中文图书书目记录上传 WorldCat 实践*

□程颖　张耀蕾　刘孝平　涂艳玲

摘要　馆藏资源书目记录上传至 WorldCat 可大力提升图书馆馆藏在全球的显示度。武汉大学图书馆已成功将 997892 条中文图书书目记录上传至 WorldCat。基于该项目的实践，文章研究了书目记录上传时面临的困难、挑战及解决对策，并分三个阶段详细介绍了项目实施过程与实践经验，重点介绍了记录格式批转换的方法，以期为其他图书馆上传书目记录提供参考与借鉴。

关键词　书目记录　格式转换　数据维护　记录上传　WorldCat
分类号　G250
DOI　10.16603/j.issn1002－1027.2021.02.008

WorldCat 是联机计算机图书馆中心（Online Computer Library Center，OCLC）管理的目前世界上最大的书目和馆藏信息数据库[1]，有 4.15 亿书目数据和 26 亿多条馆藏记录[2]。OCLC 已与世界上越来越多的图书馆签署了书目记录上传 WorldCat 项目协议，促使图书馆将馆藏资源的书目记录上传至 WorldCat。匹配成功的记录在 WorldCat 中显示所属馆馆藏，以提升馆藏在全球的显示度。但上传项目是一项复杂工程，涉及大量书目记录格式批转换，其技术难度之高，任务之繁重，对图书馆挑战颇大。

1　实践及研究现状

目前国外有英国、德国、法国和澳大利亚等国的国家图书馆，以及牛津大学、斯坦福大学、悉尼大学等多家著名大学的图书馆都将书目记录上传 WorldCat。OCLC 也重视与中国图书馆界的合作，已与国内 10 余家图书馆和机构签署书目记录上传 WorldCat 项目合作协议，旨在向全世界推广中国的文化资源。国家图书馆于 2009 年率先将 350 多万条书目记录上传 WorldCat，当年国家图书馆的国际馆际互借和文献传递量比上一年翻了近一倍。随后，中国高等教育文献保障系统（China Academic Library & Information System，CALIS）将 1987－2001 年中国出版的书刊书目记录上传 WorldCat。我们在 WorldCat 中检索得到国内 10 家图书馆和 CALIS 的书目记录上传 WorldCat 的情况（见表 1）。从上传机构看，主要为公共图书馆，高校图书馆和其他机构较少；从记录数量看，大多数机构上传量均较大，其中国家图书馆上传中文记录量和总量均最多；从记录语种看，国内图书馆上传书目记录主体是中文资源，超过一半的图书馆上传中文记录量均超过上传总量的 90%，其中南京图书馆甚至达到 99.89%。

表 1　国内图书馆或机构书目记录上传 WorldCat 的情况

编号	机构	上传记录总量	中文记录量	中文记录占比
1	国家图书馆	4231065	2935068	69.37%
2	上海图书馆	2333737	1515182	64.93%
3	广州图书馆	1732168	1572674	90.79%
4	杭州图书馆	1167212	1140910	97.75%
5	武汉大学图书馆	1083621	997892	92.09%
6	南京图书馆	639877	639170	99.89%
7	CALIS	463917	458560	98.85%
8	首都图书馆	356074	298162	83.74%
9	晋江图书馆	254209	250414	98.51%
10	清华大学图书馆	219995	61782	28.08%
11	浦东图书馆	10896	10789	99.02%

数据来源：2020 年 3 月 26 日检索 WorldCat 数据库获得。

文献调研发现，目前探讨书目记录上传 WorldCat 的文献鲜少，只有 OCLC2018 年中国区年会综述中简要介绍了 4 家公共图书馆的上传情况[2]，但暂无

* 湖北省高校图工委科研基金研究项目“数字时代图书馆资源发现系统比较研究与选择策略”（编号：2014ZD08）的研究成果之一。
通讯作者：程颖，ORCID：0000-0002-6062-147X，邮箱：00008228@whu.edu.cn。

详细介绍具体实践方面的文章。关于书目记录上传 WorldCat 项目的技术难关——CNMARC 到 MARC21 格式批转换，则有一些相关研究，如艾金勇实现了 CNMARC 到 USMARC 格式的自动转换，但未实现批量自动转换[3]；孙华等开发出 CNMARC 与 USMARC 格式自动互换系统，但必须满足条件的记录才生成 USMARC 文件[4]。国内书目记录上传项目需实现绝大多数记录 CNMARC 到 MARC21 格式的批转换，且批转换仅为上传项目中众多环节之一，故需对上传项目的整个实施过程进行深入研究。本文基于武汉大学图书馆馆藏的中文图书书目记录上传 WorldCat 的实践，详细介绍了项目实施过程，总结的实践经验可为其他图书馆提供参考和借鉴。

2 面临的困难及对策

2017 年 OCLC 与武汉大学图书馆签署书目记录上传 WorldCat 项目协议，由武汉大学图书馆将所有馆藏资源的书目记录上传至 WorldCat，以提高武汉大学图书馆馆藏的显示度。武汉大学图书馆首先确定了项目的两个目标：一是实现书目记录批量转换为 OCLC MARC21 格式，以提高格式转换的效率；二是借上传项目的契机，维护本馆书目库中记录，以进一步提升图书馆书目记录质量。

2.1 项目组织

目前国内图书馆通常采取两种方式实施上传项目：一种是与软件公司合作开发格式批转换软件。此方式花费不少经费，且仍需一些人工维护。另一种是由图书馆独立承担该项目。由馆员编写格式批转换程序，使转换规则与转换程序契合更加紧密，但图书馆需有较强的技术力量。武汉大学图书馆采取了第二种方式，并组建项目团队，主要包括三类人员：一为中文编目员，负责格式转换规则 CNMARC 部分内容并维护 CNMARC 记录；二为西文编目员，负责格式转换规则 MARC21 部分内容并维护 MARC21 记录；三为编程人员，负责编写格式批转换程序和记录批维护程序，及批量生成拼音等技术性工作，该类人员从编目员中挑选有计算机专业背景且熟悉中西文编目的人承担。三类人员发挥各自专长，紧密配合以保障项目顺利实施。

2.2 格式批转换

将中文图书书目记录从 CNMARC 格式批转换为 MARC21 格式为上传的重点及难点，挑战主要来自两个方面：一是制定格式转换规则，该规则旨在保证绝大多数书目记录都能被批转换。由于 OCLC 的 MARC21 记录增加了 880 字段以进行字段对应，因此，转换规则不仅需考虑 CNMARC 到 MARC21 字段的映射关系，还需生成 880 字段。二是依据转换规则实现批转换。格式转换非常复杂，OCLC 未提供图书馆批转换的软件，故图书馆需寻找适合的编程语言去实现批转换。

制定转换规则主要从两个角度考虑：一是从图书馆专业角度，需满足 CNMARC 到 MARC21 格式的充分融合，即 CNMARC 中的字段能正确地映射到 MARC21 字段，而不漏掉有用的字段信息；且无论 CNMARC 字段拆分或合并，均能合理组织字段内容以生成新的 MARC21 字段。二是从计算机编程角度，转换规则应具有三种特性。其一，可实现性，即不会因规则的漏洞或歧义而导致无法编程实现。其二，包容性，即非常规著录记录也可通过“包容”规则进行批转换，以尽可能转换更多记录。其三，互斥性，即各规则条件之间互斥，不存在重叠描述的情况，以避免记录同时满足多个条件而产生重复字段。

为实现批转换我们选择可扩展标记语言（Extensible Markup Language，XML）作为格式转换的过渡语言，并用可扩展样式表转换语言（Extensible Stylesheet Language Transformations，XSLT）编程。XML 为一种结构化标记语言，可用“<field>…</field>”和“<subfield>…</subfield>”结构将 CNMARC 记录表示为 CNMARC XML 格式。XSLT 为一种样式转换标记语言，可将一种格式 XML 文档转换成另一种格式 XML 文档。因此，可用 XSLT 语言编程将 CNMARC XML 格式批转换为 MARC21 XML 格式，从而实现 CNMARC 到 MARC21 格式的批转换。

2.3 数据维护

数据维护需平衡好批量维护和人工维护的关系。武汉大学图书馆编程人员全程跟随数据维护，首先从技术上考量批维护的可行性，考量依据包括问题记录的量是否达到一定规模及能否提取出问题的共性特征，满足批维护条件则制定维护规则。再通过工具、平台及编程等方法“混搭”运用来实现维护规则。当问题记录属于个性问题，或记录量较少时，则人工维

护记录。通过批量维护和人工维护的双重措施，在提升数据质量的前提下提高数据维护效率。

数据维护贯穿上传项目的整个过程。在格式批转换前，我们对武汉大学图书馆书目数据库中的数据问题进行了清理；在格式批转换中，对造成转换程序中断和报错的有问题记录进行维护；在格式批转换后，对需正式上传的记录进行最后的数据完善。

2.4 数据存储

上传项目中存储每个处理操作得到的数据尤为必要，有利于以后发现问题时可根据存储的数据追溯问题的原因。我们主要采取两种方式存储数据：一种在 ALEPH 测试服务器中，用中文临时库存储原始 CNMARC 记录及批量生成拼音后的 CNMARC 记录；用西文临时库存储正式上传 WorldCat 的 MARC21 记录。另一种在资源管理器中，将 ALEPH 之外各处理操作得到的记录文件依次编号存储。

3 实施过程

武汉大学图书馆从整体上把握项目全局，构建了中文图书书目记录上传 WorldCat 的实施过程（见图 1）。并将实施过程分为三个阶段：第一阶段为格式批转换做好各项数据预处理工作；第二阶段全力进行 CNMARC 到 MARC21 格式批转换；第三阶段将记录完善后正式上传 WorldCat。以下分阶段介绍项目的实施过程。

3.1 第一阶段：格式批转换前对数据的预处理

3.1.1 制定格式转换规则

首先由资深中文编目员和西文编目员共同商讨拟出格式转换规则初稿。再由编程人员从规则的逻辑性、严谨性及程序可实现性等方面审核，并与中文编目员和西文编目员商讨，依照初稿，程序的运行结果及带来的问题，三方共同修改规则。然后，编程人员依据规则编程测试，并分析程序运行结果能否实现规则。最终经反复修改后确定了转换规则。

项目制定了一个全字段转换规则主表和若干辅助说明附表来共同描述格式转换规则。(1)全字段转换规则主表。确定 CNMARC 到 MARC21 字段的映射关系、MARC21 字段内容的组成、字段重复性及指示符取值等，示例见表 2；(2)复杂字段转换规则附表。对于主表无法详尽描述转换规则的复杂字段，则专门增加其转换规则附表，如 CNMARC 的 200 字段到 MARC21 的 245 字段映射有多种情况，我们增加了

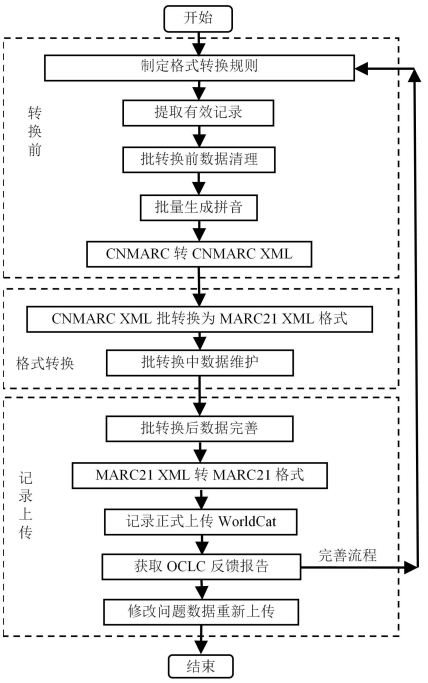

图 1 中文图书书目记录上传 WorldCat 实施过程

200 到 245 字段转换规则附表；(3)控制字段规则附表。该表指定 MARC21 LDR 字段各字符位的取值；(4)定长字段规则附表。该表确定 MARC21 定长字段各字符位的取值方法；(5)重复字段规则附表。CNMARC 很多字段可重复，而 MARC21 很多字段不可重复，该表规定了 CNMARC 重复字段的内容以符号间隔方式合并入 MARC21 字段中的方法；(6)首字母大写的子字段附表。该表汇总了所有首字母需大写的 MARC21 子字段。制定以上规则为编写格式批转换程序提供了依据。

制定格式转换规则需理清 CNMARC 到 MARC21 字段的映射关系。(1)一对一关系，即一个 CNMARC 字段只能映射到一个 MARC21 字段，如 CNMARC 的 690 字段只映射到 MARC21 的 084 字段；(2)一对多关系，即将一个 CNMARC 字段拆分为若干部分后分别映射到多个 MARC21 字段，如将 CNMARC 的 205 字段拆分后映射到 MARC21 的 250 和 880 字段（见表 2）；(3)多对一关系，即将多个 CNMARC 字段合并以生成一个 MARC21 字段，如根据 CNMARC 的 100、102 和 105 三个字段取值生成一个 MARC21 的 008 字段；(4)多对多关系，即由多个 CNMARC 字段生成若干 MARC21 字段，如根据 CNMARC 的 711 和 712 字段次序和指示符值生成 MARC21 的 110、111、710、711 及 880 字段。

表 2　全字段转换规则主表示例（以 205 字段为例）

CNMARC						MARC21						
字段名	重复性	指示符 1	指示符 2	子字段名	自动生成拼音	字段名	指示符 1	指示符 2	子字段名	重复性	子字段内容	备注
205	否	#	#			250	#	#	6	否	880－排序号	
				A	否				a	否	$aA，#B.	字段最后以点结束
				B	否							
						880	#	#	6	否	250－排序号	
				a	是				a	否	$aa，#b.	字段最后以点结束
				b	是							

3.1.2　提取有效记录

为从武汉大学图书馆书目数据库中提取有效馆藏的中文图书书目记录，项目主要采取了以下方法：

（1）分割记录文件。我们需找到合适的记录数分割量，将 100 多万条书目记录分割为若干个记录文件，以便后续分批次处理记录。确定记录数分割量必须满足两个条件：一是保证分割后的记录文件在 ALEPH 里能正常进行批量上传、批量生成拼音等批量操作，不会因为记录数过多而造成 ALEPH 系统空间爆满或停止响应。二是保证分割后的记录文件不会耗费过长时间进行记录格式的批转换。我们将书目记录按 1 万、2 万……20 万等记录量进行各种批处理操作测试，并根据以上两个条件及批操作时间，最终确定以每 5 万条记录数将武汉大学图书馆中文书目库中的记录进行分割，从而一共分割为 24 个记录文件。

（2）过滤语种。武汉大学图书馆中文书目数据库中除中文记录以外，还有日文、藏文、韩文等其他语种记录，而且有些语种 OCLC 还不兼容，故需将这些语种记录过滤。我们定义语种过滤关系式为“中文 —（日文 OR 俄文 OR 维吾尔语 OR 哈萨克语 OR 柯尔克孜语 OR 藏语 OR 韩文）”，并在 ALEPH 后台用 SQL 语句实现关系式以提取出中文书目记录。

（3）提取有效馆藏。首先清理采访记录，将 CNMARC 记录中有馆藏字段和采访字段的记录批量下载，由中文编目员审核后删除采访字段。然后，通过临时记录、屏蔽记录和删除记录的特征字段将它们排除。最后，通过过滤单册处理状态将单册状态无效的记录排除。

通过以上处理，我们提取的有效馆藏的中文图书书目记录占武汉大学图书馆中文库记录的 86.65%。

3.1.3　批转换前的数据清理

在书目记录格式批转换前进行数据清理，可为批转换扫清障碍。数据清理主要解决四个方面的问题：

（1）清理记录著录问题。首先在 ALEPH 中用 CCL 语言描述问题而得到问题记录。再制定问题记录处理规则，并依据规则用 ALEPH 的 FIX 脚本编程批处理记录，不能批处理的记录则手工维护。

（2）将 CNMARC 中全角符号和阿拉伯数字改为半角。首先汇总 CNMARC 中的全角符号和阿拉伯数字，并制定符号和数字的全半角映射规则。若全角中文符号有对应的半角符号，则直接转换，如“＜＞”改为“〈〉”。若全角符号无对应的半角符号，则依据 MARC21 著录规则修改，或用形近符号代替，如依据 MARC21 规则而删除中文书名号。对于全角阿拉伯数字，则直接替换为半角。

（3）修改 OCLC 不兼容字符。由于 OCLC 不兼容罗马字符和拉丁字符，改用形近或意近的兼容字符代替，如将罗马字符“Ⅰ”至“Ⅻ”分别用形近的英文字符“I”至“XII”代替。

（4）可提前处理的转换规则先执行。中英文字符的转换规则可先行处理，如 CNMARC 215 字段中的“页”字提前批修改为英文“pages”，这比拼音更能揭示记录原貌。

3.1.4　批量生成拼音

武汉大学图书馆中文图书 CNMARC 记录的非主要字段没有生成拼音，而 MARC21 记录除 880 字段其他字段均采用汉语拼音，故需对武汉大学图书馆百万条中文图书 CNMARC 记录自动批量生成拼音。ALEPH 系统有生成汉语拼音功能，项目选用 ALEPH 自动批量生成拼音。可在 ALEPH 测试服务器的拼音参数文件中事先设置好所有需产生拼音的字段和子字段，将中文图书 CNMARC 记录上传

ALEPH 测试服务器的过程中调用拼音例程，则可批量生成拼音。另外，还需处理汉语拼音的多音字。规定取多音字第一个拼音。在 UltraEdit 中按照多音字结构编写正则表达式，再提取表达式第一个参数，从而批量提取出所有多音字的第一个拼音。

3.2 第二阶段：格式批转换

3.2.1 格式批转换

我们在 ALEPH 中将生成拼音后的 CNMARC 记录转换为 CNMARC XML 格式，并依据转换规则编写 XSLT 程序将 CNMARC XML 格式批转换为 MARC21 XML 格式。在编程中重点解决了以下几个问题：

（1）多对多字段的映射。该映射需依据 CNMARC 多个字段的存在情况、字段内容及指示符值等多个条件组合来生成多个 MARC21 字段，据此，在实践中采用多条件语句实现。编程时，一方面按照转换规则从严到宽次序编写各条件语句，避免先运行宽松条件，而不运行严格条件的情况。另一方面，在多条件语句的最后，我们增加“其他情况”语句，以批转换非条件组合之外其他著录情况的记录。

（2）中西文内容的分割。有的 CNMARC 字段需拆分为两个部分，中文部分放在 MARC21 的 880 字段中，西文和汉语拼音部分则放在 MARC21 的西文字段中，因此，需编程分割 CNMARC 字段中、西文部分的内容。在实践中依据相应子字段能否产生拼音来判定其为中文，抑或为西文。若产生拼音，则为中文，放 880 字段中；若未产生拼音，则为西文。

（3）子字段间标识符的生成。很多 MARC21 子字段间都有标识符，如 245 的 $a、$b、$d 等子字段间均有标识符。在实践中依据 CNMARC 子字段存在情况确定对应 MARC21 子字段间标识符，然后编程用文本写入方式直接将标识符写入各子字段间。

项目编写了 6 个 XSLT 程序来逐步实现格式批转换（见图 2），其中程序二和程序五为主要程序。

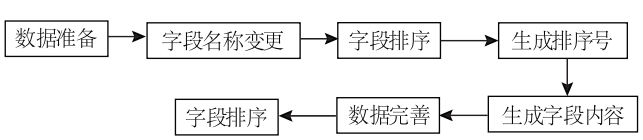

图 2 格式批转换程序执行流程图

程序一，数据准备程序。在格式批转换前对数据进行处理，使其更符合批转换的要求，如处理 OCLC 不兼容的非拉丁字符。

程序二，字段名称变更程序。将 CNMARC 字段名称变更为对应的 MARC21 字段名称。进行重复子字段合并、中西文内容分割、拼音首字母大写等 CNMARC 字段内容的处理，以及生成控制字段、定长字段及固定内容的字段。

程序三，字段排序程序。对 MARC21 数字字段从小到大进行排序。我们两次调用该程序：一次为生成连接字段排序号而排序；另一次在批转换过程末尾再次排序，使字段排列更整齐。

程序四，生成连接字段排序号程序。880 字段与对应字段通过 $6 子字段连接，$6 子字段包含“连接字段排序号”。我们在程序三字段排序的基础上，对排序位置进行自动编号，从而生成连接字段排序号。

程序五，生成 MARC21 字段内容程序。将字段拆分为 880 字段及其对应字段，生成各字段的内容，并实现多对多字段映射，及添加子字段间标识符。

程序六，数据完善程序。对前面仍未解决的问题进行最后的完善，如根据生成的 MARC21 字段修正指示符值。

3.2.2 批转换中的数据维护

当 XSLT 批转换程序中断运行时，可根据报错信息定位记录的断点，并维护有问题的记录。从报错情况看，记录问题主要有两种：一种出自指示符。当指示符值有误时，批转换程序找不到匹配的指示符，因而中断运行。可用 UltraEdit 的批量替换功能将错误的指示符批量替换为正确值。另一种问题出自重复子字段。当不能重复子字段著录为重复子字段时，批转换程序也会中断运行。在 UltraEdit 中编写正则表达式脚本批量提取出此类问题记录，再由中文编目员手工维护。批转换程序可帮助发现记录问题，有助于改善记录质量。

3.3 第三阶段：记录上传

3.3.1 批转换后的数据完善

批转换后西文编目员对 MARC21 记录进行最后的数据完善。该完善主要审核格式转换问题，并对未被正确转换的记录进行手工修改。在实践中也会根据前面批次上传的经验提前修改上传后会报错的数据，包括删除大套书、替换 OCLC 不兼容字符等。另外，也会修改自动生成拼音时产生的乱码，及未成功提取多音字第一个拼音的问题。

3.3.2 正式上传记录

书目记录文件正式上传 WorldCat 前需对其重新命名。其命名一方面要遵守 OCLC 对上传文件

的命名要求，如文件名必须包含馆藏 ID 号、OCLC 图书馆机构代码、上传日期等项，且不使用空格、“～”“!”等特殊字符；另一方面需增加本地的说明项，以便后续从 OCLC 反馈报告中快速定位有问题的记录。据此，在实践中规定上传文件名的结构为“馆藏 ID.OCLC 图书馆机构代码.来源数据库.资源类型.语种.批次.记录数量.上传日期.mrc”。项目用 OCLC 推荐的开源软件 FileZilla 上传书目记录，并可从 FileZilla 中获取上传反馈报告。根据报告修改问题记录，完善实施流程，然后再重新上传记录。

3.4 实施效果

经统计，格式批转换后手工修改格式的记录仅占总记录量的 1.72%，表明绝大多数记录能用 XSLT 程序正确地进行格式批转换，大大提高了格式转换的效率。2019 年 4－7 月项目集中将武汉大学图书馆全部馆藏的中文图书书目记录进行了格式批转换，并将 1016042 条记录上传 WorldCat。OCLC 将武汉大学图书馆上传的记录与 WorldCat 中的记录进行自动匹配，最终 997892 条记录匹配成功，占武汉大学图书馆上传记录总量的 98.21%。匹配成功的记录在 WorldCat 中添加武汉大学图书馆的馆藏代码，用户在 WorldCat 中检索时可看到武汉大学图书馆有该资源（见图 3）。

图 3 WorldCat 中显示武汉大学图书馆的馆藏

4 结语

文章总结了武汉大学图书馆中文图书书目记录上传 WorldCat 的实践经验，其主要贡献为：(1)构建了中文图书书目记录上传 WorldCat 的实施过程；(2)提供了 CNMARC 格式批转换为 MARC21 格式的方法；(3)提供了书目记录维护的方法。实践中也发现一些问题，如 OCLC 不兼容一些语种和字符，致使耗费不少人力修改 OCLC 不兼容的记录，且无法准确揭示记录原貌。OCLC 作为世界性组织，宜对各语种和字符兼容并包，应为中国的图书馆提供格式批转换软件，帮助解决上传时的技术问题。目前项目已攻克上传时的一些核心技术难关，并积累了一些重要的实践经验，下一步计划将 20 余万册馆藏古籍书目记录上传 WorldCat，以向全世界展示武汉大学图书馆馆藏的优秀历史文化资源。

参考文献

1 吴建中，吴建明.OCLC——全球在线计算机图书馆中心[M].北京：华艺出版社，2002：28.
2 穆晖.推进国际传播能力建设打造世界一流的图书馆[J].新世纪图书馆，2018(8)：21－23.
3 艾金勇，陈小莹.西文编目中的 CNMARC 到 USMARC 转换系统的设计与实现[J].电脑与电信，2014(8)：45－47.
4 孙华，陈世海.USMARC 与 CNMARC 自动转换系统[J].大学图书馆学报，2000(1)：56－58.

作者单位：武汉大学图书馆，湖北武汉，430072
收稿日期：2020 年 6 月 16 日
修回日期：2020 年 12 月 25 日
（责任编辑：关志英）

Practice on Uploading the Bibliographic Records of Chinese Books to WorldCat

Cheng Ying　Zhang Yaolei　Liu Xiaoping　Tu Yanling

Abstract: The libraries upload their bibliographic records to WorldCat, which can enhance the display of their collections around the world. Wuhan University Library has uploaded 997892 bibliographic records of Chinese books to WorldCat. Based on the practice of this project, the paper studies its difficulties, challenges and solutions, and introduces the implementation process and experiences in detail in three stages, focusing on the format batch conversion, so as to provide references for other libraries.

Keywords: Bibliographic Records; Format Conversion; Data Maintenance; Record Upload; WorldCat