“英文超级科技词表”范畴体系协作共建研究*

孙巍[1]，张学福[1]，潘淑春[1]，刘家益[1]，李嘉锐[1]，吴雯娜[2]，李军莲[3]，甄伟[4]，黄金霞[4]
（1. 中国农业科学院农业信息研究所，北京 100081；2. 中国科学技术信息研究所，北京 100038；
3. 中国医学科学院医学信息研究所，北京 100020；4. 中国科学院文献情报中心，北京 100190）

摘要：规范合理的词表范畴体系兼顾概念主题聚类、词表结构表达、本体概念映射等多方面因素，需多学科领域专家协同合作共同构建。文章分析并阐述“英文超级科技词表”范畴体系构建需求与运作机制；分析DDC类目体系的结构特点，提出以DDC为主、专业词表分类体系为辅的主干分类体系选择方法；着重研究并提出范畴体系的协作共建思路、步骤与规则；对范畴体系协作共建成果进行展示与评价。

关键词：范畴体系；超级科技词表；DDC；知识组织体系；国家科技文献图书中心
中图分类号：G254
DOI：10.3772/j.issn.1673—2286.2014.11.006

1 引言

范畴是概念的重要属性，用来说明概念所适用的学科或所归属的类；是文献信息主题聚类的重要依据，便于文献的分类组织与浏览；是科技文献信息通用本体建设的基础，有利于控制通用本体的维度和颗粒度；又是连接概念与本体的枢纽，便于建立通用本体与科技词表概念的映射关系，有利于解决因学科交叉、表达产生的维（粒）度不同、冲突和重叠等方面的问题。

鉴于范畴体系兼顾的对象广泛，以分类表、叙词表等数据源为基础来构建英文超级科技词表范畴体系的工作十分庞杂，需要考虑知识组织体系互操作规范、多学科领域专家共同协作、机器辅助人工干预相结合等多方面因素。因此，深入研究面向多学科领域的英文超级科技词表范畴体系协作共建机制与方法具有重要的理论与实践意义。

“英文超级科技词表”（以下简称“英表”）是“十二五”国家科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用研究”的主要研制内容之一，该项目由国家科技图书文献中心（NSTL）[1]牵头，由理、工、农、医四大领域相关机构专家分工协作共同完成。本文正是为了满足课题中“英表”范畴体系构建的迫切需求，分析范畴体系协作共建机制，开展NSTL“英表”范畴体系协作共建理论方法与实践研究工作。

2 “英表”范畴体系构建需求与运作机制

2.1 英文超级科技词表

英文超级科技词表（“英表”）并非传统意义上的

* 本研究得到国家“十二五”科技支撑计划“面向外文科技文献信息的知识组织体系建设与应用研究”（项目编号：2011BAH10B00）资助。

叙词表，从逻辑结构上讲，它是一个具有三层结构的科技知识组织体系，自下而上依次为基础词库、概念网络、范畴体系。基础词库层是将词汇素材层中的不同来源异构的词汇集，按一定规范进行描述，并采用统一格式进行存储而形成的词汇元数据仓储；对基础词库层中的词汇进行同义词归并，形成以概念为单位的同义词群，进而构成孤立无序的概念网络；范畴体系为概念提供分类框架，以此对无序概念进行分类类聚，在一定程度上弥补概念网络在宏观知识结构表达上的不足[2]。

2.2 “英表”范畴体系构建目标及原则

为了有效组织NSTL文献资源，提升NSTL英文文献信息服务能力，“英表”范畴体系应从主题与学科角度来实现超级科技词表概念的均衡合理分类与汇聚，词表结构的清晰表达，为后续科技文献信息通用本体建设奠定基础，有效控制通用本体的维粒度，便于建立通用本体与超级科技词表概念的映射关系。为了实现上述目标，建成一个统一结构框架体系的“英表”范畴体系，应遵循以下原则：①充分借鉴来源范畴关系原则：根据省力法则，“英表”范畴体系应充分借鉴和继承来源分类或主题体系，并以此为基础为来源范畴类目进行扩充与调整。②概念涵盖完整性原则：“英表”范畴分类体系应具备学科覆盖面广的特点，类目应尽可能涵盖理工农医等所有科技领域概念。③类目等级科学实用性原则：“英表”范畴分类体系结构应具备层级分明、等级性较强、维度层面的各等级概念分布较均衡等特点。满足专业用户的应用需求，符合用户的一般使用习惯。具备规范的注释以及分类号，可读性强。④可扩展性原则：“英表”范畴分类体系结构应具备一定的可扩展性，其等级类目及概念涵盖面可以随着概念及关系的增加而逐级扩展。

2.3 “英表”范畴体系协作共建机制

以多源异构的词汇集为基础，构建汇聚多领域概念的范畴体系，需要制定由多领域专家共同遵循的协同合作共建机制，进而解决多范畴间的不兼容性，满足不同领域范畴之间的协同操作以及范畴体系的全局调控需求，具体包括：主干类目遴选与分配机制、领域范畴自主构建机制、阶段性协同全局调控机制、交叉领域类目冲突解决机制。

2.3.1 主干类目遴选与分配机制

完全新建一个全新的范畴体系是不现实的，英表范畴体系的构建并不是从零开始，而是选择一个现有的分类体系作为主干分类体系，选择多个专业分类表、叙词表等作为辅助分类体系，根据拟构建范畴体系的功能及需求定位，对主干分类体系进行类目遴选，利用辅助分类体系对主干分类体系作相应的扩充调整。一个合理的任务分配机制能够确保多领域机构高效有序地完成范畴体系协同共建，而按学科领域所遴选的主干范畴类目是制定多领域协同共建范畴体系任务分配机制的重要依据之一。

构建一个多领域范畴体系，首先，需要按各机构的领域特征以主干分类体系为主，以专业辅助分类体系为辅来遴选各自领域的主干范畴类目（前三级），尽可能继承维系主干范畴体系等级逻辑关系，确保所遴选的主干范畴体系类目等级的连贯性，对主干范畴体系中未遴选的、综合性的，而又必设的类目需作特别遴选处理。其次，对各领域所遴选的范畴类目集作类目查重处理，得到的重复类目由遴选机构共同分析确定其最终的任务归属机构。为了避免加重后续工作量，主干类目遴选阶段产生的重复类目在任务分配时只能分配给一个领域机构。最后，对各领域机构所归属的范畴类目以及综合类目进行任务分配标识，所标识的类目集将作为“英表”范畴体系的基础主干类目，后续工作中，各领域机构必须按类目的任务分配标识来操作各自领域的主干范畴类目。

2.3.2 领域范畴自主构建机制

鉴于理、工、农、医四大部类的学科领域特点不同，各领域所遴选的包括分类表、叙词表等参考辅助专业分类体系的应用范围不同，在核心范畴体系中的分布特征也各异，即便是大致相同的应用领域，也可能因为分类思想的不同导致范畴体系间的不完全兼容性。为了提高各领域内范畴体系构建的效率，特提出领域范畴自主构建机制。即：各领域机构在遵循“英表”范畴体系整体构建原则的前提下，按照各自的学科领域特征及参考辅助专业分类体系的特点，制定各自领域的分类体系互操作规范与细则，提出各领域的核心领域范畴的扩充与调整方法，通过辅助专业分类体系与领域主干范畴类目的互操作，实现范畴类目提升与

降级、类目更名、类目拆分与合并、类目删除与新增等领域范畴体系类目扩充与调整，完成各领域范畴的自主构建。

2.3.3 阶段性整合与全局调控机制

英表范畴体系要求所涵盖的学科领域庞杂，概念主题覆盖面广，构建工作涉及的机构多，各机构的工作机制又大同小异，因此，不论从范畴体系的维度等级，还是从范畴体系类目间的逻辑关系上讲，范畴体系构建过程中均需要各领域机构分步骤、分阶段地集中对各领域范畴体系进行整合与全局逻辑调控。

整合与全局调控工作大体分三个阶段。第一阶段，对各领域范畴体系（前三级）的整合，确保其学科覆盖的完整性，类目学科主题分布态势的合理性；第二阶段，继续对理、工、农、医各领域扩充调整后的三级以上范畴体系进行整合与全局调控，本阶段工作侧重于提高主题与概念的覆盖完整性；第三阶段，从全局角度对范畴体系类目逻辑关系进行深入核查与分析，消除因学科交叉所产生的类目冲突、重叠及冗余问题，确定构建的“英表”范畴体系等级结构框架的统一性。

2.3.4 交叉领域类目冲突解决机制

针对学科交叉融合等问题所产生的领域间类目冲突、类目重叠、类目冗余等问题制定了一套解决机制。首先，各领域机构核查各自的领域范畴体系类目与其他领域范畴类目存在语义或者逻辑冲突、重叠、冗余三类类目冲突问题，并对问题类目进行冲突类型标注；其次，各领域专家针对各自存在的问题共同分析商讨，明确问题类目的最终标注类别；最后，针对类目语义逻辑冲突问题，通过调整类目等级、修改类目名称、类目融合等操作来解决；针对重叠类目考虑在其主要应用领域列类，次要应用领域则以“参见”类目形式出现；针对冗余类目则判断其类目间的冗余范围，选择直接删除较小范畴的类目。

3 DDC特点及主干分类体系选择

《杜威十进分类法》（Dewey Decimal Classification，简称DDC）[3]，是一部通用分类法，系统性强，应用较广泛，目前已被全球超过135个国家的图书馆使用[4]，且被翻译逾30种语言版本；从其类号体制看，DDC是十进制分类体系，其各级类目基本按层累计方式编号，类目体系等级分明，易于理解和使用；且DDC设有专门的维护机构持续对其进行维护和修订[5]，一直处于不断的更新与完善中。此外，DDC更能用来组织网际网络上的各种资源。

透过DDC类目，对其理学、工学、农学、医学、综合学科类目的分布特征进行粗略分析发现，理学类目主要集中在一级大类“5自然科学与数学”下，工学类目主要集中在一级大类“6技术”下，农学类目主要集中在二级类目“63农业技术”下，而医学类目主要集中在二级类目“61医学”下。由此可见，DDC基本涵盖了各学科的核心范畴类目，领域内类目分散及缺省问题可通过分类体系的局部调整与扩充来弥补。

鉴于上述DDC自身系统性强、可维护性强、易于理解、学科覆盖相对完整性等特点，本文选取DDC分类体系作为主干范畴体系，对其进行局部扩充与调整，由理、工、农、医领域机构协同共建“英表”范畴体系。

4 基于DDC的“英表”范畴分类体系协作共建

4.1 范畴体系协作共建思路

基于“英表”范畴分类体系的构建原则与协作共建机制，“英表”范畴体系的协作共建思路是：选取DDC作为主干范畴分类体系，其基本覆盖了理、工、农、医几大部类。以此为基础，理工农医各领域机构分别根据范畴体系构建目标，遵循领域范畴体系构建原则，吸收专业领域优秀范畴体系的分类思想，对主干范畴体系进行类目扩充与局部调整，既要考虑各自领域英文文献的主题分布特征，也要考虑中文用户的使用习惯，自主构建各领域范畴体系。整个范畴体系构建过程中，采取分两个阶段来交替实施“领域范畴体系自主构建”以及“多领域机构协作完成范畴整合”工作，以逐级调整扩展的方式来构建范畴体系，范畴体系经过后期的调整与完善，最终建成一个统一学科框架下的“英表”范畴分类体系。

4.2 范畴类目协作共建步骤

范畴体系协作共建工作大体分为三个阶段（如图

1）：范畴素材遴选、领域范畴构建与整合，以及范畴体系调整与完善。各阶段的主要工作内容阐述如下：

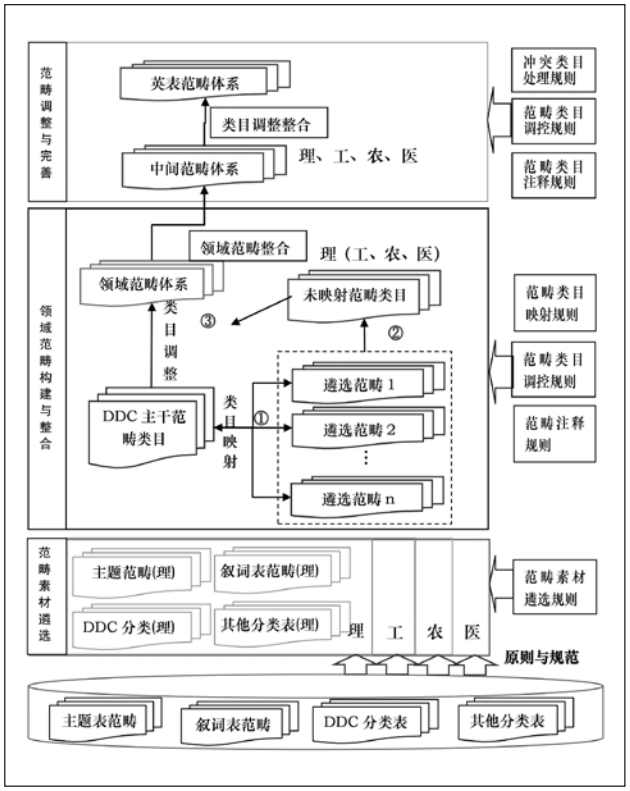

图1 “英表”范畴体系构建框架

（1）范畴素材遴选

范畴素材遴选的主要任务是从包括DDC在内的范畴素材中遴选出尽可能涵盖核心叙词概念的范畴类目，以此作为超级词表范畴体系构建的数据基础。

理、工、农、医各领域机构依据范畴遴选规则，从范畴体系的覆盖面、体系结构、范畴语言等多角度，选取DDC分类表以及具有代表性的领域范畴，即参考范畴分类体系，如领域主题表、领域分类表、领域叙词表等，作为超级词表范畴层范畴体系构建的基础数据，为“英表”范畴体系构建奠定数据基础。

（2）领域范畴构建与整合

领域范畴构建与整合的主要任务是以遴选的范畴素材（包括DDC、主题表、其他分类表等）为基础，各领域机构通过领域范畴体系间的类目互操作，自主构建生成各领域范畴体系，进而整合各领域范畴体系，并对其类目进行科学适用性调整与全局调控。

理工农医各领域机构以DDC分类体系作为主干范畴表，遴选出DDC主干范畴类目，并对其等级结构进行调整；按照各机构制定的领域范畴体系互操作具体细则，分别将DDC主干范畴类目与所遴选的专业领域参照分类表作类目映射，充分发挥专业领域优秀分类体系对DDC的扩展补充作用，生成理、工、农、医、综合领域范畴体系；依据英表范畴分类体系调控原则对所生成的分类体系进行整合与全局控制，生成中间范畴分类体系。

通常，协作共建工作可以以细分工作量的方式将复杂工作分阶段简化。由于“英表”范畴体系构建工作庞杂，为了降低工作负担，避免重复工作，此阶段我们又细分了两个步骤来还完成中间范畴体系的构建工作，即领域前三级范畴体系的自主构建与整合、领域三级以上范畴体系的自主构建与整合。

此外，在构建中间范畴体系过程中，针对前三级类目，我们通过专家辅助主干类目遴选，以及多范畴体系类目映射等方式来确保范畴体系的学科主题分布态势的合理性；而针对三级以上类目，应重点考虑范畴体系的概念分布均衡性及概念覆盖完整性。这里我们采取范畴测试的方式，对各学科领域叙词概念进行范畴类目归类，分析范畴体系等级维度上的概念分布均衡性及概念覆盖完整性，依据分析结果及范畴体系调控规则，通过概念群组归并与拆分等操作，对三级以上范畴类目进行调整。

（3）范畴体系调整与完善

范畴体系调整与完善过程的主要任务是各领域专家辅助从全局角度对范畴体系类目逻辑关系进行深入核查与分析，对当前的范畴体系中领域间重复类目、交叉冲突类目、等级关系矛盾类目等进行修正调整，进而消除因学科交叉所产生的类目冲突、重叠及冗余等问题，实现多领域范畴体系的无缝整合。范畴体系调整与完善过程主要依据范畴体系类目冲突处理规则。

4.3 范畴体系协作共建操作规则

尽管各领域机构在各自领域范畴体系构建过程中遵循自主构建原则，但为了确保建成一个统一学科体系框架下、统一风格的英表范畴体系，各领域机构在协作共建以及分别处理相似问题上应遵循一定程度上的统一化规则，规则概述如表1。

表1 范畴体系协作共建操作规则一览表

序号	协作共建操作规则	农业辅助分类体系（一级）
1	类目遴选规则	①充分继承DDC原有范畴类目及类间关系；②充分利用专业范畴体系，尽可能高度覆盖本学科领域范畴类目
2	类目映射规则	限定的映射关系类型：①相等；②包含于；③包含；④不相等；⑤相交
3	范畴类目调控规则	限定的调控操作类型：①类目更名；②类目提升；③类目降级；④类目拆分；⑤类目合并；⑥类目增加；⑦类目删除
4	冲突类目处理规则	限定的处理操作：①彻底分开，两边分别保留；②两边均保留，但有侧重；③两边均不保留，单独列类
5	范畴类目编码规则	纯数字编码，两位数字表示一个大类，以数字的顺序反映大类的序列。每三个等级用圆点符号分隔
6	范畴注释规则	操作注释的类目对象包括：①不同名同义类目间映射产生的新类目；②同名不同义；③交叉学科类目 ④类目拆分合并生成的新类目；⑤主干类目；使用注释的类目对象包括：①参见类目；②交替类目

5 范畴体系协作共建结果展示与评价

理、工、农、医各领域机构，以DDC分类体系为主干范畴体系，严格按照英表范畴体系的构建目标与需求，遵循英表范畴体系的协作共建机制与原则，协同合作共同建成了一个包含9个等级、10408个类目的“英表”范畴体系。其中，一级类目38个，涵盖了理、工、农、医、综合、通用六大部类（如表2），且经测试各大部分的核心词表概念均可归入“英表”范畴体系类目中，表明该范畴体系在一定程度上遵循了学科主题相结合的列类原则以及概念涵盖完整性原则；范畴体系类目等级整体呈规范正态分布[7]（如图2），等级性较强；从英表范畴体系的编码规则上看，该体系中的各领域范畴类目具备一定的可扩展性，可读性较强。

综上，本文构建的“英表”范畴体系在一定程度上能够有效组织NSTL文献资源，提升NSTL的文献信息服务能力。

表2 “英表”范畴体系一级类目及学科分布一览表

范畴类号	学科部类	范畴类目名称	范畴类号	学科部类	范畴类目名称
00	综合	哲学、心理学、宗教	36	医学	中国医学与其他传统医学
01	综合	社会科学	50	农学	农业基础科学
02	综合	人文与艺术	51	农学	农学
03	综合	历史、地理	52	农学	林业科学
10	理学	自然科学总论	53	农学	畜牧科学
11	理学	数学	54	农学	水产、渔业、狩猎
12	理学	物理学	60	工学	工程基础科学、通用技术
13	理学	化学	61	工学	矿业工程
14	理学	天文学	62	工学	冶金与金属工艺
15	理学	地球科学	63	工学	机械工程、汽车工程、仪器设备
16	理学	生物学	64	工学	能源、动力、电工、核工程
17	理学	植物学	65	工学	电子、通信、计算机、自动控制
18	理学	动物学	66	工学	化学工程
30	医学	医药卫生总论	67	工学	轻工业、手工业、生活服务技术
31	医学	卫生学、预防医学	68	工学	土木、建筑、水利工程
32	医学	基础医学	69	工学	交通运输
33	医学	临床医学	70	工学	航空航天、军事工程
34	医学	特种医学	71	工学	环境科学与技术
35	医学	药学	90	通用	通用概念

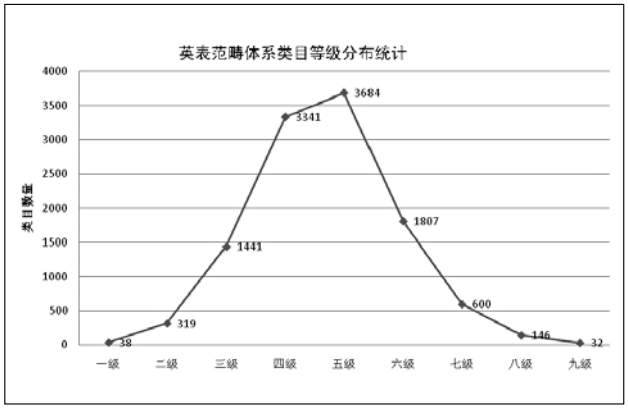

图2 英表范畴体系类目等级分布统计图

6 结束语

本文分析了“英表”范畴体系的构建需求与运作机制；制定了以DDC为主干范畴表，通过对其进行局部调整与类目扩展来构建“英表”范畴体系的整体方案；阐述了包括“范畴遴选”、“领域范畴构建与整合”、“范畴体系调整与完善”三个阶段工作的整体构建框架；从“英表”范畴体系构建目标及原则的角度对理工农医协作共建的“英表”范畴体系的适用性进行了分析与评价，进而得出“英表”范畴体系在一定程度上满足其构建需求的结论。而“英表”范畴体系在类目导航效果、类目的均衡性、实际应用中概念的涵盖率等方面的特性仍有待进一步分析与研究。

参考文献

[1] 国家科技文献中心[EB/OL]. [2014-11-23]. http://www.nstl.gov.cn/.

[2] 吴文娜,王星.基于DDC的《英文超级科技词表》范畴体系构建研究:以工程技术为例[J].图书情报工作.2011,55(22):15-21.

[3] WebDewey [EB/OL]. [2014-11-20]. http://connexion.oclc.org.

[4] OCLC. Organize your materials with the world's most widely used library classification system [EB/OL]. [2014-11-29]. http://www.oclc.org/dewey.en.html.

[5] 马张华.国外文献分类法修订维护的发展及对《中图法》的启示[J].国家图书馆学刊,2008(2):40-44.

[6] 化柏林.图书情报学核心期刊论文标题计量分析研究[J].情报学报,2007(3):391-398.

作者简介

孙巍，女，1978年生，中国农业科学院农业信息研究所副研究员，研究方向：农业知识组织与可视化分析。E-mail：sunwei@caas.cn。
张学福，男，1966年生，中国农业科学院农业信息研究所研究员，研究方向：农业知识组织与可视化分析，通讯作者，E-mail：zhangxf@caas.cn。

Research on Collaboration and Co-construction of Category System for STEST

SUN Wei[1], ZHANG XueFu[1], PAN ShuChun[1], LIU JiaYi[1], LI JiaRui[1], WU WenNa[2], LI JunLian[3], ZHEN Wei[4], HUANG JinXia[4]

Abstract: Taking into account many factors, such as the clustering of topics and concepts, expression of thesaurus structure, mapping concepts to ontology, a practical and standard category system for thesaurus should be constructed cooperatively and collaboratively by multidisciplinary experts. Construction requirement and operating mechanism of category system for “Science & Technology English Super-thesaurus” (STEST) are analyzed and elaborated; on the basis of analysis on the structural characteristics of DDC, a selection method of core classification system is proposed that supplements a focus on DDC with professional and domain classification system; and the researches are focused on the ideas, steps and rules of collaboration and co-construction of category system for STEST; the outcomes of collaboration and co-construction of category system are displayed and evaluated.

Keywords: Category system; STEST; DDC; Knowledge organization system; NSTL

（收稿日期：2014-11-20）