知识组织体系术语删除分析*

Analysis on the Term Deletion in Knowledge Organization System

张士男（中国科学院文献情报中心　中国科学院大学　北京　100190）

[摘　要]　知识组织体系的内容只有不断地调整和更新，才能准确反映知识的科学发展，及时删除"过时"术语能够使知识组织体系更加准确地揭示领域知识。不同知识组织体系在术语删除机制、术语删除原因、术语删除处理方式方面存在差异，但本质都是为了保持一致性，反映科学知识的最新进展。术语删除是检查知识组织体系质量的一种方式，因此要建立完善的术语删除机制，明确术语删除的原则，根据术语删除原因采取不同的处理方式。

[关键词]　知识组织体系　叙词表　术语删除
[中图分类号]　G254　　[文献标识码]　A

[Abstract]　With the constant adjustment and update, Knowledge Organization System (KOS) can reflect the development of science accurately, and deleting the obsolete term can make KOS more accurate in revealing the domain knowledge. There are differences on mechanism, reason and treatment of the term deletion among KOS, but the essence is all to keep the consistency of KOS and to reflect the latest progress of the scientific knowledge. Term deletion is a kind of form to check the quality of KOS, so it is essential to establish perfect term deletion mechanism, improve term deletion principle, and take different measures in different conditions.

[Key words]　Knowledge Organization System(KOS); Thesauri; Term deletion

知识组织体系的内容只有不断调整和更新，才能准确反映知识的科学发展，进而更加成功地支持数据检索和知识发现。确定一个科学分类系统内容的过程是一种组织性遗忘的过程，这就意味着一个新的分类被创建，开发者就要决定哪些内容需要从体系里删除。如果没有分类去描述一个对象或行为的含义，则信息系统无法记录这个对象或行为，没有名称的事物将在分类系统中被遗忘[1]。术语删除是检查知识组织体系质量的一种方式，如基因本体（Gene Ontology，简称GO）中删除的术语占术语总量的比例超过4.4%，这意味着每增加100个术语，就至少有4个术语被删除[2]135，建筑与艺术叙词表（Art & Architecture Thesaurus，简称AAT）可获取的主题词中有4.4%为被遗弃的主题[3]。本文在分析术语删除机制、术语删除原因、术语删除处理方式的基础上，分析术语删除中存在的问题，提出受控词表术语删除的几点建议，以期为受控词表的维护提供参考。

1　分析的数据及方法

笔者在进行数据分析时选择的词表至少满足两个条件：一是词表持续更新；二是词表提供网络服务，开放的网络环境更有利于对词表变化的系统跟踪。此外，笔者还综合考虑了词表权威性、应用程度等因素，具体如表1所示。

在词表更新维护调研过程中，笔者更加关注术语删除的相关问题，包括术语删除机制、术语删除原因、术语删除的处理方式及其对资源标引可能带来的影响。

2　术语删除现状分析

2.1　机制分析

在受控词表维护过程中，知识组织体系编制机构应该特别关注术语删除，尤其是当术语已经被使用。对于已经带有很多标注的术语一旦决定将其删除，则需要花大力气详细讨论如何处理已经使用的删除术语的标注。因此，术

*本文系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示范（STKOS）"子课题"面向外文科技文献的超级科技词表和本体建设"研究成果之一，课题编号：2011BAH10B01。

表1　术语删除分析的词表

词表名称	词　表　概　述
美国国会图书馆主题词表(Library of Congress Subject Headings,简称LCSH)	LCSH是一部国际上颇有影响的综合性主题词表。2009年,美国国会图书馆在其官方网站上公布了以简单知识组织系统(Simple Knowledge Organization System,简称SKOS)形式表达的主题词表。
国家农业图书馆农业叙词表(NationalAgriculturalLibrary'sAgricultural Thesaurus,简称NALT)[4]	NALT在20世纪90年代作为农业网络信息中心(Agricultural Network Information Center,简称AgNIC)的内部资源,第一版只有2 000个术语。1999年其面向美国农业部农业研究服务(Agricultural Research Service,简称ARS)的科研人员开通了网络访问,2002年1月起正式提供网络服务,2005年作为美国国家农业图书馆“农业在线获取”文摘索引(AGRICultural Online Access,简称AGRICOL)的资源。NALT每年发布新版本,提供SKOS、XML、PDF、Word格式。
医学主题词表(Medical Subject Headings,简称MeSH)	MeSH是一部用来标引生物医学及其相关领域的受控词表,是目前国际上最具代表性、使用最为广泛的受控医学叙词表,每年发布新版本。
建筑与艺术叙词表	AAT是一个关于艺术学、建筑学、其他文化遗产及其保护的多语种受控词汇,目前已经发布为关联数据,计划每隔两周进行AAT关联数据更新。
基因本体[5]	GO最初于1998年创建,2000年正式发布,是开放生物医学本体中时间最长的本体之一。GO是分子生物学领域的经典标准词汇,他的创建是为了解决分子生物学领域中基因产品如何分类,尤其是当分布在不同数据库的资源使用了不同的关键词集进行了描述和分类时的数据集成问题,如基因功能的跨库资源。GO包括细胞组件、生物过程、生物功能三个子本体。其通过应用准确的定义和相互关联的术语来描述生物学领域知识,主要用于标注基因产品信息,如PubMed数据库。
CAB叙词表(CAB Thesaurus,简称CABT)[4]	CABT涵盖了农业科学、生命科学、人类健康及相关主题。自1972年起其作为关键词受控清单索引CAB文摘数据库(Centre for Agriculture and Bioscience International)资源,1983年首次出版,包括48 000个优选和非优选术语,1999年起叙词表只提供电子版。大概平均两年发布新版本。

语删除要谨慎,要按照严谨的机制来执行,术语删除机制主要包括如下两项内容:

(1)由受控词表专门的维护团队或部门主导。LCSH的维护由美国国会图书馆政策与标准部(Policy and Standards Division)执行,组织方式也更加开放,最初,该部门直接通过“编目要务通报”(Cataloguing Service Bulletin)公布修改结果,自2005年起,该部门在其网站上以周报/月报(2011年8月起周报改为月报)的形式公布对修改意见的处理方式,不论意见是否被采纳,均给出了原因说明[6]。同时,最终处理结果在Cataloguing Service Bulletin公布,但并不是公布全部数据,2011年以后术语修改结果不再通过编目要务通报公布[7]。CABT主题词的变更由CABT维护团队负责,这些变更作为整部词表的主要更新每隔一至两年与整部词表发布新版本时一起出版[4]。CABT上一版本于2012年9月发布,最新版本于2014年7月21日正式在线发布。

(2)开发人员、标引人员、领域专家共同参与。GO为了使其更加成功地支持数据驱动的知识发现和数据检索,准确地表达生物知识的实体和过程,使用了术语定义,建立了术语之间的关联。GO长期维护的一个关键问题就是更新其内容以反映生物学知识的最新科学发展。GO在创建之初就意识到了这个问题,创建了反馈机制,从而GO用户能够向管理人员反馈其在自己研究领域对生物实体或过程的理解与GO中相应知识表达不一致的地方。GO的管理人员通过几种不同的机制提出候选“淘汰”术语,随后这些淘汰术语通过一个标准的门户Sourceforge tracker推出,进行14天的意见搜集,然后管理者、编辑者、领域专家共同讨论,有时本体用户也参与讨论。在没有Sourceforge tracker管理之前,GO通过邮件列表来记录和讨论淘汰术语。近60%的淘汰术语从提出到确定一般需要30~60天左右。提出淘汰术语的人员主要是GO的开发人员和标引人员。一般来讲,GO确定删除术语分为三个步骤:首先由其管理者或用户提出单个候选删除术语;其次,确定候选删除术语(一般由编辑者或标注者提出);最

后,通过会议讨论或研发人员特别兴趣小组等确定删除的术语集[2]132-138。

2.2 原因分析

不同词表内容进行调整的原因各不相同。NALT 每年进行更新,更新内容包括:将术语替换为最新术语,替换掉歧义术语,更正错误[8]。MeSH是标引生物医学及其相关领域资源的受控词表,生物医学领域经常有新概念产生,因此对于概念的增加、删除等MeSH 需要及时调整,以反映领域最新进展。MeSH 进行调整的基本类型包括直接替换、删除和增加术语,删除原因一般包括:标题使用率低;标题不能反映识别资源的优选形式或不能用于英语医学文献;标题不能与其他标题在含义上有所区分;标题与上位词明显不同但是上位词使用率低;标题具有特殊含义但是其优选形式存有歧义;标题最初拥有多种内涵,但是目前特殊含义的合适的标题令之前的标题没有必要继续存在;标题/副标题的组合可以进一步细分;层级结构中错误分类的标题[9]。AAT 数据的变化主要基于如下原因:增加新主题词或为现有主题增加新术语;反映术语及其定义在学术研究或使用中的变化;增强数据一致性;更正一些主题中的近义词或地位上升的历史数据;为历史数据建立多元层级(2001年之前词表不允许建立多元层级);更正错误,包括贡献的数据或编辑错误[10]。GO 术语删除的主要原因一是术语超出了GO 的范围,不再用于标注基因产品;二是术语被重新定义或是内涵发生变化。其本质是为了体现本体真实性[11]。LCSH 则更正了有文化偏见的主题[12]。

从上述分析可以看出,数据更新的本质是为了保持一致性,反映科学知识的最新进展。术语删除主要原因包括术语使用率低;术语歧义;术语被重新定义或内涵发生变化;术语内涵不清晰;术语内涵过于宽泛;其他问题,如将术语替换为最新术语,更正编辑性错误等。

2.3 处理方式分析

2.3.1 删除程度分析

从删除程度来看,术语删除处理方式包括临时删除和永久删除。永久删除指术语不再继续使用,被永久删除。临时删除指术语虽然被标记为淘汰(obsolete),但在受控词表文件中仍然存在,可以通过版本进行检索,被删除的词汇能够被重新启用,如GO中淘汰的术语(obsolete terms)并不是被永久删除,GO 中所有的淘汰术语都包含在新版本GO 文件中,但增加了“obsolete”标记。因此,每个子本体(细胞组件本体、分子生物功能本体和生物过程本体)都可以检索之前的淘汰术语[2]132。欧盟多语种叙词表(Multilingual Thesaurus of the European Union,简称EuroVoc)由于历史原因,将一些概念定义为“删除的概念”,并且标记为“obsolete”,但他们仍然存在于叙词表中,并且他们的URIs也被保留了[13]。

2.3.2 处理方式分析

从处理方式来看,术语删除处理方式包括以下三种。一是直接删除,并且没有指明代替被删除术语的词汇。例如,NALT在2014年的版本中直接删除了2013年版本中的术语‘environmental communication’和‘Galandromus’[8]。二是将删除术语作为替代术语的入口词。将淘汰的术语作为替代术语的同义词是最常见的处理方式。自2010年起,MeSH 在其官方网站公开发布了历年删除术语清单,并为每个删除术语指明了替代术语,同时将删除的术语作为新概念的入口词[14],如删除术语‘Nuclear Reprogramming’用“Estrogen Antagonists”代替,并将“Nuclear Reprogramming”作为‘Estrogen Antagonists’的入口词[15]。NALT 团队竭尽所能将被替换或修改的优选术语与新的优选术语建立相互参照关系,即将删除的术语作为替换他们的概念词的入口词。例如,删除术语“allopatric populations”用‘allopatry’代替[8],并将‘allopatric populations’作为‘allopatry’的入口词[16]。EuroVoc将删除的术语用新的术语代替,并且用‘use instead’作为代替标识[13]。三是指明删除术语的替换术语,但没有将删除的术语作为替换他们的概念词的入口词。NALT 直接删除了淘汰的术语,指明了替代术语,但没有将删除的术语作为代替术语的入口词,如删除“Boer”并用“Boer (goat breed)”代替,但没有将‘Boer’作为‘Boer (goat breed)’的入口词[8,17],NALT中这类处理方式多见于替换术语是在被删除术语基础上增加了限定词。

2.3.3 记录方式分析

从记录方式来看,术语删除处理方式包括以下两种。一是直接增加删除标识。知识组织体系编制机构一般在删除词汇的后面明确注明词汇为删除术语,如“obsolete term”“obsolete”“former heading”等标识。例如,GO本体为每一个淘汰的术语做了‘obsolete’标识。二是通过历史附注记录术语的删除信息。这也是最常见的处理方式。如果打开AAT印本书的任何一页,我们可以看到很多诸如

ISSN 1004-325X | CN23-1331/G2

“1992年4月删除标题”或“1993年标题改变”的历史附注。这些变化可以在历史版本中自动跟踪，当这些数据可以通过应用程序编程接口（Application Programming Interface，简称APIs）获取时，这种变化可以实时获取[10]。被删除的主题可能已经在用户数据中被使用了，为了不丢失这些数据，AAT为已经在数据中使用的这些遗弃主题的用户提供连续性信息，其中记录了有关遗弃主题的少量信息，并且将其发布关联数据，包括如下三种属性信息[3]：

- skos:prefLabel：优选名称（通常是英语形式）
- gvp:enfDate：删除时间
- dct:isReplacedBy：合并入的主题（如果已经合并到另一个主题中）

2.4 删除术语与标引

受控词表的主要作用是进行资源标引，一旦标引的术语被使用，则标注也要随之变化。CAB文摘数据库使用CABT概念标签标引资源，优选术语的任何变化都会影响成千上万条记录、大量检索文档及衍生产品。CABT整部词表发布新版本后，CAB文摘数据库标引再进行更新以反映出叙词表的变化。由于没有可供使用的开放（关联）数据，因此在开放环境下CABT的使用受限制[4]。CAB正式考虑将CABT发布为关联数据[18]。如果GO中一个术语将要被删除，则在本体文件中会被标记，同时不再用于标注基因产品。对于已经带有很多标注的过时术语，GO管理者需要花大力气详细讨论，决定现有的标注应该迁移到哪个替换的术语中。GO的一组开发人员在2008年7月召开了专门会议来讨论整体迁移事宜[2]132,141。

3 术语删除存在的问题

3.1 删除的词汇作为替换词汇的入口词容易混淆用户

概念是人脑中的思想单元，具有明确的内涵和外延，一个概念可以用不同的术语表达，同一个概念的不同术语之间是等同关系，等同术语中的一个通常被选作优选术语来表达概念，以便人进行阅读，具有等同关系的术语可能是同义术语或近义术语，也可能是反义术语。术语删除原因复杂，直接将删除的术语作为新概念的入口词，容易混淆用户的认知。如果删除的术语之前为优选术语，当删除的术语与替换他的词含义不完全一致时，将改变概念内涵。

3.2 缺少删除原因的信息记录

由于卡片目录的制作需要花费大量精力，人们往往不愿意做出过多改变。计算机技术的发展使词表维护变得更加简单，也使词表维护的机制变得更加开放。词表开始记录删除的术语及其相关信息，并且为了保持连续性，将删除的术语与替换他的术语建立了相互参照。但从上述分析可以看出，知识组织体系编制机构缺少对删除原因的记录，这也是术语删除标准不明确的表现。

3.3 对基于概念标签的标引影响更大

受控词表的主要作用是对资源进行标引和分类。术语删除会带来多重影响，可能导致概念删除或变化、词表体系结构调整。在基于概念的标引中，不同数据库的标引形式不同，一是用概念唯一标识符进行标引，这时无论概念的优选词标签如何变化，唯一标示符代表的概念内涵没有变化，则不会对标引产生过多影响。二是用概念优选标签进行标引，只要概念的优选标签发生变化，则要影响使用这一标签标引的全部资源。例如，在CABT中，当概念“climate change”的优选术语标签从“climatic change”变为“climate change”时，上千条记录需要用新的字符串进行更新并且重新标引，标引中的这种变化将会影响所有CAB文摘衍生产品。

4 受控词表术语删除的几点建议

4.1 建立完善的术语删除机制

术语删除是受控词表质量保障和维护的主要手段之一，涉及方方面面，尤其是当术语已经被使用时更要引起重视。建立严谨的术语删除机制很重要，明确哪些人员能够提出候选删除术语，通过哪个平台或方式来进行意见征集和讨论，后续问题处理、一致性检验等是知识组织体系创建机构需要考虑的重要问题。

4.2 明确术语删除的原则

删除原则是术语删除工作的指导。国际标准化组织/文献工作技术委员会（International Organization for Standardization/Technical Committees，简称ISO/TC 46）发布的报告“N 2422 Result of voting ISO/CD5127”描述了国际标准化组织/委员会阶段委员会草案《信息与文献——术语》（International Organization for Standardization/Committee Draft《Information and documentation — Vocabulary》，ISO/CD 5127）的投票结果和各国专家意见，

专家指出，当前标准中增加和删除术语的规则不明晰，删除的一些词，如“archivist”“archive science”“evidential value”“inventory”等在信息与文献领域中是有意义的，不应删除[19]。词表编辑人员应该关注所有建议和提出的候选术语，使用过于频繁或过于稀少的词语应作为删除或被修正的候选术语[20]。此外，专指性过强、内涵过于宽泛、内涵发生变化、与当前词表学科范围关联性不大、不再使用的术语等都应该被考虑。

4.3 根据删除原因采用不同的处理策略

术语删除的原因多样，知识组织体系编制机构应该根据不同的原因，采取不同的处理方式。例如，从未被使用的术语，说明并不能反映客观事实，不能记录任何一个事物或实体，可直接删除；专指性过强的术语，可以合并入其他概念；含义过于宽泛的术语，应该被删除，且拆分为多个概念等。

本文对几部主流知识组织体系的术语删除机制、删除原因及处理方式等进行了调研，分析了存在的问题，并提出了几点建议，但是分析的还不够深入，缺少系统深入的实例研究。术语删除是知识组织体系质量检验的手段之一，但不是唯一的手段，而是与其他方面的修改（如术语增加、关系调整等内容）息息相关。

参考文献：

[1] Bowker C. Star L. Sorting Things Out: Classification and its Consequences[M]. Cambridge: MIT Press, 1999: 1-377.

[2] Mayor C. The Classification of Gene Products in the Molecular Biology Domain: Realism, Objectivity, and the Limitations of the Gene Ontology[D]. London: City University London, 2012.

[3] Trust J P G. AAT Semantic Representation [EB/OL]. [2014-10-14]. http://www.getty.edu/research/tools/vocabularies/lod/aat_semantic_representation.pdf.

[4] Thomas B, Osma S. GACS: Status Quo of Three Partner Thesauri. [EB/OL]. [2014-11-20]. http://aims.fao.org/sites/default/files/GACS_Status_Quo_1%200.pdf.

[5] Mayor C, Robinson L. Ontological Realism and Classification: Structures and Concepts in the Gene Ontology[J]. Journal of the American Society for Information Science and Technology, 2014: 1-12.

[6] Summary of Decisions from the Weekly (and Monthly, as of May 2011) Editorial Meeting [EB/OL]. [2014-10-07]. http://www.loc.gov/aba/pcc/saco/cpsoed/cpsoeditorial.html.

[7] Cataloging Service Bulletin [EB/OL]. [2014-10-07]. http://loc.gov/cds/PDFdownloads/csb/index.html.

[8] Replaced Terms Between 2013 and 2014 Editions [EB/OL]. [2014-09-28]. http://agclass.nal.usda.gov/dne/replaced.shtml.

[9] Humphrey M. File Maintenance of MeSH Headings in MEDLINE[J]. Journal of the American Society for Information Science, 1984, 35(1): 34-44.

[10] Trust J P G. AAT: Frequently Asked Questions [EB/OL]. [2014-10-25]. http://www.getty.edu/research/tools/vocabularies/aat/aat_faq.html.

[11] Mayor C, Robinson L. Ontological Realism, Concepts and Classification in Molecular Biology Development and Application of the Gene Ontology[J]. Journa of Documentation, 2014, 70(1): 173-193.

[12] Knowlton S. Three Decades Since Prejudices and Antipathies: A Study of Changes in the Library of Congress[J]. Cataloging and Classification Quarterly, 2005, 40(2): 123-145.

[13] EuroVoc 4.4: Obsolete Concepts by Microthesaurus [EB/OL]. [2014-12-13]. http://eurovoc.europa.eu/drupal/?q=node/1242&cl=en.

[14] Deleted Descriptors [EB/OL]. [2014-08-28]. http://www.nlm.nih.gov/mesh/deleted.html.

[15] 2015 MeSH [EB/OL]. [2015-04-16]. http://www.nlm.nih.gov/cgi/mesh/2015/MB_cgi?term=ESTROGEN+ANTAGONISTS.

[16] Allopatry [EB/OL]. [2014-09-05]. http://agclass.nal.usda.gov/mtwdk.exe?k=default&l=60&w=192781&n=1&s=5&t=2.

[17] Boer (goatbreed) [EB/OL]. [2014-09-25]. http://agclass.nal.usda.gov/mtwdk.exe?s=1&n=1&y=0&l=60&k= default&t=2&w=Boer+%28goat+breed%29.

[18] New edition of CAB Thesaurus Published [EB/OL]. [2014-09-12]. http://aims.fao.org/community/general-information/blogs/new-edition-cab-thesaurus-published#.VDQHUlde8Y0.

[19] 刘春燕，安小米，侯人华. 术语标准研制方法及在信息与文献领域中的应用[J]. 图书情报工作，2014，58(9)：91-95.

[20] International Organization for Standardization. ISO 25964-1 Information and Documentation - Thesauri and Interoperability with Other Vocabularies - Part 1: Thesauri for Information Retrieval [S]. 2011, 87.

[作者简介]

张士男　女，1986 年生，现工作于中国科学院文献情报中心，馆员，中国科学院大学 2012 级图书馆学博士。

[收稿日期：2015-03-05]