基本形式化本体的研究进展

朱玲[1]，董燕[1]，杨峰[2*]
（1. 中国中医科学院 中医药信息研究所，北京 100700；
2. 中国中医科学院 针灸研究所，北京 100700）

［摘要］ 通过介绍基本形式化本体（BFO），给领域本体专家提供通用的顶层结构，有助于各领域专家创建的术语具有更好的互联互通性。通过文献综述的方法，在简要介绍 BFO 及其他顶层本体的基础上，对目前基于 BFO 构建的领域本体进行精炼概述，并以生物医学本体和通用医学本体为例详细阐述。目前基于 BFO 构建的领域本体达到 227 个，其中疾病相关本体 25 个，细胞相关本体 14 个，解剖相关本体 7 个，蛋白质相关本体 7 个。BFO 已经在包括生物医学领域在内的诸多范围内得到了较为广泛的应用，在领域本体的构建中引入 BFO，不但可以提升数据的质量，减少冗余的工作，而且为领域本体的构建提供了框架和基础，为未来的交互和共享提供了可能。

［关键词］ 基本形式化本体；顶层本体；交互作用；生物医学领域；通用医学科学本体
［中图分类号］ R353.11；B016；R24 ［文献标识码］ A ［文章编号］ 1005-9903（2018）02-0208-05
［doi］ 10.13422/j.cnki.syfjx.2018020208
［网络出版地址］ http://kns.cnki.net/kcms/detail/11.3495.R.20171102.1852.026.html
［网络出版时间］ 2017-11-02 18；52

Research Progress of Basic Formal Ontology

ZHU Ling[1], DONG Yan[1], YANG Feng[2*]
（1. *Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*; 2. *Institute of Acupuncture and Moxibustion, China Academy of Chinese Medical Sciences, Beijing* 100700, *China*）

［**Abstract**］ To provide a general top-level structure to the domain ontology experts by introducing basic formal ontology (BFO), which facilitated the interoperability of terminology created by experts in various fields. On the basis of brief introduction of BFO and other top-level ontologies, this paper briefly summarized the domain ontologies based on BFO, and took top ontology of bioscience (BioTop) and ontology for general medical science (OGMS) as examples. There were 227 domain-based ontologies based on BFO, 25 of which were related to the disease. BFO has been widely used in the field of biomedical field. The introduction of BFO in the construction of domain ontology not only can improve the quality of data, reduce the redundant work, but also provide the framework and foundation for the construction of domain ontology, which can provide the possibility for future interaction and sharing in knowledge.

［**Key words**］ basic formal ontology; top ontology; interaction; biomedical field; ontology for general medical science

随着计算机技术在各个领域的使用愈发广泛，如何更好地存储、管理和整合大量来自各领域的信

［收稿日期］ 20170522（011）
［基金项目］ 国家自然科学青年基金项目（81403491）；中国中医科学院基本科研业务费自主选题（ZZ090319）
［第一作者］ 朱玲，副研究员，从事中医药信息研究，Tel：13426395263，E-mail：jjzhuling@163.com
［通信作者］ *杨峰，副研究员，从事针灸文献及理论研究，Tel：13436993010，E-mail：yangfengzhuling@163.com

息已经变得迫在眉睫。本体这一概念从哲学领域引入人工智能、知识工程领域,其内涵从最初仅仅是一种研究存在的科学[1]变成了元数据及其相关关系的规范[2],或者概念模型的明确的规范说明[3],亦或是一套有关某一学科或某一领域的术语词表以及术语间关系的规范和说明[4],但本体到底是什么,本体工程、本体技术又是什么?

以本体为主题词在中国知网进行检索[5],发现2000 年后发表文章约 91 073 篇,1999 年后每年发表相关论文均超过 1 000 篇,2007 年后更是 >5 000 篇论文年增长幅度。可见,有关本体的研究与应用日益广泛。同时,不应忽视的是,有些研究对于本体的认识尚不够深刻、准确、全面,某些研究虽冠以本体之名,然察其具体内容,却与本体实质内涵差异较大。如有研究[6]基于本体构建了中医医案临证关系数据库和中医经典著作文献资源数据库 2 个本体数据库,其中古籍本体包含的功能有古籍本体导入、古籍本体查询、古籍按方查询、古籍分类查询等功能,后者显然不在本体的范畴之列。因此,对于本体应用,尤其是领域本体的构建而言,需要加以重视,深入探讨。有鉴于此,本文拟从领域本体构建原则和理念的角度,对基本形式化本体[7](basic formal ontology, BFO)的相关研究进展进行综合分析,以期为领域本体的构建提供可以共享的、能被认可的本体框架。

1 BFO 的简介

1.1 BFO 与其他顶层本体 基本形式化本体 BFO[8]是由 Barry Smith 和他的同事开发的一个形式化的本体框架,其包含了一系列不同力度的子本体,其顶层类是持续体(continuants)和发生体(occurrents),并在类与类、本体与本体之间定义了关系的一种描述世界的分类架构,以支持科研数据整合的顶层本体。其不包括类似“细胞”、“死亡”或“植物”这样属于特殊领域的具体术语,但是可以处理静态/空间特征和动态/时间特征的事实[9],后者就是属于发生体范畴,其主要用来表明在给定时间内进行的某个操作过程。主要目的是为了给领域本体专家提供通用的顶层结构,方便各领域专家创建子本体,也使得各自的子本体之间具有更好的互联互通性。

BFO,语言和认知工程领域本体[10](domain ontology for linguistic and cognitive engineering, DOLCE)和建议的顶层融合本体[11](suggested upper merged ontology, SUMO)并称为目前科学领域比较公认的三大顶层本体,三者之间互有交叉。然而,作为一个严格的顶层本体,BFO 和其他两者有所区别。与 DOLCE 和 SUMO[12]相比,BFO 不包括一些关于物理、化学、生物、心理或者其他类型实体的表达,即其自身不涉及任何具体的科学领域。因此,与其他顶层本体相比而言,BFO 体量较小,且在本体工程应用的时候更易于管理和实现,换而言之,其在支撑领域专家的本体构建中发挥着更为重要的作用。

此外,尚有日本先进科技研究所提出的另一个更先进的顶层本体[13](yet another more advanced top-level ontology, YAMATO),其理论特征主要包括质量和数量、基于事件的表达、对象、过程和事件。科学本体的最初目标就是描述具有某些共同特征的物体以及他们之间的相互关系,为科学家在个体分类时提供支持和帮助。例如科学家在实验中观察到 1 个新物质,如何为之找到合适的分类。从这个角度而言,BFO 对于世界的描述方式得到了更多科学家的认同,其支持通过构建层级关系以及关系的传递与否等来实现某种关于殊相的推理。

1.2 BFO 的层级结构 BFO[7]以持续体和发生体为顶层结构展开。持续体就是实体,是持续存在的实体。持续体必须具备以下 3 个特点,①是独立的对象和个体,比如我和你;②具有一定的质量和特征,比如某人的身高等;③在任一时间都需要占据一定的空间,比如桌子总要有个地方才能存在。持续体根据其对实体自身的依赖程度,可以分为 3 个下位类,分别为非依赖的持续体(independent continuant),比如西红柿;特别依赖的持续体(specifically dependent continuant),比如西红柿的颜色;一般依赖的持续体(generically dependent continuant),比如电脑中的某个 pdf 文档。

发生体不仅包括发现、发生、展开、发展等连续的过程,也包括过程开始或结束的边界,还有过程发生的时空。其包括 4 个下位类,分别为过程(process),过程边界(process boundary),时区(temporal region)和时空区(spatiotemporal region)。

1.3 BFO 的特点——注重对操作的描述 从内容特点来看,BFO 只是给予研究者关于本体构建的一些理论原则以及如何创建一个好的本体的策略(或理念),而不是介绍本体的构建工具,这一点与 Protégé[14](美国斯坦福大学开发的构建本体的开源工具)并不重复,且互为补充。最新的 BFO 的 owl 版本可以在 http://purl.org/obo/.owl 下载。实际

研究过程中,以 Protégé 为工具,可以很方便地导入最新版本的 BFO 以及本领域相关的术语,然后据此可以开展本领域内的本体创建工作。

一般意义上的本体主要强调领域内概念实体的层级以及实体与实体之间的关系,BFO 与之不同,其主要关注操作过程的逻辑化描述,且在进行实例与实例之间关系表述的同时也强调时间因素,其对于过程和时间两方面的准确把握为术语的全面定义以及术语间关系的准确厘清提供了可靠的方法保证,如对于中药炮制方法等涉及操作过程的领域本体构建大有裨益。

2 基于 BFO 的研究进展

2.1 已利用 BFO 的本体[8] 部分应用 BFO 的本体有阿尔茨海默病本体、副作用本体、银行本体、副作用报告本体、生物信息网络服务本体、生物医疗网格本体、血液本体、生物集合物本体、生物医学伦理本体、生物顶层本体、肿瘤细胞本体、体液本体、细胞行为本体、化学预防癌症本体、心脏疾病本体、细胞系本体、细胞周期本体、通用解剖参考本体、细胞本体、生物利益的化学实体、卡瑞尔细胞系本体、认知范例本体、概念模型本体、交换标准本体、药物相互作用本体、地球科学本体、药物本体、环境本体、情感本体、演化本体、记录行为本体、药物-药物相互作用本体、实验因素本体、电邮本体、金融报告本体、流行病本体、果蝇本体等。

截至 2017 年 3 月 30 日,基于 BFO 构建的领域本体多达 227 个,其中疾病相关本体 25 个,细胞相关本体 14 个,解剖相关本体 7 个,蛋白质相关本体 7 个,事件相关本体 4 个,已经涉及了生物医学的诸多领域,可见 BFO 已经成为其领域本体构建的主要基石。因此,新的生物医学领域本体的构建,包括中医药学领域本体的构建也应该遵循 BFO 的架构,这不但可以充分利用现有的知识和基础,如关系本体、植物学本体等,避免进行重复工作,同时也能为今后领域本体之间的交互提供便利。

2.2 生命科学领域的顶层本体[15](BioTop) BioTop 旨在为该领域的事实描述提供基础的、学界公认的、没有歧义的词汇,一般被用来在更为专业的子领域里创建新的本体顶层模型,或者被用来提升和推进现存本体。BioTop 的最初版本来自于创建一个基于形式化本体的重新设计和扩展分子生物学语料库注释本体(GENIA)的想法,GENIA 是一个应用于分子生物学语料库注释的本体。此后,BioTop 不断扩大。2011 年,简化版首次发布。2013 年,更新发布 BioTop 精简版。目前,已有精简版 2,即 BTL2,其遵循形式化的设计原则,通过 ontology web language 2(owl 2)实施,使用描述逻辑构造函数,使其可以保持自身连续性与一致性,与一体化医学语言系统(unified medical language system, UMLS)[16]相比,前者基于逻辑的推理更为可行,有学者曾经在二者之间尝试映射[17]。

BioTop 的顶层结构是某个时间节点的殊相(particular at some time),其下分状态、倾向、功能、非物质对象、信息对象、物质对象、过程、质量、角色、时间区域、价值区域,见图 1。其中过程下包括行为、生物过程实体、规范的过程实体、集合过程、不成熟的过程实体、瞬间过程实体、生命、非规范的过程实体、物理过程实体、静态过程实体等。对象属性(object property)一级类目有 26 个,分别为在一定时间、导致、伴随发生、被编码、编码、有条件、有持续时间、有起源、有参与者、有时间点、包括、被导致、是条件、被包含、是参与者、先于、是……投影、在某个时间被提及、被代表、产生于、物理上相关、物理上不相关、物理上断开、先于、投射入、代表。有学者[18]基于 BFO 和 BioTop 进行了过程属性的描述逻辑(DL)描述,进一步认可现有的上层本体论框架,在确保本体构建一致性和互操作性上具有实践意义。

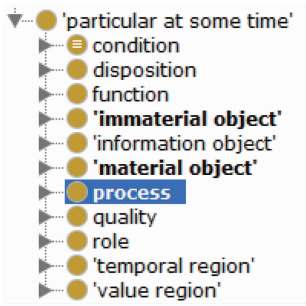

图 1 BioTop 的顶层结构

Fig. 1 Top structure of BioTop

2.3 通用医学科学本体 通用医学科学本体(ontology for general medical science, OGMS)是以 BFO 为基础构建的领域本体之一,其包含了一些临床通用的术语,如疾病、症状、诊断、疾病过程、病人、健康养生提供者等。OGMS 使用 BFO 作为顶层本体,OGMS 的范围局限于人类,但是很多术语也能被应用到其他有机体中。OGMS 提供了疾病的形式化理论架构,因此可用于特殊疾病本体的进一步阐述。从本体的视角来看,临床术语可能并不具有内在一致性,意义模糊且高度依赖上下文,因此临床医学领域本体构建相当困难。利用 OGMS 设计则可以提

供一个形式化的、清楚的、不冗余和没有歧义的临床术语表达，从而能解决上述的困难。OGMS 不是一个疾病本体，只是一个参考本体，其提供关于疾病的总体理论框架以及在临床中会使用到的术语的形式化定义，用来描述各种不同的疾病，为一系列的不同疾病和疾病家族提供本体模型的框架。Oberkampf 等[19]基于 OGMS 和其他开发的生物医学本体(OBO)本体提出了构建更利于交互的临床信息模型(a model for clinical information，MCI)。

如体质遗传疾病是遗传疾病的子类，而遗传疾病又是疾病(disorder)的子类，disorder 是有机体的特性之一，见图 2。图 3 中表示了体质遗传病，owl 表达的第 4 行 < rdfs: subClassOf rdf: resource = " &obo; OGMS _ 0000047 "/ > 表明 OGMS_0000047 是来源于 OBO 本体的。图 4 显示了体质遗传病的层级，由此可以追溯到 BFO 的物质实体以及非依赖持续体。此外 OGMS 也利用了关系本体(RO)[20]。

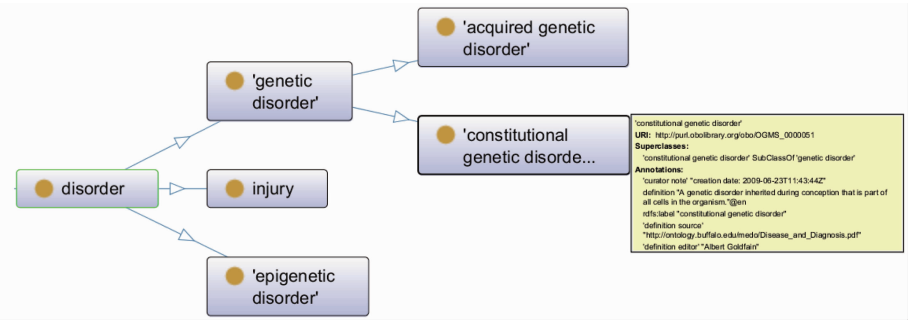

图 2　疾病的层级

Fig. 2　Level of disorder

```
< ! - - http: // purl. obolibrary. org/obo/OGMS_0000051 - - >
< owl: Class rdf: about = " &obo; OGMS_0000051" >
< rdfs: label > constitutional genetic disorder < /rdfs: label >
< rdfs: subClassOf rdf: resource = " &obo; OGMS_0000047"/ >
< obo: IAO_0000232 > creation date: 2009 - 06 - 23T11: 43: 44Z < /
obo: IAO_0000232 >
< obo: IAO_0000117 > Albert Goldfain < /obo: IAO_0000117 >
< obo: IAO_0000115 xml: lang = " en" > A genetic disorder inherited
during conception that is part of all cells in the organism. < /obo: IAO_
0000115 >
< obo: IAO_0000119 > http: // ontology. buffalo. edu/medo/Disease_and
_Diagnosis. pdf < /obo: IAO_0000119 >
< /owl: Class >
```

图 3　OGMS 的 owl 表达

Fig. 3　Expression of owl of OGMS

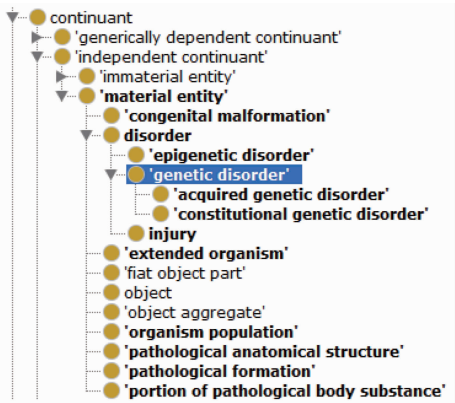

图 4　体质遗传病的层级

Fig. 4　Level of constitutional genetic disorder

2.4　一个本体和术语的真实案例——ISO　术语研究作为一个可以解决大量使用在商业、生产和运输领域的技术词汇的方式得到了蓬勃发展。术语学家们对术语的使用尤为感兴趣，特别是标准化的角度和不同技术语言之间的翻译。术语研究的焦点是概念，因为在术语学家的眼中，如果一个术语可以被翻译为另一种语言，那么在某种程度上，其就是一个概念，因为其相应领域的使用者其实是认同某一共同的观点的。国际标准组织(International Organization for Standardization，ISO)旨在建立这样一个术语基础，其追求“科技知识在概念层面的控制”。ISO 希望通过这种方式支持翻译者的工作，支持用不同语言表达的数据收集和处理者的工作。中国中医科学院中医药信息研究所牵头研制的 ISO/TS 16843-2-2015 health informatics——categorial structures for representation of acupun-cture——part 2: needling[21]则对针刺领域的概念及概念间关系进行了描述和定义，是领域本体与 ISO 结合成功的实例之一。BFO 顶层本体的引入为未来更多语义信息领域的国际通用标准的构建提供了更容易交互与共享的框架，为科研数据的交流、领域知识的共享奠定了坚实的基础。

3 讨论

概言之,BFO 是一个可以被重用的顶层本体,其强调信息系统的互操作性,并且得到了业界的广泛认可,尤其在生物医学领域已取得长足的进展,为科研数据的交流、领域知识的共享提供了坚实的基础。在领域本体的构建中引入 BFO,使用其顶层框架,不但可以提升数据的质量,减少冗余的工作,而且能够为领域本体的构建提供了框架和基础,为未来领域本体之间的交互和共享提供可能。但由于时间精力所限,本文在简单介绍 BFO 顶层分类的基础上,对以 BFO 为框架构建的生物医学领域的 2 个本体进行了简单介绍,至于如何在领域本体构建中利用 BFO,兼容 BFO,还有很多工作要开展。此外,需要注意的是,中文领域本体构建大多不采用英文作为通用语言,这对本体的共建与共享带来了一定的障碍,未来或考虑将 BFO 翻译为中文,或者在中文领域本体构建的时候同时完成英文版本,都是可行的路径。

[参考文献]

[1] 胡伟希. 自本体与对本体:中西哲学的诠释学基础[J]. 孔子研究,2005(3):4-14.
[2] 董志. 利用语义网技术实现铁路交通的地理语义查询(一)——C#中地理数据的网络获取和本体构建[J]. 电脑编程技巧与维护,2013(11):5-11.
[3] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acguisition, 1993, 5(2): 199-220.
[4] 陈建. 领域本体的创建和应用研究[D]. 北京:对外经济贸易大学,2006.
[5] 《中国学术期刊(光盘版)》电子杂志社有限公司. 中国知网[EB/OL]. http://www.cnki.net/,2017-05-05.
[6] 郑健,李其锫. 基于本体的名老中医医案研究应用系统[J]. 科技信息,2008,18(28):48-50.
[7] Arp R, Smith B, Spear A D. *Building Ontologies with Basic Formal Ontology* [M]. Commonwealth of Massachusetts: MIT Press, 2015: 85-121, 155-160.
[8] IFOMIS. Basic formal ontology[EB/OL]. http://ifomis.uni-saarland.de/bfo/,2017-05-23.
[9] Wikipedia. BFO 的定义[EB/OL]. https://en.wikipedia.org/wiki/Basic_Formal_Ontology, 2017-05-23.
[10] Borgo S, Masolo C. *Ontological Foundations of Dolce* [M]. Berlin: Springer Netherlands, 2010: 279-295.
[11] Ontolog A L. The suggested upper merged ontology[EB/OL]. https://www.researchgate.net/publication/238311522_The_Suggested_Upper_Merged_Ontology, 2008-07-09.
[12] Smith B. On classifying material entities in basic formal ontology[C]// Interdisciplinary Ontology. Proceedings of the Third Interdisciplinary Ontology Meeting: 2012 volume. Tokyo: Keio University Press, 2012: 1-13.
[13] Mizoguchi R. YAMATO: yet another more advanced top-level ontology [C] // Interdisciplinary Ontology. Proceedings of the First Interdisciplinary Ontology Meeting: 2010 volume. Tokyo: Keio University Press, 2010: 1-16.
[14] National Institute of General Medical Sciences. Protégé [EB/OL]. http://protege.stanford.edu/,2017-05-05.
[15] Schulz S, Boeker M. *BioTopLite: An Upper Level Ontology for the Life Sciences. Evolution, Design and Application* [M]. Koblenz: IOS Press, 2013: 1889-1899.
[16] U. S. National Library of Medicine. 一体化医学语言系统[EB/OL]. https://www.nlm.nih.gov/research/umls/,2017-06-15.
[17] Schulz S, Beisswanger E, Bodenreider O, et al. Alignment of the UMLS semantic network with BioTop: methodology and assessment [J]. Bioinformatics, 2009, 25(12): 69-76.
[18] Andrade A Q, Ward B, Janna H, et al. Process attributes in bio-ontologies[J]. BMC Bioinformatics, 2012, 13(1): 1-11.
[19] Oberkampf H, Zillner S, Bauer B, et al. An OGMS-based Model for Clinical Information (MCI) [C] // Interdisciplinary Ontology. Proceedings of the Fifth Interdisciplinary Ontology Meeting: 2014 volume. Tokyo: Keio University Press, 2014: 1-12.
[20] Chris M. Relational ontology [EB/OL]. https://raw.githubusercontent.com/oborel/obo-relations/master/ro.owl,2017-05-08.
[21] ISO. 针刺语义信息标准[EB/OL]. https://www.iso.org/obp/ui/#iso:std:iso:ts:16843:ed-1:v1:en, 2017-05-16.

[责任编辑 刘德文]