

基于ISO 25964的词表互操作实现探析^{*}

贾君枝

(山西大学经济与管理学院, 太原 030006)

摘要: 词表互操作是语义层面实现跨库、跨语言、跨平台检索的必经途径。随着网络知识组织工具的应用普及, 不同类型词表间的映射将是互操作实现最有效的方法。当前, ISO 25964是保证词表间映射流程规范化、提高映射质量的重要标准。本文重点分析梳理该标准的核心内容, 进一步探讨词表间互操作实现步骤及语义化描述, 从而为各机构开展互操作项目提供指导及建议, 并对其词表互操作实现的语义表述提供范例。

关键词: ISO 25964; 互操作; 词表; 叙词表

中图分类号: G254

DOI: 10.3772/j.issn.1673-2286.2016.12.002

1 引言

1997年, 网络知识组织系统 (Networked Knowledge Organization Systems, NKOS) 被首次提出, 旨在发展适应网络知识组织系统的数据模型, 在传统信息组织工具 (如同义词表、分类表、叙词表) 的基础上, 出现了新型网络知识组织工具 (如概念地图、本体、分众分类法、术语表)。在一定意义上, 这些网络知识组织工具都可统称为词表, 以概念为描述对象, 共同描述概念间的关系, 而概念的外化形式实为词汇。由于不同词表构建角度及方式不同, 所涉及的语种、涵盖领域存在差异, 应用这些词表描述网络资源形成的标引及检索结果也有所差异, 使得用户采用不同的检索方式获取所需资源。随着数据库集成应用与发展, 出现了跨库、跨语言、跨领域的操作, 由于不同数据库采用的标引词表不同, 所以会频繁地涉及不同词表间互操作的问题。目前, 解决不同词表互操作实现的最有效方法是实现词表间的映射, 即在两个或多个词表间建立词汇对应, 使词表所标引的结果间也相应地建立对应。比如, 联合国粮农组织开展的AGROVOC项目^[1]、美国国立医学图书馆开展的UMLS

Metathesaurus项目^[2]、联合国信息系统委员会开展的High Level Thesaurus Project项目^[3]都属于综合领域多语言、多类型词表的互操作项目, 但由于映射机构人员素质、词表间的差异度、映射的严格性等各种主客观条件的差异, 导致映射的质量与结果不同。

因此, 为提高映射质量, 确保映射流程的规范化管理, 国外一些机构开始制订词表的互操作标准。美国国家标准协会/美国国家标准信息协会发布Z39.19-2005定义了互操作的需求、影响互操作的因素、单语言受控词表映射成其他语言、查询、索引所采用的互操作方法^[4]; 英国编制BS 8723-4: 2007等标准均对词表互操作中方法与模型的选择、过程的实施、具体问题的应对等与词表互操作有关的内容做出规范指导^[5]。国际标准ISO 25964是国际标准化组织于2007年4月以N458文件的形式发布的叙词表国际标准修订项目计划。该计划将《单语种叙词表编制准则》(ISO 2788) 和《多语种叙词表的编制规则》(ISO 5964) 进行更新扩展形成ISO 25964。ISO 25964包含两部分, 第一部分定义叙词表编制标准^[6], 第二部分阐述叙词表与其他词表的互操作, 对不同叙词表间、叙词表和其他类型词表间建立映射及

^{*} 本研究得到山西省高校基金项目“131”领军人才工程“叙词表与其他词表的互操作标准” (编号: 2015052002) 和山西省高等学校中青年拔尖创新人才项目 (编号: 2016052002) 资助。

维护提供建议,为词表间互操作实践提供参考依据^[7]。该互操作标准更强调词表间概念层面的互操作,即语义匹配问题,而不是简单的语词形式匹配。

本文旨在分析ISO 25964-2的基础上,梳理其核心内容,探讨词表间互操作的实现步骤,从而为各机构开展互操作项目提供指导及建议,并对词表互操作实现的语义表述提供范例。

2 词表间互操作实现步骤

随着词表类型多样化发展,互操作的难度也在增加。词表间的互操作需要在明确各类型词表特点的基础上,以专业领域专家为核心,按照既定的规范步骤实施映射。现将其总结为五大步骤:明确映射目标、确定映射对象、选择映射模型、确定映射类型、实施映射。

2.1 明确映射目标

词表间互操作目的是将一种词表的表达式转换成另外一种或几种词表的对应表达式^[7]。从语义交流角度看,通过设置此类交流渠道,信息发送方和接收者能理解彼此交换、共享的信息含义,机器间也可以准确地解释及交换信息。

确立明确的映射目标,是互操作实现的首要步骤。不同的映射目标会造成映射模型及类型的差异,映射目标取决于映射结果的最终去向,属于用户需求导向。从当前的词表映射实践看,映射目标具体分为三方面。

(1) 生成新词表。为提高词表领域的覆盖范围,发挥不同词表间的互补性,通过互操作方式建立新词表,该词表与多源词表建立概念对应,使新词表在未来信息资源处理环境中发挥重要作用。如美国国立医学图书馆创建的一体化医学语言系统项目(Unified Medical Language System, UMLS)中的超级叙词表,由来自40多个生物学受控词表、术语表、分类表、专家系统中的词汇、词典及工具性词表构成,其通过概念属性的方式注明每个概念的来源词表,实则建立了与来源词表的映射关系。我国的《中分表》即为《中国图书馆分类法》与《汉语主题词表》的相互映射表,该表为标引人员同时进行分类、主题标引提供便利。

(2) 用于数据收割。由于不同数据库或系统平台采用的标引词表、元数据格式不同,导致元数据记录差异较大。数据收割需要从不同的数据库或系统平台自

动、持续地获取数据,并在此基础上提供应用服务。数据收割过程中需建立统一的元数据格式,将来源库数据与本地元数据建立映射关系,实现数据表达的一致性。欧洲图书馆采用此方式收割上千家成员馆的数据,将编码档案著录标准、都柏林核心元素集、机读目录、元数据对象描述模型等标准与概念参考模型(CIDOC CRM)建立映射^[8]。

(3) 用于联邦检索。联邦检索是同时从不同数据源检索数据,并将检索结果整合后实时提交给用户,以实现基于提问的跨库/系统检索。最关键的问题是联邦检索系统需要处理数据源之间的异构,主要体现在使用专用词汇的共享和交流上,即语义层面的异构,这将涉及用户检索提问与不同数据源匹配、数据源本身所采用的描述性词表差异的问题。因此,需要将联邦检索系统的中心词表与各数据源的词表建立映射,或者将各数据源的词表间建立映射,用于将用户检索提问式转换成各数据源需要的表达形式。

2.2 确定映射对象

确定映射对象包含三方面内容,明确来源词表和目标词表、映射方向、映射的范围和数量。

当前映射对象实际为词表,根据映射方向不同,分为来源词表和目标词表。来源词表作为起始点词表,需要在目标词表找到对应语词。

映射方向分为单向映射和双向映射,双向映射的工作量较大。对一些应用而言,如来源词表的用户要检索目标词表标引的资源,只需建立单向映射,通常采用箭头指向表示映射方向,单箭头表示单向,双箭头表示双向,由来源词表指向目标词表,确立一个方向的映射后,再进行反方向映射已有的映射可以作为参考借鉴,但不能完全借鉴,需要进一步鉴别。

在上述基础上明确映射的范围及数量。映射的范围分为全部映射和选择式映射。全部映射即将来源词表的词汇尽可能在目标词表中找到对应语词。选择式映射即部分映射,根据应用需要对来源词表的部分词汇建立映射,若来源词表与目标词表间差异较大,仅需在重叠领域建立映射;若仅为处理索引/目录数据库的数据,则只需对应用到目标数据库的部分词汇集建立映射。若仅将中文书目数据库的中图分类号与国会图书馆分类号建立对应,由于缺乏来源词表概念间明确的语义关系,映射质量会受到影响。一旦映射范围确定,

映射工作量的大小也随之确定。词表的规模越大,复杂度越高,映射的工作量越大,映射难度也随之增加。映射的数量决定是否有必要为每个概念建立多个映射。通常情况下,有精确匹配的概念不需做多个映射;在没有精确匹配的条件下,为概念建立多个映射,即建立非精确、上下位及相关映射,对后期的应用有重要的参考价值,对扩充查询词、标引词集合有很大的帮助。

2.3 选择映射模型

映射模型定义了映射采用的方法及其管理模式,通常取决于映射的目标及其条件。ISO 25964将映射模型划分为三种类型:结构统一模型、直接连接模型、中心结构模型。两词表间映射相对简单,以直接连接模型为主,两个以上词表间映射需考虑映射结构的选择。

(1) 结构统一模型,适用于所映射的词表具有完全相同的结构,仅存在表达语言、符号及其编码模式上的差异。如《汉语主题词表》中文版与《汉语主题词表》英文版,《汉语主题词表》MARC格式与《汉语主题词表》SKOS形式都属于此类结构模型。相比其他模型而言,该模型结构操作简单,仅是一种语言/符号系统表达形式的词表转换成另外一种语言/符号表达系统,映射类型单一,属于精确匹配类型,适用于跨语言检索或不同操作平台的使用。

(2) 直接连接模型,适用于所映射的词表结构差异化较大的情况(符合词表多样性条件),强调词表间两两对应的关系。对于相同或者不同类型的词表,逐一对词表的概念进行分析,建立与对应词表的映射关系。如《中国图书馆分类法》与《杜威十进分类法》(Dewey Decimal Classification, DDC)的映射属于相同类型词表的映射,LCSH与DDC的映射属于不同类型词表的映射。该模型结构是映射模型中最常用的,适用于2个或3个词表间的对应,映射工作量较大,维护成本高。

(3) 中心结构模型,适用于所处理的词表数量多,覆盖领域广且类型复杂的互操作情况。该模型将一个词表作为中心表,其他词表作为卫星表,建立中心表与卫星表间概念映射。相比直接连接模型,其优点在于:以中心表作为中介或者桥梁,将两个完全不同的词表建立间接映射,避免差异度极大的词表映射情况出现。因为词表间映射数量减少,相应也减少映射的工作量。但同时存在局限性:对中心表的要求较高,中心表的概念

集合需满足一定的覆盖度与专指度,应尽可能包含所有卫星表的概念。如缺乏符合条件的中心表,可构建一个新词表;各卫星表间为间接映射,“卫星表1”通过中心表建立与“卫星表2”的映射,形成的连续映射结果质量较低;如连接的相关映射类型会无限放大关联范围,最终可能得到相关映射,但事实上是无关的。

2.4 确定映射类型

映射类型指定义两个词表之间、概念之间的映射关系。映射类型应依据映射目标及词表类型对应择取,比如本体与本体的映射类型通常多于传统分类表、叙词表的映射类型,在于其本身的语义关系较丰富。当前,常用的映射类型主要与叙词表内部概念关系一致,有等同映射、等级映射、相关映射,等同映射是三种映射中最常用的类型。通常以相关度大小确定映射类型的选择顺序,首选等同映射,其次是等级映射,最后为相关映射。为提高映射精确度,每一种类型还可进行细分。

(1) 等同映射。来源词表与目标词表的概念范围相同时,建立等同映射,通常采用EQ(equivalence)表示。根据目标词表概念组合程度主要分为简单等同和复合等同。简单等同仅与目标词表中一个概念建立对应关系,复合等同则与目标词表的多个概念的组合形式建立对应关系,交叉复合等同是与两个概念的交集建立对应关系,并列复合等同则是与两个概念的并集建立对应关系。另外,当目标词表的概念不足以完全表示来源词表的概念时,则需增加其他限定词表(如复分表)共同表示,这种情况也属于复合等同。除此之外,根据词表间概念的相似程度,等同映射还可分为精确等同、非精确等同及部分等同。精确等同表示两个概念范围完全相等,可以彼此替代使用。由于词表间差异较大,非精确等同占比较高,部分等同相对判断难度较大,因此以建立等级映射为优先考虑。

(2) 等级映射。来源词表与目标词表的概念范围具有包含关系,建立等级映射。从下位到上位的等级映射通常采用NM(narrower mapping)表示,从上位到下位的等级映射采用BM(broader mapping)表示。根据概念间语义含义差别,可进一步细分为属种关系、实例关系、整体和部分关系^[9]。属种关系表示类与子类的关系,采用NMG/BMG表示,类是具有相同特点的个体集合,通过类与子类的关系创建类的层次结构,通常表

现为大小范畴间的关系,如学生与中学生;实例关系表示类与成员的关系,采用NMI/BMI表示,成员通常对应现实世界中存在的具体个体;整体和部分关系是一个实体对象由许多部件(部分)构成,实体所属的整体对象与实体部分对象间的关系,采用NMP/BMP表示,如手与指头。严格区分后的等级映射能提高映射数据的质量,无论应用于标引或查询扩展,都可在一定程度上提高查准率。

(3) 相关映射。来源词表与目标词表的概念范围具有部分重叠,有一定语义程度的关联,建立相关映射,通常采用RT (related mapping) 表示。与其他映射类型相比,相关映射的语义关联程度最弱,其相关性不易把握,需依赖具体应用环境进行判断,如用户的兴趣、系统检索相关文献的能力与数量等。因此在实际互操作项目中,应根据实际情况对相关性进行明确界定,防止将语义关联性较弱的概念纳入映射集合,增加标引和检索的噪音数据。

2.5 实施映射

映射项的识别,关键在于明确概念的内涵与外延。单从词汇的外在表达形式并不能识别词汇本身的真正含义,词汇表中的词汇有明确的上下文环境,因此在判断对应概念时,需检查其相应的语义关系,通过词表内部概念的等级关系、等同关系、相关关系,及注释判定其内在含义,并在此基础上完成映射。如果是双向映射,需先完成一个方向的系统映射,再实施反方向映射,通常情况下,已完成的映射结果可被进一步使用;尤其是等同映射,具有可逆性,其在反方向映射中同样有效;但其他类型映射,需做进一步审核才可使用。

映射的完成,通常采用人工映射(专家)和自动映射(计算机)两种方法实现。人工映射的准确率较高,需要专家熟悉词表的结构及专业词汇,涉及工作量较大;计算机自动映射效率高,采用概念匹配或共现映射的方式,形成候选映射集并进行排序输出,但相对准确

率低,需要专家对映射结果进行审核。

映射完成后需将映射词表的数据存入管理系统,存储映射的基本信息包括来源词表、目标词表、映射结果表、映射的注释信息等;此外,还需定期对映射数据进行维护,及时修订映射数据,如果来源词表、目标词表发生变化,应及时对数据进行修改,确保映射数据的质量。

3 映射数据的SKOS表示

为适应语义网的要求,尽可能采用简单知识组织系统表示映射结果数据。SKOS定义了一个通用数据模型,能将数据表示为机器可读数据,用于计算机应用软件间的数据交换和以机器可读的格式实现网络发布。负责ISO 25964的机构ISO TC46/SC9/WG8工作组发布《ISO 25964与SKOS/SKOS-XL模型的对应》,以适应ISO 25964-1所定义的UML模型^[10]。

SKOS映射关系属性包括skos:closeMatch、skos:exactMatch、skos:broadMatch、skos:narrowMatch和skos:relatedMatch^[11]。这些属性用于声明不同概念体系中SKOS概念间的映射链接,分别对应不同的映射类型。skos:exactMatch对应等同映射;skos:broadMatch和skos:narrowMatch对应两概念间的等级映射;skos:relatedMatch属性对应两概念间的相关映射;skos:closeMatch表示两个概念相似,对应非精确等同。为充分表达两个概念的重合程度,还定义majorMatch、minorMatch属性,分别表示重合较大及重合范围较小两种情况;将概念作为类,允许概念类之间进行集合的交并补运算,以表达复合等同的情况。SKOS语义表示中,需先定义来源词表、目标词表,再确定来源词表中的被映射词汇,对其基本信息(如词汇名、类名、优选词、非优选词等)进行描述,最后描述与目标词表所对应词汇。如将DDC中类号“363.34”与LCSH中叙词“Pollution”建立等同映射,与“Natural disasters”建立下位映射,SKOS表示如下。

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
  <skos:Concept rdf:about="#363.34">
```

```

<skos:prefLabel xml:lang="zxx">363.34</skos:prefLabel>
<skos:altLabel xml:lang="en">Disasters</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdli.strath.ac.uk/hiltm2m/schemes/DDC.xml"/>
<map:exactMatch>
  <skos:Concept rdf:about="#sh 85038303"/>
</map:exactMatch>
<map:narrowMatch>
  <skos:Concept rdf:about="#sh 85090214"/>
</map:narrowMatch>
</skos:Concept>
<skos:Concept rdf:about="#sh 85038303">
  <skos:prefLabel xml:lang="en">Disasters</skos:prefLabel>
  <skos:inScheme rdf:resource="http://hiltm2m.cdli.strath.ac.uk/hiltm2m/schemes/LCSH.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#sh 85090214">
  <skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
  <skos:inScheme rdf:resource="http://hiltm2m.cdli.strath.ac.uk/hiltm2m/schemes/LCSH.xml"/>
</skos:Concept>
</rdf:RDF>

```

4 结语

随着网络资源异构性增加,词表互操作是当前网络资源检索面临的重要问题。ISO 25964对推动词表互操作实践具有重要作用。目前,我国基于ISO 25964正在制订适合中国的标准,以期提高中文词表间的映射质量,并实现与国外词表的互联互通。本文分析梳理ISO 25964的核心内容,进一步探讨词表间互操作实现步骤及语义化描述,旨在为各机构开展互操作项目提供指导及建议,对其词表互操作实现的语义表述提供范例。由于互操作本身是一项智力劳动,需要各方面力量的投入,为提高效率,如何充分利用技术能力实现计算机自动映射将是未来研究的关键问题。

参考文献

- [1] AIMS.AGROVOC Multilingual agricultural thesaurus[EB/OL].(2016-07-2016-09-07).<http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>.
- [2] U.S.National Library of Medicine. Unified Medical Language System (UMLS)[EB/OL].(2016-07-02).<https://www.nlm.nih.gov/research/umls/>.
- [3] University of Strathclyde Glasgow.Subject-based interoperability:issues from the high level thesaurus(HILT)project[EB/OL].(2007-01-15)[2016-07-10].<http://strathprints.strath.ac.uk/2315/>.
- [4] NISO.ANSI/NISO Z39.19-2005(R2010)guidelines for the construction, format, and management of monolingual controlled vocabularies[EB/OL].(2010-05-13)[2016-07-20].http://www.niso.org/apps/group_public/project/details.php?project_id=46.
- [5] BS 8723-4:2007 Structured vocabularies for information retrieval.Guide. Interoperability between vocabularies[EB/OL].(2016-07-23).<http://www.standardscentre.co.uk/bs/BS-8723-4-2007/>.
- [6] ISO.ISO 25964-1:2011 Information and documentation-Thesauri and interoperability with other vocabularies Part 1:Thesauri for information retrieval[EB/OL].(2011-08-15)[2016-03-20].http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657.
- [7] ISO.ISO 25964-2:2013 Information and documentation-Thesauri and interoperability with other vocabularies Part 2:Interoperability with other vocabularies[EB/OL].(2013-03-04)[2016-03-20].http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53658.
- [8] OLENSKY M. Semantic interoperability in Europeana.An examination of CIDOC CRM in digital cultural heritage documentation[J/OL].TCDL Bulletin,2010,6(2).[2016-09-01].<http://www.ieee-tcdl.org/Bulletin/v6n2/>

- Olensky/olensky.html.
- [9] ALEXIEV V, ISAAC A, LINDENTHAL J. On the composition of ISO 25964 hierarchical relations (BTG, BTP, BTI)[J]. International Journal on Digital Libraries, 2016, 17(1): 39-48.
- [10] NISO. ISO 25964—the international standard for thesauri and interoperability with other vocabularies[EB/OL]. [2016-07-02]. <http://www.niso.org/schemas/iso25964/>.
- [11] W3C. SKOS mapping vocabulary specification[EB/OL]. [2016-04-02]. <https://www.w3.org/2004/02/skos/mapping/spec/2004-11-11.html>.

作者简介

贾君枝, 女, 1972年生, 山西大学经济与管理学院教授, 博士生导师, 研究方向: 信息组织, E-mail: junzhij@163.com。

The Realization of Interoperability across Vocabularies Based on ISO 25964

JIA JunZhi

(School of Economics and Management, Shanxi University, Taiyuan 030006, China)

Abstracts: Interoperability across vocabularies is necessary to realize information retrieval across database, languages, and systems. Mapping among vocabularies is an efficient way of interoperability with the application of network knowledge organization. The standards of ISO 25964 play an important role in guaranteeing the formal management of mapping procedures and improving the quality of mapping. The paper emphasizes on analyzing the core content of standard and discusses the realization steps of interoperability and semantic description. It will give reference and advice of projects of interoperability and provide paradigms for semantic description of interoperability across vocabularies.

Keywords: ISO 25964; Interoperability; Vocabularies; Thesaurus

(收稿日期: 2016-09-07)

■ 书 讯 ■

《科技报告体系构建研究》

为推进我国科技报告制度建设, 强化科技报告资源共享服务, 贺德方研究员率领中国科学技术信息研究所科技报告研究团队, 进行了国家社会科学基金重点项目“中国科技报告资源体系构建”(11ATQ006)研究, 并对20多年来中国科学技术信息研究所相关研究和实践进行了归纳、凝练、整理和补充, 最终形成了《科技报告体系构建研究》一书。

本书作为国家社会科学基金重点项目的研究成果, 总结了科技报告产生发展的管理历程、凝练了科技报告制度的建设路径、制订了科技报告资源的整合方案, 提出了科技报告体系的构建模式, 归纳了科技报告实践的操作过程。本书对各级科技计划管理人员强化科技计划项目过程管理具有借鉴作用, 对科研人员撰写高质量科技报告具有指导作用, 对各类科研机构做好科技报告呈交、推进科技项目的规范管理和机构知识库建设具有参考价值, 对图书信息机构做好科技报告深层次加工和收藏利用具有引导作用, 也可供高校信息管理、科技政策与管理等专业研究生学习参考。

《科技报告体系构建研究》于2014年12月由科学技术文献出版社出版, 定价78.00元。