[文章编号] 1004-325X(2011)02-0052-05

开放会议资源的元数据研究*

——以重要开放会议资源采集与服务系统为例

Research on the Metadata of Open Conference Resources

—— Taking Acquisition and Service System of Important Open Conference Resources as an Example

柴苗岭（四川大学公共管理学院 四川 成都 610041）
朱 江 陈漪红 姜恩波（中国科学院国家科学图书馆成都分馆 四川 成都 610041）

[摘 要] 以外文学术开放会议资源为研究对象，对会议网站/数据库、目标资源进行调研和分析发现，开放会议网站的会议内容有限，揭示资源的深度和层次较浅，资源结构相对简单。根据开放会议资源特色，遵从繁简适度的原则，通过对会议资源进行描述而形成的《会议元素名称、释译、描述规范》、《单次会议及会议录元素名称、释译、描述规范》和《会议文献元素名称、释译、描述规范》，可促进以网络为基础的开放资源的公开性、实用性、组织性，实现开放资源共享和跨仓储的无缝查找，有助于辅助用户有效获取有价值的网络会议资源。

[关键词] 开放会议资源 会议录 都柏林核心元数据 采集与服务系统
[中图分类号] G250.73 [文献标识码] B

[Abstract] The analysis of conference website or database and target resource with foreign academic open reference resources as object shows that the conference content of open conference website is limited, the depth and level of revealing resources is shallow, resource structure is simple. According to the characteristic of open reference resources, obeying modest principles to describe conference resources to form ''conference element name, explanation and translation, description standard'',''single conference and proceeding element name, explanation and translation, description standard'',''conference literature element name, explanation and translation,description standard''. It could promote public, practicability, organization of open resources based on network, realize open resources sharing and seamless search of cross database, help users acquiring valuable network conference resources.

[Key words] Open conference resource; Proceeding; Dublin Core Metadata; Acquisition and service system

近年来，随着网络技术的发展，学术研究模式及学术成果发布方式都在一定程度上有所改变。因为物理载体的形式约束，学术文献出版的速度和出版市场的分布都受到限制，进而影响了学术成果的传播和应用。大量的学术协会、会议组织机构将其会议进程、会议内容等资源以免费的OA（Open Access，开放存取）方式在网络上公开发布。我国也针对这些资源进行了研究，并制定了相关的规范，如科技部科技基础条件平台《数字规范与标准建设》项目下的子项目成果《网络资源描述元数据规范》等，但如何将开放的会议资源的非结构化文本进行分析与标引，目前还没有统一的规范。

为了有效研究开放会议资源元数据，本文以2009年中国科学院国家科学数字图书馆二期先期启动项目“重要开放会议资源采集与服务系统”的服务内容——外文学术开放会议资源为研究对象，对相关的元数据进行了研究。该系统主要对开放会议及会议文献进行评价、采集、长期保存并将其提供给用户使用。重要开放会议资源元数据研究即针对开放会议和会议文献进行元素的定义、描述，即将

* 本文系中国科学院国家科学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”的研究成果之一。

各类会议网站发布的会议文献及会议文献本身进行基于形式及语义的理解与识别，抽取其中的会议文献题名、责任者、摘要、全文链接等信息[1]。

1 开放会议资源元数据规范现状

从国内元数据研究现状来看，2002-2006年，国家科技部科技基础条件平台《数字规范与标准建设》项目下的子项目成果——《会议论文描述元数据规范》[2]171-182和《网络资源描述元数据规范》[2]216-236分别在DC（Dublin Core）元数据的基础上对会议论文和网络资源的元数据作出定义。其中《会议论文描述元数据规范》将描述的对象定位为以印刷版和电子版方式出版的会议论文;《网络资源描述元数据规范》将描述的对象定位为网络上可以公开访问的、具有网络标识的资源。此外，2008年国家科技图书文献中心（National Science and Technology Library，简称NSTL）为了提高数据库建设的标准化和规范化水平而修改并细化的《NSTL数据加工规范》对会议论文和会议录进行了专门的描述。

从国际上的网络资源元数据研究现状来看，最具有代表性的元数据收割协议是OAI-PMH（Open Archives Initiative Protocol for Metadata Harvesting，开放存取的元数据收割协议），OAI-PMH在解决异构数据库的元数据互操作上具有容易实现、可多个角色互换等明显优势。而OAI-PMH支持的元数据格式是DC元数据，适合于对网络信息资源进行描述，且DC元数据的开放性决定了其可在15个基本元素的基础上进行扩充，从而增加网络信息资源的特色描述元素。

从现有会议资源和网络资源元数据规范看，国内外还没有可直接供开放会议资源使用的元数据规范，因此以会议资源为目标的元数据研究和元数据规范的建立有其必要性和现实意义。

2 开放会议资源调研及分析

开放会议资源调研主要可分为两部分：（1）现有的会议网站/数据库的调研，即对信息检索入口进行分析和归纳；（2）目标资源的调研，即对会议资源的种类、格式、页面组织方式和深度、资源描述的层次和深度等内容进行调研。

2.1 会议网站/数据库的调研

ISI Proceedings（即ISTP的网络版，ISTP的英文全称是Index to Scientific & Technical Proceedings，科技会议论文索引数据库）、NSTL等国内外会议资源网站/数据库在用户的使用习惯和用户需求上具有较强的代表性，通过对这类网站检索入口的分析，可以了解用户对会议资源的需求及不同种类资源提供者的标引深度。调查的网站/数据库主要有以下4类：

2.1.1 信息资源服务网站

以NSTL[3]为例，NSTL是经国务院批准的基于网络环境的科技信息资源服务机构，其网站收录的会议资源可分为中文会议和外文会议，以印本文献和现有电子会议文献的检索为主，在检索入口词的内容设计上较传统，查询范围包括全部记录、含文摘记录、含引文记录和可提供全文记录。该网站属于综合性网站，在内容上缺少对会议及会议文献的深入解析。

2.1.2 开放获取会议网站

以Proceedings of WASET（World Academy of Science，Engineering and Technology，世界科学和工程技术学院）[4]为例，该网站收录了人文、生物和生命科学、自然科学和应用科学三大种类的会议召开信息和会议论文，发布的会议资源均经过同行专家评议。该网站没有提供检索入口，会议和论文的检索主要通过会议论文目录实现。从会议信息来看，Proceeding of WASET提供了论文提交资料、专题会、会议录、注册费用等6个方面的内容。会议论文提供PDF格式的全文内容。

2.1.3 综合性会议数据库

以ISI Proceedings[5]为例，ISI Proceedings收录了世界上最新出版的会议录资料，提供综合性的会议论文资料，包括学术领域内的会议、座谈会、研讨会以及其他各种会议和会议录文献。该网站的最大特点是检索入口较多，包括主题、标题、文献类型、基金资助机构等共13种检索途径。该网站在检索途径的设置上，除设计了传统的会议文献检索入口外，还对会议文献进行了分类， 增加了包括书评、年表、乐谱、评论等在内的共37类文献。

2.1.4 专业会议文献数据库

以SPIE（The International Society for Optical Engineering，国际光学工程学会）Digital Library[6]为例，SPIE Digital Library会议录汇集了光电方面7个领域的研究成果，是国际著名连续出版的会议文献出版物。会议文献收藏部分以会议录全文、会议引文和文摘为主。检索入口有12项，也可以通过年代、专题会、卷期号等直接检索。

从SPIE Digital Library提供的会议录、文摘和引文来看，会议文献的学科专指度较高（涵盖7个领域），并配以专门的卷号，在传统的会议文献检索词基础上，提供技术、卷期号、引文作者等检索入口，会议文献的检索深度在整体上强于综合性的会议库。

从上述网站检索入口的调研情况来看，检索入口的设计与可提供资源的来源、学科范围与资源标引深度有重要关系。首先，传统的印本/电子会议资源受其载体形式和标引深度的影响，检索途径少、内容传统，可以满足用户对信息的基本需求，但在会议的收录范围和时间维度上有所欠缺。其次，基于全文/全文开放的会议资源利用了现代信息技术，因此标引深度较传统会议资源要强。再次，从学科范围看，综合性学科的会议网站/数据库在学科分类和内容专指度上不如专业性的会议网站/数据库，专业性会议网站/数据库在内容维度上要强于综合性会议网站/数据库。

2.2 目标资源分析

2.2.1 资源类型分析

重要开放会议资源按内容可分为：会议资源、会议录资源和会议文献资源。其中，开放会议文献资源主要包括会议全文、会议文摘、会议陈述、会议声频、会议视频等，多以PDF、WORD、TXT、HTML、MP3、MP4等格式提供给用户。

2.2.2 页面组织方式和深度

开放会议资源列表以HTML、PDF等形式存在，其页面组织方式大致可以分为目次型、日程表型和整本会议录型，从会议资源列表到会议文献全文的揭示又分为多个层级：

(1)目次型　所有会议文献以目录形式排列，从目录到达文献全文可能存在多级链接，如存在从目录直接进入全文页面，或从目录进入文摘再进入全文页面等情况。

(2)日程表型　会议文献以标题、著者、摘要等形式嵌入到会议日程安排中，从日程列表到达全文可能有多级链接。

(3)整本会议录型　某次会议文献收录在一个文件内，从目录到达文献全文可能存在多级链接。

2.2.3 资源描述的层次和深度

会议资源描述包括会议（包括单个会议和连续会议）、会议录、会议文献3个层次的描述。从组织内容上看，因为每个开放会议网站的会议内容有限，虽然提供全文，但揭示资源的深度和层次较浅，资源结构相对简单。

3 重要会议开放资源元数据描述集设置

简单地说，《重要开放会议资源元数据规范》就是对非结构化的文本格式进行识别，抽取题名、责任者、会议名称、会议日期、会议地点、会议主办者等信息形成元数据，以便与遵循OAI-PMH的元数据格式整合，为用户提供高级检索。而元数据的制定需要在设计原则上确定著录对象。

3.1 重要会议开放资源元数据设计原则

3.1.1 兼顾信息的粒度和纬度

信息的粒度反映信息详尽程度，从会议资源描述内容看，主要包括会议资源概括性信息和会议资源具体信息。从信息维度看，会议资源描述要兼顾时间、内容和形式。

3.1.2 简单性与重要性原则

元数据设置时要兼顾简单性与重要性，对用户来说就是标引内容易于理解，采用的元数据又可以完全揭示资源的特性和特征。

3.1.3 元数据的通用性和著录深度适当原则

在选择元数据时要考虑到用户在使用印本会议文献时的使用习惯，对网络上已经揭示了但对资源揭示没有意义或者过于专深的元数据要酌情采用。

3.1.4 用户需求原则

元数据的设计主要是向用户揭示资源特征，便于用户使用，在设计过程中要从用户角度建立元数据元素，以用户需求为目标。

3.2 资源著录的范围和特征

根据资源调研的结果将著录对象分为会议、会议录、会议文献，资源著录具有以下特征：

3.2.1 会议与会议录之间的关系特征

会议可能是连续性的，有其固定的组织机构，每届会议由同一或不同机构组织。因此，一个会议可能由多级组成，如一个会议会包含多个卫星会议。因此，一个会议可能有多个会议录，一个会议录可能对应到多个会议。一个会议录可能在网络上以开放形式出现，也可能同时以电子/印本期刊、会议录的形式出现。

3.2.2 会议文献的格式特征

从目前已检索到的开放会议文献情况看，会议文献的目次形式包括整本文集汇集和日程表两种，会议内容以单篇或文献集的形式存在于PPT、PDF、WORD、HTML等各类声音和音像文件中，在会议文献中存在单篇文献隶属于多个会议录/多个会议的情况。会议文献的获取可能来自会议录，也可能来自机构仓储。

3.3 元素集的设置

在元素集设计上，遵从繁简适度的原则，根据资源调研结果和目标资源的形式特征，在DC元数据、《NSTL文献数据加工元规范》[7]、《会议论文描述元数据规范》和《网络资源描述元数据规范》的基础上对会议资源进行描述，形成《会议元素名称、释译、描述规范》、《单次会议及会议录元素名称、释译、描述规范》和《会议文献元素名称、释译、描述规范》。重要开放会议资源元数据描述集中包含3个元素集，这3个元素集之间的关系如图1所示。

图1　重要开放会议元数据描述集中各元素集之间的关系

其中，会议元素集是单次会议、会议录元素集的上位集，会议文献元素集是单次会议、会议录元素集的下位集。为了更好地揭示开放会议的规模和连续性，在会议元素集设计中采用了会议频次、会议起始时间和会议召开总届数等元素（见下页表1），便于用户对会议进行评价。

在单次会议及会议录元素集中，收录了具有通用性的核心元素，并兼顾单次会议和会议录信息的揭示。如下页表2所示，单次会议、会议录元素集在包含了传统的会议检索元素之外，根据网络上开放会议文献特征增加了会议类型，即该会议属于学术研讨会（seminar）还是工作研讨会（workshop）。

表1　会议元素集

元素修饰词	编码体系修饰词	必备性	可重复性
会议正式名称		是	否
会议其他名称		否	是
会议缩写名称		否	否
会议主办机构		是	是
会议自由关键词		是	是
会议机构描述		是	是
首次会议举办时间	W3C-DTF	是	否
会议地址	URI	是	否
会议语种	ISO 639-2	是	是
会议地理等级范围		是	否
会议举办频次		是	否
会议举办界数		是	否
会议权限申明	会议地址中出现的会议权限管理声明	是	否

表2　单次会议及会议录元素集

元素修饰词	编码体系修饰词	必备性	可重复性
会议正式名称		是	否
会议其他名称		否	是
会议缩写名称		否	否
会议录题名		是	否
会议主办者		是	是
会议录作者		是	是
会议录主题词		是	是
会议自由关键词		是	是
会议主题分类	中国科学院图书馆图书分类法	是	是
会议录分类号	其他项目组的分类体系方法	是	是
会议摘要		是	是
会议录摘要		是	是
会议类型		是	否
会议录出版者		是	是
会议日期	W3C-DTF	是	否
会议录 ISSN	ISSN	否	否
会议录 eISSN	eISSN	否	否
会议录 ISBN	ISBN	否	否
会议录 eISBN	eISBN	否	否
会议来源	URI	是	否
会议录来源	URI	是	是
会议录语种	ISO 639-2	是	是
主会议(is part of)	URI	否	否
卫星会议(has part of)	URI	否	是
相关会议录名/会议文集名	URI	否	是
会议地点		是	是
会议空间范围		是	是
会议权限申明		是	否
会议录权限申明		是	否

会议文献元素集 （见表3）包括文献类型、格式、与会议和会议录元数据的连接、文献获取权限等内容。此外，为了与中国科学院国家科学数字图书馆其他项目更好地融合，在编码体系中拟采用其他项目组的研究成果，如“会议分类”，从而实现项目间的无缝连接。

表3　会议文献元素集

会议资源元素修饰词	编码体系修饰词	必备性	可重复性
题名		是	否
交替题名		否	是
会议文献作者		是	是
会议文献作者机构		是	是
会议文献作者E-mail		是	是
会议文献关键词		否	否
会议文献自由关键词		否	否
会议文献分类号	其他课题组确定的分类法	是	是
会议文献摘要		否	否
会议文献系统摘要		否	否
会议文献发布版本		是	否
出版地		是	是
会议文献发布日期	W3C-DTF	否	否
会议文献类型		是	是
会议文献格式		是	是
会议文献统一标识符	URI	是	否
会议文献页数		否	否
会议文献语种	ISO 639-2	是	是
会议组成部分(is part of)	URI	是	是
会议录组成部分(is part of)		是	是
会议连续出版物(is part of series)	ISSN	是	是
会议文献访问限制		是	是

4　结　语

网络资源的动态性、异构性和开放性限制了会议资源的采集和利用。研究和创建开放会议资源元数据标准、规范有助于促进以网络为基础的开放资源的公开性、实用性、组织性，实现开放资源共享和跨仓储的无缝查找，有助于辅助用户有效获取有价值的网络会议资源。但如何建立准确、完整的元数据规范，还有待大量数据检验和后期的完善。

参考文献：

[1] 朱　江，尚玮娇，姜恩波，等.会议文献开放资源采集与服务系统的研建[J].情报理论与实践，2010(7):117-119.

[2] 肖　珑，赵　亮. 中文元数据概论与实例[M]. 北京:北京图书馆出版社，2007.

[3] 国家科技图书文献中心成都镜像站[EB/OL]. [2010-06-05].http://beta-cd.nstl.gov.cn/facade/search/searchByDocType.do?subDocTypes=C01,C02&name_chi=会议.

[4] World Academy of Science, Engineering and Technology [EB/OL]. [2010-06-05].http://www.waset.org/proceedings.php.

[5] ISTP [EB/OL].[2010-06-05].http://www.isiknowledge.com/?

[文章编号] 1004-325X(2011)02-0056-04

全国图书馆联合编目中心书目数据质量控制

Quality Control of Bibliographic Data from Online Library Cataloging Center

周小敏 （广东省立中山图书馆 广东 广州 510110）

[摘 要] 全国图书馆联合编目中心的书目数据存在着部分分类号和主题词错误、责任方式错误等质量问题。这些质量问题的产生原因是没有统一的著录细则、数据监控不完善、数据质量把关不严、编目员和校对员的综合素质和职业水平良莠不齐。全国图书馆联合编目中心应采取如下策略提高数据质量：制定统一的中文图书机读目录数据处理细则；建立奖惩制度，加强对成员馆的质量监控；培养优秀的编目员、校对员，加强审校机制；建立月结报送制度，定期向联编中心提交反馈的数据质量问题。

[关键词] 全国图书馆联合编目中心 书目数据 质量控制

[中图分类号] G254.362 [文献标识码] B

[Abstract] There are some quality problems with the bibliographic data from Online Library Cataloging Center(OLCC), such as errors of classification number, subject headings and responsible manner. Reasons for these problems are lack of uniform bibliographic details, perfect data monitoring, strict data quality control, accordant overall quality and professional level of catalogers and proofreaders. OLCC should take following strategies to improve data quality: formulating a unified Chinese MARC data book details; establishing clear reward and punishment system, strengthening the quality control of member libraries; training excellent catalogers and proofreaders, strengthening the review mechanism; establishing monthly billing and reporting working procedure, referring the feedback of data quality issues to OLCC regularly.

[Key words] Online Library Cataloging Center (OLCC); Bibliographic data; Quality control

全国图书馆联合编目中心（以下简称联编中心）的建立不仅使全国图书馆的书目数据资源共建共享，同时也使各图书馆结成一个全国范围的文献保障体系，为馆际互借、文献传递提供了平台。但这一切都以统一的、规范的和高质量的书目数据为依托的。笔者在实际工作中发现，联编中心书目数据库中存在很多有问题或有错误的书目数据，这会造成很多不良的影响：①就联编中心而言，错误的书目数据或由多卷册与丛书的著录不同造成的重复数据给数据用户使用带来困惑和不便，使用户需要仔细检查才能鉴别清楚，这不仅给数据用户的工作带来了麻烦，也不利于

DestApp=ISIP&locale=zh_CN.

[6] SPIEDigitalLibrary[EB/OL].[2010-06-05].http://www.spiedl.org/.

[7] 张建勇．文献数据库数据加工规范[M]．北京:知识产权出版社，2009.

[作者简介]

柴苗岭 女，1978年生，四川大学公共管理学院硕士研究生，现工作于中国科学院国家科学图书馆成都分馆，研究方向为开放资源建设。

朱 江 男，1968年生，硕士，现工作于中国科学院国家科学图书馆成都分馆，研究方向为资源建设和服务，已发表论文50余篇。

陈漪红 1963年生，本科，现工作于中国科学院国家科学图书馆成都分馆，馆员，研究方向为学科服务，已发表论文20余篇。

姜恩波 1972年生，本科，现工作于中国科学院国家科学图书馆成都分馆，研究方向为网络信息组织，已发表论文10余篇。

[收稿日期：2010-08-30]