Date : 24/06/2008

交叉语词索引：术语映射及其对信息检索的影响

Philipp Mayr & **Vivien Petras**

（GESIS 社会科学信息中心，波恩，德国）

中文翻译：刘华梅（中国国家图书馆）

Chinese Translator:

LIU Huamei (National Library of China)

Meeting: **129. Classification and Indexing**

Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

World Library and Information Congress: 74th IFLA General Conference and Council

10-14 August 2008, Québec, Canada

http://www.ifla.org/IV/ifla74/index.htm

摘要：德国联邦教育与研究部资助了一项术语映射方案，该方案已于 2007 年结题。术语映射方案的任务是组织、创建和管理受控表（叙词表，分类法，标题表）之间的“交叉语词索引”，这些词表以前是集中在社会科学领域的，但是很快地扩展到了其它学科领域。该方案已建立了 50 多万种语词关系的 64 种交叉。在项目的最后阶段，还在一个信息系统环境下进行了测试和评估词表映射效果的评价工作。本文就交叉语词索引的工作和评价结果进行报告。

1. 引言

在德国，一项具有挑战性的关于一站式学术检索的项目是 vascoda 门户[1]，它是德国联邦教育与研究部和德国研究基金会的联合项目。Vascoda 为多个学科和跨学科的数据库（如，标引和文摘服务，图书馆目录，最新题录数据库，全文文章数据库等）以及网络资源集合（见 Depping 2007 年的一个回顾）提供了一个通用的检索界面。到 2007 年，Vascoda 已经是全球科学门户 WorldWideScience.org[2]的合作伙伴。

Vascoda 门户的主要观念是把 40 多个供应者提供的高质量的信息资源构造和合并在一个检索空间（大约 8100 万篇文献）。这个检索空间是按分布式的学科门户（“虚拟主题图书馆”）组织的，并且每一个被整合的数据集归入 Vascoda 的主题组（工程和物理科学；医学

[1] http://www.vascoda.de
[2] http://worldwidescience.org/

和生命科学；法律，经济和社会科学；人文科学；宗教/文化领域；跨学科的数据集）。

Vascoda 门户包括很多精心开发和构造的信息集合，还包括在个别的集合层次用来描述和组织文献内容的成熟的主题元数据框架（标题表，叙词表或分类法）。然而，通用的检索界面，只提供对所有元数据字段的自由文本检索，而不用考虑最初打算供这些信息集合使用的精确主题词获取工具。因为信息集合的分布式特点和多种多样的主题词获取方法，在一个检索界面用同样强大的但是错综复杂的主题获取工具整合所有这些不同种类的信息资源主要是技术上和管理上的难题。

同时，大规模的网络检索应用为它们增加了新的集合和高级主题获取等扩展功能（如 Flickr 中的聚类和视频检索）。另一个突出的例子是语义网的应用[3]，它来源于整合本体和其它语义数据的高级推理机能。如果大多数的当代信息组织工作能够像语义网那样努力提供更多信息内容的结构和语义分解，那么怎样才能使数字图书馆的高级界面正好能按比例缩减回到同一论题？

在 2004 年，德国联邦教育与研究部在波恩的 GESIS 社会科学信息中心（GESIS-IZ）资助了一项术语映射方案（KoMoHe 项目[4]），该项目在 2007 年末已结题。此方案的任务之一是组织、创建和管理集中在社会科学领域的主要受控词表之间的“交叉语词索引”（交叉映射），但是很快扩展到了其它学科领域。项目的主要目标是建立、实施和评价一个术语网络，使得在一个典型的数字图书馆环境中不同种类资源能语义集成，进而实现检索。

语义集成的目的是通过主题元数据连接不同的信息系统——使得几个信息系统和个别数据库提供的高级主题获取工具能够实现分布式检索。通过不同主题术语的映射，获得一个能够检索到所有信息集合的“语义协议”。术语映射——一个受控词表的词和词组映射到另一个受控词表的词和词组——从而实现数字图书馆世界的一个数据库检索到分布式检索场景的无缝转换。

这篇文章描述了术语映射项目 KoMoHe 及其有关的词表和数据库，在检索和结果中运用开发的交叉语词索引，并且提出一个广泛的信息检索评价体系来分析术语映射对检索中的检全率和检准率的影响。

2．语义异构

通常在数字图书馆中有两种主要方法来处理语义异构问题：人工方法和自动方法。术语映射中所有工作的实质是要保留和不改变不同术语间的差异。没有人能够单独负责转换异构集合，主要是质量和成本原因。另外重要的是在数据库层次的双向操作方法。Krause（2003）也介绍了语义异构方法必须要相互补充和共同协作。

- 受控词表间的交叉语词索引：通过用户上下文分析不同的概念体系，并人工地使这些概念建立联系。这种想法不应该被超级叙词表的结构所困惑。建立交叉语词索引时，没有人试图使现有的概念世界标准化。交叉语词索引只包括部分联合的现有的术语系统，也包括有问题的转换中剩余的静态部分。这种语词索引主要提供同义词或近义词/等级关系（见表 1 和 2），也包括演绎规则关系的映射。
- 定量统计方法：转换问题通常可以模拟成两种内容描述语言之间的不明确问题。

[3] 见 http://www.w3.org/2001/sw/ 开始部分
[4] http://www.gesis.org/en/research/information_technology/komohe.htm

为了处理信息检索不明确的语词（如用户查询的词和数据集合中的词），提供的不同自动操作（概率程序、模糊方法和神经网络）都能够用在有问题的转换上（Hellweg et al., 2001）。个别文献能够标引成两种概念格式，或者以此使两个不同的和不同标引的文献能放在一定的关系中。这些类型的程序都需要训练数据。同样的文本能够翻译成两种语言从而实现多语言信息检索。

当在一个分布式检索场景中执行语义异构（如交叉语词索引）处理时，用户可以用自己熟悉的主题元数据框架来检索各种信息集合系统。术语映射能够以多种方法支持分布式检索。首先，它能够实现不同主题元数据系统对数据库的无缝检索。另外，它还能作为词表扩展的工具，因为它表现了词表网络的相等、上位、下位和相关语词关系（见表 1 和 2 中的语词例子）。第三，语义映射的这种词表网络也能够用于查询扩展和再造。

术语映射服务不仅能在精确检索中构造查询式，还能在应用于数字图书馆其它信息集合中的所有不同术语间自动转换查询式。一个检索者可以在不同的信息资源间无缝转换，因为使用的不同术语间已自动实现了语义转换。

如果映射词表可以利用的话，对于各学科间的信息系统，语义集成不仅增加了用不同主题元数据框架实现分布式检索的概率，而且还为检索者提供了进入不同学科框架和专业领域语言的窗口（见图 1）。

语义映射在外文数据库的转换方法中起重要的作用。映射可以在不同数据库或不同学科的受控词表间创建，也可以以传统的翻译方式：例如表 1 从一个德语术语到一个英语术语。

表 1 列出了德国社会科学叙词表（TheSoz）的两个来源语词（左栏），以及经过人工建立关系的映射词表（目标词表）的目标语词。词间关系类型将在表 2 中进一步解释。

TheSoz 开始语词	**关系**	**目标语词**	**目标词表**
Weiterbildung engl: "further education"	=	Weiterbildung	Psyndex, STW, Infodata, SWD, BISp, DZI
	∧	Berufsfortbildung	FES
	=	Further education	CSA-ASSIA
	=	Continuing education	CSA-PEI
	=	Adult Education	CSA-SA
	<	Education	CSA-WPSA
	=	Erwachsenenbildung	IBLK

Meinungsforschung engl: "opinion research"	0		Psyndex
	^	Einstellungsforschung	IAB
	=	Opinion Polls	CSA-ASSIA
	=	Opinions + Research	CSA-SA
	<	Research	CSA-PEI
	=	Public Opinion Research	CSA-WPSA
	=	Public Opinion Polls	ELSST
	=	Meinungsumfrage/Meinungs -forschung	IBLK

表 1　TheSoz 词表中的开始或来源语词和选择的目标语词（语义映射）

近几年，不同机构都开始努力为信息系统提供语义集成。在美国，OCLC 发起的术语服务[5]项目（Vizine-Goetz, 2004, 2006），该项目提供像 DDC，LCC，LCSH 或 MeSH 等不同受控词表间术语映射网络服务。在欧洲，数字图书馆优秀网络 Delos2 投入一个工作组（WP5）来应对知识提取和语义互操作问题（Patel et al., 2005）。英国 JISC 被任命提交一份回顾英国术语服务的报告（Tudhope, Koch et al., 2006）。另一个项目是 CRISSCROSS 项目[6]，它是由德国国家图书馆和科隆大学应用科学学院创建的一个基于叙词的多语种的研究词表，它集成了标题规范文档（SWD）和 DDC 符号（Panzer, 2008）。联合国粮农组织 FAO[7]的农业信息管理标准部门也在研究各种术语映射方案（Liang & Sini, 2006）。另一个例子是高层叙词表项目（HILT[8]），是由斯特莱斯克莱德大学用长词开发术语映射技术的项目（Macgregor et al., 2007）。

3．GESIS-IZ 的术语映射方法

语义互操作可用不同的方法实现。有关不同的术语映射方法和映射项目总结可参见 Zeng & Chan (2004, 2006a, 2006b)，Doerr (2001, 2004)和 Hellweg et al. (2001)等的文章。

KoMoHe 项目的重点是交叉语词索引。我们定义交叉语词索引是人工创建的两种受控词表语词间相等、等级和相关的交叉关系。

特别说明，词表关系将是双向的，也就是说，词表 A 的语词和词表 B 的语词建立交叉语词索引的同时，词表 B 的语词也和词表 A 的语词建立了交叉语词索引。双向关系不是必须对称的。例如，系统 A 中的“Computer”一词映射系统 B 中的“Information System”一词，但是系统 B 中的“Information System”一词也可映射到系统 A 的另一个词“Data base”。

我们的方法允许 1:1 或 1:n 的关系存在：

- 相等（＝）表示完全相同的词，同义词和准同义词
- 等级（上位词<；下位词>）
- 关联（^）表示相关的词
- 一个特例是空（0）关系，表示一个词不能映射到另一个词（见表 2 中的第 4 个词）

[5] http://www.oclc.org/research/projects/termservices/
[6] http://www.d-nb.de/eng/wir/projekte/crisscross.htm
[7] http://www.fao.org/aims/
[8] http://hilt.cdlr.strath.ac.uk/

另外，每一种关系必须标记一个相关度（高，中和低）。相关度对于调整关系的质量是次要的也是不可靠的手段，目前我们还没有应用它。

表 2 列出了词表 A 和词表 B 的单向交叉语词索引。

序号	词表 A	关系	词表 B	描述
1	hacker	=	hacking	相等关系
2	hacker	^+	computers + crime	2 个相关关系(^)
3	hacker	^+	internet + security	语词组合(+)
4	isdn device	0		空关系，语词太专指，概念不能映射。
5	isdn	<	telecommunications	下位关系
6	documentation system	>	abstracting services	上位关系

表 2　交叉语词索引的例子（单向）

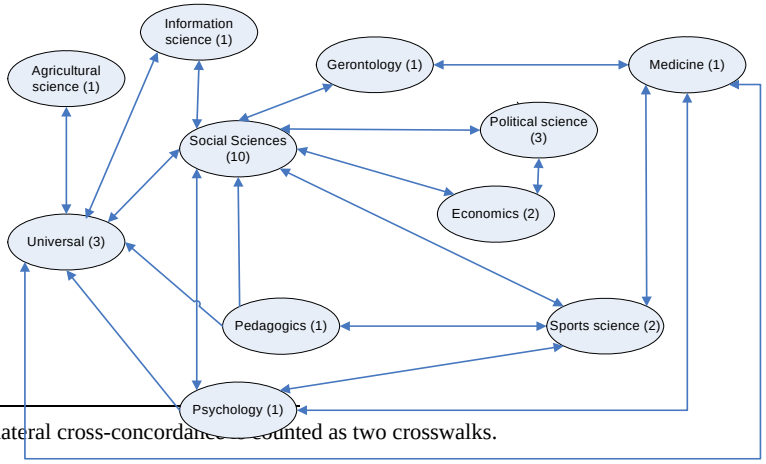
KoMoHe 项目的映射包括词表的全部或大部分。开始映射前要先分析词表的论题和句法的重叠程度。所有的映射都是由研究人员或术语学专家创建的。一个成功的映射的实质是理解语词的含义和语义以及有关词表的内在关系。这包括词干的句法检查和语义知识来查找同义词和其它相关词。

映射过程基于一套实用规则和指导方针（见 Patel et al., 2005）。在语词的映射过程中所有叙词表内的关系（包括范围注释）都要被考虑。建立的关系的检全率和检准率都要在相关的数据库进行核实，这点对于组合语词（1:n 的关系）特别重要。一对一（1:1）的语词关系是首选的，要始终进行词组和相关性调整。

最后，映射的语义要经过专家校对，并且样例要经过实证检验文献的检全率和检准率。所有的事情都考虑了，然而只用交叉语词索引构建术语网络仍然是一项高成本和费时的工作。

3.1 映射方案的结果

到目前为止，从 11 个学科和 3 种语言（德语、英语和俄语）选择的 25 种受控词表已经集成，每个词表的映射语词范围从 1000 到 17,000 个（见图 2 的映射细节）。从 64 种交叉（30 种双向[9]和 4 种单向）中产生了多于 513,000 条语词的关系。图 1 描述了按学科建立的交叉语词索引。



[9] A bilateral cross-concordance counted as two crosswalks.

图 1　KoMoHe 项目的术语映射网。括号里的数字包括在一个学科映射的受控词表数。

该项目主要在如下的受控词表中创建的交叉语词索引（叙词表、叙词列表、分类表和标题表），这些词表都在 vascoda 的主题集合中起着重要的作用。有一些交叉语词索引是从以前的项目 CARMEN[10]和 infoconnex[11]中并入的。

KoMoHe 项目中的词表大部分是德语、英语（8 个）、俄语（1 个），或者是多语种的（如：AGROVOC, IBLK, DDC）。一些词表的语词有英语或德语的翻译（如：THESOZ，PSYNDEX，MESH， INION，STW）。

映射的叙词表（16 个）：

- AGROVOC 叙词表（AGROVOC Thesaurus，AGROVOC）：农业领域的一个词表，包括大约 39,000 个语词，映射到 SWD。
- CSA 应用社会科学索引及文摘叙词表（CSA Thesaurus Applied Social Sciences Index and Abstracts，CSA-ASSIA）：社会科学领域的一个词表，包括大约 17,000 个语词，映射到 THESOZ。
- CSA 叙词表 PAIS 国际标题表（CSA Thesaurus PAIS International Subject Headings，CSA-PAIS）：政治科学领域的词表，包括大约 7,000 个语词，映射到 IBLK。
- CSA 体育教育索引叙词表（CSA Thesaurus Physical Education Index，CSA-PEI）：运动科学领域的词表，包括大约 1,800 个语词，映射到 THESOZ。
- CSA 政治科学索引语词叙词表（CSA Thesaurus of Political Science Indexing Terms，CSA-WPSA）：社会和政治科学领域的词表，包括大约 3,100 个语词，映射到 THESOZ。
- 欧洲语言社会科学叙词表（European Language Social Science Thesaurus，ELSST）：社会科学领域的词表，包括大约 3,200 个语词，映射到 THESOZ。
- INFODATA 叙词表（INFODATA Thesaurus，INFODATA）：信息科学领域的词表，包括大约 1,000 个语词，映射到 THESOZ 和 SWD。
- Psyndex 语词（Psyndex Terms，PSYNDEX）：心理学领域的词表，包括大约 5,400 个语词，映射到 THESOZ， SWD，BISP，MESH 和 BILDUNG。
- 经济标准叙词表（Standard Thesaurus Wirtschaft，STW）：经济学领域的词表，包括大约 5,700 个语词，映射到 THESOZ，SWD，IAB 和 IBLK。
- 教育叙词表（Thesaurus Bildung，BILDUNG）：教育学领域的词表，包括大约 50,000 个语词，映射到 THESOZ， SWD，PSYNDEX 和 BISP。
- 国际外交关系和区域地理叙词表（Thesaurus Internationale Beziehungen und Länderkunde，IBLK）：政治科学领域的词表，包括大约 8,400 个语词，映射到 THESOZ，STW，TWSE 和 CSA-PAIS。
- 社会科学叙词表（Thesaurus Sozialwissenschaften ，THESOZ）：社会科学领域的词表，包括大约 7,700 个语词，映射到 GEROLIT，DZI，FES，CSA-WPSA，CSA-ASSIA，CSA-SA，CSA-PEI， ELSST，IAB，IBLK，STW，SWD，BILDUNG，PSYNDEX，INFODATA 和 BISP。

[10] http://www.bibliothek.uni-regensburg.de/projects/carmen12/index.html.en
[11] http://www.infoconnex.de/

- 经济和社会发展协会叙词表（Thesaurus für wirtschaftliche und soziale Entwicklung，TWSE）：政治科学领域的词表，包括大约 2,800 个语词，映射到 IBLK。
- 社会学索引语词叙词表（Thesaurus of Sociological Indexing Terms，CSA-SA）：社会科学领域的词表，包括大约 4,300 个语词，映射到 THESOZ。
- 德国社会问题研究所叙词表（Thesaurus of the Deutschen Instituts für soziale Fragen，DZI）：社会科学领域的词表，包括大约 1,900 个语词，映射到 THESOZ。
- 德国老年人问题中心叙词表（Thesaurus of the Deutschen Zentrums für Altersfragen，GEROLIT）：老年医学领域的词表，包括大约 1,900 个语词，映射到 THESOZ 和 MESH。

映射的叙词列表（4 个）：

- 联邦体育科学研究所叙词列表（Descriptors of the Bundesinstitut für Sportwissenschaft，BISP）：运动科学领域的词表，包括大约 7,400 个语词，映射到 THESOZ，MESH 和 BILDUNG。
- 弗里德里希－埃伯特基金会叙词列表（Descriptors of the Friedrich-Ebert Stiftung，FES）：社会科学领域的词表，包括大约 4,000 个语词，映射到 THESOZ。
- 劳务市场与职业研究所叙词列表（Descriptors of the Institut für Arbeitsmarkt- und Berufsforschung，IAB）：社会科学领域的词表，包括大约 6,800 个语词，映射到 THESOZ 和 STW。
- 俄罗斯科学院社会科学科学信息机构叙词列表（Descriptors of the Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences，INION）：社会科学领域的词表，包括大约 7,000 个语词，映射到 THESOZ。

映射的分类法（3 个）：

- 杜威十进分类法（Dewey Decimal Classification，DDC）：综合词表，包括数千类号，映射到 RVK。
- 经济文献杂志分类系统（Journal of Economic Literature Classification System，JEL）：经济领域的词表，包括大约 1,000 个类号，映射到 STW。
- 雷根斯堡联合分类法（Regensburger Verbundklassifikation，RVK）：综合词表，包括数千类号，映射到 DDC。

映射的标题表（2 个）：

- 医学标题表（Medical Subject Headings，MESH）：医学领域的词表，包括大约 23,000 个语词，映射到 PSYNDEX，GEROLIT，BISP 和 SWD。
- 标题规范文档（Schlagwortnormdatei，SWD）：综合词表，包括大约 650,000 个语词，映射到 THESOZ，MESH，STW，AGROVOC 和 INFODATA。

图 2 给出了 64 种交叉的总览。THESOZ 是对内和对外映射最多的词表，所以在整个网状结构中显示在最中心的位置。其它的词表像 SWD 或 PSYNDEX 在向其它领域转换中也起了中心作用。DDC-RVK 的映射是唯一没有被连接起来的交叉语词索引。也许，CRISSCROSS 项目所做的映射 SWD 到 DDC 的术语工作可以应用来连接分离的这一对。JEL-STW 的映射是一个单向（一个方向）交叉语词索引的例子，只从 JEL 到 STW。

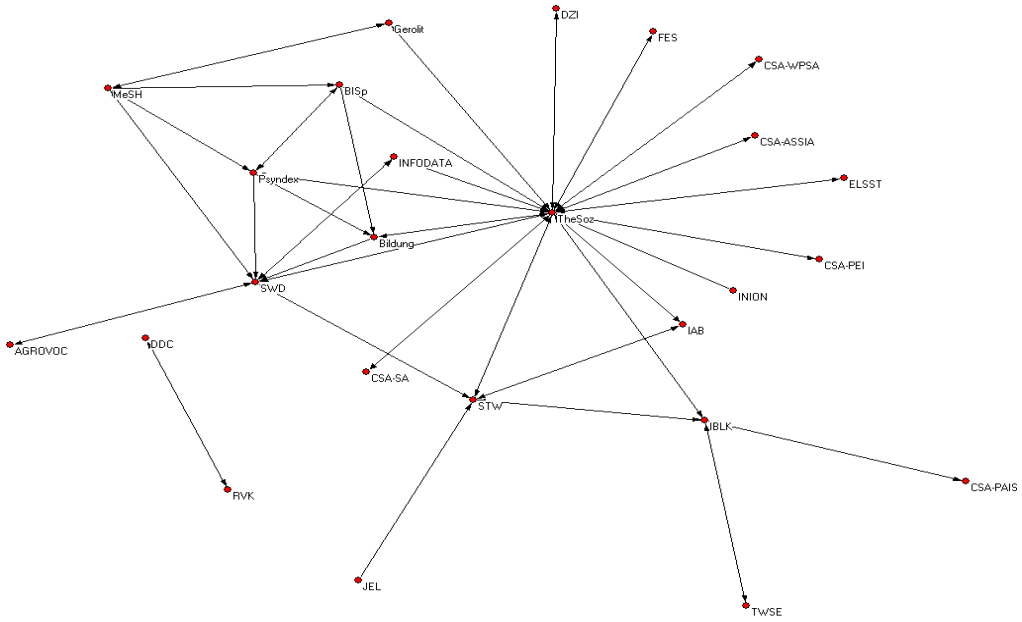


图 2　KoMoHe 项目中的映射词表网

在我们的交叉语词索引数据库中有 513,000 种语词关系可以利用，还有多于 181,000 个唯一的可以检索的概念（唯一的受控叙词或叙词组合或符号）。平均（每一种交叉语词索引）6,500 个开始语词可以映射到目标词表的 3,600 个语词（每个词 1.2 种关系）。

图 3 列出了项目中的关系类型的分布情况（与表 2 对比）。相等关系（大约 45%）是两个语词间最常用的关系。只有 12%的关系是‘空关系’（没有可以映射的语词）。

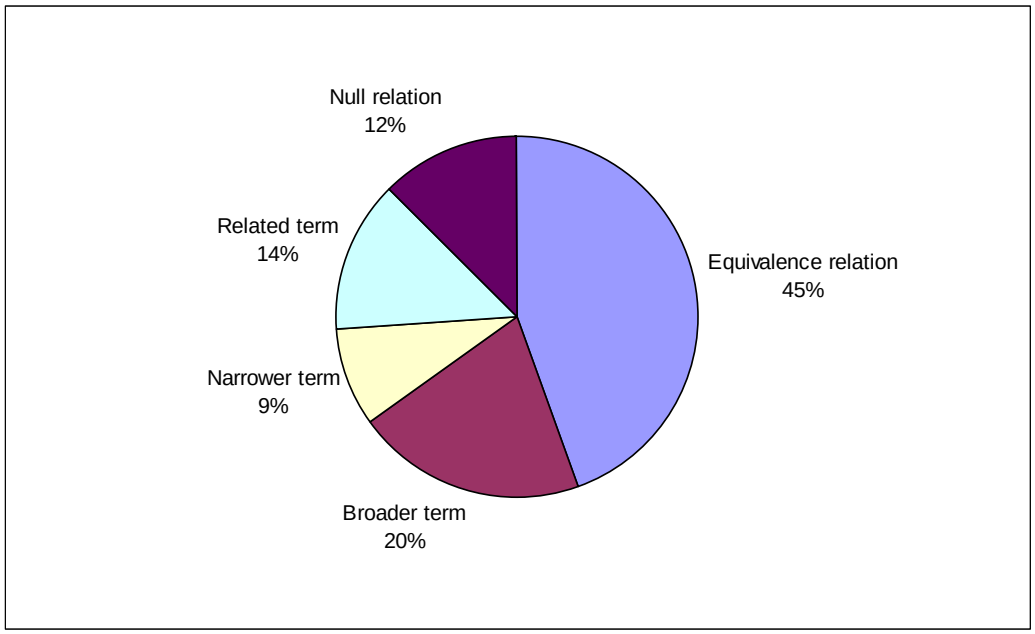


图 3　所有交叉语词索引关系类型的分布

3.2 交叉语词索引的应用

创建一个关系数据库来存储交叉语词索引供以后使用。这些关系结构可以恰当地获取不同

受控词表、语词、语词组合和关系。词表和语词以列表的形式表现，彼此独立而不用关注词表的连结结构。受控词表语词的正确拼写和大写字母都要规范化。语词组合（如，computers + crime 和 hacker 相关）也按分离的概念存储。

从数据库查找和检索术语数据，建立一种网络服务（叫作异构服务或图 4 中的 HTS，见 Mayr & Walter, 2008）来支持交叉语词索引检索，包括检索单个开始语词、映射语词、来源词表和目标词表还有各种不同类型的关系。一种应用是利用相等关系在受控词表语词列表中查找检索词，然后自动为查询添加可利用词表中的所有相等语词。如果受控词表使用不同的语言，异构服务将提供来源词到选择的受控词语言的翻译。如果来源查询包括布尔命令，进行查询扩展（如，每一个词单独扩展）后仍将完整保留。因为考虑到性能，交叉语词索引查询扩展并没有区分不同数据库和它们首选的某一概念的受控词表语词，而是将所有相等的语词都添加了。原则上，术语网络扩展查询应该用来源查询词的同义词或准同义词。

4．交叉语词索引评价

4.1 一般问题

虽然对术语映射的需要是一些团体和很多正在承担的映射项目普遍公认的，但是项目成果的实际效率和用途很少进行严格地评价。在术语网络创建的映射中有很多问题被提出：

- 一个概念可以找到多少种表达方式？
- 哪些概念是相关的？
- 词表的上位和下位只是范围上的吗？
- 哪些术语彼此最相似？
- 哪些学科/主题是临近的或相差很远？
- 在一个特定项目中不同数据库或受控词表间存在多少重叠？

最重要的问题是，大部分的映射已经创建了，但在实际检索中这些映射是怎样发挥效率和作用的。在一个有很多数据库的信息门户中，交叉语词索引能否实现分布式检索是至关重要的问题。它们能在不同语言间搭建桥梁从而能用同样的查询式跨不同的数据库实现无缝检索吗？

当评价术语映射时，出发点的分析是很重要的。什么是要检查的：映射本身的质量或相关检索的质量？映射的质量是改进检索质量的先决条件。在 KoMoHe 项目的交叉语词索引中，每一个映射都是经过合作机构的学科专家核对的。人工的创建和仔细的核对保证了映射正确、恰当和一致。

交叉语词索引的本质特征（和它们对检索的影响）不同程度地依赖映射的受控词表和交叉语词索引创建过程中的外界因素。例如，交叉语词索引的创建日期能影响每一个开始语词的关系数量。项目的早期，建立的关系比较少。交叉语词索引是一批专家从早期的一个项目（CARMEN）中讨论并经过很多选择建立的。受控词表或标引实践的变化都能影响交叉语词索引的质量。其它的差异如下：

- 来源/目标术语的规模
- 词表先组和后组的不同
- 关系的数量
- 映射目标语词的数量（覆盖率/重叠率）

- 关系的分布（相等，上位词，下位词，相关词，空关系）
- 相关度的分布（高，中，低）
- 同样语词的映射
- 专指度的不一致（如，词表在范围上有的很宽泛有的很专指）
- 映射语词的组合（映射是由多个目标语词组合的）

定量分析能够洞察交叉语词索引的基本特征，但是也不能决定使用特定映射在检索中能获得质量改进。我们设计了一个信息检索试验，目的是在一个真实世界的检索场景中来评价交叉语词索引的应用。

4.2 信息检索试验设计

在检索中，一些因素在评价术语映射的质量时发挥作用：交叉语词索引本身，有关的数据库的内容、内容的覆盖率或重叠率、检索界面、或检索排序的利用。目标是评价交叉语词索引的影响，实际的检索条件（界面、排序方法等）尽可能的保持稳定。

利用交叉语词索引的基本思想是将检索词转换成利于实现跨数据库和术语检索的其它术语。交叉语词索引的杠杆作用应该能扩大检索空间，纠正不明确和不精确的查询式，从而为特定查询找到更相关的文献。

在检索中应用交叉语词索引也提出一个警告：它们会影响检索过程本身的速度和易用性。术语映射的技术应用前提是它们的利用并没有引起检索者的注意。映射应该改进检索经验而不增加信息系统用户的工作。在评价过程中必须是通过使用一种严格地自动方法引入交叉语词索引，而不是以个人重新构建查询式的形式来进行人工干涉。

在检索中设计了两种信息检索试验来评价交叉语词索引的质量：

试验 1：应用语词映射能改进无转换的主题（如受控词表）检索吗？

在试验 1 中，一个查询词被转换成受控词表（A）的语词，然后来检索用受控词表（B）标引的书目数据库的受控词字段。检索在 A→B 交叉语词索引的帮助下重复进行，将来源受控词表的检索词转换成目标数据库的受控词。图 4 给出了这个过程的图形显示。检索结果是经过比较的。

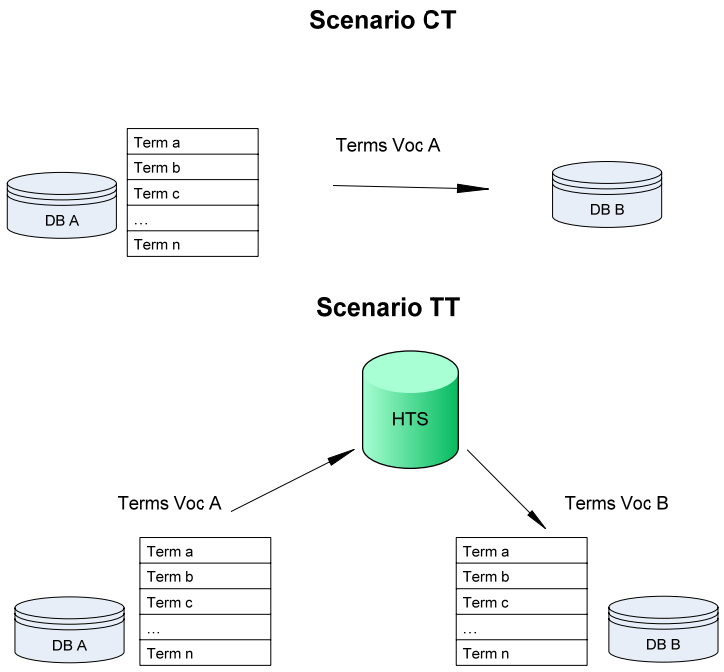


图 4　交叉语词索引信息检索评价体系

Scenario CT = 受控词检索，Scenario TT =语词转换检索，HTS =异构服务

如果两种检索的结果一样，则交叉语词索引应用的语词转换并没有起作用。如果检索结果更遭，则交叉语词索引有负面影响，如果检索结果改进了，则认为交叉语词索引的使用对检索有积极的作用。

试验 2：应用语词映射能改进自由文本检索吗？

在试验 2 中，最初的查询是在自由文本检索场景（大多数的检索者不能使用恰当的受控词表）。自由文本检索的查询词并不在受控词汇字段，而是在题名和文摘字段。在试验中，最初的查询词首先在数据库的自由文本字段进行检索。第二，查询词在存在语词映射的交叉语词索引中查找。从交叉语词索引中得到的新词将加到最初的查询词里。

举例说明试验 1 和试验 2 的不同：用自然语言查询有关“family relations”的文献。“family relations”已经是词表 A 中的一个受控词汇，不需要转换成受控词表语词，所以首先用它来检索数据库 B 的受控词字段：*试验 1 CT：Family relation*。

A→B 的交叉语词索引，实现了词表 A 中的词组“family relations”到词表 B 中的词“family”和“social relations”组合的映射。因此第二次在数据库 B 中检索：*试验 1 TT：Family AND social relations*。

对于试验 2，最初的查询词检索是在数据库 B 的自由文本字段（题名，文摘和受控词）：*试验 2 FT：Family relations*。

因为查询词在词表 A 中出现，所以在交叉语词索引中能找到查询词的映射语词。这些词加到最初的查询中并再次在数据库 B 中进行自由文本字段的检索：*试验 2 FT+TT：Family relations OR (Family AND social relations)*。

试验 1 只检索受控词字段，而试验 2 也检索其它的查询词可能出现的字段（题名或文摘）。

试验 2 通常是一个不可靠的试验，因为对查询添加了映射词，而不是像试验 1 那样取代原来的查询词。因为试验 2 也检索了试验 1 的字段，试验 2 可以认为包含了试验 1。然而在试验 2 中，因为不是所有的查询词都出现在最初的受控词表，并且只能找到很少的映射词，所以有很少的语词同时出现。

对于查询式构建，测试数据库的生产者或主机需提供帮助确保进行真实的查询。它们需根据日常的经验提供 3-10 个查询词（平均 6-7 个），这些词要转换成测试数据库使用的受控词汇。自然语言的自由文本查询每个查询式包括大约 1-3 个词，然而对于受控词的布尔查询包括大约 2-6 个词。对于所有信息检索试验中的语词映射，只有相等的关系被使用。信息系统中列出的所有文献都要被检索直到达到 1,000 篇相关文献。最后，每一次试验的文献结果集用来评价提问的相关性。

要评价交叉语词索引的效果，信息检索的传统测量方法，基于检索文献相关度评价的检全率和检准率都要用到。下面是测量方法的分析：

- Retrieved 检索到的文献：检索到文献的平均数量（所有检索类型）
- Relevant 相关的文献：检索到的相关文献的平均数（所有检索类型）
- Rel_ret：特殊检索类型检索到的相关文献的平均数
- Recall 检全率：检出的相关文献占所有相关文献的比例（一种检索类型所有查询式的平均）
- Precision 检准率：检出的相关文献占所有检出文献的比例（一种检索类型所有查询式的平均）
- P10：检准率>＝10 的检索文献
- P20：检准率>＝20 的检索文献

P10 和 P20 的计算是为了表现一个真实的检索场景，因为用户通常不会去看最后一页或两页的检索结果。对于不按相关度排序而按年或作者排序的检索系统来说，P10 和 P20 是没有意义的。

4.3 交叉语词索引和数据库试验

对于两种试验，交叉语词索引是按学科（学科内和学科间）和语言（单语或双语）划分的。学科内交叉语词索引的词表范围大部分是在社会科学领域，同时项目中创建的交叉语词索引也都固定在那些学科。学科间交叉语词索引的映射词表都在经济，医学，政治科学，心理学和社会科学领域。单语种的交叉语词索引包括德语词表；双语种的交叉语词索引包括德语和英语词表。表 3 给出了每次试验测试的交叉语词索引数量的一个总览：

试验 1：受控词检索	
学科内交叉语词索引	5 （1 个双语）
学科间交叉语词索引	8
试验 2：自由文本检索	
学科内交叉语词索引	6 （1 个双语）
学科间交叉语词索引	2

表 3　交叉语词索引试验的数量

交叉语词索引这种划分模式的背景是假设同一学科（学科内）词表间的交叉语词索引有很多重叠部分，并且包含很多相同的语词，因此对检索结果的影响比学科间的交叉语词索引要小。而不同自然语言词表间（如英语→德语）或不同分类系统间（如 DDC → LCC）的语词映射将有较大的影响，因为相同的语词或重叠部分同时出现更不可能。

试验的数目数据库包含 7 万到 1600 万的文献，大部分是在德国制作和维护的。它们中有文摘和索引数据库，也有图书馆目录。表 4 给出了测试数据库和相关词表的一个总览：

词表	学科	数据库	数据库中的文献
TheSoz – 社会科学叙词表 (GESIS-IZ)	社会科学	SOLIS	345,086
DZI – 德国社会问题研究所叙词表	社会科学	SoLit	151,925
SWD – 标题规范文档	综合（社会科学摘录）	USB Köln Sowi OPAC	72,729
CSA – 社会学索引语词叙词表（剑桥科学文摘）	社会科学	CSA Sociological Abstracts	294,875
Psyndex – Psyndex 语词	心理学	Psyndex (ZPID)	Ca. 200,000
STW – 经济标准叙词表	经济学	Econis (ZBW Kiel)	Ca. 3,000,000
IBLK – 国际外交关系和区域地理叙词表（欧洲叙词表）	政治科学	World Affairs Online WAO (SWP Berlin)	643,420
Mesh – 医学标题表	医学	Medline (Dimdi)	Ca. 16,800,000

表 4　KoMoHe 信息检索试验中的词表和数据库

很多交叉语词索引和它们各自的数据库能在家里测试，通过开放资源信息检索系统 Solr[12]，对每个数据库用同样的过程和排序模块标引文献。对于不能在家里使用的数据库，请求主机为每个预定的查询提供排序结果列表。

对于大部分的数据库，语词映射测试是双向的，从词表 A 到词表 B（A→B）和从词表 B 到词表 A（B→A）。因为它们组成不同的检索（不同的查询词依赖来源词表）和不同的数据库，因此它们是相互独立的。

5. 评价结果

5.1 试验 1：受控词检索

试验 1 评价用词表 B 的受控语词（经过语词映射转换）（TT）取代词表 A 的语词（CT）在数据库 B 检索能否改进检索效果。如果语词映射不精确或不明确，或者词表重叠多，那么从原始查询转换到映射查询可能会给查询式构建带来噪音，这将妨碍检索的质量。

表 5 给出了 13 个测试的交叉语词索引的平均结果的一个总览。最后一行列出了两种不同检

[12] http://lucene.apache.org/solr/

索类型的差异百分比。

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	156.5	144.8	42.0	0.3152	0.2214	0.1987	0.1748
TT	325.4	144.8	88.2	0.6047	0.3391	0.3052	0.2848
				91.8%	**53.2%**	**53.6%**	**62.9%**

表 5　试验 1 中所有交叉语词索引的评价结果（N=13）

利用语词转换检索到的文献数量加倍，包含查询词的更多文献被检出。检全率增加了几乎 100％，检准率增加了 50％多。在特定检索中利用交叉语词索引与没有语词转换检索相比，不仅能检索到更多相关文献（检全率）而且也更准确（检准率）。

然而，这种巨大的改进主要是因为双语交叉语词索引中英语和德语的翻译。反之，单语种的语词映射不会如此有效，因为映射语词是一样的，而不像翻译映射的情况。表 6 列出了从试验中除去双语种交叉语词索引后的检索结果：

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	169.6	141.2	45.5	0.3415	0.2399	0.2153	0.1894
TT	320.5	141.2	87.6	0.6113	0.3431	0.3126	0.2877
				79.0%	**43.1%**	**45.2%**	**51.9%**

表 6　试验 1 中所有的单语种交叉语词索引的评价结果（N=12）

因为语词重叠问题，跨越两个学科（学科间的）或同一学科领域内的（学科内的）交叉语词索引的检索结果应该不同。如果按学科区分试验结果，我们可以看到检索结果会有重大的变化。对于学科内的交叉语词索引，检全率和检准率增加的并不多。通常在信息检索中，检全率和检准率是彼此互逆的关系（如果检全率增加，则检准率降低），所以检准率小点或负数变化是实际期望的。

表 7 列出了所有单语种学科内的交叉语词索引的检全率和检准率的平均值。对于单语种学科内的交叉语词索引，检准率和检全率也增加了但是比所有的交叉语词索引要少得多。

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	126.6	101.3	36.2	0.3726	0.2491	0.2002	0.1637
TT	238.9	101.3	59.9	0.5189	0.3335	0.2784	0.2352
				39.3%	**33.9%**	**39.1%**	**43.7%**
单语种（Monolingual）							
CT	158.2	79.7	45.3	0.4657	0.3113	0.2503	0.2046
TT	202.9	79.7	51.0	0.5174	0.3441	0.2939	0.2315
				11.1%	**10.5%**	**17.4%**	**13.2%**

表 7　试验 1 中学科内交叉语词索引的评价结果（N=5）

学科间交叉语词索引的应用对于检全率和检准率有特别的改进。检全率和检准率的增加比平均的交叉语词索引明显的多（见表 8）。

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
CT	175.2	171.9	45.6	0.2794	0.2041	0.1978	0.1817
TT	379.4	171.9	105.9	0.6583	0.3426	0.3220	0.3157
				135.6%	**67.8%**	**62.8%**	**73.7%**

表 8　试验 1 中学科间交叉语词索引的评价结果（N=8）

利用交叉语词索引对于受控词检索还是有很大的积极作用的。结果集不仅更大了也更精确了。最大的影响是跨越更多学科的交叉语词索引。

5.2 试验 2：自由文本检索

试验 2 评价在自由文本检索（FT）中加入受控词能否改进检索效果，这些受控词是通过映射自然语言查询词到数据库的受控词表（FT-CK）而获取的。对于试验中的个别查询，查询词是没有变化的因为没有找到匹配的受控词。表 9 列出了所有 8 个试验的交叉语词索引的检索结果。

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	155.3	106.4	56.2	0.6026	0.4551	0.4101	0.3682
FT-CK	266.8	106.4	72.8	0.7273	0.3934	0.3203	0.3083
				20.7	**-13.6**	**-21.9%**	**-16.3%**

表 9　试验 2 中所有交叉语词索引的评价结果（N=8）

结果表明不仅而且更多的相关文献被找到。平均检全率只增加了 20%。通常，受控词简单地加到查询中也能改进检索结果。然而，可以看到检准率下降了，不过没有检全率增加的多。

表 10 列出了同一学科内的交叉语词索引映射语词的检索结果，表 11 列出了 2 个跨学科的交叉语词索引的检索结果：

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	163.8	115.8	60.2	0.5934	0.5025	0.4635	0.4090
FT-CK	244.9	115.8	77.1	0.7096	0.4449	0.3826	0.3681
				19.6	**-11.5%**	**-17.5%**	**-10.0%**

表 10　试验 2 中学科内交叉语词索引的评价结果（N=6）

	Retrieved	Relevant	Rel_ret	Recall	Precision	P10	P20
FT	129.9	78.2	44.3	0.6303	0.3129	0.2500	0.2459
FT-CK	332.4	78.2	59.8	0.7805	0.2388	0.1333	0.1292
				23.8%	**-23.7%**	**-46.7%**	**-47.5%**

表 10　试验 2 中学科间交叉语词索引的评价结果（N=2）

与试验 1 的分析相反，学科内和学科间的交叉语词索引差别不是很大。对于两个结果集，检全率比单一的自由文本检索都增加了然而检准率都下降了。学科间的交叉语词索引的检全率增加的少一点而检准率也降的少一点。可能是因为只评价了 2 个学科间的交叉语词索引，这会歪曲结果，可能不足以表示一种趋势。

如果将受控语词和自然语言语词更好的结合在一起构造查询式，而不只是简单地添加在一

起，这样自由本文检索试验的结果可能会有更多的改进。一种可能性是推理地转换自然语言语词到受控语言语词（自动地），这样才能为不同受控词表和数据库间的映射提供更好的机会（see Mayr et al., 2008 for one approach）。

总之，信息检索试验表明了在异构数据库中运用交叉语词索引检索的积极作用。所有交叉语词索引的检索结果都改善了，然而，学科间的交叉语词索引对检索结果有更高（更积极）的影响。两个试验场景中，与没有使用交叉语词索引的查询相比，所有的交叉语词索引都找到了更多相关的文献；在特殊情况下，检索结果集甚至更精确（检准率也增加）。

6. 结论和展望

在展示了交叉语词索引对检索结果的积极影响后，我们计划在 vascoda 门户中应用它们。我们已经在德国社会科学信息门户 sowiport[13]的检索中利用了很多交叉语词索引，sowiport 提供书目和其它信息资源（包括 15 个数据库，10 个不同词表和大约 250 万书目参照）。

在 sowiport 中的应用，也像试验一样，只利用相等关系在受控词表中查找检索词，然后自动将所有可利用的映射词表中的相等词添加到查询式中。因此和试验 2（自由文本检索）中的应用方法一样。布尔命令功能和查询词组间的分隔符在查询扩展（如，每个查询部分单独扩展）后都将完整保留。语词映射是自动的，检索者是看不见的，只用一个小的图标来表示转换（点击这个小图标列出可添加的查询词）。

为了进一步的研究和存储，要创建一个关系数据库来存储交叉语词索引。要从数据库查找和检索术语数据，建立一种网络服务（叫作异构服务，见 Mayr & Walter, 2008）来支持交叉语词索引检索，包括检索单个开始语词，映射语词，来源词表和目标词表还有各种不同类型的关系。然而，数据库本身也可以被查询。术语映射数据和网络服务可用作研究目的使用。一些映射已经在特殊领域开始使用，如在 CLEF（交叉语言评价系统）中检索会议资料（Petras, Baerisch & Stempfhuber, 2007）。交叉语词索引其它的作用和应用像转换，互相作用或互操作都将在以后的研究中探索。

另一种方式是基于 SKOS（简单知识组织系统）[14]语义网络的存储和查询。SKOS 标准是从 W3C 的工作草案形成的一项标准，它的目的是支持受控词表在语义网络中的应用。草案包括词表映射这一部分。一旦 SKOS 标准稳定了，我们将使我们的映射数据以这种格式应用。

值得注意的一点是如果有的资源没有可用于直接映射的方法，可通过一个支点词表创建映射。如果词表 A 映射到词表 B，词表 B 映射到词表 C，或许可以通过支点词表 B 的映射信息创建 A→C 的映射。我们希望随着表述和交换标准的开发，更多的映射和词表能应用到以后的研究中。

致谢

这个项目是由 BMBF 资助的，授权号：01C5953。我们衷心感谢所有项目合作者的协作。我们还要特别感谢我们的同事 Anne-Kathrin Walter 和 Stefan Baerisch 的帮助，他们实现了大部分的技术基础设施并帮助测试评价。

[13] http://www.sowiport.de/
[14] http://www.w3.org/2004/02/skos/

参考文献

Chan, L. M., & Zeng, M. L. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part I: Achieving Interoperability at the Schema Level. D-Lib Magazine, 12(6).

Depping, R. (2007). vascoda.de and the system of the German virtual subject libraries. In A. R. D. Prasad & D. P. Madalli (Eds.), International Conference on Semantic Web & Digital Libraries (ICSD 2007) (pp. 304-314). Bangalore, India: Documentation Research & Training Centre, Indian Statistical Institute.

Doerr, M. (2001). Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 1(8).

Doerr, M. (2004). Semantic interoperability: Theoretical Considerations.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., et al. (2001). Treatment of Semantic Heterogeneity in Information Retrieval. Bonn: IZ Sozialwissenschaften.

Krause, J. (2003). Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution. Paper presented at the IFLA 2003, World Library and Information Congress: 69th IFLA General Conference and Council, Berlin.

Liang, A. C., & Sini, M. (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. New Review in Hypermedia and Multimedia, 12(1), 51-62.

Macgregor, G., Joseph, A., & Nicholson, D. (2007). A SKOS Core approach to implementing an M2M terminology mapping server. Bangalore, India.

Mayr, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed Digital Libraries: treatment of term vagueness and document re-ranking. Library Review, 57(3), 213-224. URL: http://www.emeraldinsight.com/10.1108/00242530810865484

Mayr, P., & Walter, A.-K. (2008). Mapping Knowledge Organization Systems. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte der Wissenorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 80-95). Würzburg: Ergon.

Panzer, M. (2008). Semantische Integration heterogener und unterschiedlichsprachiger Wissensorganisationssysteme: CrissCross und jenseits. In H. P. Ohly, S. Netscher & K. Mitgutsch (Eds.), Fortschritte in der Wissensorganisation, Band 10. Kompatibilität, Medien und Ethik in der Wissensorganisation (pp. 61-69). Würzburg: Ergon.

Patel, M., Koch, T., Doerr, M., & Tsinaraki, C. (2005). Semantic Interoperability in Digital Library Systems.

Petras, V., Baerisch, S., & Stempfhuber, M. (2007). The Domain-Specific Track at CLEF 2007, Cross Language Evaluation Forum Workshop (CLEF) 2007. Budapest.

Tudhope, D., Koch, T., & Heery, R. (2006). Terminology Services and Technology: JISC state of the art review.

Vizine-Goetz, D., Hickey, C., Houghton, A., & Thompsen, R. (2004). Vocabulary Mapping for Terminology Services. Journal of Digital Information, 4(4).

Vizine-Goetz, D., Houghton, A., & Childress, E. (2006). Web Services for Controlled Vocabularies. ASIS&T Bulletin, 2006 (June/July).

Zeng, M. L., & Chan, L. M. (2004). Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. Journal of the American Society for Information Science and Technology, 55(3), 377-395.

Zeng, M. L., & Chan, L. M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. D-Lib Magazine, 12(6).