

中文名称规范文档与虚拟国际规范文档的共享问题研究

贾君枝 石燕青

摘 要 随着规范控制工作的国际化,国际名称规范文档共建共享成为必然趋势。本文对影响名称规范文档共享的要素——名称规范文档遵循的编目规则、采用的描述格式和数据元素构成的差异进行分析,基于虚拟国际规范文档(VIAF)匹配思路,通过中国高等教育文献保障系统管理中心(CALIS)、香港中文名称规范数据库(HKCAN)和VIAF的数据匹配实验,分析当前国内规范文档存在的问题:名称规范文档记录不全,检索结果输出不按相关度排列,国内各机构对中国人名的拼音标目不规范。本文在控制规则、附加信息、数据模型等方面提出相应对策,以提高文档的共享效率,希望能为推动我国规范控制工作的国际化进程提供参考。图2。表4。参考文献19。

关键词 名称规范文档 中文名称规范文档 虚拟国际规范文档 数据匹配

分类号 G254

Research on Sharing between Chinese Name Authority File and Virtual International Authority File

Jia Junzhi & Shi Yanqing

ABSTRACT With the internationalization of authority control, international co-construction and sharing of name authority files become an inevitable tendency. In this paper, we analyzed the variance that consists of different factors, which affect the sharing of name authority files. These factors include the catalogue of rules for name authority, the format of description and data elements. Based on study of the methodology of Virtual International Authority File matching and the matching experiments between CALIS and VIAF, HKCAN and VIAF, we investigated the current issues that exist in domestic authority files such as name authority file records are incomplete, the output is not arranged by relevance and non-standardization of pinyin form of Chinese names in domestic institutions. Moreover, we proposed certain relevant solutions to improve the efficiency of authority files sharing from the aspects of authority control rules, additional information and data model. It aims to provide suggestions to accelerate the internationalization process of authority control in China. 2 figs. 4 tabs. 19 refs.

KEY WORDS Name authority files. Chinese name authority file. Virtual international authority file. Data matching.

0 引言

名称规范文档是名称规范记录集合,它根据一定的规范控制规则将名称相关信息按照统

一的标目形式展现,达到规范控制的目的。名称规范文档的建立不仅能提高书目数据的质量,还能为用户查询名称实体提供便捷有效的途径。随着信息技术的发展,海量的信息使各国编目机构耗费大量人力财力,名称规范编

目工作越来越繁重,用户需求的日益深化也凸显出资源建设的不足,名称规范建设走向共建共享成为必然趋势。

国外在名称规范文档的共享方面已经有多年经验。美国国会图书馆于1977年推出的“名称规范合作计划”(Name Authority Cooperative Program, NACO),采用的是联合规范档模式^[1],到2009年已有500多家机构参与,取得了显著的成绩。欧洲多国于2001年进行的“国际规范资料库合作建制计划”(Linking and Exploring Authority Files, LEAF),同样采用联合规范文档模式,有十余家机构参与。2003年由德国国家图书馆、美国国会图书馆和OCLC开发的“虚拟国际规范文档”(Virtual International Authority File, VIAF)采用虚拟联合规范文档模式,运用文档虚拟集中、实际连接的方式,目前有35家机构为其贡献数据,其中9家仍在测试阶段^[2]。

国内对于中文名称规范文档的共享仍处于探索阶段。2003年由中国国家图书馆(NLC)、中国高等教育文献保障系统管理中心(CALIS)、香港地区大学图书馆协作咨询委员会(JULAC)和台湾汉学研究中心联合成立图书馆合作组织中文名称规范联合协调委员会(CCCNA),建立了“中文名称规范一站式查询系统”,除了能查询以上数据库的数据外,还能查询美国国会图书馆的数据。2008年委员会在第六次会议上决定建立“中文名称共享数据库”,系统能实现多种检索功能、多种数据格式实时转换,促进了国内名称规范数据的共享^[3]。

我国中文名称规范工作处于起步阶段,在共享实践中面临许多挑战,目前国际上对书目数据和名称数据的规范控制没有制定统一的标

准,在国际文档中如何处理中文字符还没有有一个成熟的方案,再加上国内名称规范工作建设不足,影响了数据共享的质量。针对此种情况,本文通过讨论国内外名称规范控制规则以及名称规范文档描述格式的异同,分析目前国内外在名称规范控制及共享方面的情况,并通过VIAF与国内名称规范库的匹配实验,探讨影响共享的问题所在。研究这些问题不仅能推动我国名称规范文档的建设工作,对于促进名称规范文档的资源共享以及名称规范文档与其他领域的互联也具有重要意义。

1 影响名称规范文档共享的要素分析

不同机构的名称规范文档共享,实际上是在数据匹配的基础上有效地实现数据整合。由于各机构名称规范文档遵循的编目规则、采用的描述格式以及数据元素构成存在差异,导致共享效率低下,以下对这三方面的要素进行详细分析。

1.1 编目规则

编目规则规定了名称标目选取原则,以确保标目在检索款目及名称系统中的一致性和稳定性。在国际合作编目的环境下,许多机构开始在规范文档中加入中文字段,这对于中文规范文档在世界范围内共享具有重要意义。目前,各个国家的编目机构依据各自不同的规则对数据进行规范处理,中文字段在文档中的位置和作用也不同,使得数据共享存在一定困难,表1列出了国内几家主要机构和美国国会图书馆使用的编目规则。

表1 不同机构使用的编目规则

机构名称	编目规则
中国国家图书馆(NLC)	中文文献编目规则、西文文献著录条例
中国高等教育文献保障系统管理中心(CALIS)	中文文献编目规则、西文文献著录条例、英美编目条例
香港中文名称规范数据库(HKCAN)	英美编目条例
台湾汉学研究中心	中国编目规则(台湾)
美国国会图书馆(LC)	英美编目条例、东亚资料的描述著录规则

国家图书馆和 CALIS 的中文名称标目采用《中国文献编目规则》(以下简称《规则》),国家图书馆的西文名称标目依据我国编制的《西文文献著录条例》(以下简称《条例》),CALIS 西文编目同时参照《条例》和 AACR2。HKCAN 使用 AACR2,其创建的规范记录更多地参考了美国国会图书馆规范数据库(LCNAF)中的记录。台湾汉学研究中心的编目规则采用台湾出版的

《中文编目规则》,该规则参考 AACR2,向国际标准靠拢^[4]。LC 编目主要依据 AACR2,《东亚资料的描述著录规则》对如何处理东亚语言资料的问题进行了补充。

尽管以上编目规则的内容大体一致,但在应用细节上仍有一定差异,表 2 对有关责任者项和个人名称标目的一些规则进行比较。

表 2 三种主要编目规则在责任者项方面的比较

	中文文献编目规则	西文文献著录条例	英美编目条例(AACR2)
著录款目	取消了主要款目,增加了通用款目	保留了主要款目	保留了主要款目
题名与责任者项信息源	书名页或版权页、封面	题名页	(代)书名页
题名与责任者项次序	正题名 并列题名 副题名及题名说明文字责任说明	正题名 一般文献类型标识 并列题名 其他题名信息 责任说明 无总题名文献	正题名 一般文献类型标识 并列题名 副题名 责任说明 无总题名文献
名称标目形式	主体部分(姓氏、本名、表字、别 名、笔名、法名、封号等)+附加 部分(生卒年、朝代、性别、学科、 职业、籍贯等)	主体部分(姓名、笔名、 称呼、教名、中文译名 等)+附加部分(生卒年、 贵族头衔、荣誉称号等)	主体部分(姓氏、本名、前 缀、教名、头衔等)+附加 部分(生卒年、贵族头衔、 荣誉称号等)
名称的选择	最为人所知的名称	最为人所知的名称	最为人所知的名称

著录款目上,《规则》取消了主要款目,考虑到主要款目产生于西欧国家多款目的字典式目录,而我国目录体系中书名、著者、主题各自分离,款目之间无主次之分^[5],而《条例》和 AACR2 照顾到西方编目的传统习惯,仍保留主要款目。对于著录项目或书目信息的信息源选择上,AACR2 对信息源的要求更加严格,对优先选择和替代顺序有明确的规定,保证了著录顺序的唯一性和确定性。《规则》考虑到我国的编目传统及中文图书版权页含有丰富的信息,将版权页与书名页并列为主要信息,并未明确其优先顺序。名称标目形式方面,都包含主体部分和附加部分,由于不同国家人名名称表示方法的差异,其中附加信息部分差异较大,《规则》中定义了名称的社会属性,如职业、学科等。

1.2 描述格式

描述格式是基于编目规则定义数据记录的计算机可读形式,以方便规范记录的交流与共享。采用描述格式描述实体,就是根据编目规则建立名称规范文档的过程。目前国内外图书馆采用的描述格式都是 MARC,但是有不同版本,中国大陆通用 CNMARC,中国台湾地区使用 CMARC,国外机构多使用 MARC21。

国际合作的环境下,一个中文名称规范记录应同时包含中西文标目,为了与国际文档建立连接,中文名称规范文档不仅要符合中文文档的特点,也要考虑国际标准。不同机构使用不同的描述格式,对不同语种文字使用的标目方式也不同。例如,CALIS 主要使用 CNMARC 格式,将中文简体、中文繁体、西文、日文等作为

并列规范标目(200 字段);HKCAN 和 LC 使用 MARC21 格式,以英文名称作为规范标目(100 字段),将对应的中文名称作为连接标目(700 字段);台湾汉学研究中心使用 CMARC/MARC21 格式,以英文名称作为规范标目(200/100 字段),对应的中文名称作为参照标目(400 字段)。了解不同机构的描述方式有助于我们研究数据的匹配问题。

CMARC 与 CNMARC 格式所采用的标准都是以 UNIMARC 为依据编制,记录结构基本一致,主要字段所揭示的文献内容信息也无太大差别,这为 CNMARC 和 CMARC 的整合奠定了基

础。CNAMRC 与 MARC21 的记录结构相似^[6],但在字段描述方面存在一些差异。针对名称规范文档,MARC21 除了有书目格式(MARC21/B)外,还有专门的规范格式(MARC21/A),详尽介绍了建立名称规范记录所用的字段。而 CNMARC 只有书目格式,对名称记录用到的相关字段没有进行细分。两种格式的有些字段难以找到合适的对应,实现转化存在一定困难,需要认真分析两者之间的异同,以避免转换时出现数据丢失的情况。表 3 以个人名称标目为例,列出了 MARC21 和 CNMARC 中描述人名规范记录的常用字段,并对两种格式进行简单对应。

表 3 MARC21 与 CNMARC 部分字段的对应

MARC21		CNMARC
100 主 要 款 目 个 人 名 称	\$ a 个人名称	701 人名-等同责任者 \$ a 主标目 200 题名与责任者说明 \$ a 正题名
	\$ b 世次(与皇室成员、教皇或其他宗教首领的名称连用的罗马数字)	701 人名-等同责任者 \$ d 世次(罗马数字)
	\$ c 与名称相关的其他文字	701 人名-等同责任者 \$ c 名称的附加成分
	\$ d 生卒日期	701 人名-等同责任者 \$ f 年代(包括朝代)
	\$ f 创作时间	200 题名与责任者说明 \$ f 第一责任者说明
	\$ l 作品语种	101 作品语种
	\$ q 名字全称	
	\$ t 作品名称	200 题名与责任者说明 \$ a 正题名
	\$ x 通用细分	600 个人名称主题 \$ x 主题复分
	\$ y 年代细分	600 个人名称主题 \$ z 年代复分
	\$ z 地理细分	600 个人名称主题 \$ y 地区复分
370-相关地点		701 人名-等同责任者 \$ p 任职机构/地址 701 人名-等同责任者 \$ c 名称的附加成分 300 一般信息附注 \$ a 附注内容
371-地址		
373-相关团体		
375-性别		
368-个人或团体的其他属性		
374-职业		
372-活动领域		
376-家庭信息		
377-相关语种		

1.3 数据元素构成

数据元素是指与名称实体相关的除了名称

以外的其他信息,在 FRAD(规范数据的功能需求概念模型)中被定义为实体的属性。当规范

记录达到一定规模时,同名记录会大量出现,造成混乱,数据元素就被作为名称标目的重要组成部分,在数据匹配时常作为名称的附加成分,是区别不同人物的重要匹配点。

数据元素主要分为两类:一类是生卒日期、性别、民族、籍贯等自然属性,另一类是学科、地址、专长、职业等社会属性。对于同名的实体,选择附加信息对于匹配的准确度有很大影响。目前对于如何选择附加信息尚没有统一标准。一般而言选取自然属性作为除名称之外的匹配点较好,因为相对于社会属性而言,自然属性更稳定、更准确。目前国际上基本达成共识,认为生卒日期应作为首要的匹配点,无法获知或者生卒日期难以确认的情况下可以使用其他信息。如对于“孔子”,史学界对其生卒日期意见不一,导致不同机构的规范文档中记录不同,可能造成错误匹配。为了进一步提高匹配准确率,也需要选取其他重要的附加信息与生卒日期搭配使用。

实际工作中,国家图书馆和 CALIS 常会添加题名关键词或学科名称以达到区分同名人物的目的,例如李莉(电路)、李莉(电子技术)、李莉(法律)、李莉(翻译)、李莉(俄语翻译),但在区分人名的同时带来了其他问题。首先,对于关键词的选择具有很大的为人因素,很多由编目员推断而来,对每个学科的名称表示不够准确,同一领域的相似学科可能会混淆;其次,添加了关键词的人名有时也不能确定其具体身份,不仅可能造成书目数据的连接错误,也给用户的查找浏览增加工作量。数据元素的选取是区分同名人物的关键,直接影响名称规范文档的质量,但目前针对附加信息元素的规范仍不够完善,需要进一步研究。

2 中文名称规范文档与 VIAF 数据匹配实验分析

2.1 VIAF 的匹配思路

虚拟国际规范文档(VIAF)项目由 OCLC、

美国国会图书馆和德国国家图书馆于 2003 年提出,并于 2012 年正式移交给 OCLC,其目标是链接各国的名称规范文档,并允许国家和区域规范形式的变体同时存在,实现世界范围内的文档共享^[7]。目前,VIAF 的数据量已颇具规模,并实现关联化,与维基百科、ISNI、SNAC 等项目建立了合作,在关联数据环境中扮演了重要角色。

2.1.1 VIAF 的数据处理流程

VIAF 根据各机构提供的数据为名称数据创建增强型规范记录,作为 VIAF 的数据库。增强型规范记录是在 VIAF 中最终用来进行名称匹配的规范记录,是由原始名称规范记录和书目记录转换而成的衍生规范记录组成的(见图 1)^[8]。由于增强型规范记录包含与来自书目记录名称相关联的附加信息,它可以支持比规范记录本身更严格的匹配过程^[9]。

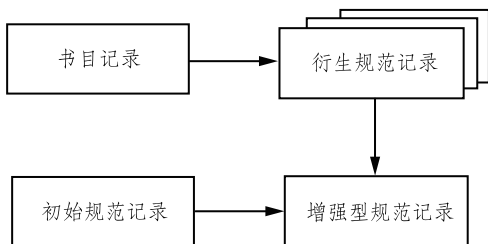


图 1 创建增强型规范记录

VIAF 处理数据的流程分两个步骤。先根据各成员机构提供的书目记录,为 100 和 700 字段中涉及的每一个人物名称手工创建衍生规范记录。衍生规范记录是从书目记录中的信息抽取而来,除了名称,衍生规范记录中还包括从有关书目记录中获取出版资料的概要说明^[10]。一般会抽取两类书目信息^[11]:一类是能够用来唯一确定一部作品的信息,如控制号;另一类是可以应用到同一作者的多部作品中的信息,如两个作者合作过不止一次,合著者就可以作为要抽取的信息。从书目记录中衍生出的信息都用 9XX 字段来描述,并根据 NACO 规则进行规范化处理,如大写字母一律变为小写字母,大部分变音符号和标点符号被删除,时间的表示精确到十年(如 1986 表示为 198X),学科被重新分类,每个学科

涵盖范围更广,并用学科编号表示。

衍生规范记录与相应的名称规范记录合并就形成了增强规范记录,将增强型规范文档作为 VIAF 的匹配文档。OCLC 开发匹配算法对这些文档进行匹配,匹配成功的文档会作为 VIAF 的最初版本。匹配过程中,可能会产生一些中间结果,表明另外一些信息可以作为规范文档中的信息^[12],因此,匹配过程也是一个不断修正规范文档的动态过程。

2.1.2 VIAF 的匹配过程

VIAF 采用开源软件 Hadoop 和分布式编程

工具 MapReduce 批量处理数据。MapReduce 是云计算中常用的一种技术,通过 Map 过程和 Reduce 过程完成数据的处理。应用 Map 过程把大量数据切分成小的数据块,在 Reduce 过程中完成数据的分布式处理,大大简化了对大数据的处理过程。使用 MapReduce 进行数据匹配时,首先在 Map 过程中将各机构提供的数据切片处理,根据 (key, value) 方式将 key 值相同的待匹配字段的数据归并,发送给 Reduce 过程。Reduce 接收到数据后,采用 OCLC 开发的算法进行数据匹配,其原理如图 2^[13] 所示。

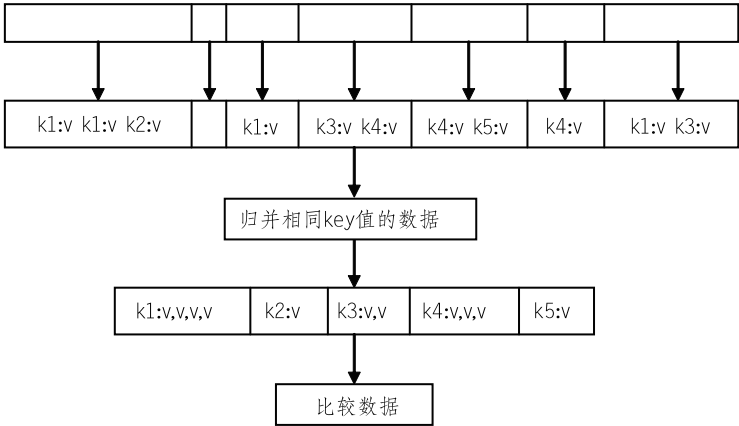


图 2 MapReduce 工作原理

由于增强型规范记录包含书目记录的附加信息,可以进行比原始规范记录更加准确的匹配。除了名称,文档中的其他信息(如生卒日期和书目信息)也常作为附加信息区别名称相似的人。

每一个用来匹配的标目都作为一个匹配点,VIAF 将这些信息分为强匹配点和弱匹配点。强匹配点有:书的题名、控制号 (ISBN、ISSN、LCCN),人物的生卒日期,合著者和作为主题词的人名。如果一个强匹配点匹配上了,就可以基本确定这两个名称是同一个人。弱匹配点有:生卒日期其中之一,作品的主题领域,作品的出版形式(书、电影、乐谱等),出版日期、国家、语种、出版社,部分题名匹配等。多个匹配点的组合可以确定一个匹配。

2.2 匹配实验

本文以国内较具代表性的编目机构 CALIS 和 HKCAN 为对象,研究这两家机构对于同一实体的不同规范表示形式与 VIAF 中数据的匹配情况。选取 500 个人名作为实验数据,包含从《中国人名大词典(历史人物卷)》^[14]和《中国人名大词典(当代人物卷)》^[15]中选取的 200 个中国历史名人和 200 个中国现当代名人,以及从《世界人物大辞典》^[16]中选取的 100 个外国名人。希望通过研究与 VIAF 的匹配情况,发现中文名称规范文档中对人名数据描述存在的问题。

1979 年起联合国决定采用《中国拼音方案》作为拉丁字母文字中中国名称的标准,1997 年 LC 也采用汉语拼音作为中文名称的规范标目,

因此本文中文名称的罗马化形式也采用汉语拼音。HKCAN 中名称的中文标目有简、繁体两种形式,可通过简繁体对应表进行自动转换,检索时只采用中文简体。

匹配实验大致分为三个步骤:根据本地数据库中人名规范标目形式在本地库中进行检索;以相同的标目形式作为检索点在 VIAF 中进行检索;手工检查 VIAF 中该目标实体的所有记录是否都被检索到,以及是否被正确匹配。

2.3 实验结果分析

VIAF 是一个连接现有文档的系统,它将各个机构提交的表示同一个人实体的记录聚合起来,分配唯一的 ID 号,作为一个 VIAF 记录。实验发现,500 条人名数据中,有 272 条(占 54%) 在 VIAF 中无记录,其中 244 条为中文名称,包括古代人名 128 条,现代人名 116 条。剩余的 228 条数据中,102 条匹配成功,即采用中文简体、罗马形式(中文名称用拼音,外国名称用英文)作检索式,在 VIAF 中都显示一条正确数据。

有 44 条数据在 VIAF 中有同一人名的不同记录集合,需要合并;有 41 条数据在 VIAF 中用中文简体检索不到,只能用西文形式检索到,即记录中没有中文简体的标目形式;错误匹配的数据(包括中文简体和西文)共 41 条。

表 4 列出了在 VIAF 有记录的 228 条数据在匹配过程中出现的三种情况,并统计这 228 条数据出现问题的次数。“VIAF 中未归并的记录”是指用中文简体或英文形式分别在 VIAF 中检索时,出现不同记录代表同一实体的情况;“规范标目形式检索不到”是指使用中文简体或英文形式在 VIAF 中检索时无记录显示;“检索结果不准确”是指检索到的记录与用户的检索需求不一致,需要指出的是,在 VIAF 中用某种形式检索时,如果检索结果中有正确的记录,但不是出现在第一条记录中,也将其归为“检索结果不准确”。例如用“Yao Ming”作为检索式时,出现的第一条记录是“周,作人,1885-1967 (VIAFID 76463787)”,而正确结果“Yao, Ming, 1980- (VIAFID 119316736)”出现在第 18 条记录中。

表 4 CALIS、HKCAN 规范文档与 VIAF 的匹配情况

存在的问题	频次		举例
	中文简体	西文	
VIAF 中未归并的记录	13	52	检索“Mao Zedong”时,出现“Mao, Zedong, 1893-1976 (VIAFID 51699962)”和“Mao, Zedong (VIAFID 191378497)”
规范标目形式检索不到	41	0	检索“比尔·盖茨”时,无记录。检索“Gates, Bill”时,显示“Gates, Bill, 1955- (VIAFID 102370574)”
检索结果不准确	7	39	检索“巴金”时,第一个记录是“Herzen, Aleksandr, 1812-1870 (VIAFID 295141994)”

实验匹配中,发现我国中文名称规范文档存在一些问题:名称规范文档记录不全,在 VIAF 有记录的 228 条数据中,有 67 条在 CALIS 中无记录;检索结果输出不按相关度排列,例如检索“马克思”时,输出“蔡文杰(马克思主义)”“陈大安(马克思研究)”等,明显将检索式与附加字段匹配的记录排在前面;中国人名的拼音标目不规范,例如“冰心”,CALIS 中的拼音规范标目是“Bing Xin”,HKCAN 是“Bingxin”,各机构对于

笔名的拼音形式没有遵循一致的规定,有可能在连接各国规范文档时造成匹配错误。

3 提高国内外名称规范文档共享效率的对策分析

从以上分析来看,国内名称规范控制工作尚存在许多不足,世界范围内的名称规范文档共享也处在建设阶段,大量中文名称还没有被

记录到国际文档中。如何促进中文名称规范文档在世界范围内共享,对于中文名称规范文档的发展具有重要意义。

3.1 提高中文名称规范文档中西文标目的规范度

规范名称标目中的各种描述形式,以免造成标目的混乱。国内对西文标目普遍采用拼音形式(有英文名的除外),但目前对于拼音标目的表示仍存在问题,不能完全遵循规则中的具体规定,导致描述形式的多样性。例如对于“鲁迅”,CALIS记录中拼音规范标目的名称为“Lu Xun”,变异标目的名称为“Luxun”。而对于“冰心”,CALIS的记录中拼音标目只有“Bing Xin”,VIAF的记录中不同机构对“冰心”的拼音标目分别为“Bing, Xin”和“Bingxin”等形式。对于中文人名的拼音形式,《中国人姓名汉语拼音字母拼写法》规定姓在前名在后,分开且姓和名的首字母大写,国际上1979年《联合国秘书处关于采用“汉语拼音”的通知》亦采用我国汉语拼音拼写法,例如曹雪芹(Cao, Xueqin),对于笔名、艺名、谥号、庙号等拼写方式在《规则》中都有明确规定,因此我国需积极参与国际编目机构的标准制订,倡导中文名称标目的规范与统一。

各机构中对外国人的中文译名也存在差异,例如对于“Schwab”,CALIS的中文译名为“施瓦布”,国家图书馆为“斯威布”,针对这种情况,各机构应参照权威资料,例如《世界人名翻译大辞典》、《不列颠百科全书》,以确保中文译名的准确和统一。

3.2 附加信息作为匹配点的处理

对人名附加信息的处理应该制定更加详细的规范。由于名称部分和附加信息都作为匹配点,可能出现附加信息与名称相匹配的情况。例如检索“马克思”时,出现“郝静之(马克思理论)”,针对此情况,可以为不同的匹配点设置匹配优先顺序,名称部分优先度要高于附加信息。根据VIAF的匹配算法,附加信息有强匹配点和

弱匹配点之分,但除了公认的生卒日期作为首要附加信息外,对其他附加信息没有制定优先度的规定,可以根据附加信息的重要程度为其设置权重,以合理利用附加信息,提高匹配的准确率。如何设置权重值仍需进一步研究讨论。

对于以学科名称作为区分同名人物的记录,例如王军(化学工程学)、王军(艺术设计),需进一步规范学科名称表,以保证描述的一致性。比如学科限定词选用《中图法》所列类目^[17],保证数据统一的同时,也具有可识别性。对于一些添加了附加成分也难以确认身份的记录,将其归为“不可区分的规范记录”,将不能确定的名称不加区分地放在一条记录里,并在注释中说明该规范记录可能对应很多不能确定的个人^[18],以确保不产生无意义的冗余数据,该类型在MARC21和CNMARC中都有定义,在VIAF中也有使用。

3.3 寻求更利于共享的数据模型

不同机构使用的描述格式差异,造成数据之间的共享困难。以MARC21和CNMARC为例,二者结构的相似性和字段的对应关系使其转换成为可能,但实际操作中情况复杂,在可变数据区,MARC21共定义97个字段,而CNMARC只定义了60个字段,有许多字段无法对应。例如MARC21的100字段设置的“名字全称”和008字段的“语种代码”,在CNMARC中找不到对应字段,这种情况可能会造成数据转换之后信息缺失,影响文档中信息的共享效率。

国际图联规范记录的功能需求与编号(FRANAR)工作组于2008年提出了规范数据的FRAD概念模型,将书目记录里的实体通过名称或标识符识别,将规范记录数据与用户相关联^[19]。FRAD认为规范记录就是关于实体信息的集合,它将实体关系以树状形式展现出来。相比MARC使用字段、子字段描述记录的方式,FRAD将实体的“名称”、“标识符”、规范记录中的“检索点”、描述规范记录的“规则”和“机构”定义在不同的逻辑层上,一改以往MARC格式对书目关系的扁平化描述,实现了逻辑分组。

对记录的描述是通过实体属性来实现的,不同的属性被定义在不同的逻辑层次上,不必考虑不同格式间字段的对应,更有利于实现规范数据的共享。

除了详细定义各种实体,FRAD 还定义了实体间、实体与属性间的各种规范关系^[19]。如将名称附加信息中的出生地作为地名实体与人名实体建立“出生于”关系,将参照标目中名称与规范标目中的名称建立“别名”关系,与出版社、国家、合著者都能建立相应关系。这种数据模型应用到名称规范文档的构建中,不仅方便各机构间的数据共享,也能进一步明确用户需求,方便用户在整个书目世界中进行查找。国内机构已着手研究 FRAD 模型,将现有的规范文档进行有效转换,为规范文档的进一步国际共享奠定基础。

3.4 VIAF 项目匹配算法的改进

从实验结果可以看出,虽然中文标目已经加入到国际规范文档中,但在实际匹配时仍存在很多问题,除了各机构本身的规范控制工作存在差异外,VIAF 记录合并和维护工作也需要改进。首先,VIAF 中对于记录的聚类存在一定的问题,有很多需要合并的重复记录;其次,在 VIAF 中检索时出现一些错误匹配的情况,例如用“Mao Dun”检索时,出现的第一条记录是

“Mao,Zedong,1893-1967(VIAFID 51699962)”,第二条记录才是“茅盾,1896-1981(VIAFID 97121854)”,这说明 VIAF 的匹配算法不够精确,需针对中文规范文档的特点,改进匹配算法。此外,建议为不同匹配点设置不同的优先级别,对于匹配结果应该按照相关度排序。这些工作不仅需要参与机构提供更完整的记录信息来帮助确认和匹配,也需要 OCLC 开发更优质的匹配和聚类算法。

4 结语

中文名称规范文档加入国际行列已经成为必然趋势,目前我国的规范控制工作仍存在许多不足,国际虚拟规范文档(VIAF)已经取得了明显成就,我们应深入地研究国际标准和国际规范控制工作,以指导我国名称规范工作的开展。本文对影响名称规范文档共享的要素进行分析,并通过对 CALIS、HKCAN 与 VIAF 中的数据进行匹配试验,分析中文名称规范文档存在的问题以及影响共享的原因,针对规范文档共享提出一些对策,以期对中文规范文档的开放共享提供参考。未来需进一步拓展研究名称实体资源的应用范围,研究名称规范文档与其他领域的互操作,使名称规范文档在语义网中发挥更大的作用。

参考文献

- [1] 林巧敏,林淑芬. 中文名称权威控制合作发展初探——国外案例之借鉴[EB/OL].[2013-07-19].<http://www.docin.com/p-134151972.html>. (Lin Qiaomin, Lin Shufen. A study on the collaboration of Chinese name authority control: lessons from NACO, VIAF and LEAF[EB/OL].[2013-07-19].<http://www.docin.com/p-134151972.html>.)
- [2] Online Computer Library Center.VIAF(The Virtual International Authority File)[EB/OL].[2013-06-24].<http://www.oclc.org/research/activities/viaf.html>.
- [3] 喻爽爽. 两岸三地中文名称规范合作与共享[EB/OL].[2014-07-25].<http://olcc.nlc.gov.cn/download-wdxz.html>. (Yu Shuangshuang. Chinese name authority files cooperation and sharing of the three geographic areas[EB/OL].[2014-07-25].<http://olcc.nlc.gov.cn/download-wdxz.html>.)
- [4] 戴祖谋. 台湾图书编目规则管窥[J]. 图书馆,1988(4):21-24. (Dai Zumou. A view of bibliographical cataloguing rules in Taiwan of China[J]. Library,1988(4):21-24.)
- [5] 黄俊贵. 关于《中国文献编目规则》的问答[J]. 中国图书馆学报,1997(2):56-60, 50. (Huang Jun-

- gui. Questions and answers about the *Chinese Bibliographical Cataloguing Rules*[J]. *Journal of Library Science in China*, 1997(2): 56-60, 50.)
- [6] 李燕, 杜薇薇, 郭华. MARC21元数据与 CNMARC 元数据的分析比较[EB/OL]. [2014-08-03]. http://www.nlc.gov.cn/old2008/service/fuwudaohang/conf2006/conf2006_liyan.htm. (Li Yan, Du Weiwei, Guo-Hua. Analysis and comparison between MARC21 metadata and CNMARC metadata[EB/OL]. [2014-08-03]. http://www.nlc.gov.cn/old2008/service/fuwudaohang/conf2006/conf2006_liyan.htm.)
- [7] VIAF: The virtual international authority file[EB/OL]. [2014-07-20]. <http://www.oclc.org/research/projects/viaf>.
- [8] Hickey T. The virtual international authority file[EB/OL]. [2013-12-25]. <http://www.Oclc.org/content/dam/research/presentations/hickey/ala-2009-cdg-wess.ppt>.
- [9] Tillett B B. A virtual international authority file [EB/OL]. [2014-02-20]. http://www.nl.go.kr/icc/download/060813_3.pdf.
- [10] Hickey T. The virtual international authority file[EB/OL]. [2014-02-20]. <http://www.oclc.org/content/dam/research/presentations/hickey/ala-2009-acig.ppt>.
- [11] 顾森. 虚拟国际规范文档——连接德国国家图书馆和美国国会图书馆的规范文档[J]. *国家图书馆学报*, 2006(4): 87-90. (Gu Ben. Virtual international authority file; linking Die Deutsche Bibliothek and Library of Congress name authority files[J]. *Journal of The National Library of China*, 2006(4): 87-90.)
- [12] Bourdon F, Boulet V. VIAF : un pivot pour l'accès multilingue à diverses collections[EB/OL]. [2014-02-21]. <http://conference.ifla.org/past-wlic/2011/79-bourdon-fr.pdf>.
- [13] Hickey T. Matching names in parallel[EB/OL]. [2014-02-21]. <http://www.oclc.org/content/dam/research/presentations/hickey/access2006.ppt>.
- [14] 廖盖隆, 罗竹风, 范源, 等. 中国人名大词典——历史人物卷[M]. 上海: 上海辞书出版社, 1990: 1-10. (Liao Gailong, Luo Zhufeng, Fan yuan, et al. Who's Who in China; historical figures[M]. Shanghai: Shanghai Lexicographical Publishing House, 1990: 1-10.)
- [15] 廖盖隆, 罗竹风, 范源, 等. 中国人名大词典——当代人物卷[M]. 上海: 上海辞书出版社, 1992: 1-23. Liao Gailong, Luo Zhufeng, Fan yuan, et al. Who's Who in China; contemporary figures[M]. Shanghai: Shanghai Lexicographical Publishing House, 1990: 1-10.)
- [16] 潘裕然, 胡冰如, 等. 世界人物大辞典[M]. 北京: 国际文化出版公司, 1990: 1-11. (Pan Yuran, Hu Bingru, et al. Who's Who in the world[M]. Beijing: International Culture Publishing Company, 1990: 1-11.)
- [17] 秦静. 同名标目附加成分存在的问题及其应对[J]. *图书馆建设*, 2010(10): 59-61, 64. (Qin Jing. Issues and related countermeasures of additional components of homonymy headings[J]. *Library Development*, 2010(10): 59-61, 64.)
- [18] 顾森. 中国国家图书馆中文名称规范工作的思考[J]. *国家图书馆学报*, 2007(3): 39-44. (Gu Ben. Thinking about Chinese name authority work of National Library of China[J]. *Journal of The National Library of China*, 2007(3): 39-44.)
- [19] IFLA. 规范数据的功能需求[EB/OL]. [2014-08-10]. http://www.ifla.org/files/cataloguing/frad/frad_2009-zh.pdf. (IFLA. Functional requirements for authority data [EB/OL]. [2014-08-10]. http://www.ifla.org/files/cataloguing/frad/frad_2009-zh.pdf.)

贾君枝 山西大学经济与管理学院教授, 博士生导师。

通讯地址: 山西省太原市坞城路 92 号。邮编: 030006。

石燕青 山西大学经济与管理学院硕士研究生。通讯地址同上。

(收稿日期: 2014-08-27; 修回日期: 2014-10-10)