

●司莉 徐丽晓 吴钢 陈红艳

OCLC 术语服务研究:背景、进展与启示^{*}

摘 要 OCLC 术语服务是互操作领域的一个研究项目。术语服务可以理解为涉及各种类型知识组织资源(包括规范文档、主题词表系统、网络分类、分类表等)的 Web 服务。该项目的目标是通过在多个受控词表之间有选择地建立映射,帮助文献收藏机构为馆藏创建相互兼容的元数据,通过任何基于网络的元数据编辑器获取术语资源,实现对术语资源的一站式获取。我们可以借鉴 OCLC 的做法,对国内外词表间映射进行研究。表 4。参考文献 3。

关键词 术语服务 受控词表 匹配 兼容 映射

分类号 G254

ABSTRACT OCLC Terminologies Service is a research project in the field of interoperability. Terminologies Service can be regarded as a Web service relating to various knowledge organization resources, including authority files, subject thesauri, network classifications and classification schedules. The goal of the project is to selectively establish mappings among different controlled thesauri, help document collection institutions to create compatible metadata for their holdings, get terminology resources through any network-based metadata editors, and realize the one-stop access to terminology resources. The authors recommend learning from OCLC and doing researches on different mappings among thesauri. 4 tabs. 3 refs.

KEY WORDS Terminologies service Controlled thesaurus Matching Compatibility.

CLASS NUMBER G254

1 OCLC 术语服务研究项目的背景

受控语言是知识组织系统的主要组成部分,即使在网络环境也有广泛的应用空间,它可以提高信息检索的效率和准确性。受控语言虽然有助于情报检索系统内部的一致性,但却降低了系统之间的兼容性^[1]。检索语言的兼容转换是实现网络信息资源整合、跨库浏览与检索的重要技术方法,也是优化信息组织、实现资源共享、满足一站式获取的前提。而现有知识组织系统间兼容转换还未能引起足够重视。

1994 年,OCLC Research 项目组开始关注该问题。OCLC Research 项目组主要从事数字图书馆相关技术的开发与研究工作,研究范围包括元数据管理、知识组织、内容管理、管理智能化、互操作、系统和交互设计等领域,其任务是方便用户获取和利用全世界的信息资源,降低图书馆成本^[2]。

OCLC 术语服务是互操作领域的一个研究项目,术语服务可以理解为涉及各种类型知识组织资源(包括规范文档、主题词表系统、网络分类、分类表等)的 Web 服务,其目标是为人和计算机获取和理解

知识组织资源中的概念以及概念间的关系提供方便,帮助软件开发者研制工具以改进搜索引擎性能,提高检索效率。该项目的目标是通过在多个受控词表之间有选择地建立映射,帮助图书馆、档案馆、博物馆等为馆藏创建相互兼容的元数据,通过任何基于网络的元数据编辑器获取术语资源,实现对术语资源的一站式获取^[3]。

2 术语服务研究项目进展

1994 年,OCLC 开始建立 LCSH 和 DDC 之间的映射,DDC/LCSH 映射产生于 OCLC WorldCat 记录中的 DDC 分类号和 LCSH 词汇。OCLC 的词表映射研究项目主要采用两种映射实现方法:直接映射(direct mapping)和同现词映射(co-occurrence mapping)。前者是指在不同的受控词表之间或词表与分类号之间建立等同词联系。后者是指在相同的元数据或编目记录中,对来自不同词表而同时出现的词建立映射。OCLC 在 Web Dewey 与 LCSH 之间创建了大约 9 万个同现词映射。

OCLC 的词表映射研究项目已经扩展到其他分

^{*} 本文系国家社会科学基金项目(05BTQ010)的研究成果之一。

类表、主题词表系统, 如 ERIC Thesaurus, LCSHac, DDC 有选择地在一些类之间建立了直接映射, MeSH, NLMC 等。OCLC 词表映射研究项目中的词表 GSAFD 中的小说和戏剧部分类型词分别与 LCSH, 联系和映射途径(直接映射、同现词映射、同时使用 LCSHac 词汇建立了直接映射。这两种方法), 如表 1 所示。其中, LCC/DDC, NLMC/

表 1 OCLC 词表的相关映射

从	到							
词表	DDC	ERIC	GSAFD	LCC	LCSH	LCSHac	MeSH	NLMC
DDC(杜威十进分类法)				直接	直接、 同现词	直接、 同现词	直接	直接
ERIC Thesaurus(美国教育资源数据库叙词表)					直接			
GSAFD(个人作品主题检索指南)					直接	直接		
LCC(国会图书馆分类法)	直接							
LCSH(国会图书馆标题词表)	直接、 同现词	直接	直接	同现词			直接	
LCSHac(国会图书馆儿童标题词表)	直接、 同现词							
MeSH(医学标题词表)	直接				直接			
NLMC(国家医学图书馆分类法)	直接							

注: 本表引自 Terminology Services. <http://www.oclc.org/research/projects/termservices.default.htm> (2006-03-10 查询)

如表 1 所示, OCLC 的许多映射研究项目都选择与 LCSH 进行映射。主要原因是: (1) LCSH 是迄今为止使用最为广泛的主题词表, 是国际范围内应用的受控词表, 被世界上许多国家翻译使用。(2) LCSH 是最大的综合性的英文标引词表, 涵盖了所有主题领域的词汇。它也被推荐用来规范 DC 主题元素的编码表, 一些著名项目就是基于 LCSH 利用 DC 元数据建立主题元素的, 例如 Colorado Digitization Program, Dspace 和 EPrints UK 等。

OCLC 研究项目中, 词表编码有多种标准: 对于规范文档、标题词表和叙词表, 采用 MARC21 规范数据格式; 对分类数据, 采用 MARC21 分类数据格式。一些著名词表都有 MARC 格式, MARC 规范格式支持词表间联系, 因此采用 MARC 格式作为编码标准。MARC 规范格式可以为许多受控词表元素提供详细编码, 在 MARC 中 1XX 字段描述选用词, 150 字段描述主题词, 4XX 字段描述非选用词, 5XX 字段描述广义词、狭义词、相关关系词, 7XX 字段描述同一词表或不同词表中等同关系词之间的关系。

3 词表映射的实现过程: 以 ERIC Thesaurus 与 LCSH 映射为例

ERIC Thesaurus 是由美国教育部赞助教育资源信息中心创建的进入教育文献资料库的公共入口词表, 它是一个规范的叙词表。

3.1 将 ERIC 转换为 MARC 规范数据格式

被映射的词表即源词表首先要转换为结构化的 MARC21 规范格式, 以 ERIC 中的 Tutors 一词转换为 MARC 记录格式为例加以说明(如表 2 和表 3 所示)。在表 2 和表 3 的例子中, 广义词(Broader Terms)转换为含有 \$w 子字段的 MARC 550 字段, 相关词(Related Terms)转换为多个不含 \$w 子字段 MARC 550 字段, 替代词(Used For)字段转换成 MARC 450 字段, “Coaching Teachers”一词曾经用作正式有效词, 在 450 字段, 增加子字段 \$wa 说明这个词曾经是正式有效词。ERIC 记录对“Coaching Teachers”进行说明, 并给出这个词的生命周期, 当这个数据转换成 MARC 时, 就为这个数据创建了 688 字段(应用历史阶段说明)。通过 MARC 规范格式对源词表和目标词表进行编码, 可以对相似信息的描述进行标准化处理, 提高词表匹配的能力。

表2 Tutors 的 ERIC 记录实例

Record Type:	Main
Scope Note:	Persons engaged, often privately, to instruct an individual or small group in a particular subject (Note: Use "College Faculty" for British postsecondary tutors -- prior to Mar80, this term did not carry a Scope Note)
Category:	Students, Teachers, School Personnel
Broader Terms:	Teachers;
Narrower Terms:	n/a
Related Terms:	Tutor Training; Tutorial Programs; Tutoring;
Used For:	Coaching Teachers (1966 1974);
Use Term:	n/a

表3 Tutors 的 MARC21 规范格式记录 (部分)

150	\$a tutors
450	\$wa\$a Coaching Teachers
550	\$a Tutor Training
550	\$a Tutorial Programs
550	\$a Tutoring
550	\$a Teachers\$wg
680	\$iPersons engaged, often privately, to instruct an individual or small group in a particular subject (Note: Use "College Faculty" for British postsecondary tutors -- prior to Mar80, this term did not carry a Scope Note)
688	\$a Coaching Teachers (1966 1974)

3.2 建立词汇匹配

通过一系列计算机程序,将源词表 ERIC 和目标词表 LCSH 中的所有选用词和非选用词进行匹配。匹配过程中,空格、字母大小写、标点符号的差异将被忽略。下面的词被认为是匹配的:

<u>ERIC Thesaurus Term</u>	<u>LCSH Term</u>
Alzheimers Disease	Alzheimer's disease
Nurses Aides	Nurses'aides

目前,复数与单数形式的词、带与没带括号的词、带与没带以逗号引出修饰词的词,都不能进行匹配。例如:

<u>ERIC Thesaurus Term</u>	<u>LCSH Term</u>
Echolocation	Echolocation (Physiology)
Crack	Crack (Drug)
Radiology	Radiology, Medical
Rh factors	Rh factor

共有 3797 个 ERIC 词匹配到 LCSH,可以根据匹配类型将它们分为 4 类:(1)PT/PT 匹配:源词表中的选用词和目标词表中的选用词之间的精确匹配。(2)PT/NPT 匹配:源词表中的选用词和目标词表中的非选用词之间的精确匹配。(3)NPT/NPT 匹配:源

词表中的非选用词和目标词表中的非选用词之间的精确匹配。(4)NPT/PT 匹配:源词表中的非选用词和目标词表中的选用词之间的精确匹配。

3.3 进行匹配评价

OCLC 术语研究项目组应用美国西北大学的 LCSH/MeSH 映射项目中建立的指标体系,对 4 种匹配类型分别进行匹配评价。其原则是:(1)匹配的两个词表中映射词概念应有同样的外延。(2)源词表中的词只能映射到 LCSH 主标题词,但当源词表中的选用词与目标词表 LCSH 中的副标题词完全相同时,则可以与 LCSH 中的副标题词匹配。(3)一对一的映射是优选的,但源词表中的词可以映射到目标词表中的表达等同概念的多个词。

表4 PT/NPT 匹配与 NPT/PT 匹配举例

ERIC	LCSH	Match Type	Valid Mapping
PT: Adolescents	PT: Teenagers NPT: Adolescents	PT/NPT	Yes
NPT: Adolescence	PT: Adolescence	NPT/PT	No
NPT: Teenagers	PT: Teenagers	NPT/PT	Yes

表4的例子中, Adolescence 的 NPT/PT 匹配是一个无效映射, 因为 Adolescence 在 ERIC 和 LCSH 中分别代表不同的概念, ERIC 中的 Adolescents 指 13~17 岁的年轻人, LCSH 中的 Adolescence 指青春期的生理发展、心理发展和社会发展。ERIC 中的 Adolescent Development 与 LCSH 中的 Adolescence 是较好的匹配。

映射的维护也是映射评价的一个方面, 可以在追踪词表记录软件的帮助下实现。当词表发生变化时, 需要通过评价来决定原有映射是否仍然有效。

3.4 实现词表间的链接

词表间的链接在 MARC 的 7XX 字段中描述, 可对下列内容进行编码: (1) 目标词表的名称或代码;

(2) 源词表的映射词; (3) 目标词表中映射词的控制号; (4) 映射组织的身份。ERIC 中的映射词与目标词表中的映射词进行链接, 就在 ERIC MARC 的规范记录中增加 750 的入口字段。在下面的例子中, 前两个 750 字段是 ERIC/LCSH 映射, 第一个 750 字段, ERIC 中的词与 LCSH 中对应词 Eidetic Imagery 的 LC MARC21 主题规范记录进行链接; 第二个 750 字段, ERIC 中的词与 LCSH 中对应词 Photographic memory 的 LC MARC21 主题规范记录进行链接; 最后一个 750 字段是 ERIC/MeSH 映射, ERIC 中的词与 MeSH 中对应词 Eidetic Imagery 的 MeSH 规范文档记录进行链接。

750#0 \$aEidetic imagery\$0(DLC)sh 85041379 \$5OCoLC-O
750#0 \$aPhotographic memory\$0(DLC)sh 00009368 \$5OCoLC-O
750#2 \$aEidetic Imagery\$0(DNLM)D004538

链接到LCSH中对应词的LC MARC21 主题规范记录
链接到LCSH中对应词的LC MARC21 主题规范记录
链接到MeSH中对应词的MeSH 规范文档

在上面的例子中, 750 字段两个指示符, #表示未定义, 0 表示目标词表为 LCSH/名称规范文档, \$0 子字段表示映射词的控制号, \$5 子字段表示提供映射数据的组织, 如代码 OCoLC-O 是 MARC 为 OCLC 研究办公室设定的机构代码。LCSH 词汇通过 OAI-Cat 框架同 LC 规范记录相联系。用户通过浏览器和机器通过 OAI-PMH Web 服务机制可以获取这些记录。

4 OCLC 术语服务研究项目对我们的启示

从对 OCLC 术语服务研究项目的分析中可以看出: 该研究项目的重点在于研究词表映射匹配以及在此基础上实现词表之间的互操作。词表映射中采用 MARC 格式记录作为中介进行转换, 词表的匹配易于实现, 能够提高匹配效率和准确性。该项目注重研究词汇匹配的效果和质量, 采用多种方法对匹配效果进行评价。

OCLC 术语服务的研究眼界很开阔, 涉及了各种类型的多种词表之间的映射匹配。我国目前已有众多的分类表、叙词表、元数据应用项目, 一些机构还编

制了自己使用的专业词表, 如深圳证券信息有限公司就研制了自己行业的证券信息主题词表。我们可借鉴 OCLC 术语服务的研究思路和方法, 如通过 MARC21 规范文档或中介词典等, 对国内词表间以及国内外词表间映射进行研究, 实现词表间的互操作, 提高检索效率, 为用户的跨库浏览和检索提供条件。

参考文献

- 1 戴维民. 信息组织. 北京: 高等教育出版社, 2004: 156
- 2 [2006-05-10]. OCLC Research. <http://www.oclc.org/research/about/default.htm>
- 3 [2006-03-10]. Terminology Services. <http://www.ocic.org/research/projects/termservices/default.htm>

司莉 武汉大学信息管理学院图书馆学系副主任, 副教授, 博士。通信地址: 武汉。邮编 430072。

徐丽晓 河南南阳师范学院工作, 现为武汉大学信息管理学院 2005 级在读研究生。通信地址同上。

吴钢 陈红艳 武汉大学信息管理学院 2005 级在读研究生。通信地址同上。(来稿时间: 2006-05-31)