Foreword to Chinese Translation

Ontology: Tool for Broad Spectrum Knowledge Integration

Barry Smith

BFO: The Beginnings

This book was first published in 2015. Its primary target audience was bio- and biomedical informaticians, reflecting the ways in which ontologies had become an established part of the toolset of bio- and biomedical informatics since (roughly) the completion of the Human Genome Project (HGP). As is well known, the success of HGP led to the transformation of biological and clinical sciences into information-driven disciplines and spawned a whole series of new disciplines with names like ‘proteomics’, ‘connectomics’ and ‘toxiocopharmacogenomics’.

It was of course not only the human genome that was made available for research but also the genomes of other ‘model organisms’, such as mouse or fly. The remarkable similarities between these genomes and the human genome made it possible to carry out experiments on model organisms and use the results to draw conclusions relevant to our understanding of human health and disease. To bring this about, however, it was necessary to create a controlled vocabulary that could be used for describing salient features of model organisms in a species-neutral way, and to use the terms of this vocabulary to tag the sequence data for all salient organisms. It was with the purpose of creating such a vocabulary that the Gene Ontology (GO) was born in a Montreal hotel bar in 1998.[1]

Since then the GO has served as mediator between the new genomic data on the one hand, which is accessible only with the aid of computers, and what we might think of as the ‘old biology data’ captured using natural language by means of terms such as ‘cell division’ or ‘mitochondrion’ or ‘protein binding’. Very soon the success, and the utility, of the GO led to the creation of other ontologies covering domains not covered by the GO, such as cell types, proteins, anatomy and diseases, grouped together under the heading ‘OBO’ for ‘Open Biomedical Ontologies’. Beginning in around 2003, however, the problem arose of determining what should serve as the common formal architecture that would bind these ontologies together, and it was in this context that Basic Formal Ontology was born.

The result, which is documented at length in the pages that follow, is the Open Biological and Biomedical Ontologies (OBO) Foundry, a suite of life science ontologies all of which descend from BFO and rest on the same set of principles of good practice for ontology development, principles designed to ensure interoperability. The OBO

[1] https://www.ncbi.nlm.nih.gov/pubmed/10802651

Foundry ontologies[2], like BFO, are created and maintained manually. It is human beings with corresponding scientific and ontological expertise who are responsible for populating the ontologies, formulating definitions and responding to requests from users of the ontology who identify errors or gaps. The underlying idea is that ontologies should be developed in a process of coordinated evolution, each ontology being used in a variety of other co-developed ontologies, for instance through use of its terms in writing definitions.

The OBO Foundry principles for coordinated evolution have brought considerable success. Where many ontology initiatives, because they are too closely tied to specific local projects and research groups, enjoy short lives, the aggressive reuse of OBO Foundry ontologies to highly diverse projects and use cases has led to an increasing influence of the OBO Foundry and with this came also the increasing influence of BFO, illustrated not least in the publication of this book.

BFO in China

As this volume demonstrates, BFO and the OBO Foundry have been discovered, too, in China, where BFO is already being used as top-level ontology by ontologies such as the Traditional Chinese Drug Ontology (TCDO)[3] and Chinese Plant Species Diversity Domain Ontology.[4] The newly initiated OntoChina program[5] aims to support collaborative ontology development, application and distribution across the life sciences and beyond. It is members of the OntoChina group who are responsible for this translation, for the translation into Chinese not only of the BFO ontology itself, but also of other BFO-based ontologies, such as the Information Artifact Ontology (IAO)[6] and the Ontology for Biomedical Investigations[7]. OntoChina also plans to train future ontologists in China using BFO and the BFO-based ontology ecosystem as foundation.

BFO in the OBO Foundry

It was not until 2016 that BFO was adopted as the official top-level ontology of the OBO Foundry and it is worth examining in this light one core principle of the Foundry suite, namely the principle of orthogonality. Two ontologies are 'orthogonal' in Foundry terminology, if they do not overlap – which means: if they share no terms in common. This principle was adopted in support of the idea that there should be exactly one ontology for each domain – so that each term should belong to exactly one ontology. The rationale for this principle is that if all those working in a given domain use the same

---

[2] Barry Smith, et al., "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration", *Nature Biotechnology*, 25 (11), 2007.
[3] Yan Zhu, Lihong Liu, Bo Gao, Yongqun He. TCDO: a community-based Traditional Chinese Drug Ontology. The 11th International Biocuration Conference (Biocuration-2018), Crowne Plaza Hotel Shanghai Fudan, Shanghai, China, April 8-11, 2018.
[4] http://manu44.magtech.com.cn/Jwk_infotech_wk3/CN/abstract/abstract4189.shtml
[5] http://ontochina.org
[6] Werner Ceusters and Barry Smith, "Aboutness: Towards Foundations for the Information Artifact Ontology", *Proceedings of the Sixth International Conference on Biomedical Ontology* (ICBO), (CEUR 1515), 2015, 1-5.
[7] Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen,et al. "The Ontology for Biomedical Investigations", *PLoS ONE*, 11(4): e0154556, April 29, 2016.

ontology, then the annotations that they create when they tag their data with ontology terms, will be consistent with each other, and thus cumulate in an orderly fashion. Such cumulativity would allow also the identification of conflicts between different sets of annotations. (Compare the way in which the SI System of Units allows not only the cumulation but also the comparison of *quantitative* data.)

If, however, another core principle of the OBO Foundry requires the use of BFO as common upper level for all Foundry ontologies. This means that BFO terms belong to each Foundry ontology, which conflicts with the idea that each term should belong to exactly one ontology. Terms not only from the BFO top-level ontology but also from mid-level ontologies such as IAO and OBI are re-used in multiple domain ontologies at lower levels. In like fashion, terms from more general domain ontologies such as the Cell Ontology are reused in ontologies such as the Cell Line Ontology and the Cell Cycle Ontology at lower levels.

How, then, are do we save the principle that each term should belong to exactly one ontology? The answer is that orthogonality should be interpreted in such a way that each term should have its source – the original home where the term and its definition were manually crafted – in some one specific ontology, marked through the fact that the identifier for this ontology is included in the relevaxnt term IRI. This original term IRI should then travel with the term when it is reused in other ontologies, thereby preserving the link to the original home and at the same time networking the ontologies together, through reuse of terms not only in multiple ontology is-a hierarchies, but also in multiple term definitions.[8]

The hierarchy of ontologies reflects the hierarchy of sciences

The way in which OBO Foundry ontologies are organized on successive levels represituation is analogous to what obtains in the organization of science. At any given stage of its development prior knowledge obtained through scientific research is organized hierarchically. Very general knowledge pertaining to general laws of physics or chemistry provides high-level starting points for scientific subdisciplines at successively lower levels of, for example, electrodynamics, quantum electrodynamics, organic chemistry, petroleum chemistry, neurochemistry, and so forth. The nodes in this (somewhat idealized) hierarchy then serve as starting points for huge numbers of further disciplines and subdisciplines, and also determine which bodies of prior knowledge are encapsulated in textbooks and which new results are reported in which journals. It is the titles of such textbooks and journals which tell us where we can find both prior knowledge and new results, just as it is the names of ontologies such as the Vaccine Ontology or the Infectious Disease Ontology which tell us where corresponding terms and definitions are to be found.

[8] For example along the lines illustrated in https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-513

Implications for deep learning

All of this, perhaps surprisingly, has implications for contemporary artificial intelligence research. This is because one major shortfall of today's dominant AI paradigm of autonomous deep learning is its inability to deal with prior knowledge. This is a problem, in areas such as biology and medicine, where huge amounts of prior knowledge already exist. It is a problem, too, for the public acceptance of the results of deep learning, since the use of neural networks creates what is in effect a black box, which means that we cannot understand how given results were achieved. This means, also, that we cannot assign responsibility – on the side of either the machine or its human programmers – when things go wrong.

How, then, to supplement neural networks with prior knowledge in areas such as biology and medicine. To create representations inside the computer of entire disciplines, or of entire suites of disciplines and sub-disciplines, is of course an impractical goal. But something more modest is not only achievable but is in fact already being achieved. This is the idea of using ontologies as the terminological skeleton of a body of knowledge – each ontology being seen as a network of nodes and edges but contained within a much larger hierarchically organized network constructed in the manner of the OBO Foundry.

One example of work demonstrating how the prior knowledge of human experts can be incorporated into the machine learning process is provided by SciMiner strategy applied by Oliver He and his team to support mining of the vaccine literature.[9] Another example is Onto2Vec approach to learning feature vectors for biological entities on the basis of their annotations to biomedical ontologies.[10] These and many further examples show how – where successful ontologies such as the Gene Ontology and BFO have thus far been the product of manual development – the border between manual ontology development and the development of useful ontology-based artifacts drawing on the autonomous workings of the machine is gradually breaking down.

The future of BFO

The world of BFO ontologies has developed in significant ways since 2015 also in virtue of the fact that the BFO approach to building suites of ontologies for the tagging of large, heterogeneous bodies of data is now being extended to a variety of other fields, including military intelligence, systems engineering, and digital manufacturing[11]. The approach is being used by the United Nations Environment Programme[12], and also in

---

[9] Junguk Hur, Arzucan Özgür, Yongqun He, Ontology-based literature mining of E. coli vaccine-associated gene interaction networks Journal of biomedical semantics, 2017/12. Junguk Hur, Arzucan Özgür, Yongqun He, Ontology-based literature mining and class effect analysis of adverse drug reactions associated with neuropathy-inducing drugs, Journal of biomedical semantics, 2018/12, 9.

[10] See Fatima Zohra Smaili, Xin Gao, Robert Hoehndorf, "Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations", Bioinformatics, 2018 - DOI:10.1093/bioinformatics/bty259

[11] Development of a manufacturing ontology for functionally graded materials

[12] DOI 10.1186/s13326-016-0097-6

the Industrial Ontologies Foundry initiative.[13] Existing ontologies are being re-engineered to be BFO-conformant,[14] and new applications of BFO are leading also to new questions as to how BFO may need to be extended to new types of entities in the future. BFO has been applied to space objects,[15] biobanking,[16] organizations,[17] functions,[18] and causes,[19] human behavior change,[20] and military doctrine.[21] One major BFO-based suite of ontologies designed for general purpose use across a wide variety of domains is the Common Core Ontologies (CCO),[22] which has in turn served as the starting point for a the growing set of extension ontologies listed in Table 1.

Aircraft Ontology	Mission Planning Ontology
Airforce Aircraft Maintenance Ontology	Occupation Ontology
Army Universal Task List Ontology	Outer Space Ontology
Airforce Aircraft Maintenance Ontology	Physiographic Feature Ontology
Army Universal Task List Ontology	Sensor Ontology
Emotion Ontology	Skills Ontology
Hydrographic Feature Ontology	Space Object Ontology
Legal and Criminal Act Ontology	Transportation Infrastructure Ontology
Military Operations Ontology	Undersea Warfare Ontology
Mission Planning Ontology	Watercraft Ontology

Table 1: Extension Ontologies of the Common Core

The goal of ontology is to advance the communication and sharing of data. It is thus a truly welcome step that the world of BFO ontologies should be extended in this way to include the new community of ontologists in China.

---

[13] http://industrialontologies.org.

[14] Thomas J. Hagedorn, Barry Smith, Sundar Krishnamurty, Ian Grosse, "Interoperability of disparate engineering domain ontologies using Basic Formal Ontology", *Journal of Engineering Design*, June 2019, https://doi.org/10.1080/09544828.2019.1630805.

[15] Alexander P. Cox, Christopher K. Nebelecky, Ronald Rudnicki, William A. Tagliaferri, John L. Crassidis, Barry Smith, "The Space Object Ontology", *19th International Conference on Information Fusion* (FUSION 2016), Heidelberg, Germany, July 5-8, 2016.

[16] OBIB-a novel ontology for biobanking

[17] Information Architecture for Organizations: An Ontological Approach
Utecht J, et al. OOSTT: a Resource for Analyzing the Organizational Structures of Trauma Centers and Trauma Systems.

[18] Andrew D. Spear, Werner Ceusters, Barry Smith, "Functions in Basic Formal Ontology", *Applied Ontology*, 11 (2), (2016), 103-128.

[19] Barton, A., Jansen, L., & Ethier, J.-F. (2018). A taxonomy of disposition-parthood. FOUST II: 2nd Workshop on Foundational Ontology, 1-10, Galton, A. & Neuhaus, F. (Eds), In: Proceedings of the Joint Ontology Workshops 2017, CEUR Workshop proceedings, Vol. 2050.

[20] The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation

[21] Peter Morosoff, Ron Rudnicki, Jason Bryant, Robert Farrell, Barry Smith, "Joint Doctrine Ontology: A Benchmark for Military Information Systems Interoperability", *Semantic Technology for Intelligence, Defense and Security* (STIDS), 2015, CEUR vol. 1523, 2-9.

[22] https://www.cubrc.org/index.php/data-science-and-information-fusion/ontology

中文版序

本体：广谱的知识集成工具

BFO：起源

本书于2015年首次出版，主要面向生物信息和生物医学信息领域的读者。本书反映了（大致）从人类基因组计划（Human Genome Project, HGP）完成以来，本体如何成为生物信息学和生物医学信息学工具集的一个既定部分。众所周知，HGP成功推动了生物学和临床科学向信息驱动学科的转变，并促进了“蛋白质组学”“连接组学”和“毒性药物基因组学（toxiocopharmacogenomics）”等一系列新学科的诞生。

显然，科学研究的对象不只限于人类基因组学，也涉及小鼠、果蝇等“模式生物”的基因组学。而这些基因组与人类基因组之间惊人的相似性，使得我们能基于模式生物的实验结果得出相应的结论来理解人类健康和疾病。但要实现这一点，必须创建一个受控词表，以物种中立的方式描述模式生物的特征（salient features），并用词表中的术语来标记重要生物的生物序列数据。1998年诞生于蒙特利尔一家旅店酒吧的基因本体（Gene Ontology，GO）[①]，就是基于这样目的建立的词表。

自此，GO成为新老基因组数据之间的中介——新基因组数据是指只有通过计算机才能访问的数据，而我们所认为的“老生物数据”则是基于自然语言使用诸如“细胞分裂”“线粒体”或“蛋白质结合”等术语来表示。很快，GO所取得的成功及其所具备的实用性，也推动了GO未涵盖领域的本体创建，如细胞类型，蛋白质，解剖学和疾病等，这些本体都归属于开放生物医学本体（Open Biomedical Ontologies，OBO）名下。然而，从2003年左右开始，发现需要确立一个共同的形式化结构来将这些本体整合到一起。于是，基本形式化本体（Basic Formal Ontology，BFO）应时而生。

这导致开放式生物和生物医学本体工场 [Open Biological and Biomedical Ontologies（OBO）Foundry] 的一整套生命科学本体，都基于BFO构建，并遵守

① https://www.ncbi.nlm.nih.gov/pubmed/10802651

同一套本体开发良好实践原则，以确保本体之间的互操作性，本书后续将详细介绍。和 BFO 一样，OBO 工场[①]的众多本体也需要人工创建和维护。只有具有相应学科和本体专业知识的人才能负责增添本体内容、制定定义并回应发现错误或漏洞的本体使用者所发送的请求。基本思想就是，本体应该以共同开发的方式开发，而且每个本体都可被用于其他共同开发的本体中，例如在编写定义时使用其他本体的术语。

OBO 工场的共同开发原则成效显著。虽然许多本体项目因与特定工程和研究团队的耦合过于紧密而寿命短暂，但 OBO 受到了积极复用且 BFO 影响力持续扩大，本书的出版尤其说明了这一点。

BFO 在中国

正如这里所说的，中国的研究者也发现了 BFO 和 OBO 工场，BFO 已经被传统中药本体（Traditional Chinese Drug Ontology，TCDO）[②]和中国植物物种多样性领域本体[③]等用作顶层本体。新启动的 OntoChina 项目[④]旨在支持跨生命科学和其他领域的协作本体开发、应用和发布。OntoChina 小组的成员负责本书的翻译，他们不仅把 BFO 本身翻译成了中文，还翻译了其他几个以 BFO 为基础的本体，如信息工件本体（Information Artifact Ontology，IAO）[⑤]和生物医学研究本体（Ontology for Biomedical Investigations，OBI）[⑥]。此外，OntoChina 进一步计划以 BFO 和基于 BFO 的本体生态系统为基础，在中国培养未来的本体学家。

直到 2016 年，BFO 才被正式作为 OBO 工场的顶层本体。这值得我们来审视 OBO 工场的一个核心原则，即正交性原则。如果两个本体之间没有重叠，也就是说：如果它们没有相同的术语，那么它们在术语层面上是“正交的”。应用这一原则是为了保证每个领域只有一个本体，从而每个术语就只属于一个本体。而这个原则的合理性在于，假定在特定领域工作的所有人都使用同一

① Barry Smith, et al. The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature Biotechnology, 25(11), 2007.

② Yan Zhu, Lihong Liu, Bo Gao, Yongqun He. TCDO: a community-based Traditional Chinese Drug Ontology. The 11th International Biocuration Conference (Biocuration-2018), Crowne Plaza Hotel Shanghai Fudan, Shanghai, China, April 8-11, 2018.

③ http://manu44.magtech.com.cn/Jwk_infotech_wk3/CN/abstract/abstract4189.shtml

④ http://ontochina.org

⑤ Werner Ceusters, Barry Smith. Aboutness: Towards Foundations for the Information Artifact Ontology. Proceedings of the Sixth International Conference on Biomedical Ontology (ICBO), 2015, CEUR 1515:1-5.

⑥ Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, et al. The Ontology for Biomedical Investigations. PLoS ONE, 2016, 11(4): e0154556.

本体，那么当他们用本体里的术语标记数据时所创建的注释也将彼此一致，这样数据就能有序地累积。同时，这种累积还能使我们甄别出不同注释数据集之间的冲突（就像采用国际单位制不仅可以进行累加，还可以进行定量数据的比较）。

然而，如果 OBO 工场的另一个核心原则要求把 BFO 用作所有工场本体的通用上层，这就意味着 BFO 术语属于每个工场本体，与每个术语应该只属于一个本体的观点相冲突。BFO 顶层本体和诸如 IAO 和 OBI 等中层本体的术语，都被复用到很多低层次的领域本体中。与此类似，一些较为通用的领域本体，例如细胞本体（Cell Ontology）中的术语又被复用在细胞系本体（Cell Line Ontology）和细胞周期本体（Cell Cycle Ontology）等更低层次的本体中。

那么，我们如何保证每个术语都只属于一个本体这一原则成立呢？答案是将正交性理解为每一个术语应有其源头—术语及其定义被人工创建的初始地。对于某些特定的本体来说，可以通过相关术语的国际资源标识符（Internationalized Resource Identifier，IRI）所确定的标识符来完成。当术语在其他本体中被复用时，应该附带其原始术语的 IRI，这样就能保留其初始链接，并通过在多个本体中 is-a（是一种）的层次结构及不同的术语定义中复用术语，将本体连接成网络。

本体的层次结构反映了科学的层次结构

OBO 工场本体按照逐层级表征的方式来组织，类似于科学的组织方式。在其发展的任一阶段，通过科学研究获得的先验知识都是以层次结构方式来组织。关于物理学或化学一般规律的非常普遍的知识，是层级依次降低的分支学科（如电动力学、量子电动力学、有机化学、石油化学、神经化学等）的高层次起点。这个（有些理想化的）层次结构中的节点随后又进一步作为其他学科和分支学科的起点，并决定哪些先验知识是在教科书中介绍的，哪些新结果在哪些期刊里发表。课本和杂志的名称能告诉我们哪里可以找到先验知识和新成果，就像一些本体的名字，如“疫苗本体”或“传染性疾病本体”那样，告诉了我们可以找到相应术语和定义。

对深度学习的启示

有些令人惊奇的是，所有这些都对当代人工智能研究具有启示意义。这是因为，目前主流的人工智能范式在深度自主学习方面存在一个重大缺陷，即无法处理先验知识。这在生物和医学等领域就是一个问题，因为这些领域已经存在了大量的先验知识。而对于公众接受深度学习的结果来说，也会是一

个问题，因为使用神经网络所产生的是黑箱，即意味着人们无法理解给定的结果是如何实现的。也就是说，当出现问题时，我们不能厘清是机器还是程序员的原因。

那么，如何用生物学和医学等领域的先验知识来弥补神经网络的缺陷呢？在计算机内部对所有学科或整套学科和子学科进行表征，显然是好高骛远的。而相对更实际的办法则不仅可以实现，而且实际上正在被实现。即以本体作为知识体的术语框架——每个本体都是一个包含若干节点和连边的网络，同时它又被包含在一个以 OBO 工场的方式所构建的更大的层次结构网络中。

何勇群团队使用 SciMiner 工具来支持疫苗文献的挖掘，就是将人类专家的先验知识整合到机器学习中的一个例证[①]。另一个案例是 Onto2Vec 方法，它通过对基于生物医学本体的注释数据，来学习生物实体特征向量[②]。上述及其他更多的案例，展示了如何借助机器自动化，逐步打破人工构建本体和基于本体的工件开发之间的边界。但是，目前一些成功的本体如 GO 和 BFO 仍来源于人工创建。

BFO 的未来

自 2015 年以来，随着 BFO 及相关本体的蓬勃发展，使用基于 BFO 构建的系列本体来标记大型、异构数据的方法，已经应用到越来越多的领域，如军事情报、系统工程和数字制造[③]等。联合国环境规划署[④]和 NIST 工业本体工场项目也都实施了该方法。一些现有的本体也正在进行重构使其符合 BFO 的规范，而 BFO 在新领域的应用也带来了前所未有的挑战，即未来如何对 BFO

① Junguk Hur, Arzucan Özgür, Yongqun He. Ontology-based literature mining of E. coli vaccine-associated gene interaction networks. Journal of biomedical semantics, 2017/12. 8.
Junguk Hur, Arzucan Özgür, Yongqun He, Ontology-based literature mining and class effect analysis of adverse drug reactions associated with neuropathy-inducing drugs. Journal of biomedical semantics, 2018/12, 9.

② Fatima Zohra Smaili, Xin Gao, Robert Hoehndorf. Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics, 2018 - DOI:10.1093/bioinformatics/bty259.

③ Development of a manufacturing ontology for functionally graded materials.

④ DOI 10.1186/s13326-016-0097-6