型及其发布网站的特点，并具体介绍开放会议文献资源的采集、跟踪和发现方法。

2 开放会议文献及其网站内容分析

2.1 开放会议文献资源的类型内容及特点

根据开放获取的定义，开放会议文献资源大体分为 OA 类资源和普通免费资源两大类（见表 1）。OA 类资源主要是指不仅免费，而且有专门机构负责维护、提供长期稳定免费服务的资源；普通免费资源主要指不完全满足 OA 定义，即缺乏专门机构的维护，不能保证长期稳定服务的免费资源。从表 1 可以看出 OA 类资源出版发行一般有固定模式和周期，稳定性更好，正式出版的 OA 资源还有 ISSN 或 ISBN 号，所以对其采集保存，提供服务不会存在太大问题，而普通免费资源就不具备这些特点，这也就使得普通免费资源成为资源采集、保存的研究重点，这部分资源也是本文采集、跟踪研究的重点。

* 本文为中国科学院数字图书馆二期先期启动项目“重要会议开放资源采集与服务系统的建设”、十二五重点建设任务“开放资源服务系统建设”、中国科学院“西部之光”人才项目“四川省重点科研机构科技文献资源开放获取集成系统的研究与建设”的研究成果之一。

表 1 开放会议文献类型内容及特点

	分类	内 容	形式特征	特 点
开放会议文献资源	OA 类资源	正式出版的 OA 资源	有 ISSN 或 ISBN	稳定性更强，有比较固定的出版、发布形式，能得到及时的维护，并能长期、稳定地提供服务
		非正式出版的 OA 资源	没有 ISSN 或 ISBN	
		自存储资源	作者自行提交到各类机构仓储库、学科仓储库中的会议论文资源	
	普通免费资源	互联网开放会议资源	机构网站、会议网站提供的文献资源等	稳定性较差，随意性大，缺少专门的维护，资源生命周期相对短暂，随时都有可能出现不可用或消失的情况
		Web2.0 环境下的会议开放资源	采用会议 WIKI 的形式发布会议文献，科研人员在自己的个人网页、BLOG 上发表的会议文献等	

表 1 的资源分类依据会议文献的发布形式和特点，按照开放会议文献的出版主体还有更详尽的分类形式（见表 2）。

表 2 开放会议文献资源按照出版主体的划分类型

类型	内 容	出版主体	实 例
OA 类资源	正式出版的 OA 资源	OA 出版机构	BMC Proceedings
		商业出版商	Elsevier Procedia
		学协会	IOP Publishing
		学术机构	PoS
	非正式出版的 OA 资源	会议录	USENIX
		高校数据库	斯坦福大学的 Electronic Conference Proceedings Archive
	自存储资源	机构仓储库、学科仓储库	ResearchSpace@ Auckland
普通免费资源	互联网开放会议资源	机构网站、会议网站	ISCRAM
	Web2.0 环境下的会议开放资源	个人主页、blog 等	王应宽博客

从表 2 可以看出开放会议文献资源按照出版主体可以划分为 9 种类型，下面分别从具体实例包含内容说明开放会议文献资源出版方式，以期对其资源有详尽了解，这样才能有针对性地提出资源采集、跟踪的方案。

（1）BMC Proceedings 由 Biomed Central 出版，是一个经过同行评议的开放获取期刊，出版经过评审的会议论文全文和会议摘要合集。内容并不局限于生物医学领域内的某一个具体的学科，所有的会议论文都可以无障碍地出版及获取。它为作者提供了出版大量数据集、插图和动态图画的机会，具有很大的灵活性。同一年出版的会议录为一卷，从 2007 年开始到 2012 年共有 6 卷，每一卷包括不同的 supplement，每个 supplement 是一个会议录，目前包括了 6 卷 20 个会议录[1]。

（2）Elsevier Procedia 是专门出版高质量会议录的在线服务平台。会议组织者可以通过它在 Science Direct 中出版会议论文，Procedia 能使这些论文在世界范围内得到快速传播并免费获取，科研人员可以及时跟踪其研究领域最新的研究动态和发展。所有录用的会议论文都会在 6 个星期内在 Science Direct 中在线出版，它为作者和会议组织者提供了一个快速且经济的出版途径[2]。Procedia 平台涵盖 9 大学科门类的会议录，即：能源科学、物理学、社会行为学、化学、工程学、计算机科学、环境科学、疫苗学、地球与行星学，但其中计算机科学、环境科学两大门类尚未收录任何会议论文。以物理学 Procedia 为例，它为一个 OA 期刊，有独立的 ISSN，出版了 33 卷 42 个会议的

会议录，每卷都可以通过 Science Direct 查找到全文[3]。

(3) IOP Publishing 是英国皇家物理学会（Institute of Physics，IOP）下属的非盈利性出版机构，IOP Science 是 IOP Publishing 的在线期刊服务平台，出版 60 多种电子期刊，其中包括 3 个开放出版会议录：Journal of Physics：Conference Series（JPCS）、IOP Conference Series：Materials Science and Engineering（MSE）、IOP Conference Series：Earth and Environmental Science（EES）。这 3 个会议论文开放出版平台提供了快捷、灵活、经济的会议录出版服务，拥有出版会议论文的高水平团队，所录用的会议论文能在 4 到 6 个星期内在线出版并可以在世界范围内获取，每一年出版若干卷，每一卷是一个会议录。JPCS 从 2004 年开始到 2012 年 4 月已有 370 卷[4]，MSE 从 2009 年开始到 2012 年 4 月共有 35 卷，EES 从 2008 年到 2012 年 4 月共有 14 卷。

(4) PoS 是由 SISSA 支持，SISSA 是意大利第一所提供博士学位的大学，它目前是意大利一流研究机构的代表之一。PoS 是一个出版科学会议论文的平台，充分利用了 SISSA 在在线出版方面的丰富经验[5]，发展了 Journal of High Energy Physics（JHEP）的会议录部分。它的目标是为会议文献的出版提供多功能的、快速的、经济合理且开放的服务，对所有读者免费开放，出版的费用出于非赢利的目的尽可能降到最低。PoS 收录了包括物理学、数学、生物学、计算机科学、医用物理学等在内的所有基础和应用科学。目前已出版了 139 个会议录，有 7 个会议录正在出版中，20 个会议录即将出版。

(5) USENIX[6]，会议录是 USENIX 几大出版物中的一种，USENIX 从 1993 年到现在的会议录全都可以在线获取。任何人都可以免费获取所有在线的会议论文。可以按会议名称或会议举行年份来查看会议录。

(6) 斯坦福大学的 Electronic Conference Proceedings Archive[7] 收录的是高能物理领域及其相关领域的学术会议的会议录，并提供基于网络的电子形式的保存服务。在目前的阶段，该站点由斯坦福线性加速中心（SLAC）的技术信息服务小组来维护，它作为一项免费的服务，为科学团体和科研人员服务，目的是聚合电子形式出版的会议录。

(7) ResearchSpace@ Auckland[8] 是奥克兰大学的数字资源仓储库或者说是在线存储库，包括论文全文和其他研究成果。它包括了 18 个不同集合，其中 B1 – Academic staff – research online 又包括了 6 个不同的收录集合，其中就有 conference papers（open access full text）。在这一会议论文集合中，可以按照会议举行的时间先后、作者、题名和主题来进行浏览，它实际为该大学会议论文的仓储库，因此，针对某一会议的收录情况缺乏系统性，只提取会议录中作者单位是奥克兰大学的论文。

(8) ISCRAM（Information Systems for Crisis Response and Management）[9] 是一个国际性的非盈利性组织。从其网站中可以获知，ISCRAM 从 2005 年起，每一年都要举行一次会议，2007 ~ 2012 年的会议录都可以免费获取，2005 年和 2006 年的会议录则需要登录才可以获得。

(9) Web 2.0 环境下的会议开放资源。王应宽在其科学网博客上发布了一篇名为《首届 OA 学术出版会议在瑞典召开，精彩报告全部视频开放》[10] 的文章，文中给出了整个会议的链接，并将会议报告的内容以超链接的方式呈现给读者，点击可直接观看。同时附上了会议情况简介，将此次会议以博文的形式提供给读者。

2.2 网站内容分析

互联网上的会议开放资源主要分为遵循 OAI – PMH 协议的结构化资源和未遵循 OAI – PMH 协议的非结构化资源。前者可利用支持 OAI – PMH 协议的元数据收割软件来采集、识别，后者则需利用专门的采集软件来采集、分析。后者也是本项目研究的重点。下文将重点对非结构化的会议开放文献资源网站进行分析。

2.2.1 网站及会议的数量对应关系

开放会议网站的形式多样，数量众多，每个网站上发布一个或多个会议及其会议录信息。

仅发布一个会议及其会议录信息的网站多是单届会议的会议网站，这些专门的会议网站以发布会议日期、招稿通知、报到注册通知、会议日程为主，部分网站也发布网络版会议录或大会发言 PPT，如 The Eleventh International World Wide Web Conference[11]。

集中发布多个会议及其会议录信息的网站多是机构网站或会议门户网站。一些机构将本机构举办的各种学术会议的会议录或大会发言 PPT 集中在本机构的网站或专门的会议门户网站上发布，如 International Towing Tank Conference[12]，1978 年以后的会议录都可以免费在线获取。

2.2.2 资源组织方式

开放会议资源列表以 HTML、PDF、其他文档格式的形式存在，每种形式的资源组织方式大至可以分为目次

型、日程表型和整本会议录型。

(1) 目次型。会议文献以目录形式排列，从目次到达文献全文可能存在一级或多级链接。如有的网站点击目录上的题名、责任者或全文图标、链接即可直接查看全文；有的网站点击目录上的题名、责任者，先看到的是文摘，再点击相应的链接后才能查看全文；个别网站需要点击3次或3次以上才能查看全文，如ISCRAM 2008[13]。

(2) 日程表式。会议文献以标题、著者、摘要等形式嵌入会议日程表中，从日程表到达全文也存在一级或多级链接。如Sixth BIS Annual Conference: Financial System and Macroeconomic Resilience[14]，其资源组织方式为典型的日程表形式。

(3) 整本会议录形式。会议文献全文或大会发言PTT收录在一个文件内，从第一个链接到最后一个链接也存在一级或多级链接。如Proceedings of the International Congresses on Education in Botanic Gardens[15]，既提供了整个会议录的链接，又给出了会议录的详细列表。

2.2.3 资源包含的文档格式

绝大多数网站只提供单篇论文一种格式的全文文档（论文全文以PDF格式为主，大会发言以PPT格式为主）；部分网站提供单篇论文的多种格式的全文文档（如PDF、PPT、音频或视频）；个别网站既提供单篇论文的一种或多种格式的全文文档，又提供整本会议录的PDF文件或压缩文件；极个别的只有整本会议录的PDF或压缩文件。

2.2.4 版权要求

绝大多数是没有明确版权要求的，但个别网站（如CEUR Workshop Proceedings等）会有明确的版权要求，版权归会议录编者和著者所有；只能进行个人和学术研究为目的的复制，禁止商业性复制或使用；整卷或单篇论文的转载（再出版，Re-publication）需取得版权所有者（编者、著者或两者同时）的同意；禁止对CEUR-WS.org站点或站点内部分内容的镜像[16]。

3 资源采集方案

3.1 资源采集原则

在尊重知识产权的前提下，拟订不同来源、不同类型开放会议文献元数据和全文资源的采集规则与采集方法，设计合理规范的采集流程和检验、审核等质量控制体系，确定不同来源的相同资源的查重与判定规则。同时制定相应的开放会议文献资源描述、服务描述和资源登记元数据规范。

3.2 采集对象

根据初步研究，网络上开放会议资源的采集对象可分为三类。

(1) 未遵循OAI-PMH协议的开放会议资源，包括未遵循OAI-PMH协议的会议举办或承办机构的机构网站、会议网站上的普通免费资源、正式出版或非正式出版的OA资源。

(2) 遵循OAI-PMH协议的开放会议资源，包括遵循OAI-PMH协议的会议举办或承办机构的机构网站、会议网站上的普通免费资源、正式出版或非正式出版的OA资源、自存储资源等。

(3) Web2.0环境下的开放会议资源，包括个人网页、BLOG等Web2.0环境下的开放会议资源。

3.3 采集方案

(1) 未遵循OAI-PMH协议的开放会议资源，主要利用自行开发的采集软件来采集、分析、标引。

(2) 符合遴选标准，遵循OAI-PMH协议的开放会议资源，主要利用二次开发的OAI-PMH元数据收割软件来采集。

(3) 通过用户参与机制，接受会议代表提交的重要会议文献资源、学协会会员提交的和会员专享的重要会议文献资源（需特别注意评估知识产权风险，并提供尊重知识产权条件下的合理使用途径）。

(4) Web2.0环境下的开放会议文献比较零散，尚未探索出实用的资源跟踪和采集方法，拟暂不采集。

(5) 在尊重知识产权的前提下，对不同版权要求的资源（包括会议代表和会员专享的资源）采用不同的采集方式和服务方式，避免发生侵权。

3.4 采集参数

根据开放会议文献资源的网站结构、页面层级、文档格式等属性，确定开放会议献资源采集参数，主要包括起始网页、域名限定、层级、文档格式等，其中层级、文档格式与链接类型、对象的关系见表3。

表3 开放会议献资源采集参数表

链接类型	链接对象	层级和文档	采集参数
只有单篇论文的链接	单篇论文的链接	一层一种文档	一层，一种文档格式
		一层多种文档	一层，多种文档格式
		两层一种文档	两层，一种文档格式
		两层多种文档	两层，多种文档格式
既有单篇论文的链接，又有整本会议录的链接		多层一种文档	多层，一种文档格式
		多层多种文档	多层，多种文档格式
	整本会议录的链接	一层	一层，一种文档格式
只有整本会议录的链接		两层	两层，一种文档格式
		多层	多层，一种文档格式

PDF格式的整本会议录须拆分后再分析，以便形成单篇论文的全文和著录信息。

3.5 会议开放文献资源信息的采集途径

单届会议及会议录信息：通过用户调查、用户推荐、资源调研等方式获取，再经过工作人员的补充和审校，经过遴选后以重要开放会议资源列表的形式提供，并上载到系统，形成单届会议及会议录信息元数据集，元数据项包括会议名称、会议名称缩写、会议日期、会议地点、会议主办者、会议学科分类、会议地域范围、会议录名称、会议录出版/发布机构等内容，向下与该届会议的每篇论文形成链接，向上与该届会议的会议系列形成链接。

会议系列信息：通过用户调查、用户推荐、资源调研和重要开放会议资源列表形成，经过工作人员的补充和审校后上载到系统，形成会议系列元数据集，元数据项包括去掉届次或年代的会议名称及缩写、会议沿革、主办者、已举办届次、会议频率、下届会议预计举办时间及网址等内容，向下与该会议系列的各届会议及会议录信息形成链接。

会议论文信息：会议论文的题名、责任者、责任者机构、关键词、摘要等信息在资源采集过程中通过HTML页面分析和PDF文本分析方式获取，其中优先选择PDF文本分析获得的题录信息，如会议论文无PDF全文或PDF文本分析失败，则以HTML页面分析结果为准。考虑到部分会议论文格式不规范，如缺少摘要，则用论文的第一自然段或第一节（Introduction）代替，如缺少关键词字段，则空缺（读者用关键词检索时，系统自动从题名和关键词两个字段中检索）。

4 资源跟踪、发现方法

4.1 现有开放会议的新一届会议文献资源的跟踪

4.1.1 在特定网站上连续、成系列发布的开放会议文献资源的跟踪

部分开放会议文献门户网站、学术机构网站、出版机构网站会定期、不定期地发布新近召开的学术会议上宣读或发表的论文，有的甚至连续编号，这类资源的跟踪相对简单，只需定期对这类网站进行扫描，将采集到的会议信息与已采集会议库信息进行比较，或按卷号比较，即可将已有会议的新一届会议或新会议准确识别出来，从而实现跟踪。

4.1.2 零散发布的开放会议文献资源的跟踪

部分开放会议文献的发布比较零散，有的是每一届会议创建一个独立的网站，设立会议通知、召投稿、食宿交通信息、会议论文、PPT、音视频发布等栏目；有的则是每一届会议挂在不同承办机构的机构网页上等等。这类零散发布的开放会议文献资源的跟踪，较难实现自动跟踪，目前根据会议频率，推算下一届会议的举办时间，在该时间段后借助搜索引擎和人工辅助判断来确定下一届会议的发布网址。

4.2 新开放会议文献资源的发现

在特定平台上发布的新开放会议文献资源，可通过4.1的方法跟踪，再与现已掌握的开放会议资源库比对，即可比较准确地发现。

零散发布的开放会议文献资源较难发现，目前主要通过以下途径来发现，并结合人工判断来识别。

（1）综合科技资源登记系统。这是中国科学院国家科学图书馆建设的一个开放的综合科技资源集合元数据登

记系统，目标是针对重要国家、重要领域和重要机构与组织的科学数据、特色资源、教学资源、软件、计划与项目、学术组织、学术会议等资源，规范描述资源、所属机构以及资源访问方式等信息，建立系统化发现、规范化遴选、知识化描述、集成化揭示的综合科技资源服务体系[17]。对开放会议资源的跟踪发现，则可利用其对学术会议信息的登记描述信息，通过计算机扫描或人工分析过滤，对可以实现开放获取的会议文献进行跟踪、采集和保存。

(2) 会议门户网站。会议门户网站也会定期、不定期地发布新近召开的学术会议相关信息，只是相关学术成果、会议论文可能不是完全开放获取的。对这类门户网站信息跟踪和发现还是相对容易的，只是会议成果可能不太容易获得，需要定期扫描加人工识别方法提取可用信息，例如部分免费会议成果或是论文题名信息等。

(3) 学术搜索引擎。学术搜索引擎上资源一般以学术论文、国际会议、全文期刊、学者为主。所以定期对各个学术搜索引擎资源进行扫描，可从学术论文的出处、会议相关记录、学者动态中发现最新会议消息，再借助人工力量鉴别完成对开放会议文献资源的跟踪和采集。

(4) 动态监测系统。采用中国科学院国家科学图书馆开发的机构信息动态监测系统也可对机构主办或承办的学术会议及其开放会议录进行监测和跟踪。

(5) 科研人员和学科馆员推荐。通过网络技术手段获取不到的资源信息，可采取科研人员和学科馆员推荐的方式。不同科研领域研究人员和学科馆员对其擅长领域的相关信息具有敏感性，且对该学科领域前沿动态有一定认识和把握，对其对口专业领域召开的会议信息了解更多，所以建立了资源推荐模块，不定时接受科研人员和学科馆员的推荐。

5 结束语

开放会议文献资源数量庞大、类型多样，发布组织方式复杂，尤其是在网络环境下，随着信息技术的发展，其文档格式、发布方式将更加复杂和多样化，因而必须对开放会议文献资源的类型和发布方式不断探索，并制定适宜的跟踪、采集方案，以保证对各类开放会议文献资源的及时跟踪、采集和保存。

注释

[1] BMC Proceedings . http: //www. biomedcentral. com/bmcproc/, 2012 - 04 - 20

[2] Procedia. http: //www. elsevier. com/wps/find/electronicproductdescription. cws_ home/718334/description#description, 2012 - 04 - 20

[3] PhysicsProcedia. http: //www. elsevier. com/wps/find/journaldescription. cws_ home/714716/description#description, 2012 - 04 - 20

[4] Journal of Physics: Conference Series. http: //iopscience. iop. org/1742 - 6596, 2012 - 04 - 20

[5] What's PoS . http: //pos. sissa. it/POSwhat. html, 2012 - 04 - 20

[6] USENIX . http: //www. usenix. org/publications/, 2012 - 04 - 22

[7] Electronic Conference Proceedings Archive. http: //www. slac. stanford. edu/econf/index. html, 2012 - 04 - 22

[8] ResearchSpace@ Auckland. http: //researchspace. auckland. ac. nz/, 2012 - 04 - 22

[9] ISCRAM. http: //www. iscram. org. /, 2010 - 04 - 20

[10] 王应宽. 首届 OA 学术出版会议在瑞典召开，精彩报告全部视频开放. http: //www. sciencenet. cn/m/user_ content. aspx? id = 259482, 2012 - 04 - 22

[11] The Eleventh International World Wide Web Conference [EB/OL]. http: //www2002. org/CDROM/ROM/, 2012 - 04 - 22

[12] International Towing Tank Conference. http: //ittc. sname. org/proceedings. htm, 2012 - 04 - 22

[13] ISCRAM2008 Proceedings. http: //www. iscramlive. org/portal/node/2236, 2012 - 04 - 25

[14] Sixth BIS Annual Conference: Financial System and Macroeconomic Resilience. http: //www. bis. org/events/brunnen07. htm, 2012 - 04 - 25

[15] Proceedings of the International Congresses on Education in Botanic. Gardenshttp: //www. bgci. org/education/edu_ proceedings/, 2012 - 04 - 25

[16] CEUR Workshop Proceedings (CEUR - WS. org) . http: //sunsite. informatik. rwth - aachen. de/Publications/CEUR - WS/, 2012 - 05 - 01

[17] 综合科技资源集成登记系统 (IRSR) . http: //irsr. llas. ac. cn/aboutus/aboutus. jsp, 2012 - 05 - 01

朱 江 姜恩波 刘春江 杨 宁 中国科学院国家科学图书馆成都分馆。
张春玲 中国科学院国家科学图书馆成都分馆，中国科学院研究生院。